Massaro, D. W. & Light (2003). Read My Tongue Movements: Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/. In *Proceedings of Eurospeech (Interspeech), 8th European Conference on Speech Communication and Technology.* Geneva, Switzerland.

Read My Tongue Movements: Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/

Dominic W. Massaro and Joanna Light

University of California, Santa Cruz
Santa Cruz, CA 95060 U.S.A
massaro@fuzzy.ucsc.edu

Abstract

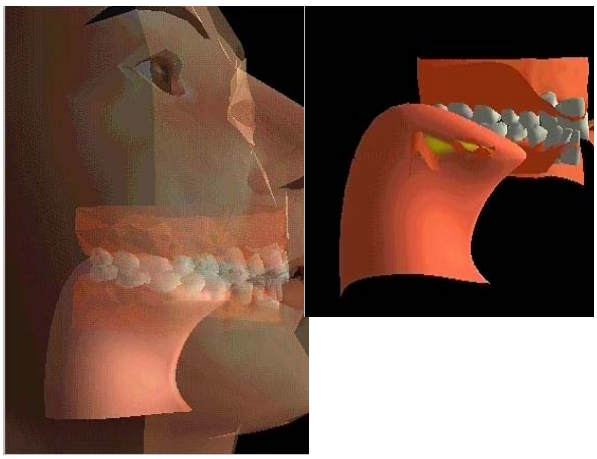
This study investigated the effectiveness of Baldi for teaching non-native phonetic contrasts, by comparing instruction illustrating the internal articulatory processes of the oral cavity versus instruction providing just the normal view of the tutor's face. Eleven Japanese speakers of English as a second language were bimodally trained under both instruction methods to identify and produce American English /r/ and /l/ in a within-subject design. Speech identification and production improved under both training methods although training with a view of the internal articulators did not show an additional benefit. A generalization test showed that this learning transferred to the production of new words.

1. Introduction

All humans have the unique ability to acquire the phonological system of a first language with ease; however, once that phonetic system is established, it is challenging to acquire the phonetic system of a subsequent language. Part of this difficulty stems from the fact that different languages utilize different subsets of phonetic contrasts and show subtle differences within the same phonetic category. Because there is not a universal mapping between phonological features and phonetic parameters (Lindau & Ladefoged, 1986), speech production and perception reflect strong influences of the phonological system of a person's first language.

One of the most well-documented cases of difficulty with a second language is the English contrast /r/ vs. /l/ by speakers of Japanese. This limitation in discrimination and production most likely reflects the lack of a contrast between /r/ and /l/ in Japanese phonology, which causes them to poorly discriminate and produce the /r-l/ contrast in English. Numerous studies have shown that discrimination of non-native contrasts can be improved with auditory training (Lively, Logan & Pisoni, 1993; Werker & Logan, 1985). Furthermore, Hardison (2002) found somewhat better learning of /r/ and /l/ by Japanese and Korean speakers when training involved a frontal view of the talker than simply auditory speech. The bimodal advantage for identification performance was larger for the more difficult test items. There was also an indication that bimodal training improved speech production more than auditory training but it was not clear whether these differences were significantly different. Extending Hardison's study, we test the hypothesis that both perception and production of these segments can be improved with bimodal speech training in which movement of the internal articulators is illustrated. Baldi, our computer-animated talking head aligned with auditory speech, is more capable than a human in demonstrating articulatory processes (Massaro, 1998). The skin of our talking head can be made transparent or eliminated so that the inside of the vocal tract is visible, or a cutaway view of the head along the sagittal plane can be shown (see Figure 1). The inside articulators can be displayed from different vantage points so that the subtleties of articulation can be optimally visualized as well. There is also highlighting (by changing color) of the areas where the tongue hits the palate and teeth. This study tests whether instruction revealing the internal articulatory processes of the oral cavity is more effective than instruction with just a normal frontal view of the tutor's face. It is hypothesized that the unique properties of our program would help Japanese native speakers perceive and produce English speech more accurately. Other issues that this report addresses include: 1) how learning occurs over the course of the study and 2) how learning differs for different minimal word pairs.

The type of training method employed on each day varied (inside articulators present (A) vs. no inside articulators present (NA)). The order of training was dependent on the participant's group (A-NA vs. NA-A). One group with 5 participants (A-NA) received training involving visible internal articulatory movements for the first half of the study (days 1-3), and no visible internal articulatory movements for the last half of the study (days 4-6); the other group with 6 participants (NA-A) received the opposite training sequence, with no visible internal articulation training during the first half of the study and visible internal articulation training for the last half of the study. We expected that the A training method to give better performance accuracy than the NA training method for both



halves of the study.

2. Method

The effects of training were assessed using a pretest posttest procedure closely modeled after the method used by Strange and Dittmann (1984). Each participant's performance was analyzed by looking at performance before and after a given training procedure on each day.

