TR-H- 212 0002

Effects of Extended Training on English /r/ and /l/ Identification by Native Speakers of Japanese.

Reiko AKAHANE-YAMADA, Yoh'ichi TOHKURA, Scott E. LIVELY (Indiana Univ.), Ann R. BRADLOW (Indiana Univ.) and David B. PISONI (Indiana Univ.)

1997.3.26

ATR人間情報通信研究所
〒619-02　京都府相楽群精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
Telephone: +81-774-95-1011
Fax　　: +81-774-95-1008

Effects of extended training on English /r/ and /l/ identification by native speakers of Japanese

Reiko Akahane-Yamada [1]
Yoh'ichi Tohkura
ATR Human Information Processing Research Laboratories,
2-2, Hikaridai, Seika, Soraku, Kyoto, 619-02 Japan

Scott E. Lively[2]
Ann R. Bradlow [3]
David B. Pisoni
Indiana University, Bloomington, IN 47408

[1] e-mail: yamada@hip.atr.co.jp

[2] Current affiliation: Customer Interface and Human Factors Division, Ameritech, Hoffman Estates, Illinois

[3] Current affiliation: Auditory Neuroscience Laboratory, Northwestern University, Illinois

1 Abstract

Adult foreign-language learners often have remarkable difficulty in learning the distinction of certain phonetic categories which do not occur in their native language. Recent studies have revealed that the appropriate laboratory training can improve one's perceptual ability significantly even for the most difficult categories. This paper studied effects from the amount of laboratory training by expanding the study of Lively et al. (1994). Lively and his colleagues trained native speakers of Japanese to identify American English /r/ and /l/ in 15 training sessions covering 15 days (i.e. 1 session per day). In our first experiment, we examined the effect of massed versus distributed training on the training result. Japanese speakers were trained in 15 training sessions covering 5 days (3 sessions per day). Their accuracy in an identification test improved from 70% in the pre-test to 80% in the post-test, whereas it improved from 65% to 77% in Lively et al.(1994). No siginificant difference in training effect was found in terms of the number of sessions per day between both studies. In our second experiment, Japanese speakers were trained in 45 training sessions covering 15 days. An additional 30 trials were found to significantly improve the subjects' ability to identify /r/ and /l/: The accuracy improved from 70% in the pretest to 83% after 15 sessions, and to a further 87% and 89% by 15 and 30 additional sessions, respectively. These results suggest that the amount of training compensated for the difficulty of developing proper internal representations for the new phonetic categories. Methodological implications for the training of phonetic contrasts in a second language will be discussed.

2 Introduction

Human speech perception and production are modified into a language-specific ones during the course of development (aging). Humans acquire the phonological system of their first language (L1) without difficulty. However, it is not always easy for them to develop the phonetic system of another language after once establishing this L1 phonetic system. When one learns a non-native language as a second language (L2), one's L2 perception is strongly affected by the L1 system (Best, 1995; Flege, 1981). Therefore, in order to perceive and produce L2 sounds adequately, he/she has to overcome the effect of L1 by learning, either through daily exposure to the L2-speaking environment or through intensive training. The age of the learner also affects the learning by interacting with the L1 effect: In general, younger learners have less difficulty (Flege, 1995; Yamada, 1995). Thus, adult learners sometimes have difficulty in learning new phonetic categories that do not occur in their native language; some of such categories are extremely difficult. For example, native speakers of Japanese have remarkable difficulty in perceiving and producing the English /r/-/l/ contrast (e.g. Goto, 1971; Miyawaki et al., 1975; Yamada et al., 1994; Yamada, 1995).

Aren't adult learners of a foreign language able to develop novel phonetic categories? Training studies of the past decade that tried to train adult subjects to acquire the perception of non-native speech contrasts, demonstrated that laboratory training can improve this perceptual ability, if the adequate training paradigms are applied. Then, what types of training paradigms are appropriate and effective for developing non-native phonetic categories? This question has been a long-standing issue in speech perception studies, but has yet to be clarified. The answer should be of important theoretical value, and the new techniques and training methods from it for developing new categories should offer the possibility of studying the larger issue of speech perception mechanisms through laboratory training experiments. Hence, many laboratory training studies have explored this issue; some of them have been successful, while others have met with only limited success (Strange 1995 and Jamieson 1995 for reviews). Such training studies have used various training methods, providing a variety of knowledge for understanding the effects of training. At one time, however, the difference in training methods made comparing studies a difficult task, because the trainings differed in multiple variables which deeply interacted with each other.

Here, we propose five types of variables for consideration: 1) target distinction, 2) stimulus material, 3) task, 4) stimulus sequence, and 5) amount of training.

2.1 Target Distinction

Previous studies had subjects trained to perceive different distinctions using different populations of the subjects: Examples include Thompson's ejective velar and uvular distinction (/ki/-/qi/) on English speakers (Werker and Tees, 1984); Hindi retroflex and dental distinction on English speakers and on Japanese speakers (Pruitt, 1995); voiced and voiceless distinction between English /θ/ and /ð/ on francophones (Jamieson & Morosan, 1986); English /r/-/l/ distinction on Japanese speakers (Strange, & Dittmann, 1984; Logan et al., 1991; Lively et al., 1994); distinction between moraic structures of Japanese vowels on English speakers (T.Yamada et al., 1994); distinction of English word-final /t/-/d/ on Mandarin speakers (Flege, 1989), etc.

2.2 Stimulus Material

The acoustic variations of the training stimuli differed among the studies: Some studies used synthetic stimuli (e.g., Strange & Dittmann, 1984; Jamieson & Morosan, 1986), while others used natural tokens (e.g., Tees & Werker, 1984; Logan et al., 1991; Pruitt, 1995). The former studies focused on the important cues for their distinction, but the stimulus set was less varied in acoustic properties than the latter studies. In addition to this difference (i.e. synthetic tokens vs. natural tokens), they differed in the number of talkers (e.g. productions by a single talker vs. multiple talkers), and in the number of phonetic environments of the target phoneme (e.g. target in a single phonetic environment vs. in multiple phonetic environments; Logan et al., 1991; Lively et al., 1992).

2.3 Task

Previous studies used different tasks in the training. Some studies used a discrimination task (e.g., Strange & Dittmann, 1984), while others used an identification task (e.g. Jamieson & Morosan, 1986; Lively et al., 1994).

2.4 Stimulus Sequence

The structure of the stimulus sequence varied among the studies. Some of the studies used a fading technique in which the training started with a stimulus set consisting of a small number of easy stimuli. Then, depending on the performance of the trainee, the number of stimuli in the training stimulus set was possibly increased and more difficult stimuli were introduced (e.g. Jamieson & Morosan, 1986; Pruitt, 1994). Other studies, in contrast, did not change the stimulus set throughout the training (e.g. Logan et al, 1991; Lively et al., 1993; T.Yamada et al., 1994).

2.5 Amount of Training

The amount of training differed among the studies. The number of total trials considerably varied across all studies as well as the number of training days. For example, Strange & Dittmann (1984) had 4212 trials over 18 days, Jamieson & Morosan (1986) had 720-1920 trials over 3 days, Logan et al. (1991) and Lively et al. (1994) had 4080 trials over 15 days, T.Yamada et al. (1994) had 2160 trials over 8 days, etc.

It is important to discuss the effects of these variables for optimizing the training method. However, the above results cannot be compared directly, because the variables interacted with each other. Instead, subsets of those studies having similar characteristics in some variables should be compared.

Regarding the 1st variable of the target distinction, the differences both in the pretest level and in the difficulty of learning among the studies should be taken into account. One of the important comparisons is Lively et al. (1994)

vs. T.Yamada et al. (1994). These two studies used identical identification tasks by using the same software (called RLtrainer and developed at ATR Human Information Processing Research Lab., Japan) and similar training stimulus sets (natural tokens varying in a phonetic environment produced by multiple talkers). Lively et al. trained Japanese speakers to identify English /r/ and /l/, and the identification score of the subjects improved from 65.1% in the pretest to 77.3% in the post-test through 4080 trials of training. In contrast, T.Yamada et al. trained English speakers to perceive the distinction between a Japanese short vowel, long vowel and moraic germinata (short vowel plus germinata of the following consonant), and the identification score improved from 65.4% in the pretest to 89.7% in the post-test only through 2160 trials of training. The pretest scores of the subjects in these 2 studies are almost equal, but English speakers (moraic structure training of Japanese vowels) improve much faster than Japanese speakers (training of the /r/-/l/ distinction) [24.3% improvement over 2160 trials in T.Yamada et al.; 12.2% improvement over 4080 trials in Lively et al.]. From this comparison, it is obvious that the training effect differs in terms of the target distinction to be trained.

