The Effectiveness of Visual Cues in L2 Perception

Midori IBA

Abstract

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. However, this training is typically long and effortful and does not exploit a source of information that has typically been of great benefit to native speaking listeners in poor listening conditions: visual information. Two studies carried out at University College London suggested that the effect of visual information on L2 perception was in fact quite weak. The present experiment is designed as a contrastive study with the two just mentioned. I examine whether or not a marked audio-visual (AV) benefit exists in L2 consonant perception by comparing other sets of consonants. Further experiments are needed to reach conclusions about the effectiveness of the AV benefit. One possibility, however, is that the AV benefit varies according to the manner of articulation of each sound and individual preferences in visual information.

Introduction

This study examines whether or not L2 learners can extract phonetic information from visual cues in the perception of novel phonemic contrasts. The procedure of the experiment is based on work done by Sennema, Hazan and Faulkner at University College London (UCL) in a project I joined when on sabbatical leave. In their previous experiment, 92 Japanese learners of English were tested on their perception of the /l/-/r/ contrast in audio, visual and AV modalities. Overall identification rates in audio and AV conditions did not differ significantly and few individual listeners showed evidence of any AV benefit. In the present follow-up experiment, I chose the consonants /v/, /b/ and /p/ to see whether the AV benefit in the perception of the L2 depends on specific consonants or not. 47 Japanese university students were trained and took the test for this experiment. The results show a considerable difference from those of the work done at UCL. In this study, firstly, the UCL studies (Study 1 and

Study 2) will be introduced and then the present experiment (Study 3) will be described in the second part of this paper. In Study 3, I have compared the results of the experiment of Japanese students with those of Spanish students done by Hazan at UCL. Finally I will consider all the results and illustrate why the results vary between the two experiments. The effectiveness of AV information will be also examined.

1. UCL Experiment on the /l/-/r/ Contrast (Study 1)

1.1 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 clusters with the additional consonant being /k/ and /f/, appearing in the structure CV, CCV, VCV and VCCV.

1.2 Speaker and recording procedures

To prepare the test items a female speaker of South Eastern British English was recorded. Recordings were made on 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 fts, 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*290. Three tokens 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 tokens.

The study involved 92 Japanese learners of English as a foreign language. 53 of these were students of Kochi University and tested in Japan, 20 students were attending a summer course in 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 group of 7 native listeners judged the test items in the two blocks of the video alone condition.

1.3 Experimental task

A closed-set identification task was built using the CSLU toolkit[1], and a conversation agent 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 tokens

were presented in three conditions (audio alone, visual alone 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. Tokens were presented at a comfortable listening level via binaural headphones.

