Acoust. Sci. & Tech. **28**, 4 (2007)

PAPER

Identification of English /r/ and /l/ in noise: The effects of baseline performance

Kazuo Ueda[1,*], Reiko Akahane-Yamada[2,†], Ryo Komaki[2,‡] and Takahiro Adachi[2,§]

[1]*Department of Applied Information and Communication Sciences, Kyushu University, 4–9–1, Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan*
[2]*ATR-Promotions/ATR Cognitive Information Science Laboratories, 2–2–2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0288 Japan*

(*Received 3 July 2006, Accepted for publication 6 November 2006*)

Abstract: The effects of baseline performance on identification of English /r/ and /l/ in noise by Japanese listeners were examined. Japanese and American listeners' perception of /r/ and /l/ was measured under various signal-to-noise ratios. Each Japanese listener's baseline performance was also observed with an identification test with a large set of stimuli. Generally, the signal-to-noise ratio had a similar effect across Japanese listeners of different baseline performances, although the overall accuracy was dominated by the baseline. When fifteen-days of /r/–/l/ identification training in quiet was applied to a group of Japanese listeners, the training effect generalized to identification performance in noise. However, the performance of Japanese listeners did not reach the level that the native listeners exhibited.

Keywords: Consonant identification, Signal-to-noise ratio, Word perception, Training

PACS number: 43.71.Es, 43.71.Gv, 43.71.Hw, 43.66.Dc [doi:10.1250/ast.28.251]

1. INTRODUCTION

Learners of a second language usually have difficulty in identifying phonemic contrasts that do not occur in their native languages [1–3]. It is extremely difficult to identify English /r/ and /l/ for Japanese-native listeners [2,4–9], probably because these phonemes are perceptually assimilated to one phonemic category of the Japanese language [8,10–12]. On the other hand, Japanese-native listeners can be trained to attain more than 90% accuracy in identifying the /r/ vs. /l/ contrast, when they hear these segments in quiet [13–16]. The attempt to clarify the effects of noise on the identification has just made a beginning [17]. The current article further examines identification of English /r/ and /l/ by native and non-native (Japanese) listeners in noise, focusing on baseline performance and training.

Miller and Nicely [18] investigated the intelligibility of English consonants for native listeners in consonant-vowel (CV) syllables presented in noise, and reported that the extent of noise tolerance was different from consonant to consonant (however /r/ and /l/ were excluded from the syllables). More sophisticated experiments by Benkí [19] using CVC syllables confirmed the same tendency of confusion observed by Miller and Nicely. Benkí included /r/ and /l/ in his experiment, and found that, at the onset of a syllable, /l/ was less confused with other consonants than was /r/. Adachi, Akahane-Yamada, and Ueda [20] examined the effects of the presentation level of the stimulus, signal-to-noise ratios, and the type of noise (white noise and pink noise), on consonant identification by native and non-native listeners. It was shown that a /b/ vs. /v/ contrast was more fragile than an /r/ vs. /l/ contrast or a /s/ vs. /th/ contrast, both for native and non-native listeners.

Other variables are also found to be effective for speech perception in noise. The accuracy of word recognition depended on the predictability provided by the sentence that embedded the word [21–23]. The familiarity of words is another variable that has been shown to affect word recognition in noise [24]. The amount of masking effect on speech sounds depends on the kind of noise masker [25–27]. Garcia Lecumberri and Cooke [28] found that the ranking of masking effectiveness across masking conditions was the same for the native and non-native listeners, with competing speech being the least effective masker, next speech shaped noise, and babble being the most

*e-mail: ueda@design.kyushu-u.ac.jp
†e-mail: yamada@atr.jp
‡e-mail: komaki@atr.jp
§e-mail: tadachi@atr.jp

effective. Some researchers [29–31], Simpson and Cooke among others, were concerned with the masking effect of the babble noise as a function of the number of voices (N) in the babble, and found that the function was non-monotonic: Increasing N made masking more effective up to $N = 6$, however, further increments of voices up to $N = 128$ did not change the amount of masking.

Non-native listeners find it harder to hear speech sounds in noise and in reverberation than native listeners do [28,32–42]. The non-native listeners seemed to have a disadvantage in utilizing contextual cues in sentences. This recent research seems to emphasize the importance of the predictability given by a carrying sentence. However, it does not mean that phoneme recognition is unimportant. Contextual cues conveyed through a carrying sentence originate from some easily detectable words, and the evoked cues involve previous knowledge that relates to a target word. However, if one cannot detect any words in a sentence, it is obvious that one cannot form a context from the sentence. Besides, a carrying sentence is not always useful in providing meaningful context that helps listeners with guessing a target word. Experimental evidence has shown that contextual cues had a clear blocking effect on developing phoneme identification by non-native listeners [43–46]: the listeners learned phoneme identification better when the training was done in an isolated word (without context) condition than in a meaningful context condition. Although the participants learned how to guess a target word from contextual cues, they did not develop phoneme identification skills as tested in an isolated word condition when they were trained in the meaningful context condition.

Other studies have observed disadvantages in consonant perception for non-native listeners. Adding reverberation or babble noise more greatly impairs non-native listeners' perception of English consonants than that of native listeners [32,35]. Consonant-cue-enhancement was effective for improving intelligibility in noise for both native (English) and non-native (Spanish and Japanese) listeners; however, the size of the effect was smaller for the non-native listeners [39].

Cutler *et al.* [47] examined English phoneme identification by native and non-native (Dutch) listeners. Their study provided evidence of the parallel effects of noise masking on the performance of these two groups of listeners, which means, the difficulty of phoneme identification in noise by non-native listeners, as well as by native listeners, is proportional to the amount of noise.