Furthermore, Logan et al. (1991) trained English /r/-/l/ using an almost identical training method to Lively et al.'s (1994): They trained Japanese speakers using the same identification task and stimuli as were used by Lively et al. However, the improvement was less in Logan et al. than in Lively et al. [5% vs. 12%]. The main differences between these two studies were the performance level at the pretest level [78% in Logan et al.; 65% in Lively et al.], and the subjects' language environment [living in an English-speaking environment for Logan et al.; and living in a Japanese-speaking environment for Lively et al.].

In addition, Pruitt (1995) trained the dental vs. retroflex distinction on both English-speaking and Japanese-speaking subjects using the same training. The English speakers showed a lower pretest accuracy and less improvement in post-test over pretest compared with the Japanese speakers [53.34% to 67.57% for the English speakers; 62.34% to 83.75% for the Japanese speakers]. These results suggest that the initial performance level of the subjects and their language background, i.e. the L1 system, do affect the effectiveness of the training.

Regarding the 2nd and 3rd variables, i.e. the stimuli and task, most of the training studies emplying a discrimination task of a single synthetic continuum met with limited success. For example, and as an important comparison, Strange & Dittmann (1984) met with less success than Lively et al (1994). Strange & Dittmann trained Japanese speakers on English /r/-/l/ using a synthetic "Rock"-"Lock" continuum. Whereas the training was effective on the training material, the improvement did not transfer to the non-training material, either synthetic stimuli or natural tokens: The accuracy for identification of natural tokens improved from 64.1% to 69.5%, but this difference of 5.4% was not significant. In contrast, as was introduced already, Lively et al. used an identification task in the training with natural tokens produced by multiple talkers. The effects of the training were examined by comparing the pretest and post-test performance in an identification test of /r/-/l/ minimal pairs, which Strange and Dittmann used, produced by a talker who was not used in the training. The accuracy improved significantly from 65.1% in the pretest to 77.3% in the post-test. They also demonstrated that their training effect transferred into new words (not used in the training) produced by one of the trained talkers and by the new talker. Although these two studies trained the same target distinction, with similar amounts of training [4212 trials over 18 days in Strange & Dittmann; 4080 trials over 15 days in Lively et al.], and similar initial performance levels [64.1% vs. 65.1%], the training effect was much more limited in Strange & Dittmann than in Lively et al., strongly suggesting that identification training using natural tokens is more effective than discrimination training using synthetic tokens.

Flege (1991) directly compared the effect of identification training and discrimination training. He trained Mandarin speakers to perceive the English word-final /t/-/d/ contrast. The effect of identification training transferred to a non-trained talker better than with discrimination training. This result is a piece of straightforward evidence showing that identification training is more effective than discrimination training.

A direct comparison of stimulus sets was conducted by Lively et al. (1993). They trained Japanese speakers to identify English /r/ and /l/ with the same technique as in Logan et al. (1991) and in Lively et al. (1994), but using training stimuli produced by a single talker. The subjects improved less from the pretest to post-test than in the other studies which used training stimuli produced by five talkers. Magnuson et al. (1995) expanded this study by training five groups of subjects, where each was trained by one of the five different talkers, while Lively et al. (1993) used only one talker as a training talker. Again, they trained Japanese speakers on English /r/-/l/ distinction, using the same training technique and same stimulus corpus as in Lively et al (1994). They found that the effectiveness of the single talker training differed depending on the talker; single talker training by some talkers failed to transfer the training effect to other talkers, while the single talker training by some other talkers succeeded to transfer the effect to other talkers. These results all together suggest that identification training using natural tokens produced by multiple talkers has a lower probability of failing in promoting an L2 phonetic contrast.

Unfortunately, there is no subset of studies we can use to compare with or discuss about the 4th and 5th variables. Regarding the 4th variable of the stimulus sequence, two studies, which used a fading technique with an identification task, both met with success [Jamieson & Morosan (1986), 68.44% to 79.38% with 2 days of training; Pruitt (1995), 62.34% to 83.75% with 12 days of training], implying that this technique might be effective. However, some trainings which did not use the fading technique have been successful, too, suggesting that the fading technique is not the only successful way. Regarding the 5th variable, there has been no study examining the effect of the amount of training. Since the improvement differs remarkably between the target distinctions, we need to compare the amount of training within a single target distinction.

As reviewed above, training studies in the past decade have gained an insight into the effects of laboratory training in various ways. On the other hand, there is as yet no systematic approach for studying the above-mentioned variables, even though such an approach is important for optimizing the training method. In particular, the fifth variable, the amount of training, must be studied soon, not only because it has not been addressed directly at all, but also because we need to determine whether the amount of training compensates for the difficulty of learning or not. The tentative goal of training studies should be to obtain a training method that is highly effective even for extremely difficult distinctions. Therefore, we need to address the question of what compensates the difficulty. From this standpoint, the amount of training is an important candidate which may compensate the difficulty. We should examine whether a larger amount of training helps to promote more robust perception of an L2 contrast, which is remarkably hard to learn, to the extent of easier distinction in a short period of time.

Thus, in this paper, we decided to examine the effect of the amount of training on /r/-/l/ identification for Japanese speakers. Since the /r/-/l/ distinction is one of the extremely difficult cases for Japanese speakers, and has been well-studied both in perception and in training studies, it is a good choice for conducting a systematic approach toward the optimization of the effectiveness of laboratory training for non-native speech sounds. As described in the section on the 1st variable (i.e. target distinction), English /r/-/l/ distinction for Japanese speakers improved from 65.1% to 77.3% over 4080 trials (Lively et al., 1994), wheras the distinction of Japanese moraic structures for English speakers improved from 65.4% to 89.7% only over 2160 trials (T.Yamada et al., 1994). Does the perception performance of the Japanese speakers further improve if we continue training beyond 4080 trials? The specific question we address in this paper is to determine whether extended /r/-/l/ training further improves the ability to the extent of the easier distinction of Japanese moraic structures achieved by the English speakers, i.e. around the 90% level, or the improvement reaches a ceiling and stays lower than 90% even after a larger amount of training trials.

In order to make a systematic comparison with the previous studies, we examined the effect of extended training by expanding the study of Lively et al.(1994). We ran subjects on 3 times larger training trials, and compared the results to those in Lively et al (1994). Lively et al. trained subjects for 1 session per day over 15 days. We extended the number of training trials by running the subjects over 3 times more sessions in one day. In other words, we arranged trials in the rather massed practice style (45 sessions / 15 days), compared to Lively et al's distributed (spaced) practice style (15 sessions / 15 days). In this design, there are 2 variables which may affect the result: The effect of extended number of trials, and the spacing effect between sessions (i.e. massed vs. distributed training).

Generally, learning under the distributed practice is superior to massed practice (e.g. Hintzman et al., 1973[9]). This implies the possibility that our new arrangement of sessions (3 sessions per day) slow down the learning compared to the Lively et al. (1 session per day). In contrast, the spacing effects (the effect of massed vs. distributed arrangement of trials) have been obtained with limited tasks under a limited experimental situations. Since there is no report which examined about this spacing effect in the identification training of novel speech contrasts, it is necessary to determine the effect of massed vs. distributed practice in our /r/-/l/ identification training.

In the experiment 1, we examined the effect of massed vs. distributed training by comparing the 15 training sessions oevr 5 days and Lively et al.(1994)'s results (same 15 training sessions over 15 days). In experiment

2, we examined the effect of extended training by running subjects over 45 training sessions over 15 days.

3 Experiment 1: massed vs. distributed training

3.1 Method

3.1.1 Subjects

Thirty-seven native speakers of Japanese with no experience living abroad served as subjects. All of them reported no history of hearing or speech disorder. A hearing screening performed at 15dB HL for frequencies between 250 and 8000 Hz showed all subjects to have normal bilateral hearing acuity.

Nineteen subjects were randomly selected and assigned to SES1 group, whose results were reported in Lively et al., 1994. The other 18 subjects were assigned to another SES3 group. The subjects in the SES1 group ranged from 18 to 22 years of age (average 20); and 18 to 22 in the SES3 group (average 19.5).

All of the subjects in this paper had been studying English since junior high school (12 years old), and none of them had received any special English conversation lessons. Since stress is put mainly on the acquisition of grammatical competence in most English classes at Japanese high schools and universities, all of the subjects had little experience in English conversation.

