THE EFFECT OF VISUAL TRAINING ON THE PERCEPTION OF NON-NATIVE PHONETIC CONTRASTS

Valerie Hazan[1] *and Anke Sennema*[2]

[1]Department of Phonetics and Linguistics, UCL, UK
[2]Institute of Linguistics, University of Potsdam, Germany
v.hazan@ucl.ac.uk, sennema@rz.uni-potsdam.de

ABSTRACT

Auditory and audiovisual training have been shown to be successful in increasing the discriminability of non-native phonetic contrasts in second language learners. The purpose of this study, which trained the English /l/-/r/ contrast with Japanese learners of English, was to investigate training effectiveness using visual stimuli alone. The aim was to test whether training with visual cues leads to (a) increased discriminability of the phonetic contrast, (b) an increase in visual influence in phonetic labelling, and (c) any cross-modal effects in audiovisual or auditory speech perception. Visual-alone training was successful in increasing the discriminability of the /l/-/r/ contrast in visual and audiovisual test conditions but there was no carry over to the auditory condition. There was also evidence of an increase in audiovisual advantage (AV>A) in the training group as a result of training, and of good generalisation to unknown words by the same speakers and to nonsense words by unknown speakers.

Keywords: Second-language speech perception, audiovisual processing, training.

1. INTRODUCTION

It is well known that second language learners have difficulty in acquiring certain phonetic contrasts that occur in the L2 but that do not occur or have a different phonological status in their native language [e.g., 2]. Over the last two decades, it has been shown that programmes of targeted auditory training can increase L2 learners' ability to discriminate such 'difficult' contrasts [e.g., 6], with good evidence of generalisation to new words and new speakers, and good evidence also of long-term retention of the training effects [5].

There has also been increasing interest in the perception of audiovisual speech in L2 speakers. There is a debate as to whether the degree of visual influence in phonetic categorisation is language- or culture-specific, or whether it is a more universal process. In a perception study comparing the auditory and audiovisual perception of non-native stimuli, the degree of 'visual influence' in perception was shown to depend on the language background of the learner and on the visual salience of the phonetic contrast [3]. In a study by the same researchers comparing auditory (A) and audiovisual (AV) training of the /l/-/r/ and /b/-/v/ contrasts in Japanese learners of English [4], sensitivity to visual cues for non-native phonetic contrasts was enhanced via AV perceptual training and AV training was more effective than auditory training when the visual cues to the phonemic contrast were sufficiently salient. Seeing the facial gestures of the speaker also led to a greater improvement in the pronunciation of the contrasts by the L2 learners, even for contrasts with relatively low visual salience. As with many studies of auditory training, this study reported individual variability in the effectiveness of training, and there is the possibility that AV training might not have been maximally effective for some learners, because of the greater cognitive load involved in attending to both the auditory and visual channels.

The aim of this study was therefore to extend these studies by assessing the effectiveness of visual training alone, as it was hypothesised that this would focus the learners' attention on the visual distinction (e.g., tongue movement, lip rounding) between /l/ and /r/ and, as a result, increase visual influence in their phonetic identification of these difficult sounds. Another aim of this study was to see whether there would be any cross-modal effects: would training using lipreading alone lead to the trainees being able to better integrate the auditory and visual information in audiovisual conditions? Following increased attention to articulatory differences between /l/ and

/r/, would there be any cross-modal effect on an auditory alone condition?

2. METHOD

2.1. Participants and recording procedure

A total of 26 Japanese-L1 learners of English as a Foreign Language, resident in Japan, participated in the study: 15 trainees and 11 controls. Participants were attending a two-week English phonetics summer school in an English-speaking country. Learners were approximately at a lower to lower-intermediate level of English proficiency, were aged between 19 and 40 years (median: 20.5 years) and had typically started learning English at school at the age of 12, but with a focus on written language. The majority had never lived in an English-speaking country. A further six students participated in the study but were excluded because their pre-test performance was above 90% in the AV condition.

Two men and one woman, all speakers of South Eastern British English, recorded the items for the pre/post-tests, and two further women and three men recorded items for the training materials and generalisation test. For all speakers, video recordings were made in a soundproof room, with the speaker's head fully visible within the frame. The video and audio channels were digitally transferred to a PC. Video clips were edited so that the start and end frames of each token showed a neutral facial expression. Stimuli were down-sampled post-editing (250*300 pixels, 25 f/s, audio sampling rate 22.05 kHz).

2.2. Speech materials

2.2.1. Pre/post test nonsense word materials

The two consonants /l/ and /r/ were embedded in nonsense words in initial (CV, cCV) and intervocalic (VCV) position in the context of the vowels /i, a, u/. In the cCV stimuli, the consonant clusters were /pl/, /br/, /gl/, /cr/. The test included two repetitions of each syllable produced by each of three speakers for initial singletons (total: 36 stimuli), and one repetition for initial clusters (total: 36 stimuli) and for medial singletons (total: 18 stimuli). The 90 items were randomised and presented in a single block with a pause after 50 trials.

