The role of visual cues in L2 consonant perception

Anke Sennema, Valerie Hazan and Andrew Faulkner

UCL, London

E-mail: anke@phon.ucl.ac.uk, v.hazan@phon.ucl.ac.uk, andyf@phon.ucl.ac.uk

ABSTRACT

This study investigates the extent to which L2 learners extract phonetic information from visual cues in the perception of a novel phonemic contrast. 92 Japanese learners of English were tested on their perception of the /l/-/r/ contrast in audio, visual and audio-visual modalities. Overall identification rates in Audio and AV conditions did not differ significantly and few individual listeners showed evidence of AV benefit.

Next, the relative benefit of training the perception of the /l/-/r/ contrast using auditory or audio-visual stimuli was evaluated for a subset of 41 Japanese learners of English. Both groups of listeners showed a significant benefit of training but learners with audio-visual training did not show greater improvement than listeners with auditory training. The learners' natural sensitivity to visual information was taken into account and analyses for a subset of learners revealed that initial visual awareness does not lead to more improvement over audio-visual training.

1. INTRODUCTION

The difficulty that L2 learners encounter in perceiving sound contrasts that do not occur or that have different phonemic status in their native language is well attested. Many studies have now shown that certain training techniques can lead to significant improvements in the perception of these new sound contrasts [e.g., 1]. However, this training is typically long and arduous and does not exploit a source of information that has typically been of great benefit to native listeners in poor listening conditions: visual information.

Two studies carried out at UCL suggested that the effect of visual information on L2 perception was in fact quite weak. In the first [2], Spanish learners of English were tested on their perception of a wide range of English consonants and vowels presented in low-level noise. The consonant confusions that were disambiguated by visual information tended to be those for sound contrasts that occurred both in Spanish and English. Hence, these confusions were likely to be the result of poor acoustic distinctiveness. Consonant confusions that were language-dependent –mostly errors in voicing and manner– were not reduced by the addition of visual cues. Little evidence of AV benefit was also found in a second study focusing on the perception of two contrasts /b/-/v/ and /p/-/b/ by Spanish learners of English [3]. However, conflicting evidence came from a study by Hardison [4], who evaluated the effect of visual cues, context and speaker variability on the perception of the /l/-/r/ contrast in Japanese and Korean learners of American English. Audio-visual training was more effective than auditory training in improving the identification of the /l/-/r/ contrast for these listeners. Prior to training, Hardison also found evidence of better performance when stimuli were presented audio-visually rather than auditorily.

The difference in the effect of visual information found between Hardison's and our studies could be due to differences in the salience of visual information or in the phonological status of sounds across languages. Therefore we wished to explore the sensitivity to auditory and visual cues to the /l/-/r/ contrasts in a large cohort of Japanese learners of British English (Study 1). In Study 2, a subset of learners underwent intensive training in either an auditory or auditory-visual condition to examine the effectiveness of visual information in training. We were especially interested in exploring whether training effectiveness with audio-visual stimuli benefited was dependent on a 'natural' sensitivity to visual cues.

2. STUDY 1

2.1 SPEECH MATERIAL

The two test 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.

2.2 SPEAKERS AND RECORDING PROCEDURE

To prepare the test items a female speaker of South Eastern British English was recorded. Recordings were made to a Canon XL-1 DV camcorder, using a Bruel 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

ISBN 1-876346-48-5

expression. The video appeared on the computer screen in a window of 340 x 290 pixels. 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.

2.3 LISTENERS

Study 1 involved 92 Japanese learners of English. Of these 53 were university students of Kochi University and tested in Japan, 20 students were attending a summer course of Phonetics at UCL, 10 were recruited from a School of English in London and 9 were students of a pre-academic language course at UCL. They were approximately at 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 none had lived in the UK for more than 4 months. They reported normal hearing and normal or corrected vision. A control group of 7 native listeners judged the test items in the two blocks of the video alone condition.

