AUDIOVISUAL PERCEPTION IN L2 LEARNERS

Valerie Hazan, Anke Sennema and Andrew Faulkner

Department of Phonetics and Linguistics
UCL, London
v.hazan@phon.ucl.ac.uk, a.sennema@phon.ucl.ac.uk, andyf@phon.ucl.ac.uk

ABSTRACT
This study investigates the use of audiovisual cues in the perception of sound contrasts which have a different phonemic status in the listeners’ L1 and L2. Two contrasts differing in the distinctiveness of their visual gestures (/bl/-/v/ and /pl/-/b/) were presented to Spanish learners of English in audio, visual and audiovisual modalities. Overall identification rates were not significantly higher audiovisually than in the audio alone condition for either contrast. For the /bl/-/v/ contrast, which is visually marked, listeners showed different patterns of performance. A subset of listeners appeared to have acquired the L2 /bl/-/v/ contrast and were sensitive to both the acoustic and visual cues marking the contrast. Those at an earlier stage of acquisition of the L2 contrast generally showed poor sensitivity to the visual as well as the acoustic cues.

1. INTRODUCTION
There has long been an interest in the issue of auditory speech perception in second language learners. Studies have shown that perceptual difficulties encountered by learners in a second language are dependent on a number of factors. These include the relation between the phonemic systems of their L1 and L2, the age of acquisition of the L2, the duration of immersion within the L2, the degree on ongoing use of L1. A large degree of individual variability is also seen between learners showing similar profiles in terms of language background and language use. The relation between the phonemic systems of the two languages appears to be a primary factor. A number of models (e.g. [1, 2]) have given predictions about the degree of perceptual difficulty encountered for different categories of speech sounds in a second language. Where two sounds have similar phonemic status in the L1 and L2, there are no additional problems with their perception in the L2. However, contrasts between two sounds that have phonemic status in an L2 but which are allophones of a single phoneme category in the L1 are particularly difficult to discriminate and identify as they tend to be assimilated to the same L1 category. These distinctions are therefore of particular interest in the study of L2 acquisition at the perceptual level. These distinctions are therefore of particular interest in the study of L2 acquisition at the perceptual level [3] and can lead to intelligibility problems [e.g. 4] and to added listener effort and cognitive load [5]. It has also been shown that even highly fluent second language learners, who have similar perception of intelligibility to native listeners for speech in quiet show disproportionate reductions in intelligibility when speech is degraded by noise [6].

It is well known that hearing-impaired listeners receive much benefit from the addition of visual cues, and that gains can also be seen for hearing listeners when presented with degraded speech. It would therefore be expected that L2 learners would benefit from seeing the face of the speaker. This improvement could be expected to come from the better use of general suprasegmental cues provided by the visual channel but also from segmental information that can disambiguate certain distinctions in place and manner of articulation especially. Surprisingly, audiovisual perception in second language learners has been relatively limited. Studies of the McGurk effect provide some information on the use of visual cues by non-native listeners (see extensive review in [7]). Studies with Japanese and Chinese learners of English showed that the McGurk effect was weaker for these subjects than for American English subjects [8]. However, a wider ranging study of concordant and discordant audiovisual syllables with listeners from four different L1 backgrounds showed that the influence of the visual cue was “dependent upon its information value, the intelligibility of the auditory cue and the assessment of similarity between the two cues” and was also affected by linguistic experience [9].

