Training and evaluating Korean learners' perception of the English word-initial *l/r* contrast: an SDT-based evaluation model*

Soonhyun Hong
(Inha University)

Hong, Soonhyun. 2010. Training and evaluating Korean learners' perception of the English word-initial *l/r* contrast: an SDT-based evaluation model. *Studies in Phonetics, Phonology and Morphology* 16.2. 297-317. Korean learners of English have difficulty in the non-native contrast of *l*/*r* phones despite a relatively long period of English study. This is because the Korean sound system does not have the *l*/*r* contrast. Even university-level Korean students with more than 10 years of English learning still suffer almost the same or similar level of difficulty. This paper tries to demonstrate that a short period of laboratory training can help university-level Korean learners of English improve the perception on *l* and *r* phones. Learners were forced to finish a session of training which consists of two blocks of *l*/*r* minimal pair identification training with feedback. This less-than-20-minutes-long session was repeated for five consecutive days. After pointing out a measurement problem occurring with the use of mean accuracy rates for the evaluation of learners' perception in the literature, it is proposed that the sensitivity index in the Signal Detection Theory be used as a more precise measure of a learner's identification performance on the novel *l/r* contrast. The sensitivity measures show that the training group demonstrated a significant perception improvement compared to the control group. **(Inha University)**

Keywords: perception of English, *l* and *r* perception, identification test, training effect on L2 phones, Signal Detection Theory, Sensitivity

1. Introduction

The contrast between English *l* and *r* is very difficult for Korean and Japanese learners of English to perceive even though they have spent years studying English. This is because their native language has only a single liquid phoneme subsuming both phones. Studies on this topic have amply been documented especially for Japanese learners of English.

Ingram & Park (1998) conducted an *l/r* minimal pair identification test on ten Korean and ten Japanese learners of English for three phonetic environments: Initial singleton, initial cluster, and intervocalic positions. It was found that Korean learners had trouble in *l/r* perception in all three positions. They further reported that perception in the word-initial singleton is more difficult for Korean learners than in the other environments while perception in the consonant cluster is more difficult than in the other environments. Park (2008) conducted a similar

* This work was supported by Inha University Research Fund. I would like to thank the three anonymous reviewers for their insightful comments and suggestions on this paper.

identification test of *l/r* in initial and final positions on three groups of Korean learners: 20 elementary, 22 junior high, and 22 high school students. She reported that their overall identification ability was very poor (mean accuracy=62.7%). She also showed that there was no statistical difference in *l/r* perception ability among the three groups. From the longitudinal perspective, it means that more than three years of English study did not facilitate learners' perception of the contrast. Similar difficulties in perception of some non-native phones have amply been reported in the literature and persists even after years of living in non-native environments (Best and Strange 1992, Flege and Eefting 1987, and Yamada 1995).

However, Logan et al. (1991) demonstrated that a short period of intensive minimal pair training can facilitate Japanese learners' perception of the *l/r* contrast. Six Japanese learners were trained for three weeks to identify *l* and *r* from four different phonetic environments in a naturally spoken minimal pair identification paradigm with feedback: Initial singleton, initial cluster, intervocalic, and final singleton pairs. The training effect was evaluated with a pretest-posttest design (Strange & Dittmann 1984) with natural tokens. It was found that there was an overall significant performance improvement in mean accuracy of *l* and *r* responses. However, post-training improvement for the initial singleton environment was rarely observed in their study than for the other environments.

Other training studies have confirmed that laboratory training can lead to substantial generalizable improvement in identification of the *l/r* contrast when they are trained with novel stimuli of novel talkers (Lively et al. 1993 and Bradlow et al. 1997). And the improved performance persisted for at least 6 months and affected production and identification of words with *l/r* (Bradlow et al. 1999). Laboratory training on other L2 phones was shown to be effective (Lambacher et al. 2005 and Nishi & Kewley-Port 2007 on English vowels and Hong 2009 on English fricatives and affricates).

The present paper will address two issues observed in previous studies of *l/r* training. First, most of the *l/r* training tasks in the literature have been focused on Japanese learners' post-training performance improvement. However, training studies for Korean learners have been extremely rare (Yu & Jamieson 1993 and Hardison 1997, 2003). It will be shown that even a five-day-long period of forced-choice minimal pair training with feedback facilitated Korean learners' perception of the contrast in *l/r*-initial words. Additional five-day-long period of training furthered their perception. Second, the mean accuracy rates of *l/r* responses were used as measurement for learners' performance in alternative forced-choice identification tests in previous studies as in Logan et al. (1991) and others. However, this paper will point out that mean accuracy rates cannot robustly measure learners' post-training perception due to decision bias. This paper instead introduced sensitivity (or d') in the Signal Detection

Theory (Harvey 1992, Egan 1975, MacMillan & Creelman 2005, and Wickens 2001) for bias-free measurement of learners' performance. The use of sensitivity as performance measurement index also sheds light on pronunciation-related pedagogy, as learners can check their own performance improvement after tests.

2. Methods

2.1 Subjects

Fifty-one native Korean learners of English of ages 20 to 25 years old, participated in the training experiment. The training group consisted of 31 subjects (21 females and 10 males) whereas the number of the subjects in the control group was 20 (10 females and 8 males). All of them were university students who had been learning English for at least 6 years since middle school and had no reported history of speech or hearing problems.

