AUDIOVISUAL PERCEPTION OF NATIVE AND NON-NATIVE SOUNDS BY NATIVE AND NON-NATIVE SPEAKERS[1]

Juli Cebrian
Universitat Autònoma de Barcelona
Juli.Cebrian@uab.cat

Angelica Carlet
Universitat Autònoma de Barcelona
Angelica.Carlet@uab.cat

This study examined the contribution of both auditory and visual cues in native and non-native speech perception. Three groups of listeners, a group of native speakers of English and two groups of EFL learners (elementary and advanced), were tested on their perception of auditory-only, visual-only and audiovisual English stimuli. The audiovisual stimuli included a congruent (matched auditory-visual stimuli) condition and an incongruent (mismatched auditory-visual stimuli) condition. The stimuli included consonant sounds that were common to both languages (/b, g/), non-L1 sounds (/v/) and sounds with a different status in the L1 and the L2 (/d, dh/).[2] The results indicated that whereas visual salience tends to play a role in native speakers' perception, non-native speakers' performance is more strongly influenced by the status of native vs. L2 sounds. A comparison of the two learner groups revealed a positive effect of L2 experience on the ability to perceive in a more native-like manner both auditorily and visually.

Keywords: audiovisual perception, foreign language learning, second language acquisition

1. Introduction

Most second language (L2) studies that explore how native and non-native speakers differ in their perception of L1 and L2 sounds focus on auditory information only. Yet speech communication is not limited to auditory cues. Visual cues play an important role, as demonstrated by the McGurk effect; using mismatched auditory and visual cues, McGurk and MacDonald (1976) observed that speakers of English were strongly influenced by the visual information despite contradictory information in the auditory input. This phenomenon has been reported for other languages as well, such as Japanese and Spanish (Massaro *et al*. 1993), and its effect in L2 speech has been examined in a few studies (see Hardison 2010). Sekiyama and Tokhura (1993) found that visual cues play a more important role in Japanese listeners' perception in a foreign language (American English) than in their native language, something that has been referred to as the "foreign language effect". This effect was also found with English speakers listening to Chinese speech (Hazan *et al*. 2006, but cf. Hayashi and Sekiyama 1998). This cross-linguistic difference in the degree of visual bias in speech perception may be due to a number of factors including the visual salience of sound contrasts, and differences in phoneme inventories across languages (Ortega-Llebaria *et al*. 2001; Wang *et al*. 2009; Hazan *et al*. 2010). The importance of visual cues is also supported by the outcomes of studies on L2 training (Hardison 2003; Hazan *et al*. 2005, 2006;

Aliaga-García 2010) as well as investigations on native and non-native perception in adverse conditions (García-Lecumberri & Cooke 2006, among others).

The goal of this paper is to investigate further the role of visual and auditory cues in native and non-native speech and to examine the effect of visual salience and of L1-L2 inventory differences on audiovisual perception of L2 sounds.

2. Methodology

An audiovisual perception experiment was conducted on two groups of Spanish/Catalan bilinguals and a group native English speakers tested on English stimuli.

2.1. Stimuli

The stimuli consisted of consonant + /a/ sequences involving the English stops /b, d, g/ and the fricatives /v, dh/. Apart from differences in VOT, /b, g/ are common to the L1 and the L2 in this study. The labiodental fricative (/v/) is not a phoneme of most varieties of Catalan and Spanish and is often mispronounced as one of the variants of the native sound /b/. In English /d/ is alveolar, and contrasts phonemically with the dental fricative /dh/, while in Catalan and Spanish, [d] and [dh] are in complementary distribution. In terms of visual detectability, the labial, labiodental and dental sounds are more salient than the alveolar and velar sounds. The characteristics of these sounds thus make them the ideal testing ground to explore the interaction between visual salience and L1-L2 relationships in audiovisual perception.

Test stimuli were elicited from a female native speaker of Southern Ontario English. Video recordings were made in a soundproof booth with the speaker's face fully visible set against a uniform background. A Canon Digital Video camcorder ZR950 was used to record the image onto a DAT minicassette. The signal was then transferred to computer files using Windows Movie Maker. Simultaneous with the video recording, the speaker's production was recorded with a Sound Devices 722 digital recorder connected to a DPA 4011 Cardiod Shotgun microphone to ensure the good quality of the sound. The speaker produced three repetitions of the non-sense /Ca/ syllables. Two repetitions were selected for the experiment. Three types of stimuli were prepared: visual-only (V), auditory-only (A) and audiovisual (AV) stimuli. A and V stimuli were created by extracting relevant portions of the audio recording and the video recording respectively. AV stimuli were created by combining the audio tracks with the visual files using Apple iMovie software. Two types of AV stimuli were created: congruent AV stimuli, in which the audio file and the video file were matched (e.g., auditory /ba/ and visual /ba/) and incongruent stimuli, where the video and the audio were not matched (e.g., auditory /ba/ and visual /va/).

