SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Prototyping a web-based phonetic training game to improve /r/-/l/ identification by Japanese learners of English

Adriana Guevara-Rukoz[1], *Alexander Martin*[2], *Yutaka Yamauchi*[3], *Nobuaki Minematsu*[1]

[1]School of Engineering, The University of Tokyo
[2]Centre for Language Evolution, University of Edinburgh
[3]School of Education, Soka University

aguekoz@gavo.t.u-tokyo.ac.jp, alxndr.martin@gmail.com,
yutaka@soka.ac.jp, mine@gavo.t.u-tokyo.ac.jp

Abstract

Even after years of study, language learners may have difficulty perceiving L2 sounds. For instance, Japanese listeners show difficulty differentiating American English /r/ and /l/. Previous research has shown that phonetic training may improve learners' perception of the contrast. While this training paradigm appears as a promising tool for language learning, its transition from the laboratory to the classroom needs to be facilitated. Not only does phonetic training require recording and/or manually editing many training exemplars, training sessions are also often long and repetitive. Given these obstacles, the long-term goal is to make phonetic training more applicable to real-life learning. In this preliminary study, we prototype a self-paced, web-based phonetic training program, featuring both identification and discrimination tasks as playable mini-games. Participants are trained using nonword minimal pairs (e.g., /lapu/-/rapu/), presented in isolation in clean speech. Their ability to identify the target phonemes is assessed before and after training, with stimuli also presented in noise and/or in sentences, to test perceptual robustness. We assess the effectiveness of the phonetic training game in its current form and discuss future improvements, notably in the context of using speech engineering to automate and augment High Variability Phonetic Training (HVPT) programs.

Index Terms: Computer-Assisted Language Learning (CALL), phonetic training, phonetics, autonomous learning, speech engineering

1. Introduction

One of the challenges of foreign language learning resides in the proper acquisition of the sound system of the target language. Indeed, the perception and production of said sounds may be compromised if the sound inventories of the native and L2 languages are in conflict. For instance, Japanese learners of English experience difficulties differentiating the English /r/ and /l/ liquid consonants, due to Japanese only having one liquid consonant in its sound inventory [1, 2].

Previous studies have shown that it is possible for Japanese listeners to improve their accuracy at identifying English /r/ and /l/ by participating in phonetic training. While exact details of the procedure vary according to the individual studies, the main idea is that learners complete tasks that require them to actively categorise and/or differentiate the target phonemes. Most studies have reported phonetic training to be effective at improving perception [3, 4, 5, 6, 7, 8, 9, 10]—and even production [11, 12]—of the target phonemes. In spite of its success in the laboratory, the application of phonetic training in the classroom is hindered by practical considerations: phonetic training programs, as described in the literature, may be cumbersome both to make and use.

This latter point is the main focus of our work. We built a phonetic training program consisting of short, web-based gaming sessions, to be completed in a self-paced manner without needing to return to the laboratory. We assessed the effectiveness of this training program not only by evaluating learners' identification of clean stimuli, but also in the context of commonly-encountered modifications of speech such as noise/channel distortions and embedding in sentences, providing a more ecological test of outcomes of our training program.

2. Methods

2.1. Participants

Twenty-four native Japanese listeners were tested in Tokyo, Japan (9 female, 15 male). Their ages ranged from 20 to 56 years old (median age: 22.5). Participants had been learning English for a median of 10 years (range: 4-40 years).

2.2. Stimuli

2.2.1. Raw recordings

Four sets of ten /r/-/l/ minimal pairs of disyllabic nonwords were built for the purpose of the experiment, yielding a total of 40 minimal pairs (i.e., 80 nonwords). All stimuli were recorded with stress on the first syllable, the target phoneme being in the first syllable for half of the items. In each set of 10 minimal pairs, the target phoneme was positioned in a simple onset (3 pairs, e.g., /rapu/-/lapu/), a simple coda (2 pairs, e.g., /setor/-/setol/), a complex onset (4 pairs, e.g., /groʃu/-/gloʃu/), or a complex coda (1 pair, e.g., /hazors/-/hazols/).

