Effect of intensive audiovisual perceptual training on the perception and production of the /l/-/r/ contrast for Japanese learners of English

Valerie Hazan, Anke Sennema, Andrew Faulkner

Department of Phonetics and Linguistics, University College London, U.K.
Email: val@phon.ucl.ac.uk

Abstract

This study investigates (a) the extent to which L2 learners use phonetic information from visual cues to improve the perception of a novel phonemic contrast following a course of intensive perceptual training, and (b) the impact of perceptual training on pronunciation. 62 Japanese learners of English were initially tested on their perception of the /l/-/r/ contrast in audio, visual and audiovisual modalities, and then undertook ten sessions of intensive training, before being tested again. Eighteen were trained using auditory stimuli, 25 using natural audiovisual stimuli and 19 using audiovisual stimuli combining a synthetic face with natural speech. All three groups of learners showed a significant benefit of training but learners with audiovisual training did not show greater improvement than listeners with auditory training. The ‘Auditory-training’ group improved their auditory perception of the sounds to a greater extent than the AV-training groups, and the AV (natural face) trainees improved their sensitivity to visual cues to a greater extent than those trained with auditory or AV (synthetic face) stimuli. The learners’ pronunciation of /l/-/r/ consonants improved significantly following perceptual training, with a greater improvement seen for those trained audiovisually with natural stimuli.

1. Introduction

Languages use different sound inventories, and it is known that we attune to the sounds of our own language by the age of one year [1]. As a result, our learning of the sounds of a second language can be effortless. It is strongly influenced by the relation between the phoneme inventories of the first (L1) and second (L2) language, with a tendency to assimilate L2 sound categories to our native sound categories. Research on the effect of intensive auditory training for the acquisition of new sound contrasts has shown positive results, but the number of training sessions involved has often been very high, which limits its practical applications. Also, this work has focused on auditory training alone, and only one study, to our knowledge, has investigated the effectiveness of visual cues in helping the acquisition of new sound contrasts [2]. Our aim was to investigate ways in which the effectiveness of auditory training for language learning could be increased by (a) enriching the information provided to the learner via the provision of speechreading information, and (b) combining successful training techniques with a computer-based approach using a conversational agent to interact with the learner. In our previous paper on initial results of a study of the /l/-/r/ contrast [3], we showed that, prior to training, there was little evidence of audiovisual benefit, i.e. that the difference between intelligibility scores for audiovisual stimuli did not differ significantly from scores for auditory stimuli. There were however significant individual differences in the sensitivity to visual cues by individual learners. Training with audiovisual stimuli was not more beneficial than training with auditory stimuli, either for the whole group, or when only considering learners that had shown sensitivity to visual cues prior to training.

In this study, we extended previous results in two directions. First, the /l/-/r/ study was extended to include learners trained using audiovisual stimuli with a synthetic face, and the effect of intensive perceptual training on pronunciation was investigated in a subset of learners.

2. Study 1: Perceptual training of /l/-/r/ contrast

This study investigated the use of visual cues in the perceptual training of the /l/-/r/ contrast. Three groups of Japanese-L1 learners of English underwent a period of training: one group were trained with stimuli presented audiovisually using a natural face, another group was trained with stimuli in which the natural speech signal was carefully synchronized to variants of a synthetic face (that of the ‘Baldi’ conversational agent developed by [4]) whilst the third group were trained with the same stimuli presented auditorily. For all groups, a conversational agent (Baldi) was used to give instructions and feedback to the learners. The relative effectiveness of training was evaluated by comparing consonant identification in the visual and auditory modalities in the identical pre- and post-training tests.

2.1. Speech materials

2.1.1. Pre and Post-test materials

The two consonants /l/ and /r/ were embedded in initial and medial position in nonsense words in the context of the vowels /i/, a, u/. The consonants were presented as singleton or cluster (with the additional consonant being /k/ and /f/) and appeared in the structure CV, cCV, VCV and VcCV. Three items were produced for each consonant in each syllabic and vowel context (27 initial /l/ and /r/, 27 medial /l/ and /r/), yielding a total of 108 items.

