Use of visual cues in the perception of a labial/labiodental contrast by Spanish-L1 and Japanese-L1 learners of English

Midori Iba¹, Anke Sennema², Valerie Hazan², Andrew Faulkner²

Institute for Language and Culture¹, Department of Phonetics and Linguistics²
Konan University, Japan ¹ University College London, U.K. ²
Email address: midori@center.konan-u.ac.jp

Abstract
This study investigates the extent to which L2 learners with different L1 backgrounds are sensitive to phonetic information contained in the visual cues to a novel phonetic contrast (labial/labiodental contrast), and the degree to which this sensitivity can be increased via intensive training. 36 Spanish-L1 and 47 Japanese-L1 learners of English were initially tested on their perception of the /b/p/-/v/ contrast in audio, visual and audiovisual modalities. There was a clear effect of language background with Spanish-L1 learners showing better performance overall, and much greater sensitivity to visual cues to the contrast. The Japanese-L1 group achieved higher scores in the AV than in the A test condition. In Study 2, 39 of the Japanese-L1 learners undertook ten sessions of intensive training before being tested again. 21 were trained using auditory stimuli and 18 using audiovisual stimuli. Both groups of learners showed a significant benefit of training, and learners with audiovisual training improved more than learners trained auditorily. The benefit from AV training did not depend on learners’ sensitivity to visual cues prior to training.

1. Introduction
It is well known that the perception of speech sounds in a second language 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 native sound categories [e.g., 1]. Research on the effect of intensive auditory training for the acquisition of new sound contrasts has shown successful outcomes. However, previous work has not exploited the full range of cues available to the listeners, as visual cues contained in speechreading have mostly been excluded (apart from [2]). These cues are known to strongly influence perception for native listeners in difficult conditions, and are an inherent component of traditional face-to-face second language teaching. The aim of this study was to investigate L2 learners’ sensitivity to both auditory and visual cues that mark a difficult contrast for Japanese and Spanish learners of English. The contrast is a labial/labiodental contrast, which we investigated using a three way contrast between the English sounds /v/, /b/ and /p/. The contrasts between /b/ and /v/ and between /b/ and /p/ are difficult for both groups of learner. In Spanish, /b/ is realised as the allophones /b/ and /b/ which are in complementary distribution and the English labiodental fricative /v/ does not occur; although its voiceless counterpart /v/ does. The allophone [b] appears only initially or after nasals and the allophone [v] in other positions. Due to the presence in the phoneme inventory of /f/ and /b-p/, visual awareness of the labial/labiodental contrast is likely to be good. In Japanese, the sound /v/ does not occur in the consonant inventory but the Japanese phoneme /b/ is sometimes realised as a bilabial approximant when between vowels so that /v/ to /b/ confusions in English could be expected.

Study 1 had two main objectives. The first was to compare the correct perception of the labiodental feature (/v/ perception corrected for response bias) across conditions per L1 group. If there is influence of visual information, one would expect an AV benefit to be seen as the labiodental feature is highly salient for native listeners. The second objective was to examine more closely the influence of the characteristics of the L1 consonantial system on L2 by comparing performance by Japanese-L1 and Spanish-L1 listeners. The aim of Study 2 was to compare the effectiveness of auditory and audiovisual training for Japanese learners of English.

2. Study 1: Perception of /v/-/b,p/ contrast

2.1. Speech material
The consonants /p/, /b/ and /v/ were embedded within nonsense words with the following structure: CV, VCV, or VC, where V was one of /i,u/ Each item was recorded three times, yielding a total number of 81 tokens.

2.2. Speaker and Recording procedure
A phonetically-trained female speaker of South Eastern British English recorded the test items. The video recordings were made in a soundproof room, with two halogen lights illuminating the blue background and the face of the talker. A fully-sized image of the speaker’s head was obtained with a fully visible lower jaw drop. Video recordings were made to a Canon XL-1 DV camcorder, using a Bruel and Kjaer type 4165 microphone. Simultaneous audio recordings were made to a DAT recorder. The video channel was digitally transferred to a PC, digitized and down-sampled for editing (250*300 pixels, 25 f/s, audio sampling rate 22.05 kHz). The video track was aligned with the separately recorded digital audio track. Stimuli were edited so that the start and end frames of each token showed a neutral facial expression.

2.3. Listeners

2.3.1. Spanish-L1 listeners
Thirty-six listeners participated in the experiment in total, but four showed ceiling effects in their score and were
consequently disregarded from further analysis. Of the remaining 32 listeners, four were recruited from a School of English in London whilst the rest were university students tested in Spain (Barcelona). Selection criteria were that learners were at a lower to lower intermediate level of English proficiency, were native speakers of Castilian Spanish and not bilingual, had started learning English after the age of 10 years and not resided for extended periods in an English speaking country, and that they reported normal hearing and normal or corrected vision. A control group of 14 native English listeners judged the test items in the two blocks of the video alone condition.