2.1. Participants

The eleven native speakers of Japanese (3 male, 8 female) who completed the training ranged in age from 19 to 36 and had resided in the United States between 3 months and 6 years at the time of testing. Their experience with and mastery of English varied considerably, but all reported that they had difficulty perceiving and producing the English phonemes /r/ and /l/ and that they were eager to improve their English language skills. Each participant received both types of training.

2.2. Procedure

All testing and training was administered individually in a quiet room on a desk and a computer. All stimuli were spoken by a computer-animated talking head (Baldi) presented visually on the computer monitor and binaurally over headphones. The auditory speech aligned with Baldi's articulation was produced by the Festival text to speech synthesis system (Black, Taylor, Caley & Clark, http://www.cstr.ed.ac.uk/projects/festival/), as implemented in the Center for Spoken Language Understanding (CSLU) speech toolkit (http://cslu.cse.ogi.edu/toolkit/). The speaking rate was 153 words/minute and the pitch was 120Hz with a pitch range of 20 Hz. The training program was designed using the rapid application developer (RAD) in the CSLU toolkit. Each participant completed two training days a week, over the course of three successive weeks. The general format of the sessions remained constant. The only manipulated variable was the type of training method employed on each day (inside articulators present (A) vs. no inside articulators present (NA)). The sequence of training sessions was dependent on the participant's group (A-NA vs. NA-A).

2.3. Generalization Test (GT).

At the beginning and at the end of the experiment, sixteen minimal pairs of test words, contrasting /r/ and /l/, were drawn from Sheldon and Strange's study (1982). A participant was presented with the written text of each target word and was required to read the word aloud after a tone sounded. No attempt to correct or alter participants' pronunciation of the /r/ and /l/ phonemes in the target words was made. Each participant recorded 32 stimulus items in all – one utterance of each of the 16 pairs. On each day, the following procedure was carried out.

2.4. Pre-Tests

Table 1 shows the testing and tutoring carried out on each day.

Identification (ID). The Identification test consisted of 3 pairs of words, contrasting /r/ and /l/ in 3 contexts – word initial prevocalic long (right/light), word initial prevocalic short (rip/lip), and word initial prevocalic round, stop consonant + liquid cluster (grew/glue). Previous research has indicated that initial consonant clusters are the hardest for Japanese speakers to identify and produce, as well as the most resistant to training (e.g. Sheldon & Strange, 1982). Sixteen instances of the same minimal pair of words (8 instances of each word in the pair) were presented during each task (3 x 16 stimuli were presented in total). To motivate the trainees, feedback in the form of schematic faces (emoticons) was given after each trial. A happy face or sad face was presented after correct or incorrect answers, respectively. The next trial was presented 1 second after feedback was given.

Table 1. The sequence of tests and tutoring, which occurred on each of the three days of the two training conditions. ID = identification; PRO = production.

ID Pre-Test	3 word pairs	8 repetitions
PRO Pre-Test	1st word pair	4 repetitions
Tutoring	1st word pair	4 practice trials
PRO Post-Test	1st word pair	4 repetitions
PRO Pre-Test	2nd word pair	4 repetitions
Tutoring	2nd word pair	4 practice trials
PRO Post-Test	2nd word pair	4 repetitions
PRO Pre-Test	3rd word pair	4 repetitions
Tutoring	3rd word pair	4 practice trials
PRO Post-Test	3rd word pair	4 repetitions
ID Post-Test	3 word pairs	8 repetitions

Each of the three pairs of words was tested and then tutored. Both groups were presented with an isolated word from each minimal pair. After Baldi said each word, the participant was asked to repeat the word after they heard the tone. Two seconds after the tone, if a response could not be detected, Baldi would ask the participant to "please speak after the tone". The production ability of the participant (correct vs. incorrect response) was determined by a speech recognition algorithm in the CSLU toolkit (http://cslu.cse.ogi.edu/toolkit/). Feedback was given after each trial, by presenting a happy or sad face depending on whether the system correctly recognized the participant's production. The next trial was presented one second after feedback was given. Eight trials of the word pair were completed (4 of each word in the pair). Upon completion of the eight trials, the participant moved on to the tutoring phase for that pair of words.

2.5. Tutoring

Although all participants received both training conditions, the order of the training differed between groups. On each day, participants were trained on all three minimal pairs (right/light, rip/lip, grew/glue) under one of the two training conditions (A or NA).

In the inside articulators (A) condition, Baldi first gave the participant verbal instructions on how to produce the /r/ segment (e.g. where to position the tongue with respect to the teeth, the shape of the tongue and lips, etc.). Baldi then showed the participant how to produce the word in the test pair involving the phoneme /r/, by illustrating a view of the inside of Baldi's oral cavity during his production. Baldi asked the participant to try and produce the word on his/her own but the participant was not given feedback about his/her production ability at this time. The same procedure was carried out for the phoneme /l/.

