AUDITORY-VISUAL L2 SPEECH PERCEPTION: EFFECTS OF VISUAL CUES AND ACOUSTIC-PHONETIC CONTEXT FOR SPANISH LEARNERS OF ENGLISH

M. Ortega-LLebaria, A. Faulkner, V. Hazan,

Dept. Phonetics and Linguistics, UCL, London, UK

ABSTRACT

This study was designed to identify English speech contrasts that might be appropriate for the computer-based auditory-visual training of Spanish learners of English. It examines auditory-visual and auditory consonant and vowel confusions by Spanish speaking students of English and a native English control group. 36 Spanish listeners were tested on their identification of 16 consonants and 9 vowels of British English. For consonants, both L2 learners and controls showed significant improvements in the audiovisual condition, with larger effects for syllable final consonants. The patterns of errors by L2 learners were strongly predictable from our knowledge of the relation between the phoneme inventories of Spanish and English. Consonant confusions which were language-dependent – mostly errors in voicing and manner – were not reduced by the addition of visual cues whereas confusions that were common to both listener groups and related to acoustic-phonetic sound characteristics did show improvements. Spanish listeners did not use visual cues that disambiguated contrasts that are phonemic in English but have allophonic status in Spanish. Visual features therefore have different weights when cueing phonemic and allophonic distinctions.

1. INTRODUCTION

In the last two decades, much attention has been focused on the problems that second language learners encounter in perceiving speech sounds. One line of research has investigated performance with non-native contrasts as a function of subject and language variables. These include: the learner's length of exposure to L2, initial age of acquisition, degree of ongoing use of L1 [e.g. 1], inherent 'skill' in language acquisition, the phonological status of L2 sounds in the learner's L1 [e.g. 2], the inherent acoustic salience of L2 sounds [e.g. 3] etc. Models of L2 speech perception that invoke primarily language variables have generally been successful in predicting areas of perceptual difficulty in L2. For example, the Perceptual Assimilation Model (PAM) [2] predicted a range of auditory discrimination by English speakers of three Zulu contrasts from assimilation patterns between the L1 and L2 sounds. When the sounds of the Zulu contrast were assimilated to two English sounds, English speakers obtained excellent discrimination scores. However, for Zulu sounds that were assimilated to a single English phoneme, discrimination scores were moderate or poor according to the goodness of fit of the Zulu sounds to the English phoneme. The ability to discriminate L2 phonemic categories can be improved by lengthy periods of auditory training, as long as appropriate methods are used [4]. For example, training using identification tasks with feedback seems to be more effective than training using discrimination tasks. Some studies have also shown that enhancing difficult phonemic contrasts for L2 learners via the amplification of key regions or alterations to the duration of segments can be effective in improving perception [e.g., 5, 6].

The potential of visual cues in computer-based auditory training has received little attention, but this might be expected to be at least as important as acoustic enhancement of key regions of speech. Much face-to-face language learning or auditory training exploits information given by looking at the teacher or speech pathologist's face. It is well known from the McGurk effect that visual cues contribute to the perception of place features, and visual cues also contribute to manner perception when the auditory input is degraded. Voicing categorization has also been shown to be influenced by visual cues to place of articulation [7, 8] and to speech rate [9] However, controlled studies of the effect of visual cues in L2 training are rare [e.g. 10, 11]

The aim of our current work is to assess the effectiveness of visual cues in improving the effectiveness of auditory training. The target population is adult Spanish learners of English, and the auditory training will focus on English phonemic contrasts that are particularly difficult for L2 learners from a Spanish background. This initial study was designed to identify training targets that might be particularly appropriate to auditory-visual training.