Ueda, Akahane-Yamada, and Komaki [17] investigated American-English consonant perception in white noise by native and non-native (Japanese) listeners with systematically varied signal-to-noise ratios. However, there were two variables found to be revisited. The first one is the individual differences observed in baseline identification performance. A total of 17 Japanese-native (J) listeners participated in Ueda *et al.* However, the size of samples was not enough to analyze the effect of baseline performance differences, given that there are substantial individual differences observed in accuracy of perception of /r/ and /l/ in J listeners [9]. Thus, an experiment with a larger number of J listeners is needed to clarify the effect of baseline performance. The second one is the range of signal-to-noise ratios. Ueda *et al.* used a range of signal-to-noise ratios that encompassed down to -12 dB, however, the range was too narrow to compare the performance between J listeners and American English-native (AE) listeners: AE listeners were far more robust against noise compared to J listeners, and the performance of AE listeners deteriorated only a little (about 10%) even when the signal-to-noise ratio was -12 dB.

The purpose of our present investigation was to examine the effect of basic identification performance level on English /r/ and /l/ perception by native and non-native listeners with a larger number of J listeners (Experiment 1). To attain this purpose, we extended the range of signal-to-noise ratios down to -21 dB, and collected identification data with 70 J listeners. We administered a screening test to control the level of performance of J listeners. Moreover, we tried to assess the effect of training on identification of /r/ and /l/ in noise by J listeners in Experiment 2. A portion of the data presented in this article was reported in our previous works [48,49].

2. EXPERIMENT 1

This experiment consisted of two phases: a screening-test phase and an identification-test phase. The screening test was to select non-native listeners for the identification test. The identification test was to assess the effect of signal-to-noise ratio on performance of identification of /r/ and /l/ by native and non-native listeners.

2.1. Screening Test

2.1.1. Method

2.1.1.1. Listeners

Eighty-eight (34 females and 54 males) Japanese-native (J) listeners, aged from 19 to 30 with an average age of 22, with normal hearing—tested with an audiometer (either Dana, DA-301 or Rion, AA-77)—, participated in the test. All the listeners reported that they had no experience of living abroad for more than three months.

2.1.1.2. Stimuli

Twenty-five minimal pairs of English words contrasted in /r/ and /l/, i.e., 50 words, were selected. Two native speakers of American English (a male and a female) produced each word. A total of 100 stimuli were recorded

using 16-bit amplitude quantization and a 22.05-kHz sampling rate.

2.1.1.3. Procedure

Notebook computers (either IBM, ThinkPad 600 or IBM, ThinkPad A30p) were used to control auditory stimulus presentation, to present minimal pair words on the screen, and to record the listeners' responses. The stimuli were diotically presented to the listeners through an adapter (STAX, SRM-1/MK2 pp) and headphones (either STAX, SR Lambda Signature or STAX, SR-303) in a soundproof booth (for the 12 ATR J listeners including six females and six males, a booth specially constructed in ATR, and for the rest of the J listeners, Music Cabin, SC-3). The stimulus presentation level was fixed at a comfortable listening level. The sound pressure level (SPL) was measured at the headphones' output with an IEC coupler (Brüel & Kjær, type 4153) and a precision sound level meter (Brüel & Kjær, either type 2231 or type 2260). The meter operated in its "Fast" and "A curve" setting. The maximum level of each stimulus was about 70 dB SPL on average.

As soon as two members of a minimal pair appeared in English orthography side by side on the computer screen, one of the members was played over the headphone. The listeners' task was to identify a word they have heard, and to indicate their judgments by clicking an appropriate word on the screen with a mouse pointer. No feedback was provided on their responses.

2.1.2. Results

The number of correct responses out of 100 trials constituted the screening test score for each listener. Therefore, the highest possible score was 100. Table 1 shows the frequency distribution of the test scores. Both the mode and the mean of the distribution fall in the range from 60–69.

Table 1 Frequency distribution and means of the screening-test scores for the Japanese-native (J) listeners in Experiment 1. The highest possible score was 100. In the table, "Original" means the original population of listeners who participated in the screening test, and "Selected" means the selected listeners who participated in the identification test in noise.

Range of the scores	Frequency		Mean scores	
	Original	Selected	Original	Selected
40–59	18	15	54.89	54.20
60–69	38	32	63.71	63.66
70–79	20	14	72.90	73.14
80–100	12	9	87.00	86.67
	$N = 88$	$N = 70$	$M = 67.17$ $SD = 10.51$	$M = 66.49$ $SD = 10.55$

2.2. Identification Test in Noise

2.2.1. Method

2.2.1.1. Listeners

Fourteen American English-native (AE) listeners (6 females and 8 males, aged from 21 to 51 with an average age of 31.93) and 70 Japanese-native (J) listeners (32 females and 38 males) participated. The J listeners were selected from the listeners who participated in the screening test. The selection was made to reduce the size imbalances across the score ranges (Table 1).

2.2.1.2. Stimuli

Fifty-three English word pairs minimally contrasting in /r/ and /l/, i.e., 106 RL words, were used as materials. Each pair was produced by one of four speakers of American English (2 males and 2 females). Speakers A (male) and B (female) produced 15 minimal pairs, speaker C (male) produced 14 pairs, and speaker D (female) produced nine pairs. The recording conditions were the same as in the screening test. The maximum SPL (Fast, A) at the headphones' output was recorded for each stimulus, and then the average was taken. The amplitude of speech stimuli was shifted so that the average maximum SPL became 64.0 dB SPL (A).