2.1.2. Training materials

For the training sessions a list of 132 minimal pairs of real words containing /l/ or /r/ was compiled and recorded. In the
training list, the sounds /l/-/r/ appeared in different vowel contexts and positions: 100 pairs with consonant in initial position (55 singleton and 45 clustered) and 32 pairs with medial position (28 singleton and 4 clustered).

2.2. Speakers and Recording procedure
To prepare the test items, a female speaker of South Eastern British English was recorded for the pre- and post-test, and two further women and three men for the training materials. Recordings were made to a Canon XL-1 DV camcorder, using a Brul and Kjaer type 4165 microphone. A full-sized image of the speaker’s head was obtained with a fully visible lower jaw drop. The video was digitally transferred to a PC, digitized and down-sampled for editing (250*300 pixels, 25 f/s, audio sampling rate 22.05 kHz). Stimuli were edited so that the start and end frames of each token showed a neutral facial expression. The video appeared on the computer screen in a window of 340 x 290 pixels. For auditory test or training conditions, the audio channel of the audiovisual recordings was used. For audiovisual stimuli with a synthetic face, the speech signal obtained from the audio channel was carefully synchronized with the conversational agent using the ‘Baldi Sync’ software provided as part of the CSLU Toolkit [5].

2.3. Listeners
62 Japanese-L1 subjects participated in the study, 40 of these were students in the English Department of the University of Kochi, and were in Japan whilst the rest were attending a language or phonetics course in the UK. They were approximately at a lower to lower- intermediate level of English proficiency, were aged between 17 and 32 years, had started learning English after the age of 13 and had not lived in the UK for more than 4 months. They reported normal hearing and normal or corrected vision.

2.4. Experimental task
For the pre- and (identical) post-test, a closed-set identification task was built using the CSLU toolkit [5], and a conversational agent [4] was used to explain the task to the listener and to give general feedback on the percentage of correct responses at the end of each section of the test. The items were presented in three conditions (audio alone, visual alone and audiovisual presentation), with two blocks of 108 items per condition. Each listener therefore heard 108 repetitions of each consonant (across vowels and positions) in each test condition. The order of items was randomized within each block, and two orders of presentation of test conditions were counterbalanced across listeners. Items were presented to both ears at a comfortable listening level via headphones.

For the training, the High Variability Phonetic Training [6] procedure was used. This involved the use of multiple items produced by multiple talkers, with the test sounds presented in different syllable positions and vocalic contexts. In the training sessions, feedback was given after every trial: if the response was correct, a ‘smiley’ appeared, if it was incorrect, the word was repeated after a prompt. Listeners were told of the percentage of correct identification achieved at the end of each block.

The pretest was followed by ten sessions of training, each lasting about 40 minutes. In the training, students were first familiarized with the two test consonants uttered by the particular speaker of that session, after which the items were presented either auditorily (A-training group), audiovisually with natural face (AV nat. training group) or audiovisually with a synthetic face (AV synth. training group). Learners were assigned to a training group on the basis of their scores in the Auditory condition of the pretest, with the aim of ensuring a reasonable balance across training groups (A=18 learners, AV nat.= 25 learners; AV synth.=19 learners). The ten sessions were held over a period of two to three weeks. The training program was run individually on laptops and all sessions were carried out under similar conditions, with students working in quiet surroundings, and stimuli presented via headphones and visually on the computer screen. At each training session, listeners were trained on two blocks of test items produced by one of the five speakers: 200 tokens with /l/-/r/ in initial position and 64 items in medial position. The blocks with different positions alternated in order per day. The order of items was randomized within each block for each listener. In days 6 to 10, listeners repeated the sessions 1-5. After the ten days of training, a post-test was done, which was identical to the pre-test.

2.5. Results
The percentage of correct consonant identification in each condition was calculated for the tests carried out pre and post-training (See Figure 1).