3.1.2 Stimuli

The stimuli were identical to those used in Lively et al. (1994). English words contrasting /r/ and /l/ were used as the stimulus material. These words contrasted /r/ and /l/ in one of 5 positions; word-initial singleton, word-initial consonant cluster, intervocalic, word-final singleton, and word-final consonant cluster. These words were produced by native speakers of American English at Indiana University. The recordings were low-pass filtered at 4.8 kHz and digitized at a 10 kHz sampling frequency with 12-bit resolution. These speech files were then transferred to ATR HIP laboratories, where they were up-sampled to 22.05 kHz and rescaled to 16-bit resolution.

In the training, 136 words (68 /r/-/l/ minimal pairs; 13 initial singleton pairs, 24 initial cluster pairs, 5 intervocalic pairs, 15 final singleton pairs, and 11 final cluster pairs) produced by 5 talkers, T1-T5 (T1, T3 and T5 were male, and T2 and T4 were female), were used.

In the pretest and post-test, 24 minimal pairs used by Strange and Dittmann (1984) were used. Sixteen pairs contrasted /r/ and /l/ (4 in the initial singleton, 4 in the initial cluster, 4 in the intervocalic, and 4 final singleton). The other 8 pairs were filler pairs which contrasted phonemes other than /r/ and /l/. These words were produced by a new male talker, who was not one of the 5 training talkers, twice. As a result, 96 stimuli (24 pairs × 2 words × 2 times repetition) were used as the pretest and post-test stimuli.

In the generalization tests, 2 sets of stimuli were used. The first set consisted of 99 words (38 words had /r/ or /l/ in the initial singleton, 32 in the initial cluster, 11 in the intervocalic, 11 in the final cluster, and 15 in the final singleton) produced by one of the female training talkers. The second set consisted of 96 words (37 words had /r/ or /l/ in the initial singleton, 29 in the initial cluster, 4 in the intervocalic, 8 in the final cluster, and 18 in the final singleton) produced by a new male talker, who was different from any ot the other training and test talkers.

3.1.3 Procedure

The experimental design employed a pretest–post-test design used by Strange and Dittmann (1984). The pretest and post-test were administered before and after the training period, and generalizations to the new words and new talker were done after the training period.

On the first day of the experiment, the subjects received the pretest, and the hearing screening test. From the 2nd day, the subjects were trained to identify /r/ and /l/ in 15 sessions. The subjects in the SES1 group (Lively, et al., 1994) received 1 session of training per day, while those in the SES3 group received 3 sessions of training per day. As a result, the subjects in the SES1 group were trained over 15 days, and those in the SES3 group were trained over 5 days. On the final day, which was a separate day from the last training day, the subjects participated in the post-test, and 2 generalization tests. In these generalization tests, transfer to new words by an old talker and to new words by a new talker were tested.

All of the experiments in this paper were done at ATR HIP laboratories in Kyoto, Japan. Each subject sat in front of a CRT monitor and keyboard in a sound-proof chamber. The stimuli were presented binaurally over headphones, Stax SR Lambda Signature, at a fixed and comfortable level. The experiments were self-paced and presentations of the stimuli and data collections were controlled by a workstation, NeXT cube.

Tests In the tests (pretest, post-test, and generalization tests), 2 alternative forced choice tasks were used. In each trial, 2 members of a minimal pair were displayed on 2 buttons shown on the CRT monitor (Fig. 1). After presenting these choices for 500ms, one of the members was played over the headphones. The subjects identified the word they had heard and chose one of the alternative words by pressing a key corresponding to a button on the screen; press "1" for the left button, and "2" for the right button. The position of the /r/ word (word containing /r/) and /l/ word (word containing /l/) were randomized and counter balanced in a way that /r/ word appeared on the right side in about half of the trials, and the /l/ word appeared on the right side in the remaining trials. There was no feedback for the subjects' responses, and the next trial began after 2 s ITI.

Training In the training, the same task used in the tests was used. However, there was a feedback for the subjects' responses. If the subject responded correctly, a chime sounded and the next trial started after 2 s ITI. If the subject responded incorrectly, a buzzer sounded and a correction trial started after 2 s ITI. In the correction trial, the same stimulus as in the previous trial followed by a sound feedback (chime/buzzer) to the subject's response. This correction trial looped until the subject made a correct response. The accumulative accuracy rate was always displayed on the CRT monitor (Fig. 1). The responses in the correction trials were not counted in the accuracy. In addition, there was a graphical representation of coins. One

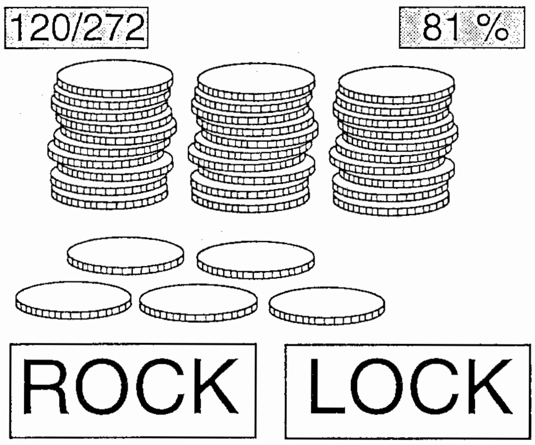

Figure 1: An example display of the CRT monitor during a training session. The response alternatives are written on the two buttons at the bottom, the accumulative correct response rate is shown in the right-top corner, the trial number is shown in the left-top corner, and the graphical coins as rewards are shown in the center of the CRT monitor. In the test session, the accumulative accuracy rate and graphical coins were not shown.

coin was added every time a subject made three correct responses except in the correction trails. After completing the training sessions in one day, the subjects were paid 1 yen (almost equal to 1 US cent) per correct response, which means 3 yen per one graphical coin, as a bonus.

Sixty-eight minimal pairs by one talker were presented twice in one training session, yielding 272 trials per session. Each session lasted approximately 30-40 minutes. Five talkers cycled from T1 to T5 session-by-session. As a result, 15 training sessions consisted of 3 cycles of 5 talkers.

3.2 Results

The overall results are shown in Fig. 2.

3.2.1 Pretest–post-test

The mean accuracy in each test for each subject was obtained. One of the subjects in SES3 showed 100% accuracy in both the pretest and post-test. This subject was excluded from all of the analyses in this paper. The rest of the data were subjected to an ANOVA, where the test (pretest vs. post-test) was the within-subjects variable. In SES3, the accuracy in the post-test was significantly higher than in the pretest [pretest: 70%, post-test: 80%, F(1,16)=33.21 p<0.005] like in the SES1 group reported by Lively et al. (1994)[pretest: 65%, post-test: 77%, F(1,18)=92.99, p<0.005].

A 2-factor ANOVA was conducted with the condition (SES1 vs. SES3) and test (pretest vs. post-test) as variables. The main effect of the test was significant [F(1,68)=24.61, p<0.005], whereas the main effect of the condition

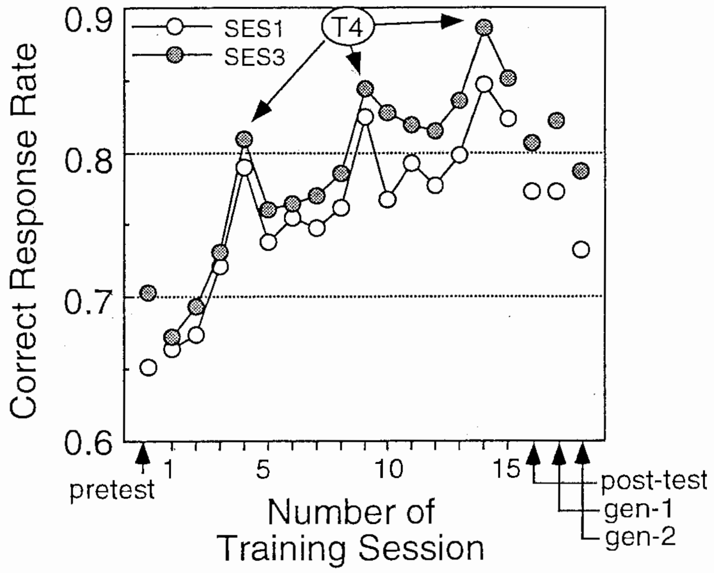

Figure 2: Accuracy in the pretest, each training session, post-test, and 2 generalization tests for the SES1 group and SES3 group. The SES1 group received 1 training session per day for 15 days, and SES3 group received 3 training sessions per day for 5 days. Gen-1 shows the generalization for new words by an old talker, and gen-2 shows the generalization for new words by a new talker. The three notable peaks marked as "T4" are the accuracies in the session of talker T4: The Japanese speakers consistently showed higher scores on the productions by T4 than on the productions by any of the other talkers.