Additionally, ten filler (FL) words contrasting /d/ vs. /k/, /d/ vs. /n/, /h/ vs. /s/, /m/ vs. /n/, and /m/ vs. /p/, produced by a male English-native speaker, were used as stimuli. These words were originally recorded using 12-bit amplitude quantization and a 10-kHz sampling rate [13]. The recorded files were digitally transferred by using 16-bit amplitude quantization and a 22.05-kHz sampling rate.

White noise was added to these speech signals to control signal-to-noise ratios. The white noise was produced with a noise generator (Brüel & Kjær, type 1049). The noise was stored in a computer and the SPL of the noise at the headphones' output was measured in the same way as for the speech sound. The amplitude of the noise was adjusted so that the signal-to-noise ratios ranged from +9 to −21 dB with five steps of 6 dB. The noise was added to the signal so that it started 200 ms earlier and lasted 200 ms longer than the signal.

2.2.1.3. Conditions

The following three variables were controlled in the experiment: (1) the signal-to-noise ratios yielded seven levels including a no-noise (quiet) condition, (2) the contrasted consonants, the /r/ and /l/ (RL) and the fillers (FL), yielded two levels, and (3) the native languages of the listeners, American English vs. Japanese, yielded two levels. The first and the second factors are within subjects factors. The FL condition also worked as a warming-up and a practice condition.

2.2.1.4. Procedure

For the AE listeners, notebook computers (IBM, ThinkPad 600X) or IBM PC/AT compatible machines

with SoundBlaster 64 Gold card were used for stimulus presentation and response acquisition. All the AE listeners were tested in a booth specially constructed in ATR.

The trials were divided into blocks by the signal-to-noise ratios. The order of signal-to-noise ratio blocks was randomized across listeners. In each signal-to-noise ratio block, 10 FL trials were given before the 106 RL trials. Within each signal-to-noise ratio block, the RL stimuli were further divided into sub-blocks by speakers. The order of the speaker sub-blocks was fixed throughout the experiment, but the order of words within each speaker block was randomized across the listeners.

The method of stimulus presentation and response acquisition was the same as in the screening test. No feedback was provided on listeners' responses.

2.2.2. Results and discussion

Correct response rate for each stimulus and listener was calculated and pooled across listener groups. Figure 1 shows the averaged results for each native-language group; Fig. 2 shows the subdivided version of Fig. 1(b), i.e., J listeners' results divided for each rank of screening-test performance.

Identification accuracy by the AE listeners was nearly 100% in the no-noise condition, whereas it was 70% at a signal-to-noise ratio of −21 dB in Fig. 1(a). The J listeners showed similar identification accuracy to the AE listeners for the FL words, whereas they performed much poorer on the /r/ vs. /l/ contrasted words (70% accuracy without noise and 55% accuracy with a signal-to-noise ratio of −21 dB, Fig. 1(b)). These correspondences of performance between native and non-native listeners were able to be observed by virtue of widening the range of signal-to-noise ratios and a larger number of J listeners, compared to those used in our previous investigation [17].

In Fig. 2, the performance curves of RL identification by J listeners are ordered in accordance with the screening-test ranks. The RL curves of J listeners seem to gradually approach the RL curve of AE listeners in accordance with screening-test performance [from Fig. 2(a) to (d)]. The RL curve of the highest performance group [Fig. 2(d)] is closest to the curve of the AE listeners.

To see the overall effects, the arc-sine-transformed data was submitted to analysis of variance (ANOVA) with three-way factorial design where signal-to-noise ratio (SNR) and consonants (C, i.e., FL or RL) were within subject variables, and native language of the participants (NL) was between subject variable. The effect of the subjects was specified as a random effect nested by the native language factor. The dependent variable was weighted according to the number of words, i.e., trials, in each condition. The correction was made for the unbalanced number of words in the following six categories: five types of RL contrasts, i.e., initial singleton, initial cluster,