2.2 Stimuli

Two groups of spoken samples were used in the experiment: One group for testing and another for training. The word tokens for testing composed of 22 minimal pairs of *l/r*-initial real words, and their spoken token samples were extracted from various electronic dictionaries (*Yahoo English-Korean Dictionary*, *MacMillan Dictionary*, *E4U CD-Rom Dictionary*, and *Collins Cobuild Dictionary*). The reason for this was that electronic dictionaries could provide ample samples of various native speakers of English (22 *l*-initial word tokens spoken by 12 females and 10 males and 22 *r*-initial word tokens spoken by 14 females and 8 males). On the other hand, the word tokens for training consisted of 37 minimal pairs of *l/r*-initial real words, the spoken samples of which were extracted from *Longman Pronunciation Coach* (37 *l*-initial word tokens spoken by 21 females and 16 males; 37 *r*-initial word tokens spoken by 30 females and 7 males). Notice that the minimal pairs for training were never overlapped with those for testing, and spoken samples for testing and for training were extracted from different sources. By separating training samples from testing samples, the transfer of a generalization effect from training to a post-training test could be checked more accurately.

All the spoken samples were recorded in a wave format at 16 kHz, 256 Kbps through Total Recorder 6.0 on a PC. They were spoken by an unidentifiable but large number of English speakers. Each of the spoken samples was carefully verified through careful listening by the author, and was normalized in Wavesurfer 1.8.5.

2.3 Procedures

2.3.1 Pretest and posttest

The experimental design employed a pretest-posttest design used by Strange & Dittmann (1984), Logan et al. (1991), and Hong (2009). Posttest was taken by the subjects in the training and the control group right after the 5-days-long training.

For pretest and posttest, an *l/r* identification protocol was built in Alvin 1.5 (Hillenbrand & Gayvert 2005). Forty four spoken samples (22 *l*-initial words + 22 *r*-initial words) were repeated two times and randomized within the same block, totaling 88 presentations. The protocol provided a computer screen showing 2 icons with target *l* or *r* inside.

Each listener in the training group of 31 subjects and in the control group of 20 subjects heard the randomized stimuli via a PC over a headphone and was forced to click on either of the *l/r* icons. After each click, a 500 ms pause was given before the next presentation. If a listener made a wrong click or changed his/her clicked decision, s/he could go back and make a readjustment click after listening to the previous presentation again. Furthermore, listeners were allowed to listen to the stimulus up to three times and the sound volume of the headphone was freely adjustable for comfortable listening.

2.3.2 Training

An *l/r* minimal pair identification protocol with feedback was built in Alvin 1.5 (Hillenbrand & Gayvert 2005). Seventy four spoken samples (37 *l*-initial words and 37 *r*-initial words) were randomized and repeated two times within the same block, totaling 148 presentations.

Each listener in the training group of 31 heard the randomized presentations of *l/r*-initial words via a PC over a headphone and was forced to click on either of the *l/r* icons on a computer screen. After each click, a feedback was given by blinking the correct answer. A 500 ms pause was given between the click and the following presentation. Furthermore, listeners could go back to listen to the previous presentation for practice. The training session for a day consisted of two repetitions of the same training protocol (148*2=296 presentations) and the listeners in the training group had five training sessions for 5 days (296*5=1480 presentations). One day's session took about less than 30 minutes.

3. Results

3.1 Pre-training

The mean accuracies of the control group and the training group were

70.68% (s.d.=13.95) and 71.77% (s.d.=13.08) for *l* identification and 72.84% (s.d.=14.55) and 72.65% (s.d.=15.05) for *r* identification. The accuracy differences both in *l* and *r* identification between the two groups were not significant (t-test: T(49)=-0.284, p=0.778>0.1 (two-sided) for *l* identification; T(49)=0.044, p=0.965>0.1 (two-sided) for *r* identification). This means that the two groups might be equivalent in performance in *l* and *r* identification.

3.2 Training

The training group completed two blocks of *l/r* identification training everyday for five days whereas the control group did not during the same period. The performance development in *l/r* identification during the training is not illustrated here. As learners were given icon-flashing feedback, they could listen to a wrong clicked presentation again for practice at his/her own will and make an adjustment click. Therefore, each result of the training blocks did not represent the performance of the learner.

3.3 Post-training results

The mean accuracies in *l* identification at posttest were 84.02% (s.d.= 14.04) for the training group and 73.86% (s.d.=12.57) for the control group, showing that the training group outperformed the control group by 10%. The training group had achieved a steep post-training performance improvement in *l* identification whereas the control group tendered almost no post-training improvement at posttest. As for *r* identification, the mean accuracies at posttest were 83.72% (s.d.=12.94) for the training group and 72.73% (s.d.=15.61) for the control group, showing that the training group outperformed the control by 11%.

4. Discussion

4.1 A response bias problem and the Signal Detection Theory

When a learner's ability to differentiate between *l* and *r* is to be evaluated, response bias to favor either *l* or *r* should always be considered for better evaluation. In Logan et al. (1991) and Lively et al. (1993), the mean value of *l/r* accuracies were used for the evaluation of subjects' perception performance. However, we are going to show that when a learner's ability of *l/r* perception is to be modeled, response bias should necessarily be filtered out. Otherwise, learners' ability could not be correctly evaluated.

The Signal Detection Theory (henceforth, SDT) (Harvey 1992, Egan 1975, MacMillan & Creelman 2005, and Wickens 2001) can address this problem. The perceptual decision-making in an alternative forced-choice *l/r* identification test takes place in the presence of some perceptual uncertainty due to categorical perceptual confusion between the two categories in the variable *l/r*-initial minimal pair stimuli in the following different vowel contexts which are naturally produced by multiple speakers.