2.2.2. Generalisation materials

The generalisation test included 20 minimal pairs of real words with /l/ and /r/ in word-initial position: ten with /l/-/r/ as singletons and ten as the second consonant in a consonant cluster. Each of the minimal pairs was produced by two different speakers from the set of speakers who produced stimuli for the training sessions (i.e. these were 'known' speakers to the trainees).

2.2.3. Training materials

For the training sessions, materials included 71 minimal pairs of real words: 38 containing /l/ or /r/ as singletons in initial position and 33 containing /l/ or /r/ in initial clusters.

2.3. Experimental task

Test and training programmes, designed using the CSLU toolkit [1], ran individually on desktop computers, with students working in quiet surroundings and stimuli presented via headphones at a comfortable listening level. In the pre-training test session, all participants carried out the nonsense-word test first and a short McGurk test which is not reported here. The post-training test was identical to the pre-test for controls but the trainees additionally did a generalisation test. The pre/post tests took around 40 minutes to complete.

In the pre/post tests, the nonsense-word items described above were presented in three conditions: 'auditory' (A), 'audiovisual' (AV) and 'visual' (V) in a two-alternative forced choice identification task, with response choices of R and L, and no feedback. Two orders of presentation (AV, A, V or A, AV, V) were counterbalanced across participants. Participants responded by clicking on the appropriate letter symbol with a mouse. The test items were identical in all conditions, but in the single-modality tasks, either the auditory or visual channels were removed. Each block of 90 items was presented once in each test condition, yielding a total of 270 responses per participant.

For the trainees, the pre-test was followed by seven sessions of training, each lasting about 40 minutes and carried out within a two- week period. All the training was carried out in a visual-alone condition, with the trainees seeing but not hearing the speaker. The training program was run individually on desktop PCs, with trainees working in quiet surroundings under the supervision of an experimenter. In the training, after each presentation, the trainee had to click on L or R on the screen. If the response was correct, a 'smiley' appeared. If the response was incorrect, a sad face appeared and the video was presented again with the correct label shown. At the end of each block, a bar chart showed the percentage of correct L and R responses, with a message of encouragement or congratulations.

At each training session, listeners first heard two blocks of test items produced by one speaker: 76 'singleton' tokens and 66 'cluster' tokens as described above. After a short pause, a second speaker was presented with again a singleton block followed by a cluster block. Over the seven days of training, trainees saw four of the speakers three times and one of the speakers twice.

Control participants carried out the pre- and post-test only, separated by the same amount of time as the trainees. The controls were also summer school students, so were receiving some English phonetics tuition over the two-week period, but no specific visual training.

3. RESULTS

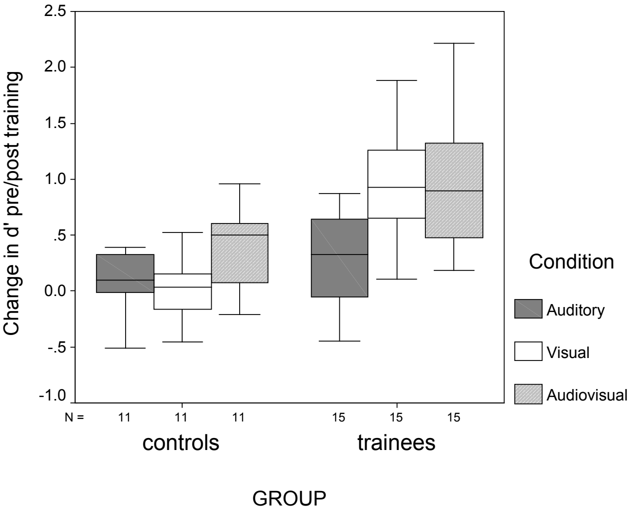
Table 1: Means and standard deviations for the score (d') representing the discriminability of the /l/-/r/ contrast in the pre- and post-test in each test condition (A, AV, V) for the training and control groups.

Group	Cond	Pre-test	Post-test	Diff post-pre
Training	A	0.56	0.84	+0.28
(N=15)		*(0.56)*	*(0.66)*	*(0.51)*
	V	0.18	1.05	+0.87
		(0.54)	*(0.38)*	*(0.60)*
	AV	0.59	1.51	+0.93
		(0.54)	*(0.58)*	*(0.60)*
Controls	A	1.05	1.3	+0.25
(N=11)		*(1.06)*	*(1.16)*	*(0.51)*
	V	0.68	0.68	-0.00
		(0.58)	*(0.55)*	*(0.30)*
	AV	1.04	1.43	+0.39
		(0.90)	*(1.08)*	*(0.38)*

3.1. Pre/post test

The pre/post test involved nonsense words by speakers that were not seen in the training. The percentage of correct consonant identification obtained in each condition was calculated. In the pre-test, over all participants (n=26), mean /l/-/r/ identification was 63.4% (s.d. 13.4) in the A condition, 57.5% (s.d. 11.2) in the V condition and 64.1% (s.d. 12.8) in the AV condition. In order to correct for any potential bias in responses, scores were converted to the signal detectability measure dprime (d'), which is calculated as the z-value of the hit-rate minus that of the false-alarm-rate. First, the pre-test results were examined to look at participants' performance in the different modalities (see Table 1). A repeated-measures ANOVA showed that the difference in discriminability across the A, AV and V conditions was not significant although there was a trend for lower performance in the V condition in the trainee group. There was therefore no evidence of 'audiovisual advantage' (AV>A) in either the trainees or controls. There was a trend for higher pre-test performance in the control group but this did not reach significance.