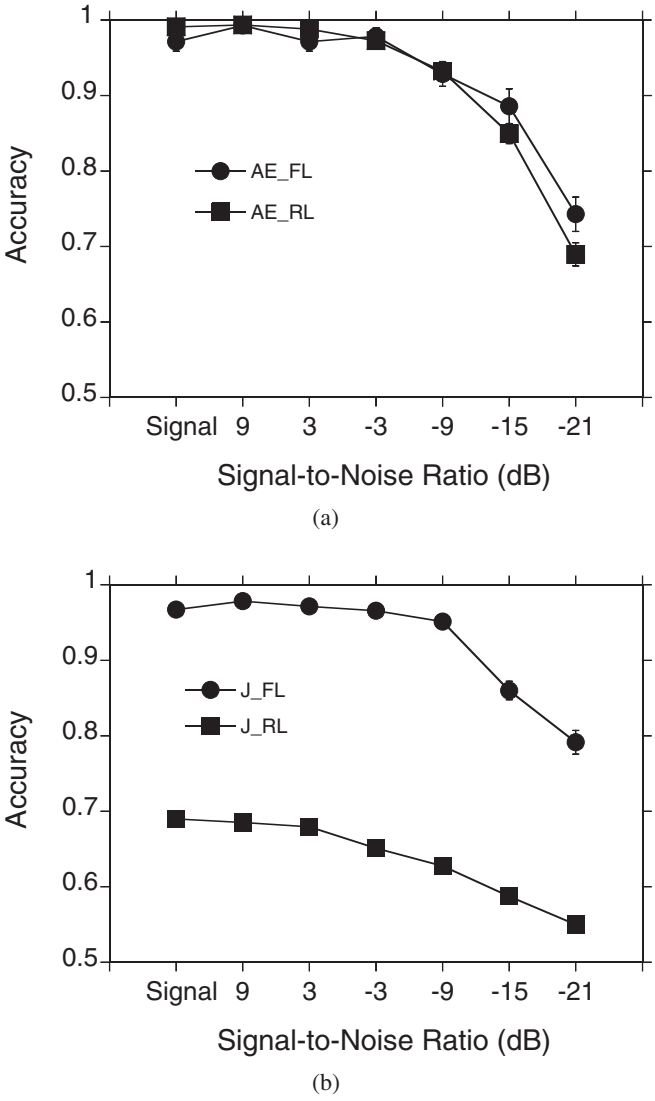


Fig. 1 Accuracy (± s.e.m.) of identification presented as a function of signal-to-noise ratios. FL: Filler word pairs that contrasted /d/ and /k/, /d/ and /n/, /h/ and /s/, /m/ and /n/, and /m/ and /p/. RL: /r/ and /l/ contrasted word pairs. (a) The accuracy given by American-English-native (AE) listeners ($N = 14$), and (b) by Japanese-native (J) listeners ($N = 70$).

medial, final cluster, and final singleton (20, 24, 16, 24, and 22 words, respectively), and FL contrast (10 words). The main effects of SNR, NL, and C, and the interaction effects of NL × C and SNR × NL × C were statistically significant [$F(6, 3418) = 69.60$, $p < 0.0001$; $F(1, 82) = 40.81$, $p < 0.0001$; $F(1, 3418) = 276.54$, $p < 0.0001$; $F(1, 3418) = 342.32$, $p < 0.0001$; $F(6, 3418) = 2.88$, $p < 0.0083$, respectively], whereas the interaction effect of SNR × C was not statistically significant.

The subdivided results for J listeners were submitted again to ANOVA. The main effects of SNR, C, and rank of screening-test performance (R), and the interaction effects of SNR × C were significant [$F(6, 2818) = 52.35$, $p < 0.0001$; $F(1, 2818) = 1258.33$, $p < 0.0001$; $F(3, 66) =$

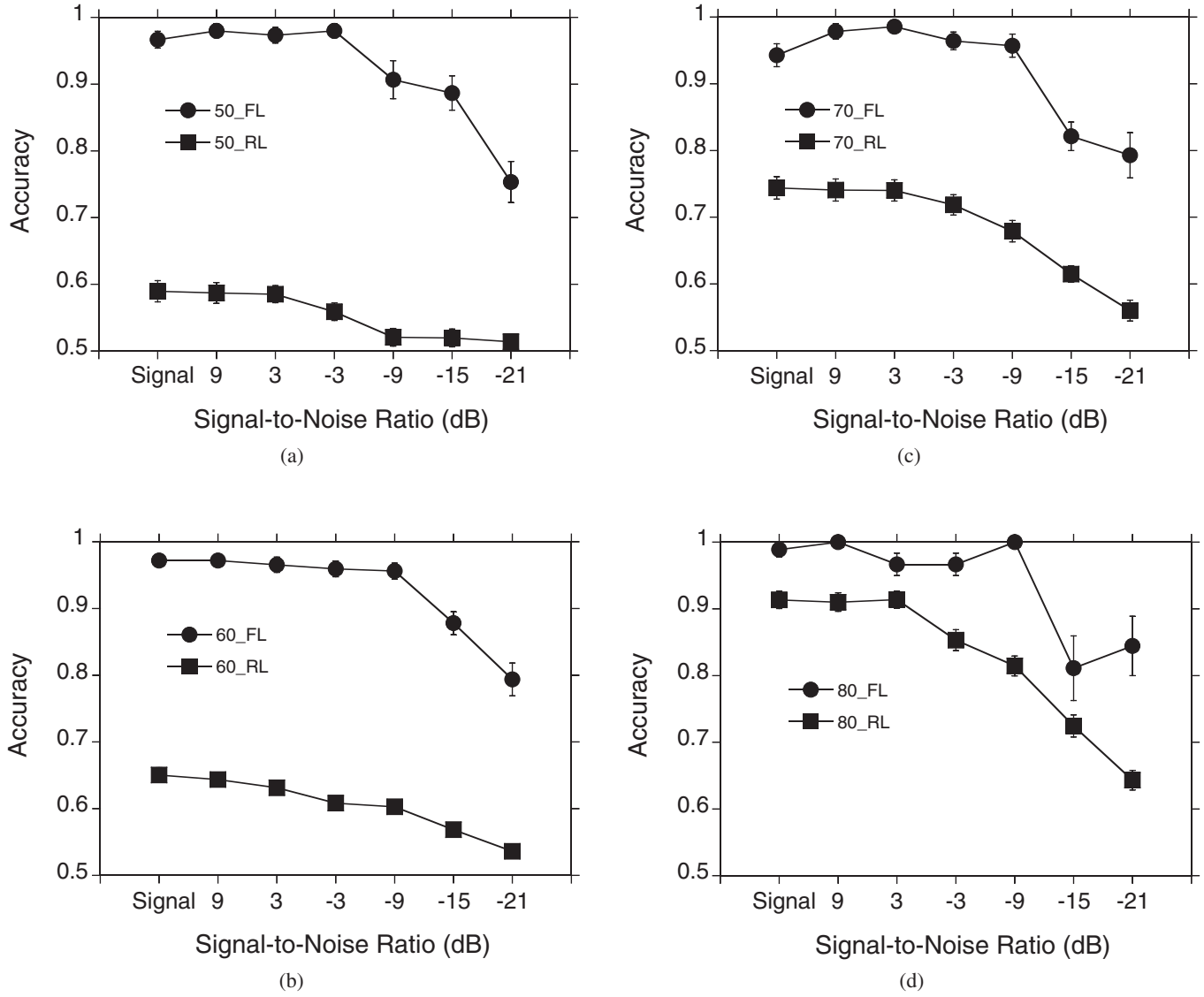


Fig. 2 Accuracy (± s.e.m.) given by J listeners, shown for each rank of screening-test performance by the listeners who scored, (a) less than 60 ($N = 15$), (b) between 60 and 69 ($N = 32$), (c) between 70 and 79 ($N = 14$), and (d) more than 80 ($N = 9$).