Five adult native speakers of American English (3 female: speakers F1, F2, F3; 2 male: speakers M1 and M2) were recorded in a soundproof room. The stimuli were recorded with a sampling rate of 44.1 kHz (16 bit), and were later downsampled to 16 kHz. Items were shown on-screen one by one, in a randomised order. Speakers produced the stimuli in isolation, as well as embedded in three carrier sentences where the target nonword was in initial position ("[nonword] is a big city."), medial position ("See you at [nonword] station!"), or final position ("Have you seen my [nonword]?").

2.2.2. Speech modifications

From these raw recordings we produced noise-embedded stimuli, using speech engineering software presented in [13]. There

10.21437/SLaTE.2019-5

were a total of ten possible noise modifications.

- Additive noise: The soundwaves of noise recordings were overlapped to those of the stimuli at signal-to-noise ratio of -1 dB. Noise recordings featured a crowded event, a train station, a street, a room with air conditioning, a mall, or babble noise.
- Reverberation: Audio recordings were transformed so as to mimic speech being produced in (1) a church or (2) a ballroom.
- Channel distortions: Audio recordings were transformed in order to resemble audio from (a) second-generation (2G) digital cellular networks used by mobile phones, or (b) radio communication devices.

Additionally, we created 3 artificial voices by modifying recordings by male speaker M1. The three voices were obtained by applying a linear transformation in the cepstrum domain to speaker M1 recordings as to slightly modify his vocal tract length (VTL) in said recordings. We used VTL ratios of 0.90, 0.84, and 0.94.

2.3. Procedure

2.3.1. Pre-training and post-training tests

Before and after completing all phonetic training sessions, participants were tested in terms of their identification of /r/ and /l/ in nonwords. Participants were tested in a soundproof room, using headphones and an audio volume that they deemed comfortable. At each trial, participants listened to a stimulus once. They were then required to choose one of the two possible transcription options appearing on-screen (e.g., *rapu* and *lapu*). They were not provided with any feedback during the tests.

For the pre- and post-training tests, only stimuli from speakers M1 and F1 were used. There were a total of four blocks, differing in the nature of the stimuli presented within each block: (1) noiseless isolated, (2) noiseless sentence-embedded, (3) noisy isolated, (4) noisy sentence-embedded. In each block, participants were asked to identify the correct transcription of all items in two of the stimuli sets. In total there were 640 trials (2 speakers ×4 blocks ×2 stimuli sets ×10 minimal pairs ×2 pair items ×2 repetitions). The procedure lasted approximately 40 minutes, including short breaks after each block.

2.3.2. Training

Participants were assigned to one of the following training conditions: (1) low natural variability (LNV), (2) high natural variability (HNV), (3) high artificial variability (HAV). They differed in the speakers used for training, as shown in Table 1. For LNV, different tokens of each stimulus, produced by the same speaker (M1), were used. For HNV, tokens produced by different speakers were used. For HAV, voices B, C, and D corresponded to speaker M1 recordings after being transformed to slightly change the VTL of speaker M1. In this initial study we did not focus on the difference in performance between these three training regimes, due to sample size-related limitations. However, they were included in order to allow for exploratory analyses for future studies[1].

Apart from the differences at the level of the voices, the training procedure was identical for all participants. It consisted

[1]We did not find any difference in performance between the three groups within the scope of this study.

	A	B	C	D
LNV	M1	M1 (token 2)	M1 (token 3)	M1 (token 4)
HNV	M1	F2	F3	M2
HAV	M1	$M1_{0.90}$	$M1_{0.84}$	$M1_{0.94}$

Table 1: *Voices used during training. Subscript numbers indicate VTL ratios, when applicable.*

of 12 training sessions, each lasting approximately $10-15$ minutes. All participants completed a maximum of one session per day within an interval of 3-5 weeks. All participants completed the sessions on their own (e.g. at home), in a self-paced fashion, with the objective of improving their perception of the /r-l/ contrast. They were instructed to do the training sessions in a quiet room, using headphones or earphones.

Participants logged into an interactive website built using jsPsych [14] to complete the training sessions. The combination of participant ID and session number determined the stimuli used during each training session. Only 3 out of 4 voices were used per session (e.g., voices A, C, D). Similarly, only one set of 10 minimal pairs was used per training session. All voices and stimuli sets were equally represented in the training overall.