1.4 Results

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

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

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

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

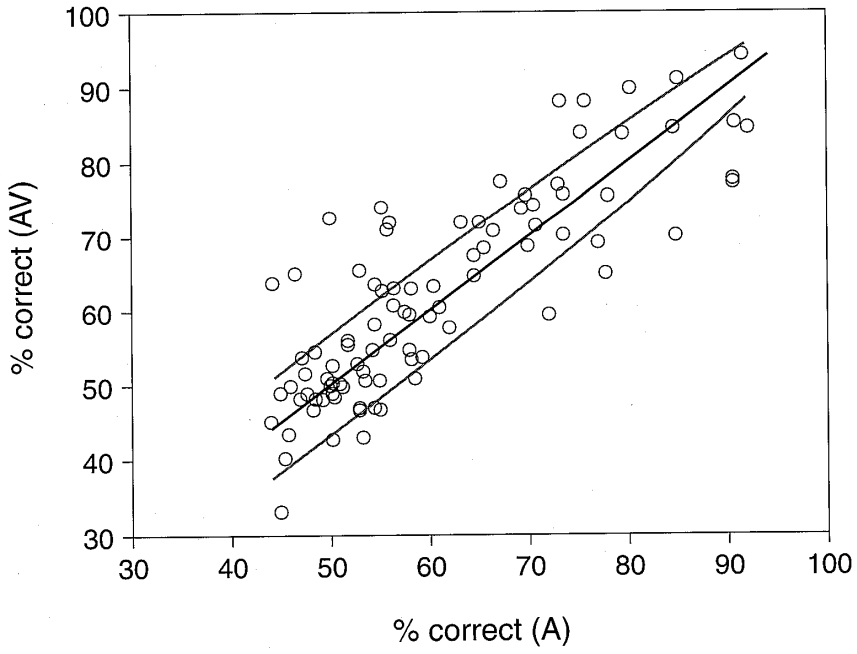

Figure 1: Scatter plot of % correct identification in the A and AV conditions. The middle line is the prediction AV = A, whilst the outer curves are individual 95% confidence limits assuming binomial distribution.

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, 8 out of the 92 subjects (8.7%) showed evidence of a positive AV benefit, 5 (5.4%) were negatively affected by the addition of visual cues and the rest (85.9% of subjects) showed no real effect when visual cues were added.

2. UCL Experiment on the /l/-/r/ Contrast (Study 2)

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 was trained with stimuli presented audio-visually whilst the other group was 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 pre-test. We were interested in whether learners who performed above chance on the visible speech condition showed the advantage of audio-visual over auditory training.

2.1 Speech material

In the pre- and post-test the same speech material was used as in Study 1 (see 1.1). 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).

2.2 Speaker and recording procedure

Two female speakers and three male speakers 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 an example for the speech sound. The recording procedure was similar to that in Study 1.

2.3 Listeners

A subset of 41 Japanese learners of English who were involved in Study 1 partici-

pated in the training. 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 in Phonetics at UCL and 4 were students in a pre-academic language course at UCL.

2.4 Experimental task

The High Variability Phonetic Training procedure was used. This involved the use of multiple tokens 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 smiling face 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 pre-test was followed by ten sessions of training, each lasting about 40 minutes. The training program was run individually on laptops. In the training, students were first familiarized with the two test consonants uttered by the particular speaker of that session, after which the tokens were presented either auditorily (for 18 learners in the 'audio' condition) or audiovisually (for 23 learners in the AV condition). Learners were assigned to each condition on the basis of their scores in the auditory condition of the pre-test, with the aim of ensuring a balance across training groups. The ten sessions were held over a period of two to three weeks. All sessions were carried out under similar conditions, with students working in quiet surrounding on laptops and stimuli presented via binaural 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 in initial position and 64 tokens 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 pre-test.

2.5 Results

The percentage of correctly identified consonants in each mode (V, A, AV) is presented in Table 2.

Table 2: Percentage of correct /l/ and /r/ identification for each test mode before and after training.

	Pre-test			Post-test		
Training mode	V	A	AV	V	A	AV
A	58.0	58.9	62.4	61.6	80.6	80.2
s.d.	*8.4*	*11.5*	*13.8*	*7.8*	*15.6*	*14.9*
AV	52.3	57.6	58.0	62.9	75.0	77.9
s.d.	*7.1*	*10.9*	*10.9*	*7.7*	*14.9*	*13.0*

A repeated-measured ANOVA tested the effects of training (pre/post), of training condition (audio/audio-visual) and of test mode (audio, visual and audio-visual presentation). 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 visual alone condition was poorer than both other conditions but that there was no significant difference in performance between the A and AV conditions. The results for the two training modalities are shown in Figure 2:

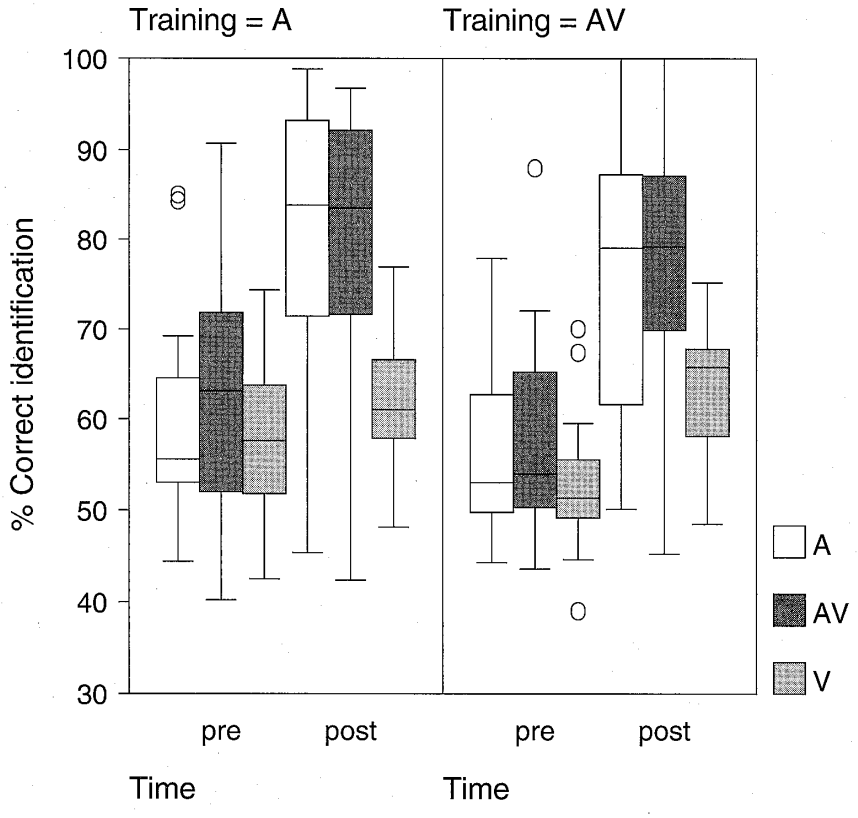

Figure 2: Box plots of identification scores for three test conditions per training modality for pre- and post-test for 41 subjects.

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 pre-test, 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 correlation between test mode and time of testing [F (2, 78) = 30.8; p = 0.0001] and a significant if weaker correlation between test mode, time and training group. This three-way correlation is due to training condition affecting performance on 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 the degree of improvement from pre-test to post-test. The fact 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. The /l/-/r/ contrast is difficult to see for L2 learners and we therefore tested for the factor 'visual awareness'. According to a binomial distribution, for 216 trials, scores of 120 or more (55.6%) are significantly different from chance at the p = 0.05 level. Estimated from the pre-test performance in the video alone condition, analyses showed that if visual awareness is estimated from pre-training AV benefit, no clear pattern of correlation emerges. This measure is unrelated to improvement in the V condition. For the A trained group, pre-training AV benefit is negatively correlated to improvement in the AV condition (R = –0.54, p = 0.02), while in the AV trained group only the A condition shows correlations between AV benefit and improvement over training (R = 0.45, p = 0.03).

3. The Experiment on the Labial-Labiodental Contrast among /v/, /b/ and /p/. (Study 3)

This study investigated the perception of the labial-labiodental contrast using a three-way contrast among the English sounds /v/, /b/ and /p/. Both the contrast between /b/ and /v/ and between /b/ and /p/ are difficult for Japanese and Spanish learners of English. For the Spanish group, given the differences between the consonantal systems in Spanish and English, the following L1-linked confusions were predicted. First, the English voiced plosive /b/ was expected to be confused with the voiceless plosive /p/ as both English /b, d, g/ and Spanish /p, t, k/ are phonetically realized as voiceless unaspirated plosives and have similar VOT (Voice Onset Time) values. In Spanish, /b/ is realized as the allophones /b/ and /v/. The voiced labiodental

fricative /v/ (written as 'u' or 'v') merged with the bilabial plosive /b/ (written 'b'). Contemporary Spanish written 'b, v' do not correspond to different phonemes. The allophone [b] appears only initially or after nasals and the allophone [v] elsewhere, so the pattern of confusion could be expected to be position-specific. In Japanese, there is a voicing contrast between /p/ and /b/, but /p/ tends to be unaspirated so that there is potential assimilation of the English consonant /b/ to the Japanese consonant /p/. The English sound /v/ does not occur in the system of Japanese consonants, but the Japanese phoneme /b/ is sometimes realized as a bilabial approximant when between vowels so that /v/ to /b/ confusions in English can be expected.