21.04, $p < 0.0001$; $F(6, 2818) = 2.99$, $p < 0.0065$, respectively]. The interaction effects of SNR × R and SNR × C × R were not significant.

The effect of native language is clearly seen in these results: J listeners showed lower performance in the RL contrast, whereas AE listeners maintained relatively and constantly high performance both for the RL and FL contrasts. The performance for RL identification by J listeners approached that for FL identification according to the rank of the screening test performance; however, the RL curve did not completely reach the level of the FL curve. Perceptual assimilation to the phonemic categories of one's native language [1,10] probably plays some role in lowering the performance for the RL contrast in noise as well as in quiet.

3. EXPERIMENT 2

In Experiment 1, the RL curve of the high-scored J listeners who scored more than 80 in the screening test [Table 1 and Fig. 2(d)] came close to the curve of the AE listeners. However, only 12 listeners out of 88 listeners who participated in the screening test were assigned in this category. Accordingly, the obtained curve looks less smooth compared to those obtained with the other classes of listeners, probably due to an insufficient number of listeners. Thus, we decided to train J listeners, whose screening-test scores were relatively low, so as to obtain more listeners in this high-scored group, and to observe the effect of training on identification of /r/ and /l/ in noise.

Some of the J listeners in Experiment 1 continued on

Table 2 Frequency distribution and means of the screening-test scores for the J listeners selected for Experiment 2.

Range of the scores	Frequency	Mean scores
50–59	3	54.67
60–69	13	63.15
	$N = 16$	$M = 61.56$ $SD = 4.03$

to participate in Experiment 2. The procedure of Experiment 2 consists of six steps: (1) administration of a screening test, (2) selecting listeners based on their performances in the screening test, (3) randomly assigning the listeners to an experimental group and a control group, (4) administration of a pretest, (5) training the experimental group and resting the control group, and (6) administration of a post-test. The effect of training was assessed with comparison of identification performance exhibited in the pretest and the post-test. Actually, the first and the fourth steps had been already done in Experiment 1. Stimuli, conditions, and the procedure of the identification test in noise are also identical to the corresponding elements in Experiment 1. Therefore, concerning the experimental procedure, the other parts of the procedure are described in this section.

3.1. Method

3.1.1. Listeners

Sixteen J listeners were selected out of the J listeners who participated in Experiment 1. They were selected because their screening-test scores were relatively poor, i.e., less than 70. Their individual performances in the screening test are summarized in Table 2.

3.1.2. Procedure

The listeners were randomly divided into two groups of eight participants: an experimental group and a control group. The pretest described in Experiment 1 was administered to both groups.

Listeners in the experimental group were trained to identify /r/ and /l/ for 15 days within three to four weeks. On each day, training lasted approximately two hours for each listener. The training was scheduled not to have three successive untrained days. The equipment was the same as in Experiment 1. The program developed by Yamada, Adachi, and ATR [50] was used for the training. It contains 576 training trials with feedback on a listener's responses. The materials were pronounced by four American-native speakers (2 females and 2 males). The training program was run twice a day for each listener and the ratio of accurate identification was recorded every time when each speaker-sub-block was finished. Immediately after all the trainings were completed, the same test of identification as in the previous section was administered as a post-test.

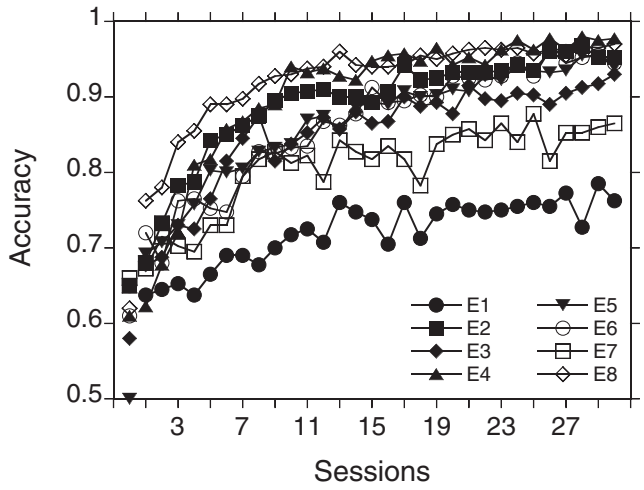


Fig. 3 Learning curves for each participants (E1–E8) in the experimental group. Screening test scores converted to accuracy were plotted at the left end.

Listener in the control group received a post-test three to four weeks after a pretest. No experimental trials were administered between the two tests.

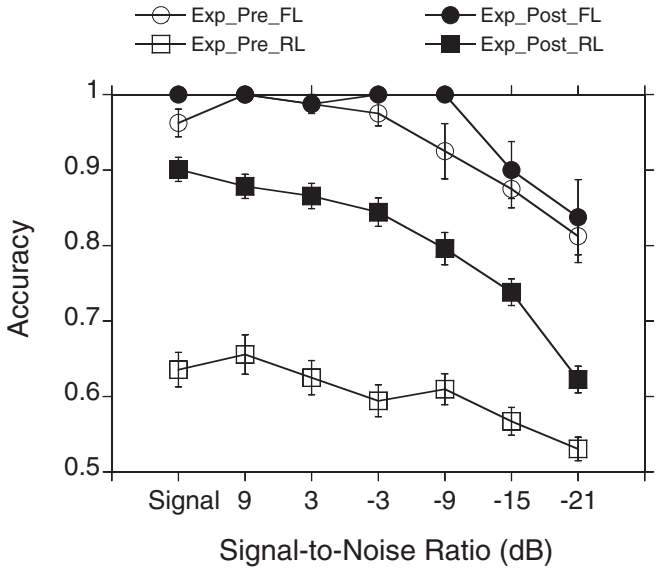
3.2. Results and Discussion

The results are shown in Figs. 3 and 4. The effect of training is clearly visible, although the improved performance at RL identification in no-noise conditions still fell short of the performance obtained in the comparable FL condition (Fig. 4).