Let us suppose that the x axis refers to the one-dimensional perceptual scale: perceptual strength for *l* decision increases rightward whereas perceptual strength for *r* decision increases leftward. Now a learner's perceptual judgment occurs at a certain point on the perceptual scale when *l/r* is presented. Since all *l* stimuli, for example, are followed by different vowels and produced by different speakers, they are perceptually variable to Korean L2 learners. The probabilities of learners' perceiving the strength of *l* and giving *l* responses when variable *l* stimuli are presented, will vary since they depend on their own perceptual ability out of their experience with the phone. Namely, not all *l* stimuli are identified as *l* by Korean learners. This is also the case with *r* stimuli. Therefore, it can be assumed that the probability distribution of *l* responses is formed separately from that of *r* responses along the x axis and the two distributions might be overlapping due to Korean learners' perceptual confusion between the two categories. Less perceptual difficulty means less overlapping of the two distributions; more perceptual difficulty more overlapping. Then sensitivity resulting from the *l/r* identification test can be interpreted as one-dimensional perceptual distance between the two phones.

Under this assumption, we can reinterpret the dichotomy of "Signal-present" and "Signal-absent" in original SDT as of *l*-present and *r*-present, respectively. The following is the decision matrix derived from a two-alternative forced-choice identification test with the same number of *l* and *r* stimuli:

Table 1. *l/r* decision matrix

		Responses	
		l	*r*
Stimuli	*l*	**Hit Rate$_l$ (Correct identification *l*)**	**False Alarm Rate$_r$ (*r* for *l*)**
	r	**False Alarm Rate$_l$ (*l* for *r*)**	**Hit Rate$_r$ (Correct identification *r*)**

The Hit Rate$_l$ (Correct identification *l*, hereafter HR$_l$) is the proportion of *l* responses given for *l* presentations and the false alarm rate$_l$ (*l* for *r*, hereafter FAR$_l$) is the proportion of incorrect *l* responses when *r* presentations are given. Now, when two learners' identification ability on *l/r* is evaluated, the measures of their respective *l/r*-accuracy rates may not be robust measures of their perception ability of *l/r*. Consider the following table:

Table 2. Example results of an *l/r* identification test

learner	**HR_l** (Correct identification *l*)	**HR_r** (Correct identification *r*)	**FAR_l**	**FAR_r**	**Mean of HR_l and HR_r**
A	.70	.95	.05	0.3	0.825
B	.825	.825	.175	0.175	0.825

When two learners' *l/r*-responses are compared, learner B has a higher score in HR_l (0.825>0.7) but lower score in HR_r (0.825<0.95) than learner A. Notice that FAR_l values show that learner B (0.175>0.05) is more willing to click on *l* whereas learner A (0.3>0.175) more willing to click on *r*. Learner B gets a high HR_l at the expense of a relatively higher FAR_l whereas learner A achieves a higher HR_r at the expense of a relatively higher FAR_r. The mean value of HR_l and HR_r (hereafter, MHH) is directly under the influence of learners' bias (the learners' delicate preference for either *l* or *r*) and cannot constitute a robust measure of learners' differentiating ability between *l* and *r*. If MHH were adopted, the two learner's performance would be evaluated to be the same. However, SDT turns out a different result: Learner A is better than learner B. It considers the relationship between HR_l and FAR_l for learners A and B.

Table 3. Decision matrixes for two learners

Learner A		responses *l*	responses *r*	total	Learner B		responses *l*	responses *r*	total
input	*l*	.70	.03	1	input	*l*	.825	.175	1
	r	.05	.95	1		*r*	.175	.825	1

In SDT, a learner's performance in *l/r* identification can be measured by sensitivity (or d'), which is a bias-free measurement of how sensitive a learner is to the difference between *l* and *r* presentations. It measures how many standard deviations there are between the means of the Gaussian probability distributions[1] of *l* and *r* responses[2]. If a learner is more

[1] We are assuming that the resulting rates for *l* and *r* responses from the identification test by the same subject are deterministic, which means that if the test is retaken by the same subject, the resulting response distributions will be the same.

[2] Sensitivity (d') is defined as the standardized difference between the means of False Alarm Rates and Hit Rates: d'=z(FAR) - z(HR). z(FAR) and z(HR) are the z scores corresponding to the right-tail p-values represented by FAR and HR, respectively.

sensitive to the difference between *l* and *r*, the difference between the means of the two probability distributions will become bigger. Sensitivity turns out that learner A's identification performance is better than learner B's: d'=2.17 (learner A) vs. d'=1.87 (learner B).

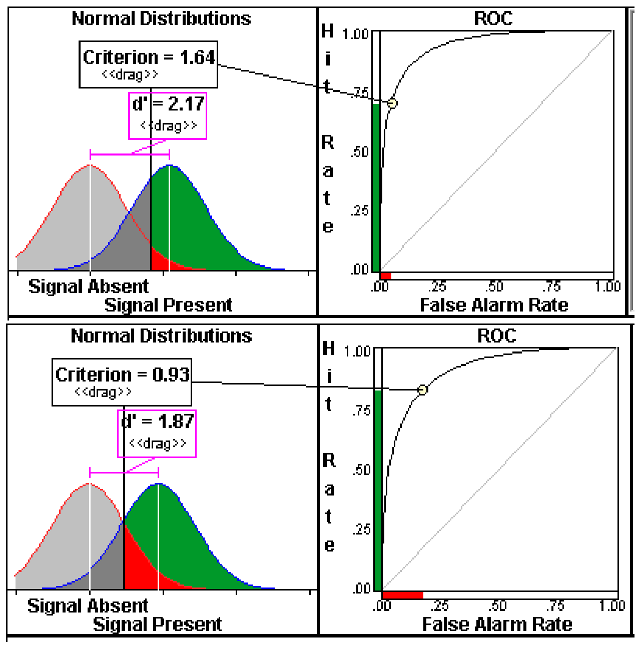

Figure 1. ROC graphs for learner A (top) and for learner B (bottom); "Signal Absent" and "Signal Present" refer to *r* responses and *l* responses, respectively (Berger 2006).