Each training session consisted of two mini-games: a two-alternative forced choice (2AFC) identification task and an ABX discrimination task. In order to allow participants to have access to an acoustically varied set of clear examples of /r/ and /l/, both mini-games were preceded by a familiarisation phase in which all stimuli from the session's set of ten minimal pairs could be heard, accompanied by their corresponding transcriptions. For each minimal pair in the set, participants were sequentially presented (ISI = 1000 ms), for each of the three voices, the /r/ item (e.g., /rapu/) with its transcription on the left (e.g., *rapu*), followed by the /l/ item (e.g., /lapu/) with its transcription on the right (e.g., *lapu*). The sides of presentation corresponded to the location of the R and L keys on the keyboard. In total, participants heard 60 stimuli in each familiarisation phase (3 speakers ×10 pairs ×2 pair items).

The 2AFC identification task was presented as a boxing game in which the participant's avatar boxer successfully hit the opponent each time the participant accurately responded to an identification trial. At each trial, participants listened to a stimulus and were then provided two possible transcriptions on-screen (/r/-item on the left, /l/-item on the right, following the position of the response keys on a Japanese keyboard). After choosing a transcription using the R and L keys, participants were given feedback and the stimulus was played again. The 30 trials (3 speakers ×10 items, one per minimal pair, with half being /r/-items and half /l/-items) were randomly assigned to one of the three boxing rounds composing the mini-game. A round was won if the number of trials answered correctly was equal or higher than the number of trials answered incorrectly within the round. Participants were deemed to have won the mini-game if they won two out of three boxing rounds.

The ABX discrimination task was presented as an ice fishing game. Participants were asked to help a polar bear avatar catch fish hiding in one of two holes in the ice. The location of the fish was indicated to the participant as follows: Participants sequentially heard three stimuli A, B, and X (ISI = 500 ms). A and B were /r/- and /l/-items (or vice versa), and X was either an /r/- or /l/-item. Participants were then asked to identify whether the third item (X) was a token of the same type as the first item

Figure 1: *Schematics of the identification (left) and discrimination (right) mini-games in a training session. From top to bottom: familiarisation, trial stimulus listening and participant response input, trial feedback, mini-game summary feedback, and overall session feedback.*

(A; choose ice hole on the left) or the second item (B; choose ice hole on the right). Participants answered using the arrow keys on the keyboard. They were provided feedback and listened to the three stimuli again, now with the corresponding transcriptions on-screen. There were a total of 40 trials, as there were 4 trials for each of the 10 minimal pairs. These four trials corresponded to various permutations: of the three speakers used in the training session, one was always used for X, with the other two speakers being used for A or B (two possibilities for A-B: *spk1-spk2* or *spk2-spk1*); also, A was either an /l/- or /r/-item, and vice versa for B (two possibilities for A-B: *R-L* or *L-R*). Item X was an /r/- (or /l/-item) for half of the trials, similarly being the same item as A (or B) half of the time. Participants were deemed to have cleared the mini-game if they had accurately responded to at least 25 out of the 40 trials. At the end of the two mini-games, participants were provided a final score, as shown in Figure 1.

2.4. Data analysis

Statistical analyses were performed with the R statistical software (version 3.4.4) [15], using Markov Chain Monte Carlo generalised linear mixed-effects models (MCMC glmm) [16, 17]. These Bayesian models sample coefficients from the posterior probability distribution conditioned on the data and given priors (in our case, parameter-expanded priors). Effects were considered statistically significant if the 95% highest posterior density (HPD) interval estimated for the explanatory variable of interest did not include zero.

In order to assess the effect of training on /r/-/l/ identification accuracy, we fitted a model with CORRECT as the binomial response variable. As fixed effects we included TEST TYPE (pre-test *vs* post-test), NOISE (noiseless *vs* noisy), SENTENCE (isolated item *vs* sentence-embedded item), CONSONANT (/l/ *vs* /r/), CLUSTER (simple *vs* complex), POSITION (onset *vs* coda). All variables were categorical with two levels, and were contrast coded using deviance coding. We also included the interaction between TEST TYPE and all other fixed effects, as well as PARTICIPANT and MINIMAL PAIR as random effects.