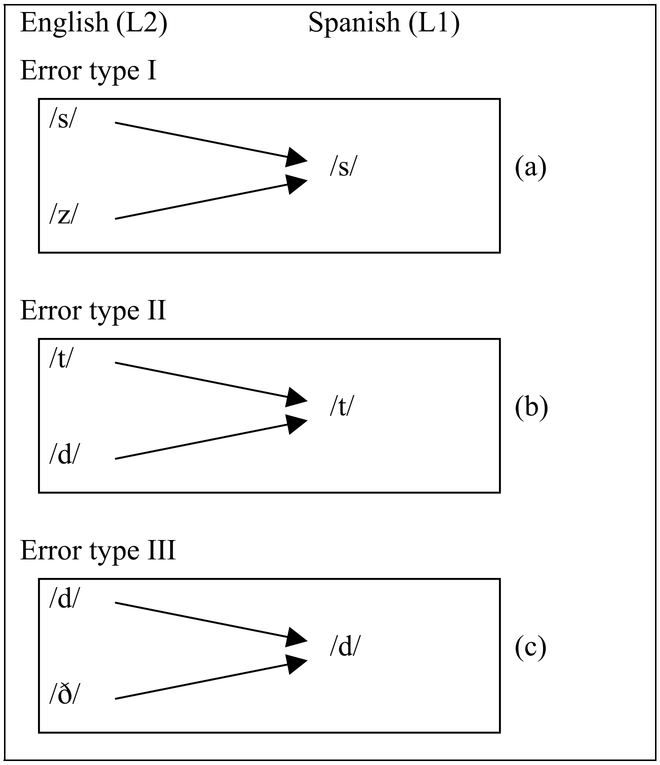

Figure 1. Patterns of assimilation of English stop and fricative phonemes to Spanish stop phonemes

The areas of major auditory perceptual difficulty for Spanish speakers of English relate to obstruent voicing. Voicing is a contextual rather than a contrastive feature for Spanish fricatives. Therefore, according to PAM, Spanish speakers of English will tend to assimilate English voiced and voiceless fricatives to Spanish voiceless phonemes, making the English fricative voicing contrast difficult to perceive (Error type I, Figure 1a). Similar patterns of assimilation are predicted by PAM between Spanish and English stops. 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 (Error type II, Figure 1b). 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 (Error type III, Figure 1c).

In this preliminary study, we are investigating whether the addition of visual cues aids identification of English consonant and vowels by Spanish learners of English without any auditory training. More specifically, our research questions were as follows:

1. How does the use of visual cues by L2 listeners compare to that of native speakers when attending to segmental differences?
2. How will visual cues influence error types I and II?
3. Will visual cues improve perception of error type III, which involve a visible place/manner contrast?

2. METHODOLOGY

2.1. Test materials

Test materials comprised 16 consonants and 9 vowels of British English. The consonants /b, d, g, p, t, k, v, z, ð, f, s, ʃ, tʃ, dʒ, m, n/ were embedded within mono- or bi-syllables. Each contained one of the consonants in the syllabic context CV, VCV, or VC, where V was one of /i, ɑ, u/. The vowels, comprising 7 monophthongs and 2 diphthongs were presented within 9 English bVd words (bad, bed, bid, bead, bud, board, bared, bide, boughed).

2.2. Speaker and Recording procedures

A female speaker of South Eastern British English recorded the test items. Four utterances of each consonant item and seven of each bVd word were recorded. Recordings were made to a Canon XL-1 DV camcorder, using a Bruel and Kjaer type 4165 microphone.

2.3. Stimuli

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. Two phoneticians selected the tokens that were most natural, 2 for each consonant in each syllabic and vowel context and 6 for each bVd word, yielding a total of 288 consonant and 54 vowel tokens. A low-level speech spectrum shaped noise (to CCITT Rec. G227) was added at a +18 dB speech-to-noise ratio to mask environmental sounds during testing, this low level of noise would not be expected to affect auditory intelligibility.

2.4. Listeners

The 36 subjects who participated in the experiment were native speakers of the Spanish dialect spoken in Las Palmas, and were staff or students in the English Department of University of Las Palmas in Gran Canaria. While 8 were highly proficient in English and 6 had lived in an English speaking country for at least a year, the remaining 28 were 1st year students who had spent less than 2 months in an English speaking country. Their ages ranged from 19 to 35 years and they reported normal hearing and vision.

Control data was obtained from a group of 12 native speakers of British English, who worked or studied at UCL. They also reported having normal hearing and vision and their ages ranged from 20 to 36 years.

2.5. Experimental task

A closed-set identification task was built using the CSLU toolkit [11]. Instructions to the listeners were explained in Spanish via Baldi [12], a conversational agent. After introducing himself and the human English talker, Baldi invited subjects to do some practice exercises in order to get familiar with the program's interface and the natural female talker. In the first set of practice exercises, listeners were presented with 16 buttons that displayed graphemes representing target consonants. Subjects were asked to play audio-visually presented natural speech tokens by clicking on each button as many times as they wanted. Since subjects had knowledge of phonetic symbols, it was very easy for the experimenter to make them aware of the two possible orthographic confusions with Spanish graphemes 'z' and 'j'. Once they were familiar with the consonant task, they repeated the task with the vowels.

A second set of practice exercises involved the identification without feedback of consonants and vowels with auditory (A) and auditory-visual (AV) presentation. Once the experimenter was sure that the listeners understood the task, the test was started.