3.1 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 the following: /i, a, u/.

3.2 Speaker and recording procedures

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. The talker was seated comfortably in front of the camera. Each test item appeared as a prompt on a monitor next to the camera, the talker then looked into the camera and said the word three times at normal speech rate and at normal intensity level. A full-sized image of the speaker's head was obtained with a fully visible lower jaw drop. In between the utterances the talker closed her mouth so that each utterance showed a neutral facial expression. The recording procedure was similar to that in Study 1.

3.3 Listeners

a. Spanish L1 listeners

Thirty-six listeners participated in the experiment in total, but 4 showed ceiling effects in their score and were consequently disregarded in further analysis. Of the remaining 32 listeners, four were recruited from a school of English as a second language in London where they were attending an intensive language course, and the rest were university students tested in Spain. They were chosen on the basis that they were approximately at a lower to lower-intermediate level of English proficiency, were native speakers of Castilian Spanish and they were not bilingual. The listeners were aged between 18 and 26 years and had all started learning English after the age of 10; only 2 had lived in an English speaking country for more than 4 months. They all reported normal hearing and normal or corrected vision.

b. Japanese L1 listeners

Forty-seven listeners participated in the experiment. All were students of Konan University in Kobe, Japan. They were from various faculties, including letters, economics, business administration, law, and science and technology. They took an elective intermediate listening course open to second year students and above. The pre-test and post-test were carried out during class hours, and the other ten training sessions were done individually in a computer library. They were aged between 19 and 23 years, had already learned English for six years at junior high school and high school before entering university, but they did not often practice listening to English sounds. All of them were native speakers of Japanese, and no one had ever lived in an English-speaking country. They all reported normal hearing and normal or corrected vision.

3.4 Experimental task

A closed-set identification task was built using the CSLU toolkit (version 2.0b2), 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 and (3) audiovisual presentation, with two blocks of 81 items per condition (three tokens for each vowel and consonant in three positions). Each listener therefore heard 54 repetitions of each consonant (across vowels and positions) in each test condition.

The experiment was run on laptops in quiet surroundings (never more than 2 listeners at a time). Items were presented to both ears at a comfortable listening level via headphones. For the AV and V conditions, the video image appeared in a window on the computer screen. The 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.

3.5 Results

a. Group results

Analyses were focused on the perception of /v/ as this would reflect the perception of the labiodental feature in the V and AV conditions. For Spanish L1 listeners, mean correct /v/ identification was 87.0% (s.d. 12.3) for the audio condition, 88.6% (s.d. 14.44) for the AV condition and 82.3% (s.d. 22.2) for the visual condition. For the Japanese-L1 listeners, mean correct /v/ identification was 64.4% (s.d. 19.1) for the audio condition, 65.7% (s.d. 19.4) for the AV condition and 51.1% (s.d. 18.75) for the visual condition. Mean scores for /v/ perception for both groups in all conditions are given in Table 3. Analyses of variance for repeated measures were then applied to look

at the within-subject effect of test condition (A, AV and V) and between-subject effect of L1 background (Spanish L1 or Japanese L1). See Figure 3.