Some evidence as to the weak effect of visual information on L2 perception came from a study by Ortega-Llebaria et al. [10]. They investigated auditory-visual and auditory consonant and vowel confusions in a mild background of noise in a study with 36 Spanish learners of English and 12 English controls. Speech material included multiple repetitions of 16 consonants and nine vowels of British English produced by a single speaker. There was a small but significant improvement in the audiovisual condition for both listener groups in the consonant test but no significant improvement in vowel intelligibility. The consonant confusions that were disambiguated by visual information tended to those that were common to both listener groups and were likely to be the result of poor acoustic distinctiveness. Consonant confusions which were language-dependent – mostly errors in voicing and manner – were not reduced by the addition of visual cues. Spanish listeners did not use visual cues that disambiguated contrasts that are phonemic in English but have allophonic status in Spanish (e.g. /bl/-/v/ and /ld/-/d/). Ortega-Llebaria et al. therefore concluded that visual features may have different weights when cueing phonemic and allophonic distinctions. However, in a study with Japanese and Korean listeners on the perception of the /hl/-/l/ contrast by Japanese and Korean listeners, Hardison found a significant improvement in intelligibility of /hl/-/l/ when visual cues were present relative to the audio condition (from 65% to 75%). There is therefore conflicting information as to the use of visual cues by L2 listeners.
An issue resulting from these studies is whether it is possible to increase the effectiveness of training via the addition of visual cues. The potential of visual cues for L2 training of phonemic contrasts has only been investigated in a small number of studies so far. Hardison [9] 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. Audiovisual training was more effective than auditory training alone in improving the identification of the /l/-/r/ contrast for Japanese listeners (68.7% pre-test to 90% post-test). Perceptual training also had an impact on the production of this consonant contrast by the Japanese learners, with greater improvements in accent ratings for those who had received audiovisual training. Similar effects of audiovisual training were obtained with the Korean learners of English.

However, as audiovisual training was found to be effective in a population which appeared to be sensitive, in the pretest, to the information provided by visual cues, it is not possible to state whether training effectiveness with audiovisual stimuli is dependent on there being a 'natural' sensitivity to visual cues for the specific contrast being trained, or whether listeners can attune to visual cues as a result of training. The aim of our study is therefore to evaluate the relative effectiveness of audio and audiovisual training for L2 perception for contrasts in which visual cues vary in 'informativeness' for the L2 population. As a first step to this work, we wished to verify and extend the findings of Ortega et al. in a more focused study. Indeed, the evidence for a lack of informativeness of the visual channel obtained in that study involved a large number of consonants. In contrast, Hardison’s evidence for a significant influence of visual cues on perception was obtained in a study focusing on a single contrast only. A new study was therefore carried out that focused on the /p/-/b/-/v/ distinction for Spanish listeners. Also, a visual alone condition was added in order to be able to evaluate the information provided by each modality and the degree of integration when both are presented together.

2. STUDY

This study investigated the audiovisual perception of the distinction between English /p, b, v/ in Spanish L2 learners of English. This three-way distinction covers two areas of major auditory perceptual difficulty for Spanish speakers of English [10]. Spanish voiced stops have [+continuant] and [-continuant] allophones. The [-continuant] allophones have shorter VOT than their English counterparts, causing Spanish speakers to assimilate both English voiced and voiceless stops to their Spanish voiceless counterparts. Moreover, voiced English fricatives tend to assimilate to the [+continuant] allophone of the Spanish voiced stop. Thus, English voiced fricatives and voiced stops tend to assimilate to a single Spanish voiced stop. Visually, the distinction between the bilabial /b/ and labiodental /v/ is highly distinctive in English and the two sounds are typically shown to belong to different visemes [11]. The distinction between /p/ and /b/ is not likely to be disambiguated by visual cues as visual cues carry little information to the voicing distinction. The main research question of the study was therefore whether Spanish listeners would be sensitive to the clear visual information provided for the /b/-/v/ contrast, and whether we would find evidence of integration in the audiovisual condition relative to the audio alone and visual alone conditions.

2.1. Speech material

The consonants /p, b, v/ were embedded within nonsense words with the following structure: CV, VCV, or VC, where V was one of the following: /l, A, w/.

2.1. Speaker and Recording procedures

A female speaker of South Eastern British English recorded the test items. Three utterances of each item were recorded. Recordings were made to a Canon XL-1 DV camcorder, using a Bruel and Kjaer type 4165 microphone. The video was digitally transferred to a PC for editing. Stimuli were edited so that the start and end frames of each token showed a neutral facial expression. Three tokens were produced for each consonant in each syllabic and vowel context, yielding a total of 81 items.