3. Results

3.1. Main effect of training

We found a significant main effect of TEST TYPE (posterior mode: 17.3, HPD interval: $[12.0, 23.4]$), indicating that, relative to their pre-training performance, participants were globally more accurate at identifying the correct transcriptions of items containing /r/ and /l/ after training (Figure 2). Importantly, participants gave a median rating of 4 out of a maximum of 5 to training (mean rating: 3.7), with regards to how entertaining they found it to be; 54% of them stating that they would be willing to play the game again.

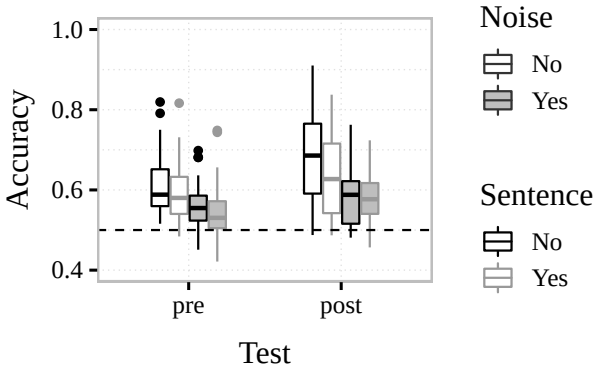

Figure 2: *Identification accuracy at pre- and post-training tests. Stimuli were either noiseless (white fill) or noisy (grey fill), with items being presented in isolation (black outline) or in carrier sentences (grey outline). Boxplots display the distribution of the scores across participants (median, quartiles and extrema).*

3.2. Item-intrinsic variation

We examined whether participants experienced more difficulty identifying correct transcriptions due to certain characteristics that were intrinsic to the items used as experimental stimuli.

We found a significant main effect of CONSONANT (posterior mode: -9.4, HPD interval: $[-14.3, -5.1]$), indicating that participants were generally less accurate at identifying /r/-items than /l/-items. The interaction between TEST TYPE and CONSONANT was significant (posterior mode: 19.1, HPD interval: $[11.5, 30.4]$); participants' identification of /r/-items appears to have benefited more from the training interval than their identification of /l/-items.

Similarly, we found a significant main effect of CLUSTER

(posterior mode: -21.9, HPD interval: $[-34.0, -8.9]$), suggesting that participants encountered more difficulties identifying phonemes when they were part of a complex cluster (e.g., /gloʃu/) that when they were not (e.g., /lona/). The interaction between TEST TYPE and CLUSTER was not significant (posterior mode: 0.6, HPD interval: $[-7.3, 10.2]$); we did not find any evidence supporting that training benefited one type of cluster (e.g., simple) more than the other (e.g., complex).

While numerically, participants were generally more accurate at identifying the target phonemes when these were positioned in the syllable coda, as opposed to the syllable onset, we did not find this difference to be significant (POSITION; posterior mode: 2.2, HPD interval: $[-9.6, 16.7]$). The interaction of POSITION with TEST TYPE was not significant either (posterior mode: 1.9, HPD interval: $[-6.7, 13.2]$).

3.3. Item-extrinsic variation

Next, we examined how modifications that are extrinsic to the items affected identification accuracy. Namely, we focused on noise (e.g., background noise, channel distortions) and sentence embedding, as listeners often encounter speech in less-than-optimal conditions when outside of the laboratory.

We found a significant main effect of NOISE (posterior mode: -24.0, HPD interval: $[-31.4, -18.9]$), indicating that participants were generally less accurate at identifying noisy stimuli than noiseless stimuli (Figure 2). The interaction between TEST TYPE and NOISE was marginally significant (posterior mode: -9.6, HPD interval: $[-18.2, 0.0]$); training seems to have been more beneficial for the identification of noiseless stimuli than for that of noisy stimuli.