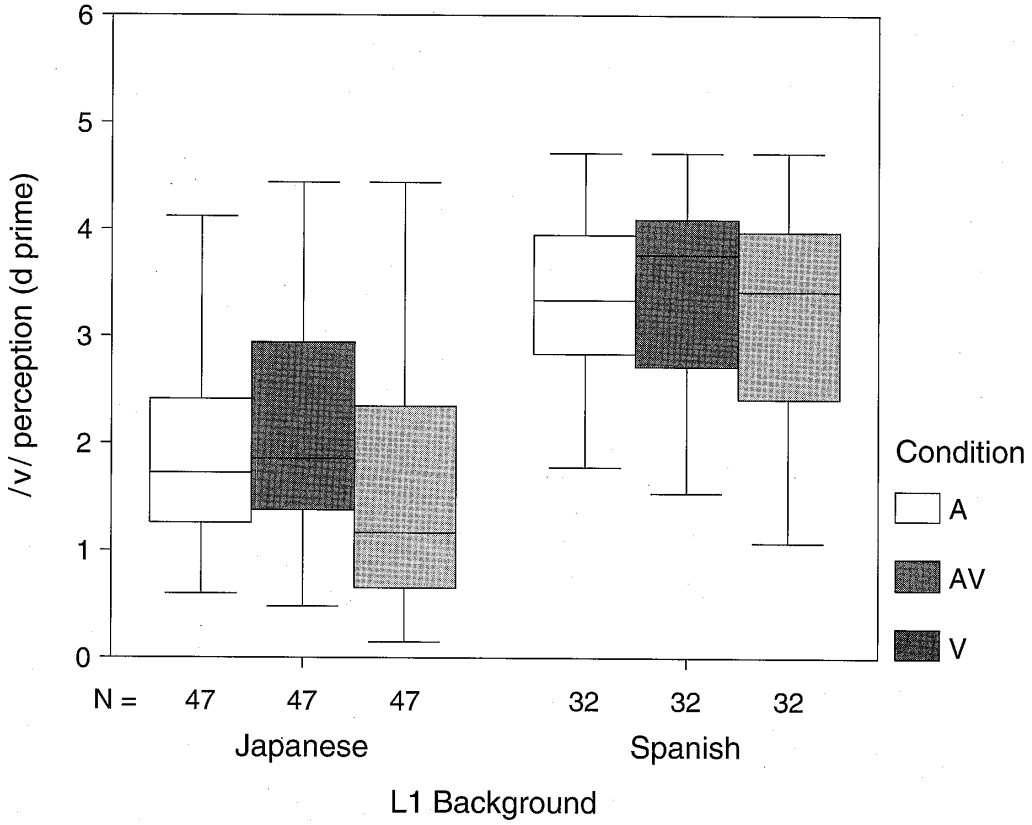

Figure 3: Within-subject effect of test condition and between-subject effect of L1 background.

Table 3: Group results. Mean scores and standard deviation of Japanese learners and Spanish learners.

Descriptive Statistics

L1 Background		Mean	Std. Deviation	N
A	Japanese-L1	1.8943	.88511	47
	Spanish-L1	3.3138	.72154	32
	Total	2.4693	1.07740	79
AV	Japanese-L1	2.1511	1.04420	47
	Spanish-L1	3.4544	.86437	32
	Total	2.6790	1.16387	79
V	Japanese-L1	1.4930	1.02728	47
	Spanish-L1	3.0913	1.11972	32
	Total	2.1404	1.32069	79

The effect of test condition was statistically significant {F (2,154) = 19.05; $p < 0.0001$} and pairwise comparisons with Bonferroni adjustments show that the /v/ perception in the visual-alone condition (V) was significantly different from perception in the audio-alone (A) and audiovisual (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. The condition of L1 background interaction was not significant.