The identification testing consisted of 4 parts, (1) vowels in bVd words with AV presentation; (2) vowels with A presentation; (3) consonants in the 288 syllables with AV presentation; (4) consonants with A presentation. The human talker spoke all test items. Order of items was randomized within each part for each listener. The order of the four parts was counterbalanced across listeners, so that there were 9 Spanish listeners per order of presentation, of whom 2 were proficient listeners and 7 were first year students. The control group took only parts (3) and (4) of the test, so there were 6 subjects in each order of presentation.

3. RESULTS

3.1. Accuracy of identification

The percentage of correctly identified target sounds (see Figure 2) indicated that AV presentation improved consonant identification in both language groups, by 3.7% for Spanish speakers from 71.4% in the auditory condition and by 5.7% for English speakers from 89.5%. Vowel identification by Spanish subjects improved by only 1.7% (from 82.3%). ANOVAS within language groups with factors of mode (AV, A) and order of presentation were performed for vowel and consonant identification. Mode was significant only for consonants, while the interaction mode*order was significant only for vowels, indicating that learning effects due to task order may have obscured any effect of the visual cues for vowels**.**

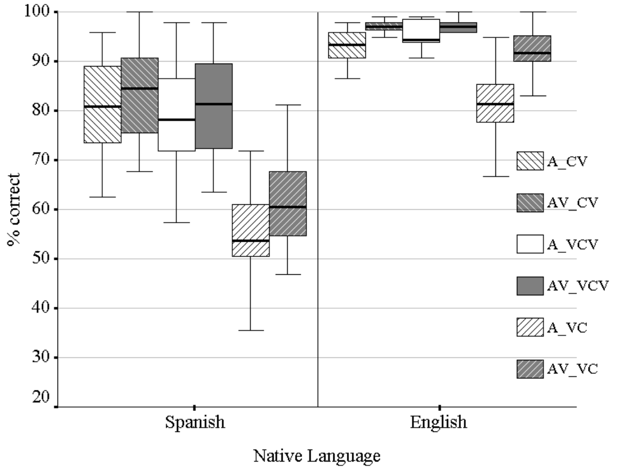

Figure 2: Percent correct identification for audio and audio-visual consonants in each syllabic context for each language group.

For consonants, ANOVAs across language groups were performed on percentage of correct identification with factors of native language (L1), presentation mode, syllabic and vowel context. The main factors of presentation mode (A or AV), syllable, and the interactions 'mode*syllable' and 'syllable*L1' were significant. Post-hoc analyses indicated that the improvement in identification due to visual cues was significant across all subjects and within each language group. There was no significant 'mode*L1' interaction, hence we have no evidence of any special advantage of the AV mode for L2 speakers. English subjects performed significantly better than Spanish speakers in each syllabic context. Within each language group, effects of syllabic context were similar in that VC syllables showed lower accuracy than CV and VCV syllables. The mode*syllable interaction arose mainly from the mode of presentation effect being stronger for VC than for CV and CVC syllables.

Confusion matrices for both groups in the A and AV conditions are shown in the appendix. Analyses of feature-level perception using both information transfer and simple percent correct scores indicated that subjects from both language groups extracted broadly similar information from visual cues. As expected, AV presentation significantly improved both Spanish and English subjects' perception of consonant place and manner. Place errors were reduced by visual cues within each of the manner classes of plosive, fricative and nasal. The reduction of manner information is at least in part likely to arise from the strong correlation of manner and place for anterior English consonants. As predicted, errors of voicing in L2 were common in the A condition. These were not significantly reduced in the AV condition.

3.2. Consonant confusions

A comparison of confusion matrices (see Appendix) showed that Spanish speakers made two groups of errors. One group, Common Error Type (CET), includes the errors that were made (1) by both language groups, (2) were not predicted by the L1-L2 phoneme assimilation patterns and related to the acoustic-phonetic sound characteristics, and (3) involved target sounds that obtained a significant improvement in the audio-visual condition. The target sounds that obtained significant improvements were the [+anterior] sounds /p/, /b/ /ð/, /m/, /n/ for English listeners, and /p/, /f/ /ð/, /m/, /n/ for Spanish listeners. These sounds were mostly confused with sounds that contrasted in place and manner features in both language groups, but also with sounds contrasting in voicing in the case of the Spanish subjects. For example, the 11.2% errors that English subjects made in perceiving target /p/ in CV syllables in the auditory condition were mostly related to confusions that involved place and manner, i.e. /k/, /t/, /tʃ/. For Spanish speakers, the 20.4% errors in the perception of /p/ included errors of place and/or manner with voicing, i.e., /d/, /g/, /v/, /dʒ/, and also pure voicing errors, i.e. /b/. The addition of visual cues reduced the errors of place and manner in both language groups.