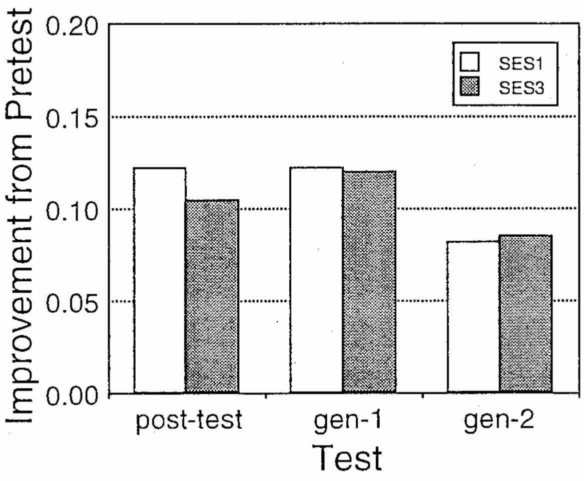

Figure 3: Improvement in the post-test and generalization tests from the pretest for the SES1 group and SES3 group. Gen-1 shows the generalization for new words by an old talker, and gen-2 shows the new words by a new talker.

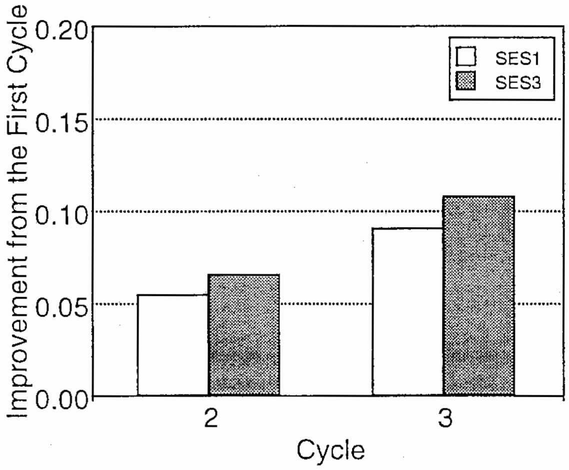

Figure 4: Improvement during training for the SES1 group and SES3 group. The improvement from the mean accuracy in cycle 1 to cycle 2, and to cycle 3 are displayed.

was not significant [F(1,68)=3.52, ns]. The interaction between the condition and the test was not significant [F(1,68)=0.15, ns].

Furthermore, the improvement in accuracy from the pretest to post-test for each subject was compared between the SES1 and SES3 groups. The left 2 bars in Fig. 3 show the improvement from the pretest to post-test. An ANOVA, which treated the condition (SES1 vs. SES3) as the between-subjects variable, showed that the improvement in accuracy did not differ significantly between SES1 and SES3 [SES1: 12%, SES3: 10%, F(1,34)=0.68, ns].

3.2.2 Generalization test

Generalization items were tested only in the post-test phase. Consequently, there is no base line to discuss the improvement in those items. However, the result of a pretest–post-test comparison showed that the pretest's performance did not differ between SES1 and SES3 significantly. Depending on this result, we assume that the two groups started from almost equivalent initial levels. This assumption allows us to compare the performance in the generalization tests between the 2 groups directly. The mean accuracy in each test for each subject was obtained. Separate ANOVAs were conducted for the 2 generalization tests. The condition (SES1 vs. SES3) was the between-subjects variable. In the generalization test of a new item by an old talker, the accuracy did not differ between SES1 and SES3 significantly [77% in SES1, 82% in SES3, F(1,34)=1.77, ns]. SES1 and SES3 also showed no significant difference in the generalization test of a new item by a new talker [73% in SES1, 79% in SES3, F(1,34)=1.89, ns].

The improvement in accuracy was further obtained by using the pretest score as a base line, i.e. ("accuracy in the generalization test" - "accuracy in the pretest") was used as a pseudo-transfer score, and it is called the "transfer rate", henceforth. The middle and right 4 bars in Fig. 4 display the transfer rate for each generalization test. Separate ANOVAs were conducted on the transfer rate in the 2 generalization tests. The condition (SES1 vs. SES3) was the between-subjects variable. In the generalization test for new item by old talker, there was no significant difference between SES1 and SES3 [SES1: 12%, SES3: 12%, F(1,34)=0.01, ns]. There was also no significant difference in the generalization test of a new item by a new talker [SES1:8%, SES3: 8%, F(1,35)=0.01, ns].

3.2.3 Training

The mean accuracy for each session of training in SES3 was subjected to an ANOVA. The cycle (cycle1, cycle2 and cycle3) was the within-subject variable. The accuracy increased with cycle significantly [74% in cycle1, 81% in cycle2, 85% in cycle3, F(2,32)=131.27, p<0.001].

In order to compare the improvement during the training between SES1 and SES3, the improvement in accuracy from the first cycle to the 2nd and 3rd cycles and from the 2nd to 3rd cycle for each subject were calculated. ANOVAs were conducted for each improvement, from the 1st cycle to 2nd cycle, from the 2nd cycle to 3rd cycle, and from the 1st cycle to 3rd cycle. The condition (SES1 vs. SES3) was the between-subjects variable. The improvement did not differ significantly between SES1 and SES3 in the 1st cycle to 2nd cycle [5.4% in SES1, 6.5% in SES3, F(1,178)=2.53, ns], and in the 2nd cycle to 3rd cycle [3.6% in SES1, 4.3% in SES3, F(1,178)=1.12, ns]. However, the improvement from the 1st cycle to 3rd cycle differed significantly [9.0% in SES1, 10.8% in SES3, F(1,178)=4.92, $p<0.05$].

3.3 Discussion

The accuracy of the trainee in the SES3 group increased during the training, and also from the pretest to post-test. In addition, a transfer to new talkers and to new words were observed. These results replicated those of Lively et al.(1994), which are shown as the SES1 group in this paper.

More importantly, there was almost no difference observed between SES1 and SES3. The accuracy of the trainee in SES1 and SES3 did not differ significantly in both the pretest and post-test. This shows that both groups started from a similar level, and reached a similar level through training. In addition, the amount of improvement from the pretest to post-test did not differ between the 2 groups, showing that the 2 groups improved by the same amount.

The accuracies in the generalization tests did not differ between SES1 and SES3. Also, the difference from the pretest to generalization tests did not differ between 2 groups. This suggests that the training transferred to new talkers and to new words to a similar extent between the 2 groups.

During the training, both groups increased in a similar way. The improvement from the 1st cycle to 2nd cycle and from the 2nd cycle to 3rd did not differ significantly. Only one difference between the 2 groups was observed in the improvement from the 1st cycle to 3rd cycle. The mean increase was 9.0% in SES1 and 10.8% in SES3, but this difference is not surprisingly large even though it is significant.

This significant difference in improvement from cycle 1 to cycle 3 during training can be considered as a marginal difference when discussing the difference of training effect in SES1 and SES3 by 2 reasons. First, the difference is not large. Actually, when this improvement was broken down into two improvement parts, from 1st cycle to 2nd cycle and from the 2nd cycle to 3rd cycle, no difference was observed between the 2 groups. Second, the evaluation should be conducted mainly with the pretest and post-test performance rather than with the improvement during training. As a result, we may conclude that the training in SES1 and SES3 did not differ substantially. In other words, the massed arrangement of sessions in SES3 (three sessions per day) did not inferior or superior to the distributed arrangement in SES1 (one session per day).

In the next experiment, we examined the effect of the amount of training on developing new phonological categories in perception. Since we learned that "three training sessions per day" does not reduce the training effect compared with "one training session per day", we expanded the amount of training by not expanding the training period. We trained subjects three sessions per day over 15 days; unlike the one session per day over 15 days for SES1, and three sessions per day over five days in SES3.

4 Experiment 2: the effect of extended training

4.1 Subjects

Thirteen native speakers of Japanese with no experience living abroad served as subjects. They ranged from 18 to 45 years of age (average 24). All of them reported no history of hearing or speech disorder. A hearing screening performed at 15dB HL for frequencies between 250 to 8000 Hz showed all subjects to have normal bilateral hearing acuity.

4.2 Procedure

A procedure similar to the Experiment 1's was used. In contrast to Experiment 1, the subjects received 45 training sessions by repeating the training phase 3 times with mid-tests in between.