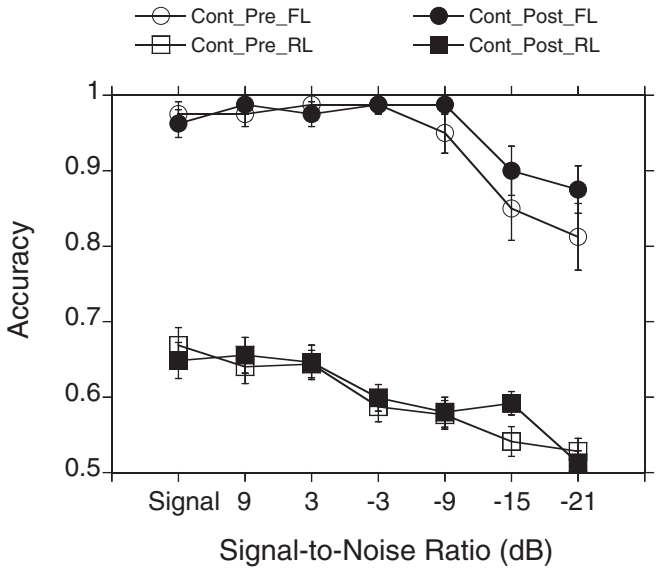
The arc-sine-transformed data was submitted to ANOVA. The accuracy ratios were weighted according to the number of words (trials) as in Experiment 1. The analysis revealed that the main effects of training (T), pretest vs. post test (PP), SNR, and C, and the interaction effects of T × PP, T × C, PP × C, and T × PP × C were significant [$F(1, 1288) = 23.53$, $p < 0.0001$; $F(1, 1288) = 35.75$, $p < 0.0001$; $F(6, 1288) = 19.15$, $p < 0.0001$; $F(1, 1288) = 745.84$, $p < 0.0001$; $F(1, 1288) = 22.15$, $p < 0.0001$; $F(1, 1288) = 19.25$, $p < 0.0001$; $F(1, 1288) = 11.68$, $p < 0.0007$; $F(1, 1288) = 16.75$, $p < 0.0001$, respectively]. The other effects were not significant.

4. GENERAL DISCUSSION

A substantial difference in the performance of identification of /r/ and /l/ between the native listeners and the non-native listeners was observed in Experiment 1: the native listeners showed higher and more robust performance in the noisy conditions, whereas the non-native listeners, on average, showed much lower performance even in the quiet condition. However, there observed no clear evidence that the signal-to-noise ratio affected the identification accuracy in an entirely different way among



(a)



(b)

Fig. 4 Accuracy of identification as a function of signal-to-noise ratios and training by Japanese-native participants. Pre: pretest. Post: post-test. FL and RL abbreviate the same as in Fig. 1. (a) The experimental group ($N = 8$), and (b) the control group ($N = 8$).

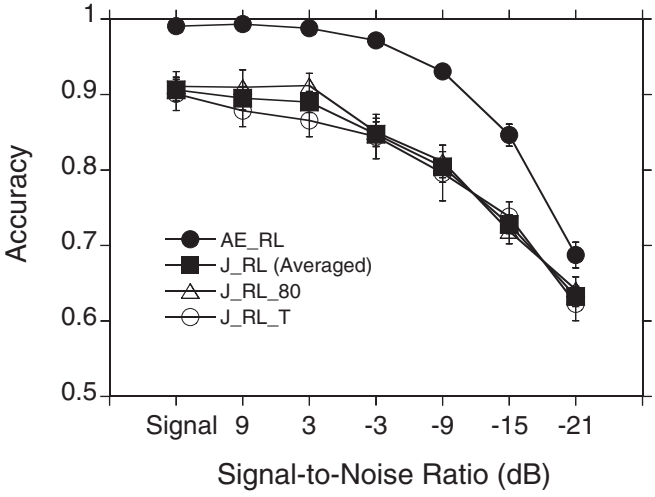


Fig. 5 Accuracy of identification of /r/ and /l/ as a function of signal-to-noise ratios and participant groups. The participant groups are consisted of the AE listeners ($N = 14$), the J listeners who scored more than 80 in the screening test ($N = 9$), and the J listeners who were trained in the Experiment 2 ($N = 8$).

the listener groups, i.e., native and non-native listeners, and high and low baseline performance groups within non-native listeners.

It has previously been confirmed with well-controlled laboratory experiments that the effect of training on English /r/ and /l/ listening by non-native listeners is generalized to other words, other talkers, and speech production [14,15]. The effect of perceptual training on perception and production was retained for at least three months [16], and possibly more.

Experiment 2 in the present research revealed that the effect of training on perceptual identification was generalized also to the other words and the other talkers in the noisy conditions. However, performance deteriorated when the signal-to-noise ratios decreased. No obvious interaction effect was observed between the two factors, the signal-to-noise ratios and the consonants, in Experiment 2.