Similarly, we found a significant main effect of SENTENCE (posterior mode: -7.1, HPD interval: $[-11.9, -3.0]$), indicating that participants were generally less accurate at identifying items embedded in sentences than when they were items presented in isolation (Figure 2). The interaction between TEST TYPE and SENTENCE was not significant (posterior mode: -2.4, HPD interval: $[-11.1, 6.4]$); we did not find any evidence suggesting that training benefited more identification of isolated or embedded items.

4. Discussion

In this preliminary work, we prototyped a web-based phonetic training game, with the goal of improving perception of the English /r/-/l/ contrast by Japanese learners of English with real-life application in mind. Participants completed 12 sessions of approximately $10 - 15$ minutes of phonetic training in a $3 - 5$ week interval (total: $2 - 3$ hours). Our main finding was that our web-based, self-paced training was effective at improving participants' identification of the /r/-/l/ contrast, akin to previously reported phonetic training studies on the same contrast [3, 4, 5, 6, 11, 12, 7, 8, 9, 10]. It is unclear, however, if the magnitude of the improvement may have been dampened due to participants doing training at home, instead of under the more strict experimental conditions proper to laboratory training. For instance, participants in [11] averaged around 80% accuracy at post-test identification. In contrast, in our study, participants achieved a post-test mean accuracy of 68% for noiseless isolated stimuli. This difference could be due to the training environment, but also to differences in length of training, number of training sessions, and even the composition of the stimuli used for training and testing. Indeed, we corroborated previous findings that all phonetic environments are not created equal: Participants were less accurate at identifying /r/ and /l/ in clusters and (numerically) in onset position. This is reminiscent of previous results where initial clusters and intervocalic phonemes were found to be difficult for Japanese listeners, contrary to singletons in coda position, which were easier [4, 5]. Increased difficulty with consonant clusters could be in part due to these phoneme structures not being allowed by phonotactic constraints of Japanese, which affects their correct perception [18]. Additionally, we found that even though participants were generally more prone to errors when identifying /r/-items than /l/-items, training proved to be more beneficial for the former. This is in contrast with what was observed by [8], but in agreement with [19], who claimed English /l/ to be more strongly assimilated into the Japanese /r/ category than English /r/, making English /r/ easier to learn for Japanese listeners.

Aside from examining the effect of phonetic training on characteristics that are intrinsic to the items (target phoneme, position, ...), we examined pre-training perception and post-training improvement for items set in more challenging, yet more realistic conditions. Namely, we assessed identification accuracy for items in degraded speech (i.e., added noise or channel distortions) and items embedded in sentences. As may have been expected, identification accuracy was lower for items in noise and/or in sentences. Unlike the greater improvement seen for /r/-items in spite of them being more challenging than /l/-items in general, we did not see greater post-training improvement for the more difficult noisy and/or sentence-embedded stimuli relative to their noiseless, isolated counterparts. Since degraded perception in noise is pervasive in non-native listeners [20, 21, 22], future work should aim to specifically target these sub-par conditions for perception in phonetic training programs, especially insofar as they represent more ecological, real-life environments. This is also true for sentence-embedding, where previous work outside of the field of L2 learning has shown that training with sentence-embedded stimuli may prove beneficial for learning to correctly perceive both embedded and isolated items after training [23, 24, 25].

5. Conclusions

In this work we prototyped a phonetic training game to enhance /r/-/l/ perception by Japanese learners of English. Contrary to most previous work on phonetic training, our aim was to build a tool optimised to be applicable to real-life learning: short, gamified training sessions were completed at home on a dedicated website, using variable equipment, in a self-paced manner. The training was effective at improving /r/-/l/ identification, yet the effect was smaller than in previous studies. Future modifications of the training program include increasing the number of training sessions, adding noisy and/or sentence-embedded tokens during training in order to target these challenging cases, and exploring the effects of high phonetic variability provided naturally (i.e., various speakers) or artificially (i.e., by automatically modifying recordings from a single speaker).

6. Acknowledgements

This research was funded by the Japan Society for the Promotion of Science through a Postdoctoral Fellowship for Research in Japan (Standard) and a KAKENHI grant (18F18724) given to International Research Fellow A. Guevara-Rukoz. The authors would like to thank the English speakers for recording the stimuli, and Vincent Dubois for assistance with website preparations. Special thanks go to all participants of the study.