Figure 1: Box-plots of the score representing the change in discriminability of the /l/-/r/ contrast (d') from the pre- to post-test in each condition (A, V, AV) for the training and control groups.



A repeated-measures ANOVA was carried out on scores obtained in the pre- and post-test to evaluate the within-group effect of time of testing (pre/post) and test condition (A, V, AV) and across-group effect of group (trainees, controls). Overall, the effect of time of testing was significant [F(1,24)= 51.23; p<.0001] and there was a significant time*group interaction with an examination of means suggesting that the trainees improved their discriminability of the /l/-/r/ contrast to a greater extent than controls. The evaluation of training effectiveness was carried out on the difference between pre-test and post-test d' scores so as not to be influenced by any differences in baseline performance levels (see Figure 1). A significantly greater difference in discriminability post-training was obtained for trainees than for controls [F(1, 24)= 14.49; p<0.005]. The effect of test condition was significant [F(2, 48)= 4.64; p<0.02]: there was a greater increase in discriminability for the AV condition than for the A condition. There was a significant test condition by group interaction [F(2, 48)= 5.25; p<0.01]: both groups showed similar improvements for the A condition but trainees showed greater increases in discriminability in both the V and AV conditions.

The data was then specifically examined to see whether there was evidence of an increase in audiovisual advantage (AV>A) as a result of training. For the control group, there was no difference in discriminability between the A and AV conditions in either the pre- or post-test. For the training group, there was no also difference in the pre-test, but, in the post-test, discriminability in the AV condition was significantly higher than in the A condition. There was also a significant difference between performance in the AV and V conditions.

3.2. Generalisation (new words by known speakers)

It should be noted that although the trainees had seen the speakers in the visual-only training, they had never heard their voices prior to the generalisation test, which was run in the usual three test conditions. A repeated-measures ANOVA was run on the set of post-test d' data for trainees (nonsense syllables and words), to look at within-group effects of test condition and effect of test material. The effect of test material was not significant showing that there was good correlation between performance on nonsense words produced by unknown speakers and 'new' words produced by known speakers. The effect of test condition was significant [F(2,28)=13.57; p<0.001] and pairwise comparisons suggested that this was due to performance in the A condition being significantly lower than performance in the AV and V conditions, which did not differ from each other. The interaction between test condition and test material was also significant [F(2,28)=12.35; p<0.001]; this was due to performance in the V condition being higher in the generalisation test with known speakers than in the nonsense-word test with unknown speakers. This suggests that trainees became particularly attuned to the speakers' visible articulatory gestures as a result of the training, even though the effect also generalised to unknown speakers.

4. DISCUSSION

This study shows that purely visual training of the /l/-/r/ contrast was successful in increasing the discriminability of this contrast, despite its relatively low visual salience. These increases were shown not only in the visual alone test condition but also in the audiovisual condition, with evidence of 'audiovisual' benefit (AV>A) in the post-test only for the trainee group. It should be noted that this visual training effect was obtained with Japanese speakers who are generally thought to be relatively insensitive to phonetic visual information in Japanese [7], although they showed greater visual influence with non-native speakers. As for the training in [4], the impact of training modality can be seen on the specific channel trained and in greater AV integration, but with minimal impact on the channel that was not trained. There is therefore little evidence of cross-modal effects of training. Information that was learned about the articulatory gestures characteristic of /l/ and /r/ did not assist the listener when decoding acoustic cues to these contrasts.

4. REFERENCES

[1] Cole, R., Massaro, D.W., de Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P., Tarachow, A., Solcher, D. 1999. New tools for interactive speech and language training: using animated conversational agents in the classroom of profoundly deaf children. *Proc. ITRW on Methods and Tools in Speech Science Education (MATISSE)*, London, 45-52.

[2] Flege, J.E., 1995. Second-language speech learning: theory, findings, and problems. In: Strange, W. (ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*. Baltimore: York Press, 229-273.

[3] Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., Chung, H. 2006. The use of visual cues in the perception of nonnative consonant contrasts. *J. Acoust. Soc. Am.*119, 1740-1751.

[4] Hazan, V., Sennema, A., Iba, M., Faulkner, A. 2005. Effect of audiovisual perceptual training on the perception and production of consonants in Japanese learners of English. *Speech Com*. 47, 360-378.

[5] Lively, S. E., Logan, J. S., Pisoni, D.B. 1994. Training Japanese listeners to identify English /r/ and /l/. III: long-term retention of new phonetic categories. *J. Acoust. Soc. Am.* 96, 2076-2087.

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

[7] Sekiyama, K. and Tohkura, Y. 1993. Inter-language differences in the influence of visual cues in speech perception. *J. Phon.* 21, 427-444