The second group of errors included those predicted by PAM’s assimilation patterns (i.e., Error type I, II, and III of Figure 1). These errors were found only in L2. Pure voicing errors refer to error type I and II, which are illustrated in the assimilation patterns of Figure 1a and b. For example, in the auditory condition Spanish speakers identified voiced obstruents with their voiceless counterparts, with errors ranging from 24.3% to 45.2% (see Appendix). The addition of visual cues did not improve these errors, which ranged from 22.7% to 46.3%.

Error type III, which was predicted by the assimilation pattern illustrated in Figure 1c, involved the confusions /v/-/b/ and /ð/-/d/ and were indeed made by Spanish listeners, mainly in CV and VCV syllables. In contrast, English listeners only occasionally labeled auditory /b/ as /v/, and they never labeled /v/ as /b/. The addition of visual cues did not reduce the rate of type III errors by the Spanish subjects. In the auditory condition, 20.8% of /v/ targets were identified as /b/ in CV syllables, and 24.5% in VCV syllables. In the audiovisual condition these confusions occurred at rates of 18.1% and 25.5% respectively. The /ð/-/d/ confusion was bi-directional. In the auditory condition, 14.4% of /d/ targets were identified as /ð/ in CV syllables, and 16.2% in VCV syllables. 33% of responses to target /ð/ were /d/ in CV, and 32.9% in VCV. Visual cues did not improve these scores, which reached 14.8% in CV and 16.7% in VCV for /d/ targets, and 29.7% and 29.4% for /ð/ targets.

4. DISCUSSION

The addition of visual cues improved the consonant perception by both native and L2 speakers. Language background had no discernible effect in this improvement. However, the amount of phonetic information affected the audio-visual improvement. In the VC context where consonants were most difficult to identify in the auditory condition, both native and L2 subjects obtained the most improvement with the addition of visual cues.

Feature analysis showed that visual cues led to a reduction of errors in place and manner of articulation for English consonants that is similar for Spanish listeners to that shown by native listeners. Since place and manner of articulation are correlated in English consonants, it is possible that subjects extracted information mainly about place of articulation from [+anterior] sounds, which have visible articulations. Moreover, since CET errors were common to both languages, they can be related to the acoustic-phonetic characteristics of the stimuli.

Errors related to language, i.e. type I, II, and III errors, offer some interesting results. Given that visual information can influence stop VOT boundaries for native speakers [7, 8 9], it is conceivable that visual cues might enable Spanish L2 listeners to switch to the use of English VOT boundaries and consequently improve type II errors. Although visual information could have helped Spanish speakers to hear these differences in VOT length, they did not use them in phoneme classification, although this may change after auditory-visual training. Further research with discrimination tasks could assess whether L2 perception of VOT can be influenced by the addition of visual cues.

Type III errors, which included the confusions /v/-/b/ and /ð/-/d/, did not show any significant lessening in the presence of visual cues, despite these involving [+anterior] targets confused in place and manner of articulation. Spanish subjects may have learnt to disregard certain visual cues to place/manner in their L1, since voiced stops have

[+continuant] and [-continuant] allophones, and may have transferred this perceptual pattern to L2. Consequently, Spanish speakers of English may have used the visual cues to place/manner as allophonic features, not as distinctive cues to a phonemic distinction. This explanation would indicate that visual features, like auditory features, can have different weights when cueing phonemic and allophonic distinctions. Learning a L2 may establish an L2 specific representation involving L2 visual as well as auditory feature weights. Therefore, L2 confusions linked to L1 allophonic relations may be an important target for auditory-visual training.

5. REFERENCES

1. Flege, J.E. "Second-language Learning: the Role of Subject and Phonetic Variables" *Proc Speech Technology in Language Learning Conf.* (25-27 May,1998, Marholmen, Sweden)

2. Best, C.T. et al. "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system". *J. Acoust. Soc. Am, 109, 775-794, 2001*

3. Werker, J.F. and Logan, J. "Phonemic and phonetic factors in adult cross-language speech perception" *Percept. Psychophys, vol. 37, 35-44, 1985.*

4. Logan, J.S. and Pruitt, J.S. "Methodological issues in training listeners to perceive non-native phonemes" *Speech perception and linguistic experience: Issues in cross-language research, (Ed.) Winfred Strange,* York Press Inc., Timonium, Maryland, 1995.