2.2. Listeners

32 listeners participated in the experiment. Four of these were recruited from a School of English as a second language in London where they were attending an intensive language course whilst the rest were university students tested in Spain. They were approximately at a lower intermediate level of English proficiency, were aged between 18 and 26 years, had all started learning English after the age of 10 and none had lived in the UK for more than 4 months. They reported normal hearing and normal or corrected vision.

2.3. Experimental task

A closed-set identification task was built using the CSLU toolkit [12], and a conversation agent was used to explain the task to the listener and to give general feedback on the level of performance at the end of each section of the test. There were three test conditions: (1) Video alone presentation; (2) Audio alone presentation (3) Audiovisual presentation with two blocks of 81 per condition. Each listener therefore heard 54 repetitions of each consonant (across vowels and positions) in each test condition. Order of items was randomized within each block for each listener. 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.

3. RESULTS

The data was analyzed to obtain percentages of correct identification of /p, b, v/ per condition for each of the 32 listeners. First, the data was examined for overall identification rates in each condition. Mean consonant identification scores were 58.7% in the video alone condition, 77.3% in the audio alone condition and 78.1% in the audiovisual condition. An analysis of variance for repeated measures showed that the effect of mode of presentation was significant \( F(2,62)=91.5; p=0.0001 \). There was also a significant consonant by mode interaction. Pairwise comparisons with Bonferroni adjustments showed that intelligibility for the audiovisual condition was not significantly higher than intelligibility for the audio alone condition. The effect of consonant was also significant \( F(2,64)=73.5; p=0.0001 \) with a mean identification score of 55.6% for /b/, 72.5% for /p/ and 86.0% for /v/.
To investigate whether the relative lack of effect of the visual information on perception is particularly relevant, it is essential to know whether the identification errors were due to voicing (/p/-/b/) or manner/place (/b/-/v/) confusions. Indeed, we would not have expected the voicing errors to be disambiguated by the addition of visual cues. The main consonant confusions were as predicted by our knowledge of the relation between the phonemic systems of the two languages. Indeed, /v/ was often perceived as /b/ and /b/ was often perceived as /p/. The numbers of such errors were calculated by extracting the relevant two cells in confusion matrices for each individual subject. These confusion rates were then used in further analyses of variance for repeated measures with within-subject factors of contrast (/v/ to /b/, /b/ to /p/), condition (A, AV, V) and position (initial, medial, final). The effect of contrast was strongly significant \[ F(1,31)=149.1; p=.0001 \] and mean error rate for the /b/ to /p/ confusions (40%) was much higher than the error rate for the /v/ to /b/ place/manner confusion (11.1%). The effect of condition was not significant and there was no significant contrast by condition interaction. There was therefore no evidence of reduced confusion rates in the audiovisual condition relative to the audio condition. The effect of consonant position was significant \[ F(2, 62)=8.2; p=0.001 \] and higher rates of confusions were obtained for consonants in initial than in medial or final positions (see Table 1).

In order to verify that data from students showing little evidence of difficulty with this contrast was not skewing results, analyses of variance were also carried out on a subset of 24 subjects who had more than 5% confusion errors for this contrast in the audio condition. The effects found were as for the total data set. Therefore, even if we only consider the data from those students who had some difficulties in perceiving the contrast, we do not see any significant evidence that they were helped by the addition of visual cues.

<table>
<thead>
<tr>
<th></th>
<th>Initial Mean s.e.</th>
<th>Medial Mean s.e.</th>
<th>Final Mean s.e.</th>
<th>ALL Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>/v/ - /b/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>17.1 3.2</td>
<td>11.5 2.3</td>
<td>3.7 1.1</td>
<td>10.8</td>
</tr>
<tr>
<td>AV</td>
<td>14.5 3.4</td>
<td>12.4 3</td>
<td>3.5 1.3</td>
<td>10.1</td>
</tr>
<tr>
<td>V</td>
<td>11.9 2.7</td>
<td>14.6 3.1</td>
<td>10.9 2.6</td>
<td>12.5</td>
</tr>
<tr>
<td>/b/ - /p/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>41.2 4.9</td>
<td>34.9 4.4</td>
<td>38.3 5.2</td>
<td>38.1</td>
</tr>
<tr>
<td>AV</td>
<td>49.8 4.9</td>
<td>31.2 4.2</td>
<td>39.1 4.3</td>
<td>40.0</td>
</tr>
<tr>
<td>V</td>
<td>51.7 2.9</td>
<td>37.2 2.9</td>
<td>40.2 3.4</td>
<td>43.0</td>
</tr>
</tbody>
</table>

Table 1: Mean and standard error of percentage error for the /v/ to /b/ (place/manner) and /b/ to /p/ confusions in each mode and each consonant position.