The more a learner's performance improves, the higher the sensitivity value becomes. And the resulting ROC (Receiver Operating Characteristic) curve[3] moves upper leftward, getting more bowed out toward the upper left corner. The ROC curve for learner A is more bowed out than that for learner B when the two graphs are compared. If perfect identification occurs (e.g. HR_l and HR_r assumed to be both 99%), the sensitivity value would be about 4.65, which is the maximum. In what follows, sensitivity index will be used as a measurement of a learner's *l/r* identification performance.

When sensitivity values are plotted as a function of mean accuracy values (MHH) from the all resulting data, the relationship is characterized as quadratic fit lines for both pretest ($F(2, 48)=2120.53$, $p<0.001$, $r^2=0.989$),

[3] ROC curve is a graphical plot of the hit rates as a function of false alarm rates, as shown in the right-hand graphs in figure 1.

and posttest ($F(2, 48)=1577.84$, $p<0.001$, $r^2=0.985$) rather than as linear fit lines (pretest: $F(1, 49)=1262.95$, $p<0.001$, $r^2=0.963$; posttest: $F(1, 49)= 647.07$, $p<0.001$, $r^2=0.930$). This quadratic relation is due to the fact that MHH is seriously affected by bias whereas sensitivity is not.

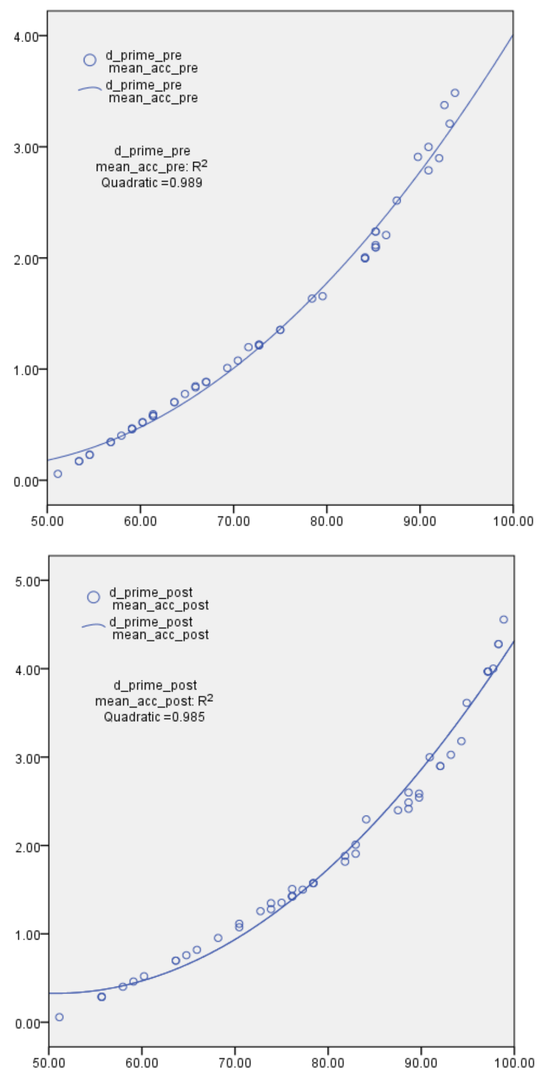

Figure 2. Sensitivity values as a function of mean accuracy rates for pretest (top) and posttest (bottom)

This means that previous studies (Strange & Dittmann 1984, Logan et al. 1991, Hazan et al. 2005) in which mean accuracy rates were used as measurement index for learners' perception performance from forced-choice minimal pair identification tests, were flawed. In SDT, response bias can be represented by the value c, which is defined as the distance between the criterion and the neutral point where neither response is favored ($\beta=1$). If c is 0, no bias; if c is negative, some bias toward *l* responses; if a c value is positive, some bias toward *r* responses. The following plots are c values as a function of mean accuracy values (HMM) for pretest and posttest:

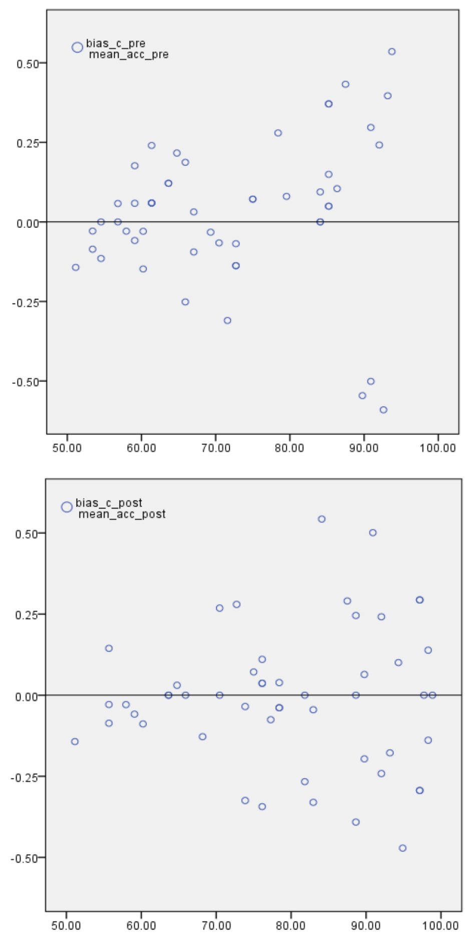

Figure 3. Bias c as a function of mean accuracy (HMM) for pretest (top) and posttest (bottom)

As illustrated, a problem of HMM is that higher HMM is more subject to bias, which distorts the HMM measurement of learner' performance. This means that HMM cannot be a robust measure index of learners' perception performance. Note that sensitivity is not affected by such bias.