5. Jamieson, D.G. and Morosan, D.E. "Training non-native contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones" *Percept. Psychophys., 40, 205-215, 1986.*

6. Hazan, V. and Simpson, A. "The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects" *Lang. Speech, 43, 273-294, 2000.*

7. Green, K.P., & Kuhl, P.K. " Integral processing of visual place and auditory voicing information during phonetic perception". *J. Exp. Psychol: HPP, 17, 278-288, 1991*

8. Faulkner, A, & Rosen, S. " Contributions of temporal encodings of voicing, voicelessness, fundamental frequency and amplitude variation in audio-visual and auditory speech perception" *J. Acoust. Soc. Am., 106, 2063-2073, 1999.*

9. Green, K. P. & Miller, L. L. "On the role of visual rate information in phonetic perception" *Percept. Psychophys., 50, 269-76, 1985.*

10. Akahane-Yamada, R., Bradlow, A., Pisoni, D.B., Tohkura, Y. "Effects of audio-visual training on the identification of English /r/ and /l/ by Japanese speakers" *J.Acoust.Soc.Am, vol. 102, 3137, 1997*

11. Davis, C. and Kim, J. "Perception of clearly presented foreign language sounds: the effects of visible speech" *Proc. AVSP '99* (Aug 7-10, 1999, Santa Cruz, CA).

12. Massaro, D. *Perceiving talking faces: From speech perception to a behavioral principle.* The MIT Press, Cambridge, MA, 1998.

Acknowledgement

Funded by EPSRC grant GR/N11148. We thank Prof. Ruth Campbell for her advice on the analysis of this data.

APPENDIX: CONSONANT CONFUSION MATRICES. Responses in columns as percentages. Consonants are ordered by place of articulation and grouped according to commoner errors within place.

English Listeners: Auditory condition. All contexts.

	m	p	b	v	f	ð	d	t	n	z	s	ʃ	tʃ	dʒ	g	k
m	**82**		1	3					13							
p		**93**						1								4
b	1	1	**82**	5		1	4								4	
v				**92**	1	6									1	
f				1	**92**	6										
ð			1	28		**65**	1			4						
d							**99**									
t						1		**99**								
n	11		1						**87**							
z										**90**	10					
s											**100**					
ʃ								1				**96**	2			
tʃ													**99**	1		
dʒ													4	**72**	24	
g															**100**	
k		1											3			**95**

English Listeners: Auditory-visual condition. All contexts

n.b. /g/ - /dʒ/ confusions here and elsewhere are likely to be orthographic in origin

	m	p	b	v	f	ð	d	t	n	z	s	ʃ	tʃ	dʒ	g	k
m	**97**		2													
p		**99**														
b		3	**96**													
v				**96**	2	1				1						
f				3	**95**	2										
ð				9		**84**				6					1	
d							**100**									
t								**100**								
n									**99**							
z										**92**	8					
s									1	1	**98**	1				
ʃ												**96**	3			
tʃ													**100**			
dʒ													2	**73**	25	
g															**100**	
k																**100**

Spanish Listeners: Auditory condition. All contexts.

Boxed cells indicate assimilation errors.
Type I - single line
Type II - double line
Type III - triple line

	m	p	b	v	f	ð	d	t	n	z	s	ʃ	tʃ	dʒ	g	k
m	**85**		1					0	13							
p		**88**	3					2						1	2	4
b		32	**57**	2	1	1	4	1							2	
v			15	**48**	24	8	3			2						
f				2	**92**	2				2	1					
ð			1	5	9	**49**	24			8	2	1				
d						10	**62**	24					1	1	1	
t						2	3	**83**					10	1		
n	10						1	0	**88**		0					1
z						1				**50**	36	11				
s						1	1			19	**74**	6				
ʃ						1				2	5	**89**	2			
tʃ						2		1				2	**86**	7	2	
dʒ							1						35	**54**	8	
g		1						1							**52**	45
k		1						1							6	**91**

Spanish Listeners: Auditory-visual condition. All contexts.

Boxed cells indicate assimilation errors.
Type I - single line
Type II - double line
Type III - triple line

	m	p	b	v	f	ð	d	t	n	z	s	ʃ	tʃ	dʒ	g	k
m	**97**								2							
p		**95**	4				1	1								
b		40	**54**	1			1	1							1	
v			15	**57**	25	2	1			1						
f				2	**97**											
ð				1	1	**61**	21			10	3	1				
d						11	**64**	23					1		1	
t						2	5	**83**					7	1	1	
n	2								**97**							
z						1				**53**	37	8				
s						1				15	**78**	6				
ʃ						1				3	4	**89**	2			
tʃ						2		1				1	**85**	9	2	
dʒ						1	1	1					35	**52**	10	
g		1						1							**52**	46
k		1						1							5	**93**
	m	**p**	**b**	**v**	**f**	**ð**	**d**	**t**	**n**	**z**	**s**	**ʃ**	**tʃ**	**dʒ**	**g**	**k**