2.4 EXPERIMENTAL TASK

A closed-set identification task was built using the CSLU toolkit [5], and a conversational agent [6] 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 and audio-visual 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. Two orders were used for the presentation of the three conditions: AV, A, V or A, AV, V, and the two orders were counterbalanced across listeners. Items were presented to both ears at a comfortable listening level via headphones.

2.5 RESULTS

The overall identification accuracy in each condition is shown in table 1:

Test mode	Auditory	AV	Visual
% correct	60.6	61.9	55.2
s.d.	12.9	13.6	8.1

Table 1: Percentage of correct /l/ and /r/ identification per test condition for 92 listeners.

The control group of native listeners scored 78.4% correct, s.d. 5.9, in the video alone condition. A repeated-measures analysis of variance examined the within-group effect of test condition and the between-group effect of 'institution'. The effect of 'institution' was not significant showing that listeners tested in the UK did not perform differently from listeners tested in Japan. The effect of test condition was significant [$F(2, 176)=10.20$; $p=0.0001$]. Pairwise comparisons with Bonferroni adjustments showed that /l/-/r/ identification rate was significantly poorer in the 'visual only' condition than in the other two conditions while identification in the AV condition did not differ significantly from performance in the audio condition.

The degree of AV benefit was calculated by evaluating whether performance in the AV condition was outside that expected from the binomial distribution of individual scores in the A condition. As can be seen in Figure 1, 17 out of the 92 subjects (18.5%) showed evidence of a positive AV benefit, 14 (15.2%) were negatively affected by the addition of visual cues and the rest (66.3% of subjects) showed no real effect when visual cues were added.

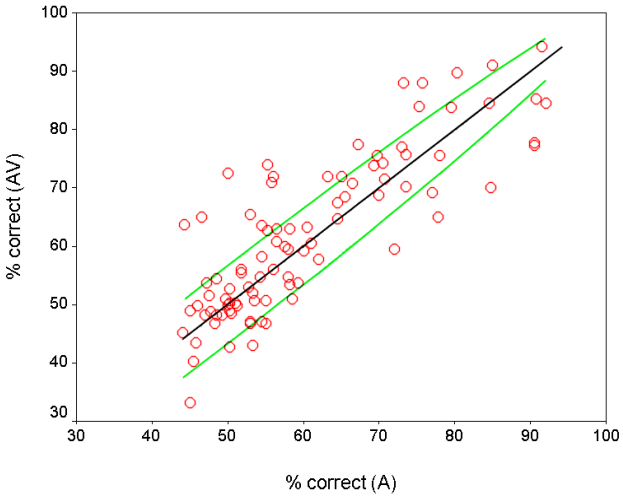

Figure 1: Scatterplot of % correct identification in the A and AV conditions. The middle line is the prediction AV = A, the outer curves are individual 95% confidence limits about the scores in the Audio condition assuming a binomial distribution.

3. STUDY 2: TRAINING

The second stage of our study investigated the use of visual cues in perceptual training of the /l/-/r/ contrast. Two groups of learners from Study 1 underwent a period of training: one group were trained with stimuli presented audio-visually whilst the other group were trained with the same stimuli presented auditorily. The relative effectiveness of training was evaluated in a post-test, in relation to the baseline use of each modality as assessed in the pretest. For learners who were not at chance on the visible speech we were interested in whether they showed an advantage of audio-visual over auditory training.

3.1 Speech material

In the pre- and post-test the same speech material was used as in Study 1 (see 2.1). For the training sessions a list of 132 minimal pairs of the /l/-/r/ contrast (real words) 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).

3.2 Speakers and Recording procedure

Two female talkers and three male talkers of South Eastern British English recorded the training items. Three utterances of each item were recorded. In addition each speaker recorded a token for /l/ and /r/ in initial and medial position which was played as example for the speech sound. The recording procedure was similar to that in Study 1.