7. References

[1] H. Goto, "Auditory perception by normal Japanese adults of the sounds L and R," *Neuropsychologia*, vol. 9, no. 3, pp. 317–323, 1971.

[2] P. Iverson, P. K. Kuhl, R. Akahane-Yamada, E. Diesch, Y. Tohkura, A. Kettermann, and C. Siebert, "A perceptual interference account of acquisition difficulties for non-native phonemes," *Cognition*, vol. 87, no. 1, pp. B47–B57, 2003.

[3] W. Strange and S. Dittmann, "Effects of discrimination training on the perception of /r-l/ by Japanese adults learning english," *Perception & psychophysics*, vol. 36, no. 2, pp. 131–145, 1984.

[4] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *The Journal of the Acoustical Society of America*, vol. 89, no. 2, pp. 874–886, 1991.

[5] 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," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1242–1255, 1993.

[6] 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," *The Journal of the acoustical society of America*, vol. 96, no. 4, pp. 2076–2087, 1994.

[7] B. D. McCandliss, J. A. Fiez, A. Protopapas, M. Conway, and J. L. McClelland, "Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 2, no. 2, pp. 89–108, 2002.

[8] 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," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3267–3278, 2005.

[9] E. M. Ingvalson, L. L. Holt, and J. L. McClelland, "Can native Japanese listeners learn to differentiate /r/-/l/ on the basis of F3 onset frequency?" *Bilingualism: Language and Cognition*, vol. 15, no. 2, pp. 255–274, 2012.

[10] Y. Shinohara and P. Iverson, "High variability identification and discrimination training for Japanese speakers learning English /r/-/l/," *Journal of Phonetics*, vol. 66, pp. 242–251, 2018.

[11] 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," *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2299–2310, 1997.

[12] 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, no. 5, pp. 977–985, 1999.

[13] H. Zhang, Y. Inoue, D. Saito, N. Minematsu, and Y. Yamauchi, "Computer-aided High Variability Phonetic Training to improve robustness of learners' listening comprehension," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, 2019.

[14] J. R. De Leeuw, "jsPsych: A Javascript library for creating behavioral experiments in a Web browser," *Behavior research methods*, vol. 47, no. 1, pp. 1–12, 2015.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[16] J. D. Hadfield, "MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package," *Journal of Statistical Software*, vol. 33, no. 2, pp. 1–22, 2010. [Online]. Available: http://www.jstatsoft.org/v33/i02/

[17] M. Plummer, N. Best, K. Cowles, and K. Vines, "CODA: Convergence diagnosis and output analysis for MCMC," *R News*, vol. 6, no. 1, pp. 7–11, 2006. [Online]. Available: https://journal.r-project.org/archive/

[18] E. Dupoux, K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler, "Epenthetic vowels in Japanese: A perceptual illusion?" *Journal of experimental psychology: human perception and performance*, vol. 25, no. 6, p. 1568, 1999.

[19] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada, and T. Yamada, "Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/," *Journal of Phonetics*, vol. 32, no. 2, pp. 233–250, 2004.

[20] Y. Takata and A. K. Nábělek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 663–666, 1990.

[21] V. Hazan and A. Simpson, "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects," *Language and Speech*, vol. 43, no. 3, pp. 273–294, 2000.

[22] D. Tabri, K. M. S. A. Chacra, and T. Pring, "Speech perception in noise by monolingual, bilingual and trilingual listeners," *International Journal of Language & Communication Disorders*, pp. 1–12, 2015.

[23] S. L. Greenspan, H. C. Nusbaum, and D. B. Pisoni, "Perceptual learning of synthetic speech produced by rule." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, no. 3, p. 421, 1988.

[24] Y. Hirata, "Training native english speakers to perceive Japanese length contrasts in word versus sentence contexts," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2384–2394, 2004.

[25] P. C. Stacey and A. Q. Summerfield, "Comparison of word-, sentence-, and phoneme-based training strategies in improving the perception of spectrally distorted speech," *Journal of Speech, Language, and Hearing Research*, 2008.