The following analysis focuses on the performance difference in RL identification between the native listeners and the non-native listeners (Fig. 5). Firstly, we analyzed the performance of the J listeners who scored more than 80 in the screening test in Experiment 1 and the performance of the J listeners who had finished the training in Experiment 2. Analysis of variance revealed that the main effect of signal-to-noise ratios (SNR) was significant [$F(6, 566) = 92.13$, $p < 0.0001$], whereas the main effect of participant groups (PG) and the interaction effect of SNR $\times$ PG were not significant. Therefore, we merged the data from those two groups in the following analysis. We call the merged listeners "high-performance J listeners."

The two curves of the AE listeners and the high-performance J listeners in Fig. 5 seem to converge towards the right-end of the graph. The analysis of variance confirmed this impression: the main effect of SNR and native language of the participants (NL) and the interaction effect of SNR $\times$ NL were significant [$F(6, 1042) = 264.45$, $p < 0.0001$; $F(1, 29) = 62.89$, $p < 0.0001$; $F(6, 1042) = 6.30$, $p < 0.0001$, respectively].

We would like to suggest two possible explanations for the interaction effect of SNR $\times$ NL. One possibility is that saturation of the performance of AE listeners at high SNR conditions caused the significant interaction with the performance by high-performance J listeners (a ceiling

effect hypothesis). Another possibility is that the high-performance J listeners, who trained in a laboratory or in the other places, somehow acquired to detect cues that are not used by the native listeners but are weakly correlated with the RL contrast, and are robust in noisy conditions (a compensational-cue-acquisition hypothesis). The following assumptions are included in this hypothesis: (1) such cues actually exist, (2) the high-performance J listeners succeeded in acquiring the cues, and (3) the high-performance J listeners acquired the correct cues along with the *wrong* cues, but learning of the correct cues was not complete. This hypothesis may explain the fact that the high-performance J listeners exhibited lower accuracy compared with the AE listeners, and that the right half of the accuracy curve of the AE listeners has a steeper slope compared with the high-performance J listeners. At present, we cannot decide which hypothesis is true.

Cutler *et al.* [47] found parallel effects of noise masking for native (English) and non-native (Dutch) listeners. They suggested that the increasing difficulty in speech perception in a noisy environment experienced by non-native listeners mainly comes from ineffective use of contextual cues rather than inaccurate and fragile phoneme recognition. It is hard to decide whether the results of Cutler *et al.* contradict our findings; Cutler *et al.* drew their conclusion from identification performance averaged over all possible American English CV and VC syllables, whereas the present authors found the interaction effect with /r/ and /l/ minimal pairs, which are extremely difficult to identify for Japanese-native listeners. There is a possibility that these two studies examined different parts of a curve, and therefore the slopes appeared differently.

Further work is needed to clarify the relationship between phoneme recognition and contextual understanding in non-native listeners.

ACKNOWLEDGEMENTS

We would like to thank Noriko Yamasaki, Rei Jitsuda, and Yoshinori Yamamura for their assistance in running the experiments for the Japanese-native listeners, and Jonathan Goodacre for his helpful comments on the draft. This work was supported in part by Grant-in-Aid for the 21st Century COE program and Grant-in-Aid for Scientific Research Nos. 14101001 and 17202012 from the Japan Society for the Promotion of Science.

REFERENCES

[1] L. Polka, "Linguistic influences in adult perception of non-native vowel contrasts," *J. Acoust. Soc. Am.*, **97**, 1286–1296 (1995).

[2] J. C. L. Ingram and S.-G. Park, "Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners," *J. Acoust. Soc. Am.*, **103**, 1161–1174 (1998).

[3] W. Strange, R. Akahane-Yamada, R. Kubo, S. A. Trent and K. Nishi, "Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners," *J. Acoust. Soc. Am.*, **109**, 1691–1794 (2001).

[4] H. Goto, "Auditory perception by normal Japanese adults of the sounds "L" and "R"," *Neuropsychologia*, **9**, 317–323 (1971).

[5] K. Miyawaki, W. Strange, R. Verbrugge, A. M. Liberman, J. Jenkins and O. Fujimura, "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," *Percept. Psychophys.*, **18**, 331–340 (1975).

[6] K. S. MacKain, C. T. Best and W. Strange, "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl. Psycholinguist.*, **2**, 369–390 (1981).

[7] M. Mochizuki, "The identification of /r/ and /l/ in natural and synthesized speech," *J. Phonet.*, **9**, 283–303 (1981).

[8] C. T. Best and W. Strange, "Effects of phonological and phonetic factors on cross-language perception on approximants," *J. Phonet.*, **20**, 305–330 (1992).

[9] R. A. Yamada and Y. Tohkura, "The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners," *Percept. Psychophys.*, **52**, 376–392 (1992).

[10] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience*, W. Strange, Ed. (York Press, Timonium, Md., 1995), pp. 171–206.

[11] C. T. Best, G. W. McRoberts and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.*, **109**, 775–794 (2001).

[12] R. Komaki, R. Akahane-Yamada and Y. Choi, "Effects of native language on the perception of American English /r/ and /l/: Cross-language comparison between Korean and Japanese," *Tech. Rep. IEICE*, SP99-45, pp. 39–46 (1999).

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

[14] R. Akahane-Yamada, "Learning non-native speech contrasts: What laboratory training studies tell us," *Proc. 3rd Jt. Meet. Acoust. Soc. Am. and Acoust. Soc. Jpn.*, Honolulu, Hawaii, pp. 953–958 (1996).

[15] 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," *J. Acoust. Soc. Am.*, **101**, 2299–2310 (1997).

[16] 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," *Percept. Psychophys.*, **61**, 977–985 (1999).

[17] K. Ueda, R. Akahane-Yamada and R. Komaki, "Identification of English /r/ and /l/ in white noise by native and non-native listeners," *Acoust. Sci. & Tech.*, **23**, 336–338 (2002).