4.2 Comparison between the control and the training group

The two ROC plots below show the plots of HR_l of *l* presentations (y-axis) as a function of FAR_l resulting from pretest and posttest. The distance between each plot and the diagonal line represents sensitivity. When the two ROC graphs are compared, it is observed that the plots from posttest results of the training group have shifted to the upper left corner from the pretest plots more than the plots from posttest results of the control group do. The training group learners have achieved higher HR_l along with lower FAR_l after training than the control group learners do. And sensitivity improvement of the training group after training (filled circles) seems to be more drastic than that of the control group (x symbols).

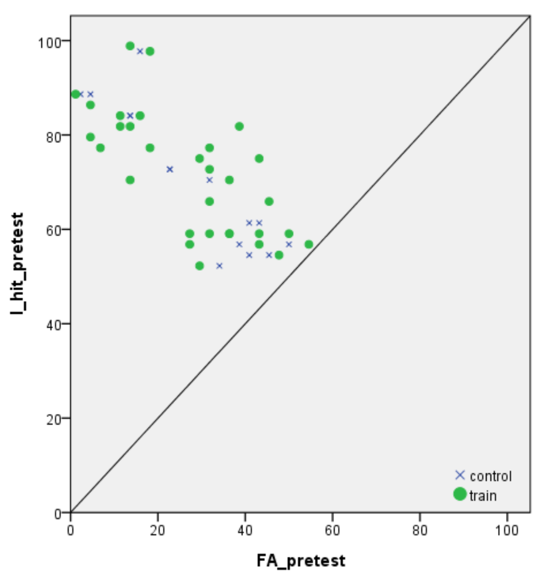

Figure 4. Scatter plots of *HR_l* as a function of FAR_l for pretest

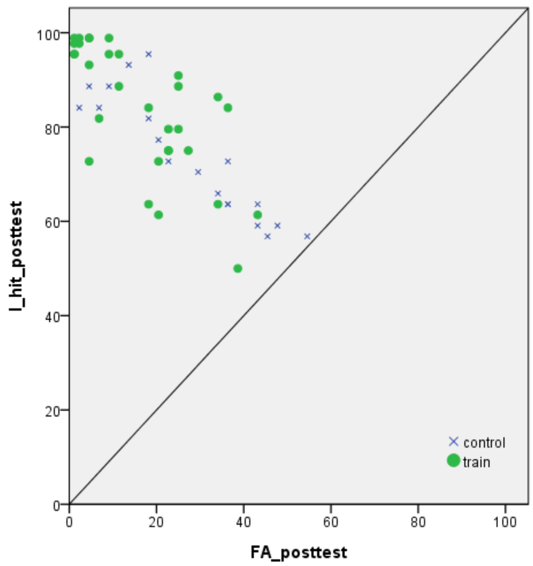

Figure 5. Scatter plots of HR_l as a function of FAR_l for posttest

Sensitivity values are calculated from pretest and posttest results of each learner of the control and the training group. And posttest sensitivity is plotted against pretest sensitivity for the control and the training group. The fit line for the training is positioned well above the fit line for the control, indicating that the training group member have achieved more sensitivity improvement after training than the control.

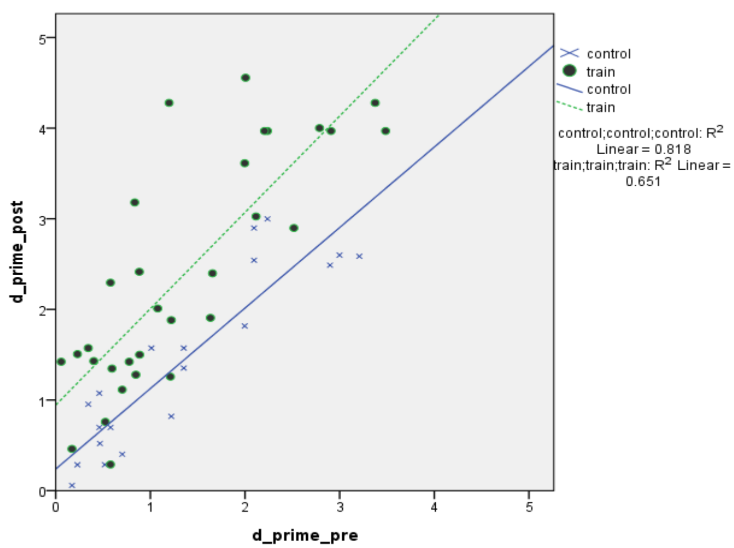

Figure 6. Scatter plots of posttest sensitivity as a function of pretest sensitivity for the control and training group

The mean difference of sensitivity between pretest and posttest were 0.091 for the control group and 1.03 for the training. If the mean sensitivity difference between pretest and posttest for the training group is statistically greater than that for the control, it can be concluded that the five-days-long training was effective. According to a mixed-measures ANOVA, estimated marginal mean sensitivity difference between pretest and posttest was significantly larger for the training group (pretest sensitivity=1.36 (s.e.=0.18) vs. posttest sensitivity=2.39 (s.e.=0.21)) than for the control (pretest sensitivity=1.32 (s.e.=0.22) vs. posttest sensitivity=1.41 (s.e.= 0.26)), as illustrated as an interaction effect between sensitivity difference (posttest d' vs. pretest d') and subject groups (control vs. training): $F(1, 49)=25.368$, $p<0.001$.