On the 1st day of the experiment, the subjects were given a pre-test, which was identical to that in Experiment 1. The subjects were then trained in 3 training phases. Each training phase consisted of 15 training sessions over 5 days, equal to the amount of training and training period of SES3 in Experiment 1. Between every two training phases, mid-tests, which were identical to the pretest and post-test, were conducted on a separate day. On the final day, the subjects participated in the post-test, and 2 generalization tests, which were identical to those in Experiment 1. Furthermore, follow-up tests were conducted three months and six months after the conclusion of the training.

4.3 Result

The overall results are shown in Fig. 5.

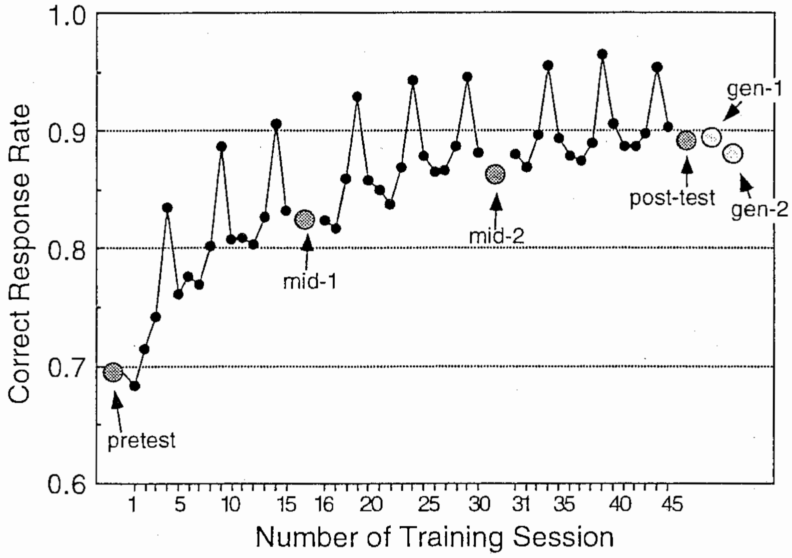

Figure 5: The accuracy in all the tests and training sessions are displayed for the EXT (extended) group. Mid-1 shows a mid-test after 15 training sessions, and mid-2 shows a mid-test after 30 training sessions. Gen-1 and gen-2 show a generalization to new words by an old talker and a generalization to new words by a new talker, respectively. Note that the notable peaks are again observed every 5 sessions corresponding to the session by talker T4.

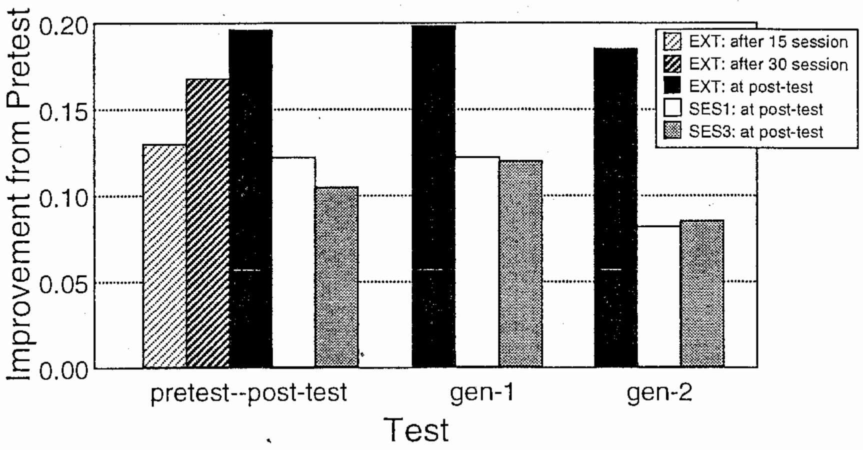

Figure 6: Improvement in the post-test, and 2 generalization tests from the pretest accuracy for each training group are displayed. Gen-1 and gen-2 show a generalization to new words by an old talker, and a generalization to new words by a new talker, respectively.

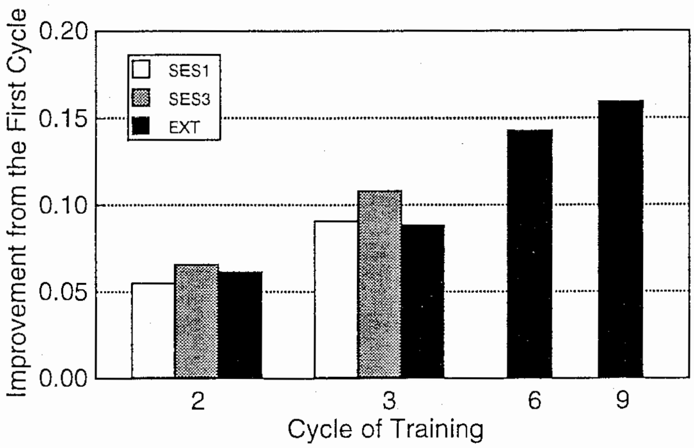

Figure 7: Improvement in the 2nd, 3rd, 6th and 9th cycle of training from the first cycle of training are displayed.

4.3.1 Pretest–mid-test–post-test

First, the mean accuracies in the pretest and post-test for each subject were subjected to an ANOVA, where the test (pretest vs. post-test) was the within-subjects variable. This analysis confirmed that the accuracy in the post-test was significantly higher than in the pretest [pretest: 69%, post-test: 89%, F(1,12)= 96.85 p<0.001]. The accuracy increased by 20% from the pretest to post-test, whereas it increased by 12% and 10% in Experiment 1.

In order to test whether this 20% of improvement was significantly larger than the improvement of 12% or 10% in Experiment 1, the improvement in each mid-test and post-test over the pretest was calculated and compared with the improvements in SES1 and SES3 (Fig. 6). The improvements in each test in this experiment were submitted to separate ANOVAs. We call the current group EXT, which stands for extended training. The condition (EXT, SES1, vs SES3) was the between-subjects variable. The improvement from the pretest to mid-test1, was 13% for EXT, and did not differ significantly from the improvement from the pretest to post-test in SES1 (12%) and SES3 (10%) [F(2,46)=0.58, ns]. The improvement from the pretest to mid-test2 (16%) also did not differ significantly from the improvement from the pretest to post-test in SES1 and SES3 [F(2,46)=3.02, ns]. However, the improvement from the pretest to post-test in EXT (20%) was significantly larger than the pretest–post-test improvement in SES1 and SES3 [F(2,46)=7.65, p<0.01].

Finally, the improvement in each mid-test and post–test over the pretest were compared. The period of the test was the within-subjects variable. The improvement in the mid-test2 (17%) and post-test (20%) were significantly higher than that in the mid-test1 (13%) [F(1,12)=13.23, p<0.01 in the mid-test2, F(1,12)=21.43, p<0.001 in the post-test]. However, the improvement in the mid-test2 (17%) and post-test (20%) did not differ significantly [F(1,12)=3.03, ns].

4.3.2 Generalization tests

In order to compare the transfer of training to generalization items in the 15 sessions of training and in the extended 45 sessions of training, the "transfer rate" (see Experiment 1) was calculated for each subject and compared to those in Experiment1. Separate ANOVAs were conducted for each generalization test. The group (SES1, SES3 vs. EXT) was a between-subjects variable.

In the generalization test of new items by an old talker, the effect of the group was significant [12%(SES1), 12%(SES3), 20%(EXT), F(2,46)=6.17, p<0.01]. Post hoc tests using Tukey's HSD procedure showed significant differences between the groups except between SES1 and SES3. In the generalization test of new items by a new talker, the effect of the group was significant [8%(SES1), 8%(SES3), 18%(Extend), F(2,47)=10.08, p<0.001]. Post hoc tests using Tukey's HSD procedure showed significant differences between the groups except between SES1 and SES3.

4.3.3 Training

The mean accuracy for each session of training was subjected to an ANOVA. They cycle (cycle1-3) was the within-subjects variable. The accuracy increased with cycle significantly [F(8,96)=69.28, p<0.001]. The accuracy constantly increased from cycle1 (75%) to the final cycle9 (91%). A post hoc test using Tukey's HSD procedure showed that the accuracy in the first 4 cycles (from cycle1 to cycle4) was significantly lower than most of the following cycles. However, the improvement from the 5th to 9th cycle was not statistically significant except for the finding that the accuracy in cycle 5 was significantly lower than in cycle 9.