[18] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, **27**, 338–352 (1955).

[19] J. R. Benkí, "Analysis of English nonsense syllable recognition in noise," *Phonetica*, **60**, 129–157 (2003).

[20] T. Adachi, R. Akahane-Yamada and K. Ueda, "Intelligibility of English phonemes in noise for native and non-native listeners," *Acoust. Sci. & Tech.*, **27**, 285–289 (2006).

[21] G. A. Miller, G. A. Heise and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *J. Exp. Psychol.*, **41**, 329–335 (1951).

[22] G. A. Miller, "Decision units in the perception of speech," *IRE Trans. Inf. Theory*, **IT-8**, 81–83 (1962).

[23] J. R. Duffy and T. G. Giolas, "Sentence intelligibility as a function of key word selection," *J. Speech Hear. Res.*, **17**, 631–637 (1974).
[24] I. Pollack, H. Rubenstein and L. Decker, "Intelligibility of known and unknown message sets," *J. Acoust. Soc. Am.*, **31**, 273–279 (1959).
[25] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, **88**, 1725–1736 (1990).
[26] J. M. Festen, "Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice," *J. Acoust. Soc. Am.*, **94**, 1295–1300 (1993).
[27] C. Liu and D. Kewley-Port, "Formant discrimination in noise for isolated vowels," *J. Acoust. Soc. Am.*, **116**, 3119–3129 (2004).
[28] M. L. Garcia Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," *J. Acoust. Soc. Am.*, **119**, 2445–2454 (2006).
[29] R. Carhart, C. Johnson and J. Goodman, "Perceptual masking of spondees by combinations of talkers," *J. Acoust. Soc. Am.*, **58**, S35 (1975).
[30] R. L. Freyman, U. Balakrishnan and K. S. Helfer, "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," *J. Acoust. Soc. Am.*, **115**, 2246–2256 (2004).
[31] S. A. Simpson and M. Cooke, "Consonant identification in *N*-talker babble is a nonmonotonic function of *N* (L)," *J. Acoust. Soc. Am.*, **118**, 2775–2778 (2005).
[32] A. K. Nábělek and A. M. Donahue, "Perception of consonants in reverberation by native and non-native listeners," *J. Acoust. Soc. Am.*, **75**, 632–634 (1984).
[33] S. Buus, M. Florentine, B. Schalf and G. Canevet, "Native, French listeners' perception of American English in noise," *Proc. Inter-Noise '86*, pp. 895–898 (1986).
[34] M. Florentine, "Speech perception in noise by fluent non-native listeners," *Trans. Tech. Comm. Psychol. Physiol. Acoust.*, H-85-16 (1985).
[35] Y. Takata and A. K. Nábělek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *J. Acoust. Soc. Am.*, **88**, 663–666 (1990).
[36] T. Hatoh, S. Kuwano, E. Costigan and S. Namba, "Listening ability of non-native language under the presence of noise," *J. Acoust. Soc. Jpn.* (*J*), **54**, 695–703 (1998).
[37] D. N. Kalikow, K. N. Stevens and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.*, **61**, 1337–1351 (1977).
[38] L. H. Mayo, M. Florentine and S. Buus, "Age of second-language acquisition and perception of speech in noise," *J. Speech Lang. Hear. Res.*, **40**, 686–693 (1997).
[39] V. Hazan and A. Simpson, "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects," *Lang. Speech*, **43**, 273–294 (2000).
[40] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *J. Acoust. Soc. Am.*, **112**, 272–284 (2002).
[41] S. J. van Wijngaarden, H. J. M. Steeneken and T. Houtgast, "Quantifying the intelligibility of speech in noise for non-native listeners," *J. Acoust. Soc. Am.*, **111**, 1906–1916 (2002).
[42] S. J. van Wijngaarden, A. W. Bronkhorst, T. Houtgast and H. J. M. Steeneken, "Using the speech transmission index for predicting non-native speech intelligibility," *J. Acoust. Soc. Am.*, **115**, 1281–1291 (2004).
[43] R. Akahane-Yamada and Y. Ikuma, "Effects of acoustic and semantic contexts on the perceptual learning of phonemes in second language," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 393–394 (2003).
[44] Y. Ikuma and R. Akahane-Yamada, "An empirical study on the effects of acoustic and semantic contexts on perceptual learning of L2 phonemes," *Annu. Rev. Engl. Lang. Educ. Jpn.*, **15**, 101–108 (2004).
[45] Y. Ikuma and R. Akahane-Yamada, "Effects of acoustic and semantic contexts on the perceptual learning of speech segments in a second language: Experimental evidence from the extended laboratory training study," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 429–430 (2004).
[46] Y. Ikuma and R. Akahane-Yamada, "Effects of acoustic and semantic contexts in learning L2 phoneme perception," *Tech. Rep. IEICE*, SP2004-35 (2004).
[47] A. Cutler, A. Weber, R. Smits and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *J. Acoust. Soc. Am.*, **116**, 3668–3678 (2004).
[48] K. Ueda, Y. Nakajima and R. Akahane-Yamada, "Speech perception in a noisy environment: The implication of auditory scene analysis," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 507–510 (2004).
[49] K. Ueda, Y. Nakajima and R. Akahane-Yamada, "An artificial environment is often a noisy environment: Auditory scene analysis and speech perception in noise," *J. Physiol. Anthropol. Appl. Hum. Sci.*, **24**, 129–133 (2005).
[50] T. Yamada, T. Adachi and ATR, *Scientific Inquiry: How to Improve English Listening Skills*, Blue Backs (Kodansha, Tokyo, 1998).