3.3 Listeners

A subset of 41 Japanese learners of English who were involved in Study 1 participated on a voluntary basis in the training. Of these 21 were university students of Kochi University and tested in Japan, 9 learners were recruited from a School of English in London, 7 students were attending a summer course of Phonetics at UCL and 4 were students of a pre-academic language course at UCL.

3.4 Experimental task

The High Variability Phonetic Training [1] procedure was used. This involved the use of multiple items with the test sounds presented in different syllable positions and vocalic contexts produced by multiple talkers, with feedback given after each response. In the training sessions, feedback was given after every trial: if the response was correct, a 'smiley' appeared, if it was incorrect, a conversational agent repeated the word after a prompt. Listeners were told of the percentage of correct identification achieved at the end of each training block.

The methodology used for the pre/post-test was as in Study 1. 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 condition) or audio-visually (AV condition). Learners were assigned to each condition on the basis of their scores in the Auditory condition of the pretest, with the aim of ensuring a balance across training groups (A=18 learners, AV= 23 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 surrounding and stimuli presented via headphones and visually on the computer screen. At each training session, listeners either saw and heard, or just heard 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. Each listener therefore heard 132 repetitions of each consonant (across positions) in each session. 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 pretest.

3.5 Results

Figure 2 shows the overall identification accuracy in each test condition (audio, audio-visual and visual presentation) per training modality before and after training. A repeated-measured ANOVA tested the effects of training (pre/post), of training modality (Audio/AV) and of test condition (A, AV, V). As in Study 1, the main effect of test mode was significant [$F(2, 78)=35.2$; $p=0.0001$] and post-hoc tests showed that performance in the V condition was poorer than both other conditions but that there was no significant difference in performance between both training groups.

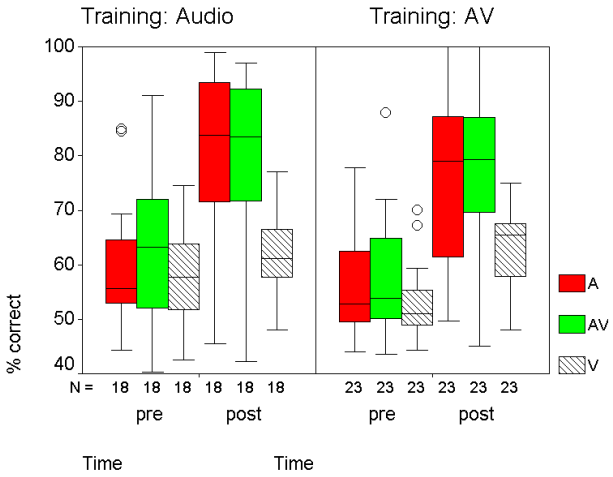

Fig. 2: Box plots of identification scores for three test conditions per training modality for pre- and post-test.

Results indicate a significant effect of training [$F(1, 39)=158.02$; $p=0.0001$] as performance was higher in the post-test than in the pretest, but there was no significant effect of training condition: both training groups improved by about 20% in the Audio and AV modes in the post-test.

There was a significant interaction between test mode and time of testing [$F(2, 78)= 30.8$; $p=0.0001$] and a significant although weak interaction between test mode, time and training group. This three-way interaction is due to training condition affecting performance in the visual alone condition only. AV training led to more improvement in visual-only test scores compared to audio only training. In Audio and AV tests, the training condition did not affect degree of improvement from pretest to post-test. The fact

ISBN 1-876346-48-5

that the AV training group improved in the visual condition even if they did not show evidence of AV benefit (i.e. difference between A and AV identification) is important as it shows that training did have a positive effect on listeners' sensitivity to visual information.

The listeners' natural sensitivity to visual cues is one factor that might account for the limited effect of AV training. Those learners who are at chance on their use of visual cues may not have been able to use the additional information provided in AV training. We therefore estimated the factor 'visual awareness' on the basis of the pretest performance in the Video alone condition (for 'visual aware': scores of >55.6% correct). Figure 3 shows the results for the emerging subgroup of 17 listeners with 'visual awareness' per training modality. The post-test performance for both training groups is on a similar level which suggests, that even those learners who did initially make use of visual cues, do not benefit more from training with audio-visual stimuli.