An analysis of variance for repeated measures was carried out on the intelligibility scores obtained in the pre- and post-test to evaluate the within-group effect of time of testing, and across-group effect of training mode. Overall, training was effective in improving the perception of the /l/-/r/ contrast \([F(1,59)=172.2; \ p<.001]\). However, the effect of training condition was not significant, with identification increasing on average by 14.9% for both the ‘auditory training’ and ‘AV nat. training’ groups, and by 10% for the ‘AV synth. training’ group. A further analysis of variance was run on the difference in pre- and post-test scores. The within-group effect of test mode was significant \([F(2,118)=45.5; \ p<.000]\); pairwise comparisons with Bonferroni adjustments showed that the increase in intelligibility post-training for the ‘visual-alone’ test mode
was lower than the increase in intelligibility for the ‘audio-alone’ and audiovisual test modes. Also, the training condition by test modality interaction was found to be significant \[F(4,118)=2.72; \ p<0.05\]. This appears to be due to the ‘A training’ group improving their auditory perception of the sounds to a greater extent than the AV-training groups, and the AV (natural face) trainees improving their sensitivity to visual cues (in the visual condition) to a greater extent than those trained with auditory or AV (synthetic face) stimuli.

In order to evaluate the degree to which the effects of training would generalize to new materials, a generalization test was completed by 28 of the learners. This generalization test included 66 real words not heard in the training produced by the talker heard in Training sessions 1 and 6. ANOVAs were carried out to compare scores obtained in pre-test, post-test and generalization in the modality in which learners were trained. Post-hoc tests showed that scores were significantly lower in the pre-test than post-test and generalization but that the generalization scores did not differ from the post-test scores. The training therefore generalized well to new items.

3. Study 2: Effect of perceptual training on speech production

In order to evaluate the impact of perceptual training on the pronunciation of the contrast being trained, learners were recorded before and after the course of perceptual training producing words containing the /l/ and /r/, and their productions were evaluated by 12 native speakers of English.

3.1. Speakers and Listeners

The speakers were 25 of the trainees from Study 1. These had completed the training study in the following conditions: Auditory training (10 students), Audiovisual training with natural face (10 students), and Audiovisual training with synthetic face (5 students).

The listeners were 12 native speakers of British English tested in London. The majority were either phoneticians (PhD, students) or trained EFL teachers. Four were PhD students from another university department who reported some experience with L2 learners, one was a judge in a debating club.

3.2. Speech materials and recording conditions

At the time of the pre- and post-test, the Japanese trainees were asked to read a list of 25 minimal pairs of words which included the sounds /l/ and /r/ in a variety of phonetic environments and positions. For this task, a subset of 20 pre/post tokens per trainee were selected, yielding a total of 500 tokens for 25 trainees. These were normalized for level.

The audio recordings were either made in a soundproof room using a Sony TCD-D10 digital audio recorder with an ECM-959DT microphone, or in case of students tested in Japan, in a classroom using a Sony MZ-N707 minidisk recorder with an ECM-T15 microphone.

3.2.1. Experimental tasks

Two independent perceptual evaluation tests were carried out in which native speakers of British English were asked to judge the Japanese pre- and posttest recordings. The tests included a minimal-pair identification task and a quality rating task. For each test, the listeners evaluated the productions of each Japanese learner, focusing on the /l/-/r/ realizations in the word rather than on a correct pronunciation of the word itself. The experiment was run on a laptop and items were presented to both ears at a comfortable listening level via headphones. The identification test took about 45 minutes, the rating task about 55 minutes.

3.2.2. Minimal-pair identification task

The listeners were asked to identify tokens in a two-alternative forced choice task. On each trial, the minimal pair appeared in writing on two buttons on the screen, listeners heard the token of a Japanese trainee and indicated their choice of the word by pressing the button. The order of the items was randomized across trials.

3.2.3. Consonant rating task

Listeners were asked to judge the realizations of /l/ and /r/. The intended word appeared on the screen, followed by the auditory prompt. Listeners rated the /l/ or /r/ in the word on a scale from 1-7 (1=bad, 7=excellent). The order of presentation was randomized across trials.