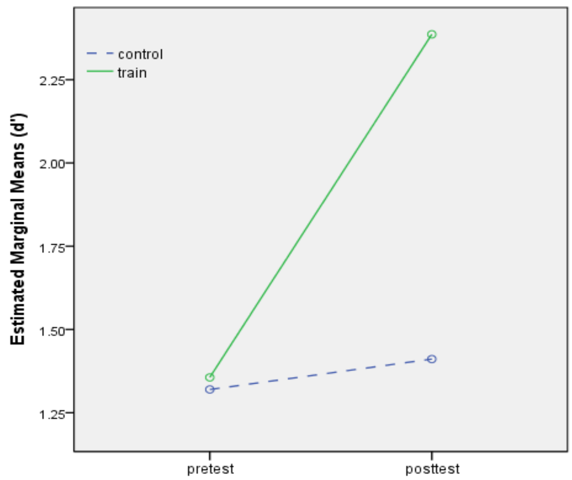

Figure 7. Estimated mean sensitivity improvement comparison between the control and the training group

At pretest, the sensitivity level was almost the same (about 1.3 out of maximum 4.65) for both control and training groups. These pretest figures were disappointing, despite the fact that university-level Korean learners of English had spent more than 6 years on English study. Now, the training group invested about total 150 minutes for five days in minimal pair identification training, listening to each of the *l/r* presentations and clicking on either of *l/r* icons. The results show that the rather short period of training has drastically enhanced their *l/r* identification performance.

4.3 Performance development patterns for the training group learners below 80% at pretest

In this sub-section, the learners in the training and control group who scored below 80% in both HR_l and HR_r at pretest were selected for comparison in perception improvement. Those learners belonging to the below-80 subgroup of the training still struggled to differentiate between *l* and *r* even after training, and their improvement did not seem to be satisfactory compared to the above-80 subgroup of the training. Despite the struggle, if training was effective, sensitivity improvement for the below-80 subgroup of the training is expected to be significantly more than for the below-80 subgroup of the control group.

When the control and training below-80 subgroups are compared in sensitivity, the training subgroup shows better sensitivity improvement than the control subgroup, as is predicted:

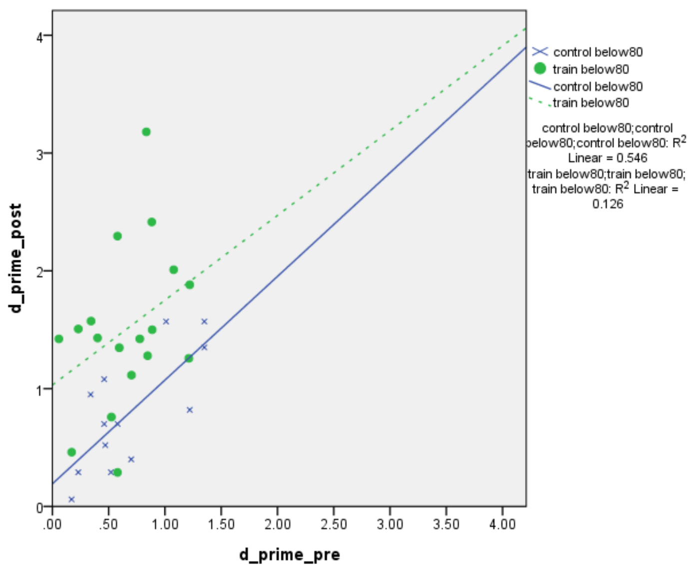

Figure 8. Scatter plots of sensitivity for two below-80 subgroups of the training and the control group

A mixed-measures ANOVA reported a significant interaction effect between the pretest_posttest sensitivity and the train_control variable: $F(1, 29)=23.33$, $p<0.001$. This means that even below-80 members of the training group showed significantly more improvement than those of the control:

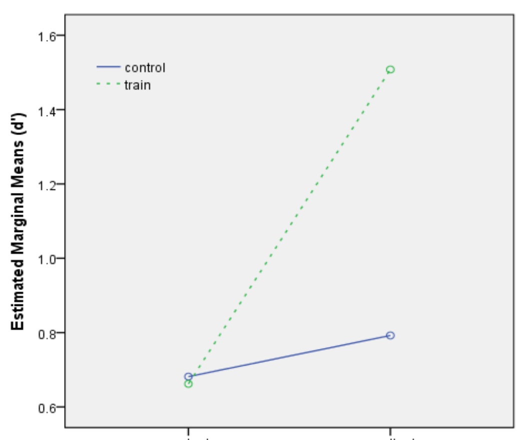

Figure 9. Estimated mean sensitivity improvement comparison between below-80 subgroups of the training and the control

Despite overall post-training performance improvement, the below-80 subgroup of the training has still shown significantly better improvement. This might suggest that further training may help them for better performance. More training beyond the 5-day period might be required for those subjects with poor post-training performance. These subjects might still suffer from a biased judgment and wrong discrimination even after training.

Unfortunately, we could not offer a second round of training to all of the subjects. Only five subjects with different sensitivity achievement volunteered for a second round.