In teaching the participant how to produce the two segments, four different views were shown: a view from the back of Baldi's head looking in, a side view of Baldi's mouth alone (static and dynamic), a second side view of Baldi's whole face where his skin was transparent, as well as a frontal view of Baldi's face with transparent skin. Each view gave the participant a unique perspective of the activity taking place during production. The order of presentation of the viewpoints was always the same. First, Baldi told the participant that they were about to see a back view of his head, and that they should imagine his oral cavity as though it were his/her own. Baldi produced the word in the pair containing the phoneme /r/ at a reduced speed rate of 63% of the normal duration, he gave the participant helpful tips about tongue positioning, etc. (e.g. "Remember to point your tongue, raise the sides of your tongue, and round your lips"), and he asked the participant to repeat this word back to him. The speech recognition module was used to recognize this articulation and feedback was provided about the participants' production ability via a happy or sad face. The same procedure was carried out for the word in the pair containing the phoneme /l/. Next, Baldi informed the participant that they were about to see the inside of his mouth from a side profile. The same procedure was carried out for this side condition, the second side conditions, and the back view condition except that static, as well as dynamic side views of Baldi's internal oral articulators were revealed. Finally, instruction used a frontal view.

In the no internal articulators (NA) condition, the exact same sequence of oral instruction was presented as in the A condition, but the standard frontal view of Baldi was seen without any views of the oral cavity. Other than this difference of how Baldi was viewed, the training procedure was the same for both groups. The amount of instruction and the practice time was equivalent in the two conditions. Baldi spoke at the same reduced speed of 63% the normal duration and the number of training presentations and tests was exactly the same as it was for the internal articulators (A) condition.

2.6. Post-Tests

Production (PRO). The production test was also administered using the same materials and procedures as the pretest . The tutoring phase for a word pair ended by Baldi saying "Okay, now let's see what you've learned". After each tutoring session was completed, each participant performed the repetition phase once again: Baldi in his normal frontal view would say a word and the participant would have to say it back to him. Feedback was given and after the eight trials were completed, the session ended.

Because the voice recognition system was not always accurate, we decided to evaluate the productions manually once the experiment was completed. Experimenter JL scored each produced word as correct or incorrect, without any knowledge of the experimental conditions. The individual sound files were not labeled for this scoring because the training condition, participant and test (pretest or posttest) were stored in a separate log. The identification of the sound files was carried out only after all scores were noted and the number of correct responses was recorded for each participant.

Identification (ID). After tutoring and repetition testing of the three word pairs was completed, the identification test was administered once again, using the same materials and procedures as for the pretest. The number of correct responses during pretest and posttest was recorded.

3. Results

3.1. Identification

A 2 x 3 x 3 analysis of variance (ANOVA) was carried out, with training condition (A vs. NA), test (Pre-Test vs. Post-Test), day of training within training condition (day 1, 2, or 3), and word pair involved in training (right vs. light, rip vs. lip, grew vs. glue) as the independent factors. Percent correct identification performance was the dependent variable.

Overall performance improved from about 87% to 95% across the 3 days of training within each training condition, $F(2,20) = 20.09$, $p<.001$. Performance was also 4% better in the Post-Test than the Pre-Test, $F(2,20) = 19.74$, $p < .002$, indicating that learning did occur. However, the interaction between test and day, $F(2,20) = 17.29$, $p < .001$, indicated that most of the learning (as measured by the difference between the Pre-Test and the Post-Test) occurred on the first day of a training condition.

Identification varied with respect to the minimal word pair involved, $F(2,20) = 17.09$, $p<.001$. The word pairs that were easiest to identify were the right vs. light and the rip vs. lip contrasts (approximately 95% correct); whereas the grew vs. glue contrast proved to be the most difficult (83%).

The main effect of training condition was not significant, nor was the interaction of training condition and word pair, showing no overall differences between the two training methods.

3.2. Production

The same analysis was done for production performance as for identification. Percent correct production performance was the dependent variable. Overall performance improved from about 54% to 60% across the 3 days of training but was not statistically significant. Performance was 4% better in the Post-Test than the Pre-Test, $F(2,20) = 6.87$, $p < .025$, indicating that learning did occur.