2.3.2. Japanese-L1 listeners

Forty-seven listeners participated in the experiment. All were students at the University of Konan in Japan taking an English course, and were tested in Japan. Learners were aged between 19 and 23 years, had already learned English for six years at high school before entering university, but they did not often practice listening to English sounds. Selection criteria were the same as those for the Spanish-L1 group.

2.3.3. Experimental task

A closed-set forced-choice identification task was built using the CSLU toolkit [3]. A conversational agent [4] was used to explain the task to the listener and to give general feedback on performance at the end of each section of the test. To test perception in each modality, and the ability to integrate auditory and visual information, stimuli were presented in three test conditions: (1) Video alone (2) Audio alone and (3) Audiovisual, with two blocks of 81 items per condition (i.e. 54 repetitions of each consonant per listener in each test condition). Listeners had to give a response of B, P or V. Item order was randomized within each block for each listener, and conditions were presented in two orders counterbalanced across listeners. The experiment was run on laptops in quiet surroundings. The video appeared on the computer screen in a window of 340 x 290 pixels, and items were presented to both ears at a comfortable listening level via headphones. The whole experiment took about 45 minutes.

2.4. Results

For this study, we were primarily interested in evaluating whether learners could perceive the difference between labial and labiodental consonants (i.e. difference in manner and place of articulation) either using auditory or visual cues or their combination. Analyses were therefore based on the percentage of correct /v/-not /v/ identification as this would reflect the perception of the labiodental feature in each modality. For Spanish-L1 listeners, mean /v/ identification was 87.0% (s.d. 12.3) for the A condition, 88.6% (s.d. 14.4) for the AV condition and 82.3% (s.d. 22.2) for the V condition. For the Japanese-L1 listeners, mean identification was 64.4% (s.d. 19.1) for the A condition, 65.7% (s.d. 19.4) for the AV condition and 51.1% (s.d. 18.75) for the V condition. The native control group achieved a score of 90.6% correct /v/ identification in the V condition. The proportion of correct /v/ responses was converted to dprime (Z value of the hit-rate minus that of the false-alarm-rate) to correct for any response bias, given that there were three possible responses (B, P, V). Analyses of variance for repeated measures were then applied to the adjusted scores to look at the withinsubject effect of test condition (A, AV and V) and between-subject effect of L1 background (Spanish-L1 or Japanese-L1) (See Figure 1).

![Figure 1: Mean /v/ perception (in dprime) for the three test conditions for each of the two learner groups.](image)

The effect of condition was statistically significant \(F(2,154)=19.05; p<0.0001\). Pairwise comparisons with Bonferroni adjustments show that /v/ perception was significantly lower in the V condition than in the A and AV conditions, and that /v/ perception was significantly better in the AV than in the A condition. The effect of L1 background was strongly significant \(F(1, 77)=53.01; p<0.0001\) with Spanish-L1 learners achieving higher scores than Japanese-L1 learners.

In order to look in more detail at performance within L1 group, repeated-measures ANOVAs were run on the data for each language-group individually. For the Japanese-L1 group, the effect of test condition was strongly significant \(F(2,92)=24.64; p=0.0001\) and post-hoc analyses showed the same pattern as above (AV>A>V). However, for the Spanish-L1 group, the effect of condition just failed to reach significance. In order to check for a ceiling effect, we also calculated the statistics on the lowest 50% of Spanish-L1 listeners when ranked on their AV performance (mean /v/-perception dprime for these 16 listeners: AV 2.77, A 2.87, V 2.73). There again, it appears that the effect of condition was not significant. Spanish-L1 listeners can therefore identify /v/ as well with visual as with audio cues but there is no evidence of AV integration as AV performance is not better than A or V performance even for lower-performing listeners.

It is informative to examine the relative use of auditory and visual cues in individual learners. Scores obtained for each modality are presented in Figure 2. It can be seen that there are individual differences in the use of auditory and visual information, especially for Japanese-L1 listeners. Although there is overlap in performance for the A condition between the Spanish-L1 and Japanese-L1 learners, only one Japanese-L1 learner (2% of the group) achieved a score of over 3 for the V condition whilst 19 Spanish-L1 learners did (53% of the group).
There was also a distinct difference in higher scores for /v/ perception were obtained for the Spanish distinction being phonemic in neither language, generally, group but not for Spanish-L1 group. Despite the /b-p/-/v/ contrast, there is evidence of AV benefit for the Japanese-L1 group, on average, /v/ perception was as good in lipreading-alone than in the audio-alone condition, which suggests a sensitivity both to the acoustic and visual cues that are marking the labiodental/labial distinction. This is likely to be due to the presence of both labial and labiodental consonants in the phoneme inventory, even if the /b/-/v/ distinction itself is not phonemic. For the Japanese-L1 group, even for those learners who were identifying the contrast well above chance level, the audio-alone perception was significantly better than the lipreading-alone perception, which suggests better use of acoustic than visual cues.