In order to compare the improvement during the first 15 sessions of training with SES1 and SES3, the improvement in accuracy from the first cycle to the 2nd and 3rd cycles in the 3 groups were submitted to separate ANOVAs. The condition (SES1, SES3, vs. EXT) was the between-subjects variable. The improvement in the 2nd cycle did not differ significantly among the conditions [5.4% in SES1, 6.5% in SES3, 6.2% in EXT, F(2,242)=1.36, ns]. The improvement in the 3rd cycle over the 1st cycle was affected by the condition [9.0% in SES1, 10.5% in SES3, 8.8% in EXT, F(2,242)=3.46, p<0.5]. However, Tukey's HSD procedure showed that there were no significant differences between the groups.

These results are partly displayed in Fig 7. The mean accuracy in the 3rd, 6th and 9th cycle of training signify the last cycle of each of the 15 training sessions. The mean accuracy in the 2nd cycle was added for the comparison with SES1 and SES3.

4.3.4 Memory Retention

The memory retention was also assessed through follow-up tests 3 months and 6 months after the conclusion of the extended training. Eight subjects returned 3 months after the training. The averaged score of the pretest, post-test, and 3-month test for these subjects were, 67.8%, 88.9% and 87.6%, respectively (Fig. 10, left panel). Accuracies in the post-test and 3-month test were significantly higher than in the pretest [F(1,7)=124.5 p<0.001; F(1,7)=101.6, p<0.001]. However, there were no significant differences between the accuracies in the post-test and 3-month test [F(1,7)=0.52, ns].

Five of them returned 6 months after the conclusion of the extended training (Fig. 10 right panel). Their scores in the post-test (86.8%), 3-month test (87.1%) and 6-month test (85.3%) were significantly higher than pretest (64.7%) [F(1,4)=176.9, p<0.001 for post-test; F(1,4)=109.4, p<0.001 for 3-month test; F(1,4)=40.2, p<0.005]. There were no significant differences among the scores in the post-test, 3-month test and 6-month test [F(1,4)=0.0 (ns), 1.5 (ns), and 0.8 (ns) for post-test vs. 3-month test, 3-month test vs. 6-month test, and post-test vs. 6-month test, respectively].

4.4 Discussion

The accuracy increased 20% from the pretest to post-test in this training group. A comparison of the performance between the 3 trainings (SES1, SES3 and EXT) showed that the accuracy in the test after 15 training sessions (12%, 10%, 13%) did not differ. However, the post-test accuracy after 45 training sessions (20%) significantly differed from the accuracy after 15 training sessions. These results showed that the extended training group was not significantly more accurate on the pretest and post-test items by the 15 training sessions than the SES1 and SES3 training groups. But the additional 30 training sessions improved the trainees' accuracy significantly. In addition, the improvement in the post-test from the pretest was significantly larger than that at the mid-test1 (20% vs. 13%). This result also supports the hypothesis that the additional 30 training sessions contributed toward improving the accuracy on the pretest/post-test items.

In the extended training, generalization tests were conducted only in the post-test phase. When the transfer rate was estimated by having the accuracy on the pretest as a base line, i.e. by measuring the difference between each generalization test and pretest in accuracy, significantly larger transfer rates than in the SES1 and SES3 groups were observed in each generalization test (12% in both SES1 and SES3 vs. 20% in EXT for new items by an old talker;

8% in SES1 and SES3 vs. 18% in EXT for new items by a new talker). This result suggests that the training of 45 sessions allowed trainees to generalize their identification ability to new items and new talkers much more than the training of 15 sessions.

Regarding the improvement during training, the subjects in EXT showed a similar amount of improvement to those in SES1 and SES3 during the first 15 training sessions. After the 16th training session, the accuracy constantly increased until the end of the training, even though the latter half of the training cycles did not show significant improvement. This result suggests that at least the additional several training sessions improved the accuracy significantly, but the trainee might have met an asymptote at the final stage of this training.

Furthermore, the subjects' identification ability stayed about 20% higher over the pretest even 3 or 6 months after the conclusion of extended training, suggesting that the present extended training was highly effective in modifying the subjects' perception of /r/ and /l/ over a long period.

Considering all of these results together, the effect of the additional 30 training sessions had a significant effect on the subjects' ability to identify /r/ and /l/ in the untrained items. Even though there was no significant effect on the improvement during the last 5 cycles (25 sessions) of training, this additional training further improved the subjects' identification ability for items in the pretest/post-test and generalization tests. We conclude that the current extended training developed more robust internal representations of /r/ and /l/ than the training with smaller sessions.

4.5 General Discussion

This paper examined the effect of the amount of training on the identification of English /r/-/l/ contrast for Japanese speakers of expanding the study by Lively et al. (1994). We obtained three main results. First, Experiment 1 demonstrated that the effectiveness of training does not differ between concentrated training [5 training days, 3 sessions per day] and diffused training [15 training days, 1 session per day]. Second, Experiment 2 showed that additional training trials significantly improve the subjects' ability to identify /r/ and /l/. The accuracy improved from 69.5% in the pretest to 82.5% after 4080 trials, and reached 86.7% and 89.1% after 8160 trials and 12240 trials, respectively. Third, this performance in perception was maintained over a 3 or 6 months period, demonstrating that the extended training produced a long-term modification of adults' phonetic categories.

In addition to the above results, we supplement here that the effects of talker and phonetic environment reported in Lively et al. (1994) were replicated. In both Experiments 1 and 2, similar effects were observed of talker on accuracy during training as reported by Lively et al. (1994). One of the training talkers, T4, was easier to identify her /r/ and /l/ than the other talkers (i.e. accuracy to this talker was higher), and this can be seen as notable peaks in Fig. 2 and Fig. 3. T3, T5, T1, and T2 were the 2nd to 5th easiest talkers to identify, respectively (Fig. 8).

Also observed were the same effects of phonetic environment on accuracy as reported by Lively et al. Overall, the /r/ and /l/ in the final singleton, final consonant cluster, initial singleton, intervocalic position, and initial consonant cluster were the 1st to 5th easiest environments, respectively (Fig. 9).

The above results offer two theoretical contributions. First, they demonstrate that the adult learners' difficulty in learning non-native sounds does not solely depend on aging. In contrast to the present results, it has been reported that Japanese speakers, especially adults, do not acquire the /r/-/l/ distinction even after several years' of living in an English-speaking environment both in perception (Yamada, 1995) and in production (Yamada et al., 1994). Similar results have been obtained in other contrasts. For example, Bohn and Flege (1990) studied German speakers' perception of the English /ɛ/-/æ/ contrast. Not all of the German speakers who had been exposed to an English-speaking environment for several years could acquire a completely native-like perception, although they did perceive the contrast better than in less experienced German speakers. These phenomena can sometimes be empirically explained by a biological factor., i.e. the critical period hypothesis proposed by Lenneberg (1967). However, another factor which relates to the learning strategy may interact with the acquisition. There is a possibility that the knowledge of a language might work as an inhibitory factor when adults or adolescents are exposed to an English-speaking environment: Contexts, such as conversational situations, sentences, words, etc., help listeners' understanding. However, these aids may inhibit the perceptual learning of phonetic segments, because the listeners do not have to depend only on sounds. In fact, not all adult learners of a foreign language fail to acquire new phonetic categories by being exposed to the English-speaking environment (e.g. Yamada, 1995; Flege et al., 1995). Large individual differences are often observed in L2 acquisition studies. These imply that the adult learners' failure in learning new phonetic contrasts does not solely depend on the loss of plasticity by aging, but also on some other factors. The present finding that adults can develop even difficult non-native phonetic categories by laboratory training reveals that failure through exposure to the L2-speaking environment for adult learners depends not only on age but also on the lack of chance to learn the language in an adequate way.

Second, they demonstrate the possibility that adequate training paradigms can promote any L2 categories. There are phonetic contrasts which are extremely difficult to learn for speakers of some specific languages (Bohn, 1995 for a review), like the present case of English /r/-/l/ distinction for Japanese speakers. The laboratory training technique can make it better, but the improvement is less than other easier contrasts. According to the "Perceptual Assimilation Model (PAM)" proposed by Best (1995 [1]), the difficulty in perceiving non-native segments are due to the perceptual assimilation to native segments. In Japanese phonological system, there is no sounds similar to English /r/ or /l/. Japanese speakers aissimilate both /r/ and /l/ for Japanese /r/. However, Japanese /r/, is usually a flap (/ɾ/), and so, both English /r/ and /l/ are discrepant for Japanese /ɾ/. Thus, in her model, English /r/-/l/ contrast for Japanese speakers can be classified as Single-Category assimilation, where discrimination is expected to be poor. Many previous reports have shown this poor discrimination of Japanese speakers (e.g. Miyawaki et al.), as well as poor identification (e.g. Mochizuki, 1981). The present results demonstrate that the amount of training can compensate for this difficulty due to the difference in phonological system between L1 and L2: Adult learners can promote the robust perception of L2 phonetic categories if an adequate amount of adequate training is provided.