4.4 Post-training2 and results

Five listeners in the training group voluntarily participated in a second round of training, the procedure of which was exactly the same as in the first round: They were forced to complete five training sessions for five days. There was a one week break between the end of posttest1 and the start of the second round of training. A repeated-measures ANOVA was run on the resulting data to check whether an additional round of training might have helped listeners improve their sensitivity.

The sensitivity values of pretest, posttest1, and posttest2 are plotted pair wise: Pretest (x-axis) vs. posttest1 (y-axis) and posttest1 (x-axis) vs. posttest2 (y-axis).

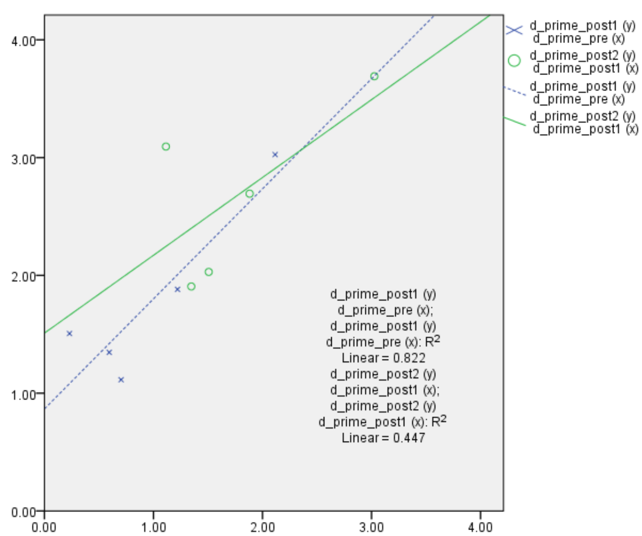

Figure 10. Sensitivity plots between pretest vs. posttest1 and posttest1 vs. posttest2

Statistically significant difference was found in sensitivity values across three tests: Repeated-measures ANOVA $F(2, 8)=33.747$, $p<0.001$. Through pair wise sensitivity difference comparison, sensitivity difference was significant both between pretest and posttest1 ($F(1, 4)=31.270$, $p=0.005$) and between posttest1 and posttest2 ($F(1, 4)=11.086$, $p=0.029$). The second training turned out to be as effective as the first.

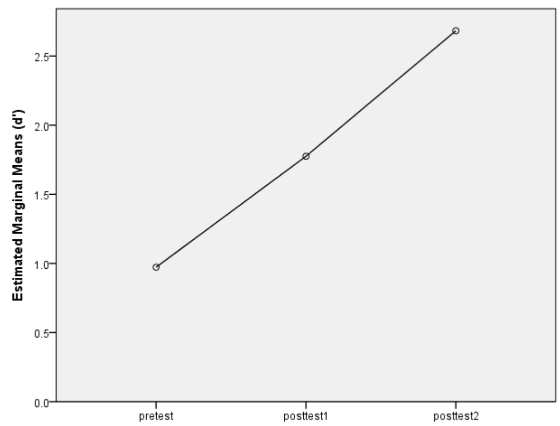

Figure 11. Estimated mean sensitivity across tests

Though the number of the learners is limited to only five, this result suggests that another round of training would still effectively help learners achieve a sensitivity improvement as much as the first round. Due to learners' individual difference in post-training2 achievement, some students with still poor sensitivity might need a third round of training to achieve a 90% or above level.

4.5 Sensitivity-based model for *l/r* identification performance evaluation

The current *l/r* training experiment based on a measure of a sensitivity index can be readily used in school environments. Learners can check their own *l/r* identification ability on the sensitivity scale: Maximum sensitivity =4.65 (perfect identification) and minimum sensitivity=0. Based on this sensitivity scale, a learner's *l/r* identification ability can be easily checked.

Let us consider four learners who went through pretest, training, and posttest. And the results are given as the following graph where the y-axis refers to sensitivity.

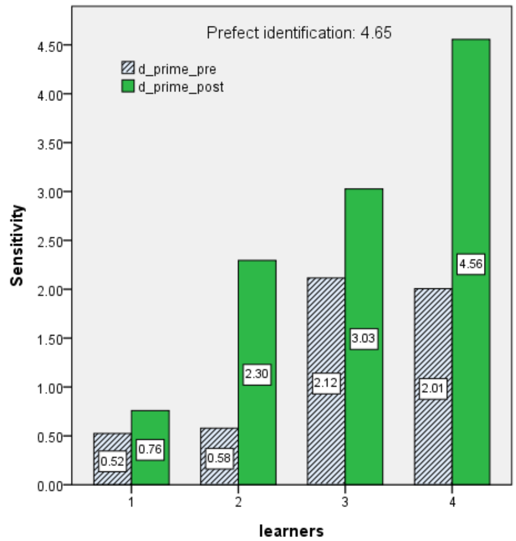

Figure 12. Sensitivity represented as bars for pretest and posttest across four learners

The two bars for each learner show the sensitivity values of pretest and posttest. The graph will let learners see their sensitivity values from both pretest and posttest. Learner1 and learner2 got almost the same sensitivity values at pretest. After training, however, learner2's sensitivity improved from 0.58 to 2.30 whereas learner1' sensitivity improved to mere 0.76. Though learner3 and learner4 had almost the same sensitivity value at pretest, learner 4 has enjoyed a drastic sensitivity improvement to 4.56, which is close to the maximum 4.65.