In order to look in more detail at performance within L1 groups, repeated-measure ANOVAs were rerun on the data for each language group individually. For the Japanese L1 group, the effect of test condition was again strongly significant [F (2,92) = 24.64; $p < 0.0001$] and post-hoc analyses showed the same pattern as above (AV > AV > V). However, for the Spanish L1 group, the effect of test condition narrowly failed to reach significance. This is probably due to a ceiling effect as many listeners achieved high scores in all three conditions. This shows that listeners could recognise /v/ (when contrasted with /b/ and /p/) using lip-reading information as well as with audio information. In order to check for a ceiling effect, we also calculated the statistics on the lowest 50% of listeners when ranked on their AV performance (mean 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/ equally well with visual and with audio cues but no evidence of AV integration as AV performance is not better than A or V performance, even for lower-performing listeners.

b. Individual results

Average results can hide quite different patterns of behaviour, and it is therefore useful to look in more detail at individual performance correlations between perception via audio and visual channels. In Figure 4, a crossplot of /v/ perception in the audio alone and visual alone conditions is presented for individual Japanese L1 and Spanish L1 listeners. High-performing individuals (mostly from Spanish L1 backgrounds) who have near perfect perception of /v/ auditorily also have near-perfect perception of /v/ in the lip-reading alone condition. The level of performance in Japanese L1 listeners is typically lower but there is a R^2 of 0.63, showing fairly strong correlation between purely auditory and purely visual perception. Despite this, there is evidence of a sub-group with fair use of auditory cues but random performance on the basis of visual cues, and a sub-group with better use of visual than auditory cues.