In contrast to the discussion about the success of laboratory training, we should note that no study has succeeded in developing completely native-like perception; accuracies have not reached 100%, showing that all trainees behave differently from native speakers even after training. In future studies, we should clarify what training method makes the subjects' perception completely native-like. In order to call the perception "native-like", the performances of several perception properties at least must be satisfied. For example, achieving no misses in the identification test of natural tokens, would show a categorical perception on the synthetic continuum, using the proper perception cue to distinguish the contrast, etc. We must figure out what perception properties should be assessed to call the perception "native-like" and examine which property is obtained by what kind of training. In other words, we should study in what way what type of training changes the trainee's perception. In such studies, a training method that improves the trainee's perception to a high performance level is required. In addition, being able to observe the learning process might be helpful, because each perception property is thought to be acquired at different stages in the learning process.

In order to train a trainee to reach a high performance level, this paper showed that the amount of training can overcome any difficulty at least when the adequate training is provided. However, more effort towards further optimization of the effective training method is necessary. Consideration on the structure of the stimulus set and its sequence attempted in the perceptual fading technique (Terrace, 1963; Jamieson & Morosan, 1986; Pruitt, 1995) may be one of the most hopeful ways. The data during the training (Fig. 9, right) demonstrated that the different phonetic environment showed different learning curve; the easier environments met asymptote (ceiling) earlier, while difficult environments kept improving all through the training period. This result implies the possibility that the consideration of the structure of the stimulus sequence may further make the current training more effective: We may be able to reduce the number of trials required to reach a high performance level by using an effective stimulus set consisting of natural tokens, edited tokens, and/or synthetic speech in combination with the fading technique. However, further examination is necessary to determine the validity of this fading method.

As for the learning process, it was observed in a usual way using the data obtained during the training. In the present study, we have assessed mid tests, which allowed us to analyze the learining process in detail. Overall, trainee improved all through the trainig. However, various patterns in the learning process was observed in the individual data (see Appendix, Figure 11). Some subjects met asymptote quickly (S53, S55, S58, S61), S63 improved in the late stage, and the others showed gradual improvement all through the training. From the theoretical point of view, the extended training procedure with mid-term probes allowed us to look at different learning strategies, and thus to begin to address the question of what is learned as a result of this type of training. For example, present result demonstrated that different phonetic environment showed different learning curve during test and training (Fig. 9, right). In the data during training, the easier environments (final singleton, and final cluster) met asymptote (ceiling) earlier, while difficult environments (intervocalic, initial singleton, and initial cluster) kept improving all through the training period. These three difficult environments seem to show similar learning curve. Similarly, at least two difficult environments (intervocalic, and initial singleton) showed similar learning curve, while easier environment (final singleton) met asymptote. Note that the number of /r/-/l/ minimal pairs used in the training (i.e. number of training trials) differed largely among environments; 11 final cluster pairs, 15 final singleton pairs, 5 intervocalic pairs, 13 initial sigleton pairs, and 24 initial cluster pairs. Comapring intervocalic, initial singleton and initical cluster in the data during training, the similar improvement was achieved with 5:13:24 ratio of training trials (words). Moreover, this relation generalized into the results in tests. If we assume that trainee develop context-specific "allophone", the environment with much number of training trials (words) would improve more than the environment with less number of training trials (words). Present data imply that the trainee develop not solely context-specific "allophone", but also context-independent abstract units. Further examination is needed to determine the nature of internal representation that trainee acuire through this type of training. In such examinations, the training with mid tests or daily probe tests will be powerful.

We believe the attempts to find effective training methods and studies using those effective methods will provide insight into the nature of speech perception development. A systematic approach towards the optimization of training L2 phonetic categories was only recengtly initiated by Jamieson and Morosan (1986). However, the results of such an approach have already been utilized. For example, Bradlow et al. (submitted), Yamada et al. (1995), and Bradlow et al. (1995) found that training in the perception domain transfers to the production domain using the present method. They trained Japanese speakers to identify English /r/ and /l/ in perception using the extended training reported in this paper. The /r/ and /l/ productions of trainees were recorded before and after the training. An evaluation by native speakers of English showed that the production ability improved from the pretest to post-test even though the training was only in perception. In addition to the contribution to speech perception and production research, a practical contribution to foreign language teaching is possible. All of these studies along with future efforts in L2 speech training studies will further make both theoretical and practical contributions possible.

References

[1] Best, C.T. "A direct realist view of cross-language speech research", In Speech Perception and Linguistic Experience, Strange, W. (Ed.), York Press, Timonium MD, pp.171-204 (1995).

[2] Bohn, Ocke-Schwen, "What determines the perceptual difficulty encountered in the acquisition of non-native contrasts?", *Proceedings of ICPhs 95*, 84-91 (1995).

[3] Flege, J.E., "Two techniques for training a novel second-language phonetic contrast", submitted to Applied Psycholinguistics",

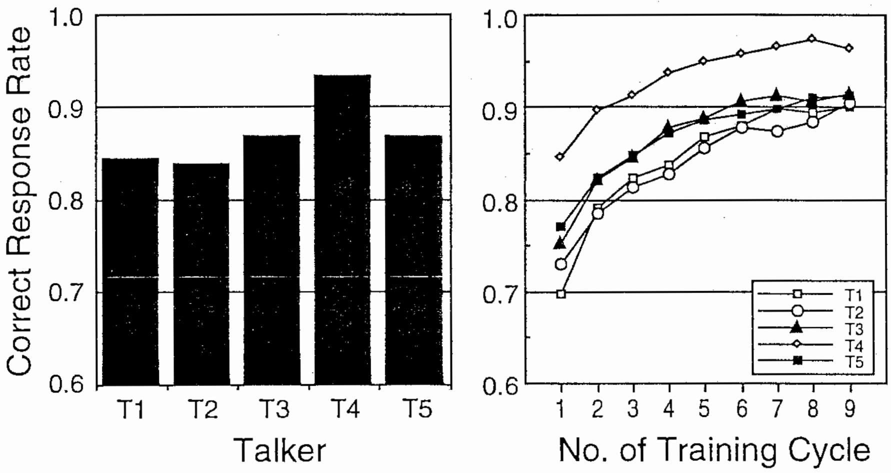

Figure 8: The effect of talker on accuracy during training in Experiment 2. The left panel shows the accuracies by the talker averaged across all training cycles. The right panel shows the accuracies by the talker and training cycle count.