3.3. Results

3.3.1. Minimal pair identification test

The percentage of /l/ and /r/ productions that were correctly identified by native listeners is shown in Figure 2. It can be seen that /l/ identification is generally much poorer than /r/ identification, which is due to the presence in the consonant inventory of Japanese of a retroflex flap which is closer to the English /r/ than the English /l/. Analyses of variance on the pre-post training scores revealed that the correct identification of the consonants produced by the L2 learners increased significantly post-training \[F(1,297)=16.8; \ p<.001\]. Evaluations of the effect of training mode were carried out on the scores reflecting the difference in identification scores pre-post training to normalize for differences in pre-training scores. An analysis of variance for repeated measures showed that the effect of training (A, AV nat or AV synt) was significant \[F(2,297)=16.8; \ p=.008\]. Post-hoc tests showed that the difference between the post- and pre-production scores was significantly greater for ‘AV nat. training’ learners than for the ‘auditory training’ learners or for the ‘AV synth. training’ learners (although this might be due to a ceiling effect in the production of /r/ for that group, as can be seen in Figure 2).
3.3.2. Consonant rating task

Ratings given by native listeners to the words produced by the learners before and after training are used in analyses of variance for repeated measures. Ratings were higher for the post-training tokens than the pre-training tokens [F(1, 297)=18.35; p<.001]. Evaluations of the effect of training mode showed that this effect was significant [F(2,297)=5.07; p=.007]; post-hoc tests revealed that ratings improved less for tokens produced by the ‘AV synth.’ training learners than other groups.

4. Discussion

Results of Study 1 showed that there was little advantage in using audiovisual stimuli for the training of the difficult /l/-/r/ contrast for Japanese-L1 learners of English. Although those trained audiovisually with a natural face became more sensitive to visual cues when tested in a visual-alone condition post-training, this sensitivity did not translate to higher scores when stimuli were presented audiovisually, compared to those who had received auditory training or AV training with a natural voice synchronized to a synthetic face. This lack of an audiovisual effect may be due to a number of reasons. First, the visual salience of the /l/-/r/ contrast in British English is relatively low, as native speakers obtained an average score of 72.1% on the lipreading alone task undertaken by the trainees. This explanation is supported by our training results with a more visually-salient contrast (/b,p/l-/v/), which show a small but significant effect of audiovisual training [7]. It is also possible that intensive perceptual training requires a focus on a single modality so that the visual modality may have acted as a distractor, at least for some learners. Finally, the background of the learner might influence the relative effectiveness of audiovisual training as languages with fewer visemes will be less informative visually than languages with a greater number of visemes [8].

A key finding of our study was that the use, in intensive perceptual training, of audiovisual stimuli with a natural face led to a greater improvement in the pronunciation of the difficult consonants compared to that for learners trained with auditorily or AV synth. stimuli. Exposure to the visible articulatory gestures involved in the production of /l/ and /r/ therefore seems to be have been effective, even without active teaching of the articulatory gestures involved in the production of the two consonants. This finding needs to be verified with a wider range of phonemic contrasts and speakers.

5. Conclusions

Results of this study confirmed that intensive perceptual training can result in improvements in the pronunciation of the sounds being trained, even without specific pronunciation training. This is important because computer-based perceptual training programmes are more reliable than computer-based pronunciation training programmes, which need to provide accurate automatic ratings of the learner’s productions. Contrary to what might be expected from speech perception studies with native speakers, audiovisual presentation of stimuli may not automatically produce gains in perception relative to auditory stimuli, especially for speech contrasts that may be visible but not highly salient.

6. Acknowledgements

Funded by grant GR/N11148 from the Engineering and Physical Sciences Research Council of Great Britain. We thank Prof. M. Taniguchi (Kochi University), Jo Thorp (London Bell School) and the UCL Language Centre for their substantial help in organizing the testing of Japanese students.

7. References

[7] Iba, M., Sennema, A., Hazan, V., Faulkner, A. This volume.