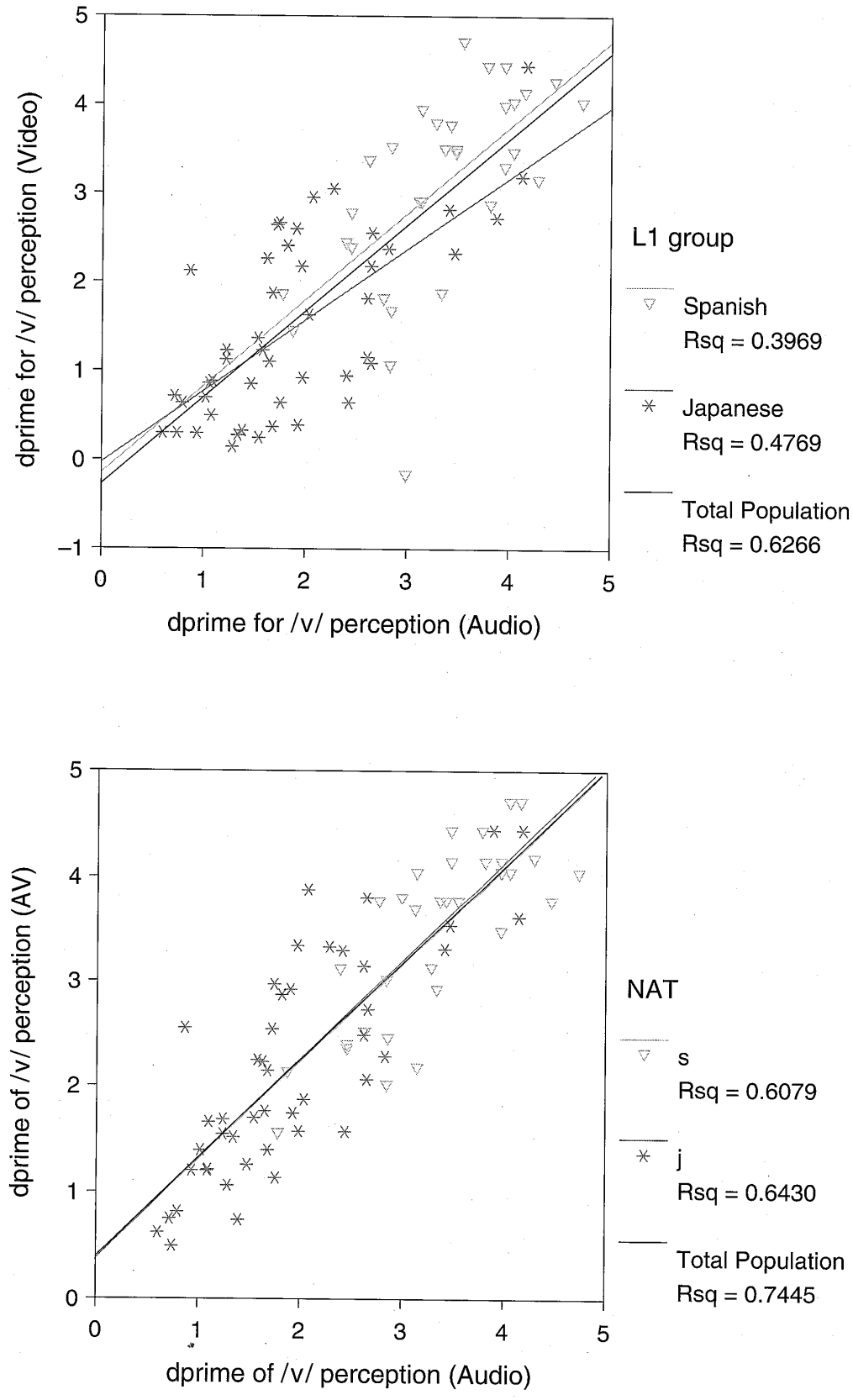

Figure 4: Scatter plot of d-prime perception in the A and AV conditions.

In Figure 4, individual performance in the audio performance is plotted against performance in the audiovisual condition to look at evidence of AV benefit i.e. performance with two modalities being better than in the dominant modality alone. The confidence bands are also included in order to highlight individual cases where audiovisual performance is significantly better or poorer than audio-alone performance.

In summary, for /v/ perception, there is evidence of AV benefit for the Japanese L1 group but not for the Spanish L1 group. However, this could be mostly due to a ceiling effect for Spanish-L1 listeners, as even lower-performing Spanish listeners were above the mean identification scores obtained for Japanese L1 listeners. Generally, higher scores for /v/ perception were obtained for the Spanish listeners. There was also a difference in the 'visual-alone' performance between L1 groups. For the Spanish L1 group, on average, /v/ perception was equally good in both lip-reading alone and in the audio alone conditions; for the Japanese L1 group, the audio alone perception was significantly better than the lip-reading-alone perception, which suggests better use of acoustic than visual cues.

4. Discussion

Many people might have predicted a AV benefit for L2 consonant perception. Yet the results of the studies mentioned in this paper show that things are not that simple. On average, mild to no AV benefit in the perception of non-native contrasts for second-language learners was found. This is true not only across three different contrasts differing in visual distinctiveness, but also across learners with different L1s and thus different relations of L2 contrasts to the L1 phonological system.

There is some evidence in support of the effect of visual salience: weak AV benefit was found for /v/ perception, for which native listeners achieve 94% identification; no AV benefit was found for the l/r contrast for which native listeners achieve 79% identification. Furthermore, some evidence of individual differences was found: Some learners show no AV benefit, others an A or V bias. For both contrasts, those achieving high scores under the A alone condition also achieved high scores on the V alone condition. This is evidence of sensitivity to visual as well as auditory cues once the contrast is acquired.

Further experiments and evidence are needed to reach conclusions about the effectiveness of the AV benefit. One possibility, however, is that the AV benefit varies according to the manner of articulation of each sound and individual preferences in visual information. Some listeners seem to have been confused by visual cues while others took advantage of them without any instruction.

Note

1) CSLU (Center for Spoken Language Understanding) is an Oregon Graduate Institute of Science and Technology Research Center that focuses on spoken language technologies. This institute provides a free on-line suite of tools, the CSLU Toolkit that enables exploration, learning, and research into speech and human-computer interaction.

References

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

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

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

Logan, J.S., Lively, S.E., and Pisoni, D.B. "Training Japanese listeners to identify English /r/ and /l/", Journal of the Acoustical Society of America, vol.89, pp. 874-886, 1991.

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.

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.