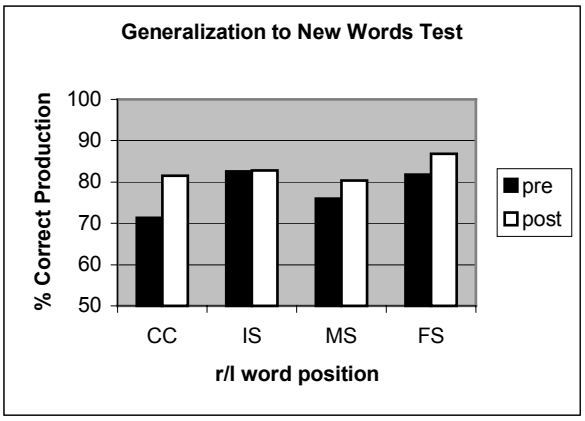
Production varied with respect to the minimal word pair involved, $F(2,20) = 132.32$, $p<.001$. The word pair right vs. light was easiest (84%), the rip vs. lip contrast intermediate (64%), and the grew vs. glue contrast the most difficult (22%).

The A training method did not prove to be more significant than the NA method. Training condition did not interact with Test. The feedback from the participants, on the other hand, seemed to support the value of A training. The NA-A subjects got excited when they were shown the internal visible speech in the second half of the study. The A-NA subjects were not pleased when the inside articulators were no longer revealed. Some expressed that they were just getting the hang of it and that was the best tutor they had ever had.

3.3. Generalization Test (GT)

In this 'generalization to new words' task, ten undergraduate judges rated the production abilities of each subject on each word in the production test, for the pretest of the first day and the posttest of the last day. Ratings ranged on a 5-point scale (1:unintelligible to 5:good/clear pronunciation). The judges

did not know anything about the conditions under which the sound files were made. Sound files for one subject were not clear and this subject had to be omitted from the analysis. A three-way repeated measures analysis of variance (ANOVA) was carried out, with test (pre vs. post), generalization word and group (A-NA vs. NA-A) as the independent factors and rating as the dependent measure. Test and word were the within subject factors, whereas group was the between subject factor. Both groups received higher ratings on the posttest than the pretest $F(1,8) = 12.26$, $p=.008$). Four separate ANOVAs were further carried out, measuring the effects of pretest vs. posttest for /r/ and /l/ word initial consonant cluster position (CC), word initial singleton position (IS), mid word singleton position (MS) and word final singleton position (FS) individually. Figure 2 gives these results. Although word initial consonant cluster position was the only set of word pairs that showed a significant improvement from pretest to post test ($F(1,8) = 6.69$, $p=.032$), performance improved for all four classes.



4. Discussion

This study investigated the effectiveness of our current technology and pedagogy (Massaro, 1998) as a language tutor for the perception and production of non-native phonetic contrasts. Specifically, we assessed whether adding visual information about inside articulatory procedures during production was more effective than simply presenting a normal view of the face. Although the perception and production of words by the Japanese trainees generally improved in both training conditions (A and NA), the training method with a view of the visible articulation (A) was no more effective than a frontal view of the tutor.

Several limitations of this experiment might be responsible for failing to find a difference between the two training methods. Only 11 participants were trained and tested in each condition, two of the three training stimuli had a ceiling effect, and the amount of training (three short sessions under each condition) was relatively short. Furthermore, there were other opportunities for learning (such as the Identification (ID) task that was present in both training conditions. These limitations should be taken into consideration when evaluating the present results and when applying this technology in future applications.

5. Conclusions

The goal of this research was to evaluate a new technique for training adults on the perception and production of non-native speech contrasts by testing whether instruction revealing the internal articulatory processes in the oral cavity was more effective than instruction with just a normal frontal view of the tutor's face. Both types of training were effective but training with our new method showing articulatory processes in the oral cavity did not produce significantly more learning. A generalization test showed that this learning transferred to the production of new words.

6. Acknowledgements

The research and writing of the paper were supported by grants from National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), and Public Health Service (Grant No. PHS R01 DC00236).

7. References

[1] Black, A. W.; Taylor, P.; Caley, R.; & Clark, R. "Festival", http://www.cstr.ed.ac.uk/projects/festival/.

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

[3] Hardison, D. (2002). Sources of variability in the perceptual training of /r/ and /l/: Interaction of adjacent vowel, word position, talkers' visual and acoustic cues. *Proceedings of the 7th International Conference of Spoken Language Processing*, ICSLP - 2002.

[4] Lindau, M. and Ladefoged, P. (1983). Variability of Feature Specifications, 464-479. In J. Perkell, & D. Klatt (Eds.), *Invariance and Variability in Speech Processes. New Jersey: Lawrence Erlbaum Associates, Inc.*

[5] Lively, S.; Logan, J. & Pisoni, D. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America. 94*, 1242-1255.

[6] Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*, 312-319.

[7] Massaro, D. W. (1998b). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, Massachusetts: MIT Press.

[8] Strange, W. & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. Perception and Psychophysics. 36(2), 131-145.

[9] Werker, J. & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37*, 35-44.