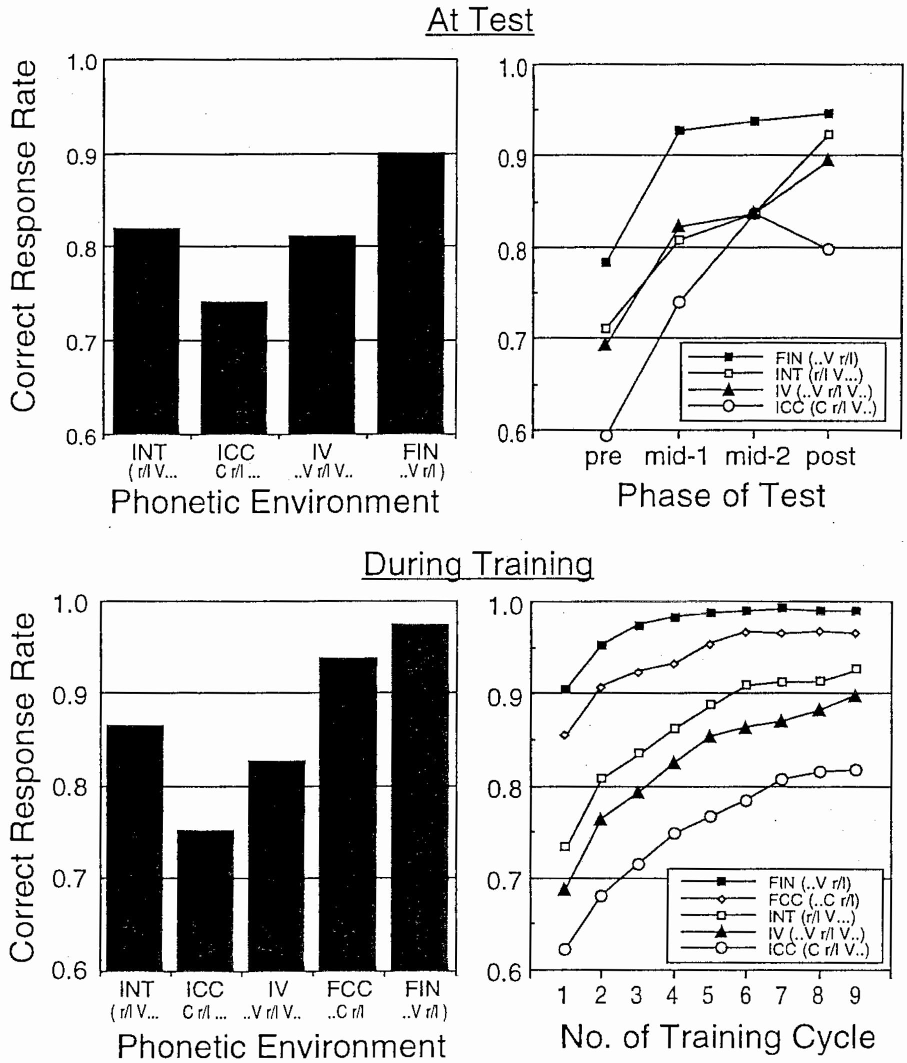

Figure 9: The effect of a phonetic environment on accuracy in the tests (upper panels) and during training (bottom panels) in Experiment 2. The left panels show the accuracies by the phonetic environment averaged across all tests or training sessions. The right panel shows the accuracies by the phonetic environment and phase of test or training cycle count.

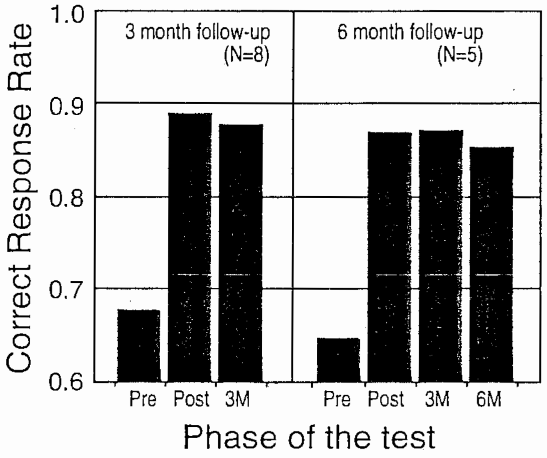

Figure 10: Memory retention of the training in the extended training. The left three bars show the accuracies for the pretest, post-test and 3-month follow-up test. The right four bars show the accuracies for the pretest, post-test, 3-month and 6-month follow-up tests.

[4] Flege, J.E. "Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training", *Journal of Acoustical Society of America*, 86,1684-1697 (1989)

[5] Flege, J.E. "Speech learning in a second language", in Ferguson, D. et al. (eds.) *Phonological development: Models, research, and application* York Press, Parkton, MD. pp. (1991)

[6] Flege, J.E. and Takagi, N. and Mann, V. "Japanese adults can learn to produce English /r/ and /l/ accurately.", *Language and Speech* **38**, 25-55 (1995).

[7] Jamieson D.G., and Morosan, D.E. "Training non-native speech contrasts in adults: Acquisition of the English /th/-/th/ contrast by francophones", *Perception & Psychophysics* **40**, 205-215 (1986).

[8] Goto, H. "Auditory perception by normal Japanese adults of the sounds "l" and "r"", *Neuropsychologia* **9**, 317-323 (1971).

[9] Hintzman, D.L. and Block, R.A. and Summers, J.J. "Modality tags and memory for repetitions: Locus of the spacing effect", *Journal of Verbal Learning and Verbal Behavior* **12**, 229-239 (1973).

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

[11] Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A.M., Jenkins, and Fujimura, O. "An effect of linguistic experience: The discrimination

of [r] and by native speakers of Japanese and English.", *Perception & Psychophysics* **18** 331-340 (1975)

[12] Jamieson, D.G., "Techniques for training difficult non-native speech contrasts", *Proceedings of ICPhs 95*, 100-107 (1995).

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

[14] Mochizuki, M. "The identification of /r/ and /l/ in natural and synthesized speech", *Journal of Phonetics* **9**, 283-3031 (1981).

[15] Sheldon, A. and Strange, W., "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception", *Applied Psycholinguistics* **3**, 243-261 (1982).

[16] Strange, W., "Phonetics of second-language acquisition: Past, present, future", *Proceedings of ICPhs 95*, 76-83 (1995).

[17] Strange, W. and Dittmann, S. "Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English", *Perception & Psychophysics*, **36**, 131-145 (1984).

[18] Werker, J.F. and Tees, R.C. "Phonemic and phonetic factors in adult cross-language speech perception", *Journal oc Acoustical Society of America* **75**, 1866-1878 (1984).

[19] Yamada, R.A. and Strange, W. and Magnuson, J.S. and Pruitt, J. S. and Clarke, W. D. III, "Production English and /l/ by native speakers of Japanese", Proceedings of the 1994 International Conference on Spoken Language Processing, Yokohama, pp.2023-2026, (1994).

[20] Yamada, R.A "Age effect and acquisition of second language speech sounds: Perception of American English /r/ and /l/ by native speakers of Japanese", In Speech Perception and Linguistic Experience, Strange, W. (Ed.), York Press, Timonium MD, pp.301-316.

Appendix

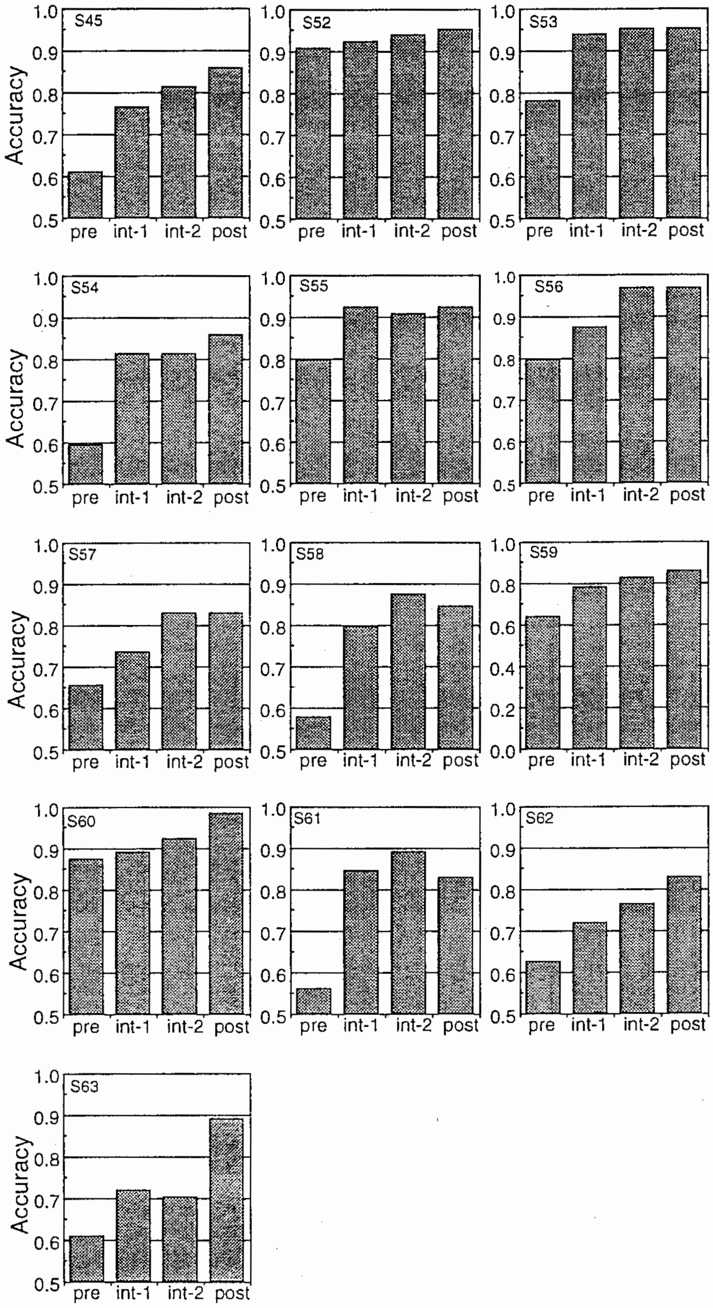

Figure 11: Appendix A: Individual data for accuracies at pretest, mid-tests, and post-test in Experiment 2.