3. Study 2: Perceptual training study

Study 1 showed that Japanese-L1 listeners had difficulty in perceiving the distinction between /v/ and /b,p/ but that some listeners at least had some sensitivity to the visual cues marking the contrast. In Study 2, we investigated (a) whether the perception of the contrast could be increased by a period of intensive perceptual training, and (b) whether audiovisual training would be more effective than auditory training for all listeners or for the subset of ‘visually-aware’ learners.

3.1. Speakers and Listeners

For the training study, three female and two male talkers of South Eastern British English were recorded. 39 Japanese-L1 listeners participated in the training. These were a subset of the Japanese listeners tested in Study 1. Twenty-one of these listeners carried out the training in the Auditory training condition, and 18 learners in the Audiovisual training 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.

3.1.1. Pre and post-test materials

The pre- and post-test materials were those described in Study 1.

3.1.2. Training materials

A number of English minimal pairs (real words) were recorded, that contrasted /b/, /p/ and /v/ in a range of varying vowel environment and in different positions and syllabic contexts. The recording procedure was as described for Study 1. A subset of 75 items was chosen for the training sessions, with 15 minimal pairs of words with /b/ and /v/ in initial and medial position and 15 minimal pairs of words with /b/ and /p/ in initial, medial and final position.

3.1.3. Training task

The High Variability Phonetic Training procedure was used [5]. 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. The training software was designed using the CSLU toolkit. A conversational agent was used to give instructions to the learner, with feedback given after every trial to guide the participants during their training. Listeners were told of the percentage of correct identification achieved at the end of each training block.

The training consisted of 10 sessions, held over a period of 4 weeks, 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 (for the ‘A training’ learners) or audiovisually (for the ‘AV training’ learners). The training program was run individually on PCs, with students working in quiet surrounding and stimuli presented via headphones. At each training session, listeners either saw and heard, or just heard five blocks of test items produced by one of the five speakers. Each student therefore perceived 60 repetitions of each consonant (across positions) in the b-v contrast, and 90 repetitions of the consonants in the /b/-/p/ contrast (across positions) in each session. The blocks with different consonants alternated in order per day, and the order of items was randomized within each block for each listener. After ten sessions of training a post-test was done, which was identical to the pre-test.

3.2. Results

The proportion of correct /v/ responses for pre-test and post-test results in each modality (A, AV, V) were converted to d’prime to correct for any response bias (See Figure 3). 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 /v/ perception [F(1,37)= 108.7; p=.000]. The effect of test mode was strongly significant [F(2, 74)= 46.8; p=.000] and post-hocs showed the following pattern: AV>A>V, thus replicating findings obtained for the pre-test.
The evaluation of the relative effectiveness of the two training modes was carried out on the difference in pre-post training scores so as not to be influenced by differences in overall levels of performance across groups. The effect of training mode was significant \( F(1,37)=10.3; p=.003 \), with a greater improvement in /v/ perception seen for the ‘AV training’ group than the ‘A training’ group. Post-hoc tests revealed that the improvement in perception was greater for the A and AV test condition than for the V test condition.

In order to see whether audiovisual training was of greater benefit to those learners with prior sensitivity to visual cues, the ‘AV training’ group was divided into ‘visually-aware’ (N=8) and ‘non visually-aware’ (N=10) subgroups based on a criterion of correct /v/ perception dprime > 1.15 in the V test condition in the pretest. An ANOVA for repeated measure failed to show any difference between the subgroups in terms of their improvement post-training although there was a trend to a greater improvement in the V condition for the audiovisually-trained group (See Figure 4).

4. Discussion

The results of Study 1 revealed that the Japanese-L1 learners did not, on average, show strong sensitivity to visual cues to the labial/labiodental contrast, although a higher performance was obtained for the AV than A test condition. However, audiovisual training was more effective in increasing the correct perception of the contrast than auditory training. In a previous study looking at the effect of intensive perceptual training on the perception of the /l/-/r/ contrast for Japanese learners of English [6], we did not find audiovisual training to be more effective than auditory training. The fact that a significant effect of training mode was obtained here suggests that the visual salience of the contrast is a key factor in determining the effectiveness of audiovisual training. Indeed, the /l/-/r/ contrast is cued visually for native listeners in that they score well above chance in a visual-alone condition for this contrast (72% correct) but the visual cues are much more salient for the labial/labiodental contrast, with native listeners scoring 91% correct /v/ perception on the visual condition of the pretest materials.

5. Conclusions

Audiovisual training was shown to be more effective than auditory training in improving the perception of the visually-salient /b/-/p,v/ contrast for Japanese learners of English.

6. Acknowledgements

Funded by grant GR/N11148 from the Engineering and Physical Sciences Research Council of Great Britain. We thank Prof. Joaquim Llisterri (UB, Spain) for his help in organizing the testing of Spanish students and Dr Marta Ortega-Llebaria for her work on developing the test and training materials.

7. References