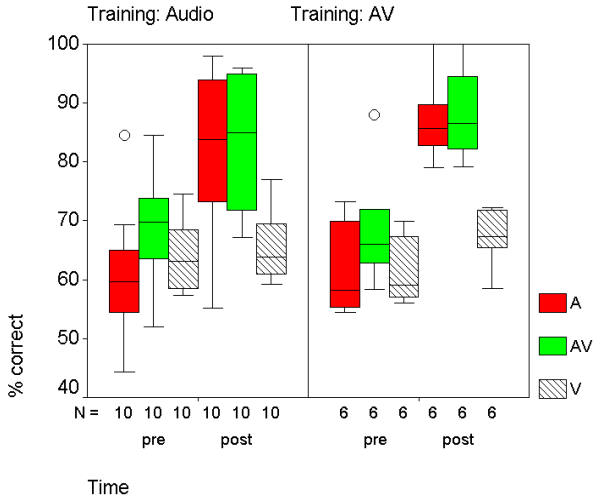

Fig. 3: Boxplots of identification score for the subgroup 'visually aware' in the two training modalities Audio and Audio-visual.

4. DISCUSSION

Our previous work on Spanish learners of English (/b/-/v/, /p/-/b/ distinction) had shown little evidence of L2 AV benefit. This lack of effect has now been replicated with a different population of Japanese L2 learners and a different phonemic contrast (/l/-/r/). Comparison of pre-training performance across modalities showed that only 18.5% showed a significant benefit of AV presentation. These results contradict those of Hardison [4], whose Japanese learners of American English showed evidence of L2 AV benefit. The level of 'visual informativeness' may be a factor in this difference in that the visual contrast between /l/ and /r/ in American English is likely to be greater than in British English.

Results of the training study reveal that there was no significant advantage in training audio-visually over the use of auditory stimuli except for purely visual identification. This was the case even if only those L2 learners who had showed themselves to be 'visually aware' were considered. Their initial advantage of being able to make use of visual cues did not translate into a better performance through audio-visual training.

The lack of AV benefit cannot be attributed to the speaker of the testing material not being 'visually informative', as close to 80% correct identification of /l/-/r/ presented as singletons and clusters was achieved by native listeners in a lipreading alone condition. It is, however, possible that the test speaker produces speech in which auditory cues are dominant over visual ones in AV presentation.

Hence, although the training has not led to the development of an L2 AV benefit, a positive effect on listeners' sensitivity to visual information could be established. The ability to make use of visual cues may be important for the everyday use in a foreign language environment.

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.

REFERENCES

[1] Logan, J.S., Lively, S.E., and Pisoni, D. B. "Training Japanese listeners to identify English /r/ and /l/", J. Ac. Soc. Am., vol. 89, pp. 874-886, 1991.

[2] Ortega-Llebaria, M, Faulkner, A., Hazan, V. "Auditory-visual l2 speech perception: effects of visual cues and acoustic-phonetic context for Spanish learners of English". Speech, Hearing and Language: UCL Work in Progress, vol 13, pp. 39-51, 2001.

[3] Hazan, V., Sennema, A., Faulkner, A. "Audiovisual perception in L2 learners", Proceedings of ICSLP, pp. 1685-1688, 2002.

[4] Hardison, D. Acquisition of second-language speech: Effects of visual cues, context and talker variability. Applied Psycholinguistics, in press.

[5] Cole, R. "Tools for research and education in speech science". Proceedings of the International Conference of Phonetic Sciences, San Francisco, CA, 1999.

[6] Massaro, D.W., and Cole, R. From "Speech is special" to talking heads in language learning. In Integrating Speech Technology in the (Language Learning and Assistive Interface, University of Abertay Dundee, Dundee, Scotland, 2000, 29-30, 153-161.