5. Conclusion

In this paper, a forced-choice minimal pair identification testing protocol was devised to test Korean learners' perception of English word-initial *l/r*, using naturally spoken tokens produced by multiple native speakers. And a forced-choice minimal pair identification protocol with feedback was used to train Korean learners on *l/r*. After a five-days-long training period, pretest and posttest results were compared based on sensitivity in SDT. It turned out that the minimal pair training was quite effective. Furthermore, it was proposed that sensitivity be a more robust measurement index to correctly evaluate L2 learners' *l/r* perception ability than the mean accuracy values used in Logan et al. (1991) and Lively et al. (1993).

The testing and training protocols on English *l/r* used in the current experiment make an independent program within Alvin 1.5 for Windows environments and can be used individually at home or in a school environment, which may help Korean students enhance their ability of perceiving the difference between word-initial *l* and *r*. However, the testing and training procedures should be strictly monitored by teachers or parents for a better result. Our experience says that the testing and training procedures are so boring and tedious that Korean students without motivation may drop out more easily than expected.

The current experiment has a crucial limitation in that only word-initial *l/r* perception training was tried due to the limited time offered by the participants. However, the proposed protocols and SDT-based sensitivity index can be extended for a full perception experiment which may additionally include intervocalic *l/r* and syllable- or word-final *l/r*.

REFERENCES

BERGER, DALE. 2006. Signal detection applet. http://wise.cgu.edu/sdtmod/signal_applet.asp. Claremont Graduate University.

BEST, CATHERINE T. and WINFRED STRANGE. 1992. Effects of phonetic factors on cross-language perception of approximants. *Journal of Phonetics* 20, 305-330.

BRADLOW, ANN R., DAVID B. PISONI, REIKO AKAHANE-YAMADA, and YOH'ICHI TOHKURA. 1997. Training the Japanese listener to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *JASA* 101, 2299-2310.

BRADLOW, ANN R., REIKO AKAHANE-YAMADA, DAVID B. PISONI, and Yoh'ichi TOHKURA. 1999. Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception and Psychophysics* 61, 977-985.

Collins Cobuild English Dictionary for Advanced Learners: CD-Rom. 2001. 4th Ed. Glasgow: Harper Collins Publishers.

E4U CD-Rom Dictionary. Seoul: YBM Sisa Co.

EGAN, JAMES. P. 1975. *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

FLEGE, JAMES E. and WIEKE EEFTING. 1987. Production and perception of English stops by native Spanish speakers. *Journal of Phonetics* 15, 67-83.

HARDISON, DEBRA M. 1997. Bimodal input in second-language speech: Focus on /r/ and /l/. In Leather, J and A. James (eds.). *New Sounds 97: Proceedings of the Third International Symposium on the Acquisition of Second-Language Speech*, 125-134. University of Klagenfurt, Austria.

HARDISON, DEBRA M. 2003. Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics* 24, 495-522.

HARVEY, LEWIS. O., JR. 1992. The critical operating characteristic and the evaluation of expert judgment. *Organizational Behavior & Human Decision Processes* 53.2, 229-251.

HAZAN, VALERIE, ANKE SENNEMA, MIDORI IBA, and ANDREW FAULKNER. 2005. Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication* 47, 360-378.

HILLENBRAND, JAMES and ROBERT GAYVERT. 2005. Open source software for experiment design and control. *Journal of Speech, Language, and Hearing Research* 48, 45-60.

HONG, SOONHYUN. 2009. Training Korean listeners to perceive American English fricatives and affricates: A listener-customized adaptive approach. Studies *of Phonetics, Phonology and Morphology* 15.1, 147-170. The Phonology-Morphology Circle of Korea.

INGRAM, JOHN C. and SEE-GYOON PARK. 1998. Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *JASA* 103.2, 1161-1174.

LAMBACHER, STEPHEN, WILLIAM MARTENS, and GARRY MOLHOLT. 2005. The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics* 26, 227-247.

LIVELY, SCOTT E., JOHN S. LOGAN, and DAVID B. PISONI. 1993. Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *JASA* 94, 1242-1255.

LOGAN, JOHN S., SCOTT E. LIVELY, and DAVID B. PISONI. 1991. Training Japanese listeners to identify English /r/ and /l/: A first report. *JASA* 89, 874-886.

MacMillan English Dictionary: CD-Rom. 2002. Macmillan Education.

MACMILLAN, NEIL A. and DOUGLAS C. CREELMAN. 2005. *Detection Theory: A User's Guide.* 2nd ed. London: Lawrence Erlbaum Association.

NISHI, KANAE and DIANE KEWLEY-PORT. 2007. Training Japanese Listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research* 50, 1496-1509.

PARK, SOON-AE. 2008. A study on the perception of English /l/ and /r/ by Korean elementary, junior high, and high school students. *Language and Linguistics* 42, 73-92. Hankook Univeristy of Foreign Studies.

STRANGE, WINIFRED and SIBYLLA DITTMANN. 1984. Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics* 36, 131-145.

WICKENS, THOMAS D. 2001. *Elementary Signal Detection Theory*. Oxford; New York: Oxford University Press.

YU, KOO AND DONALD G. JAMIESON. 1993. Training of the English /r/ and

/l/ speech contrast in Korean listeners. *Canadian Acoustics* 21, 107-108.
Yahoo English-Korean Dictionary. http://kr.dic.yahoo.com/search/eng/.
YAMADA, REIKO A. 1995. Age effect on acquisition of non-native phonemes: Perception of English /r/ and /l/ for native speakers of Japanese. In Winfred Strange (ed.). *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, 305-320. Baltimore: York Press.

Soonhyun Hong
Department of English Language and Literature
Inha University
253 Yonghyun-dong, Nam-gu, Incheon, Korea
e-mail: shong@inha.ac.kr

received: June 25, 2010
accepted: August 7, 2010