It is telling though to examine more closely the difference in performance between the audio alone and the video alone conditions for both contrasts as clear differences emerge. Pearson’s correlations (Table 2) revealed that the /v/ to /b/ confusion rates in the three modalities were quite strongly correlated (see Figure 2). A subgroup of L2 listeners had clearly acquired the ability to distinguish /b/ from /v/ and were able to do so using either acoustic or visual information. Those listeners who were struggling in making this distinction accurately were making errors when either the acoustic information or the visual information were presented. No such correlation between error rates for the audio and visual modes of presentation was obtained for /b/ to /p/ confusions (see Figure 3).

<table>
<thead>
<tr>
<th></th>
<th>AUDIO</th>
<th>VIDEO</th>
<th>AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUDIO</td>
<td>1</td>
<td>.744</td>
<td>.870</td>
</tr>
<tr>
<td>VIDEO</td>
<td>1</td>
<td>.628</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Pearson’s correlations of /v/ to /b/ confusion rates obtained in the three modalities; all were significant at p<.01.

Figure 2: Scatterplot of /v/ to /b/ confusion rates obtained by individual listeners in the audio and visual conditions.
4. DISCUSSION

In this study, we examined the perception of two L2 contrasts that varied in the 'informativeness' of the visual cues. Both the voicing contrast between /b/ and /p/ and the place/manner contrast between /b/ and /v/ lead to perceptual difficulties for Spanish learners of English but the /b/-/v/ contrast is marked by visual cues which are highly contrastive for English listeners whereas the voicing contrast is not visually marked. The first point to note is that the level of difficulty of the /b/-/v/ contrast varied substantially across listeners, as the rate of consonant confusion (/v/ perceived as /b/) in the audio condition varied from 0 to 42%, and the mean confusion rate was lower than that seen in our previous study. Our listeners were primarily students from a school of translation in Spain and were therefore probably more ‘language aware’ and linguistically experienced than listeners tested previously. However, the main question, which was whether visual cues would have a greater effect in disambiguating the /b/-/v/ than the /b/-/p/ contrast, can still be examined.

Neither for the voicing nor in the place/manner contrast did we find evidence of significantly better perception in an audiovisual than in an audio condition. However, the difference between the labial and labiodental gestures were clearly informative for a number of listeners who had very low rates of confusion between /b/ and /v/ in the video alone condition. It appears that a distinction needs to be made between those listeners who have clearly acquired the ability to distinguish the two English phonemic categories and those who still in the process of acquisition. The successful learners have acquired a sensitivity both to the acoustic cues and to the visual cues marking the distinction. However, those listeners who have not completely mastered the contrast showed higher rates of consonant confusion both in the auditory and visual modalities and did not appear to gain any further advantage from seeing the speaker. It can be concluded that in the process of L2 category acquisition, sensitivity to visual as well as to acoustic cues is acquired, but this sensitivity may not be present at early stages of the process, for certain contrasts at least. Patterns of individual performance in our study also suggest that during this process of acquisition, the relative sensitivity to acoustic and visual cues will vary across learners.

5. ACKNOWLEDGEMENTS

This project was funded by grant GR/N11148 from the Engineering and Physical Sciences Research Council of Great Britain. We wish to acknowledge the contribution of Marta Ortega-Llebaria to the early stages of the project. We also wish to thank Joaquim Llistteri from the Universitat Autonoma de Barcelona and Jo Thorp and colleagues from the Bell London School for their substantial help in organizing the testing of Spanish listeners.

6. REFERENCES