2.2. Subjects

Ten speakers of Canadian English (mean age: 30) and twenty Spanish/Catalan bilinguals participated in the study. The latter consisted of ten second-year English major undergraduate students (mean age 21; henceforth advanced-intermediate English learners or AEL) and ten elementary learners of English, who had been studying English for about seven months in a language school (mean age 34; henceforth elementary EFL learners or EEL).

2.3. Procedure

Participants were presented with congruent and incongruent AV, visual-only and auditory-only stimuli and they had to write down on a sheet of paper the syllable they heard. The test was preceded by a short practice session with congruent AV stimuli. The experimenter sat with the subjects to make sure they were looking at the screen during the experiment and also to assist them should they have questions regarding how to write down their responses. NES were tested individually in a soundproof booth at the University of Toronto, Canada, facing a Dell computer screen and positioned equidistant from two good quality loudspeakers. AEL and EEL subjects were tested in a quiet room with an Acer portable PC and loudspeakers. Use of headphones was avoided to eliminate a possible bias to the auditory signal. A subset of the AEL-EEL subjects were tested in groups of two or three at a time. Although the experimenter monitored the subjects' performance, the fact that not all the AEL-EEL subjects were tested individually makes the results for the non-native group preliminary.

3. Results

In the case of the A, V and congruent AV conditions, responses were analyzed in terms of correct identification of the stimulus. Regarding the incongruent AV condition, responses were classified as 'visual' responses (i.e., the response matched the visual cue), 'audio' responses (the response matched the auditory cue), 'fused' responses (the response was intermediate in place of articulation, such as the labiodental /va/ with respect to the bilabial /ba/ and the dental /dha/, or in manner of articulation, such as the coronal stop /d/ with respect to the labial stop /ba/ and the coronal fricative /dha/), and 'other' for combined or unrelated responses. Given the preliminary nature of the non-native data, the results were not submitted to statistical analyses at this time. Group differences will thus be discussed in terms of numerical differences in the descriptive statistics.

The results for the auditory-only condition are presented in Table 1. The native English speakers had no difficulty identifying the target consonant correctly (90-100% correct identifications). The non-native speakers were able to correctly identify the stop consonants most of the time, although often heard as voiceless. This is not surprising given the lack of prevoicing typical of English voiced stops in initial position, corroborated by spectrographic analyses of the stimuli. With respect to the fricatives, EEL tended to hear /v/ as /b/, and /dh/ as /d/ (that is, the same response as for the sound /d/). The AEL group outperformed the EEL group and obtained 55% correct identifications of /v/ and /dh/, although still notably lower than AEL's scores.

Target sound	*NES*	*AEL*	*EEL*
/b/	**100 /b/**	**65 /b/, 17.5/p/**	**75 /b/, 15 /p/**
		7.5 /dh/, 5 /d/, 5/v/	7.5 /d/, 2.5 /g/
/d/	**90 /d/**	**60 /d/, 15 /t/**	**87.5 /d/, 7.5 /t/**
	10 /dhd/	15 /v/, 7.5 /dh/**,** 2.5 /th/	5 /g/
/g/	**100 /g/**	**60 /g/, 40 /k/**	**75 /k/, 25 /g/**
/v/	**90 /v/**	**55 /v/**	80 /b/
	7 /dh/, 3 /dhv/	35 /d/	10 /t/, **5 /v/**, 5 /g/
		10 /b/	
/dh/	**93 /dh/**	**55 /dh/**	60 /d/, 10 /t/
	7 /v/	30 /b/, 15 /d/	15 /g/, 10 /b/, 5 /p/

Table 1: Percentage of responses in the auditory-only condition for native English speakers (NES), advanced-intermediate EFL learners (AEL), and elementary EFL learners (EEL). Correct identifications are presented in boldface.

Table 2 presents the results for the visual-only condition. The NES obtained high scores of identification for all sounds except /g/. The lower scores for the velar stop are not surprising given that it is the least visually salient of the five consonants. Interestingly, non-native speakers were more successful than NES at identifying /g/ as a velar stop (60-65%). The non-native's scores for /b/ were also high (92.5%), but their results for the other consonants were lower. The non-native listeners provided more voiceless responses, probably due to their perception of some stimuli as voiceless in the AV and A conditions. Regarding the fricatives, AEL subjects outperformed the EEL group again, yielding more correct identifications for the voiced fricatives /v/ and /dh/. The EEL group tended to identify the labiodental as /f/ and the interdental as /d/, showing a clear influence of the native language inventory.

Target sound	*NES*	*AEL*	*EEL*
/b/	**97.5 /b/**	**42.5 /b/ 55 /p/**	**70 /b/, 22.5 /p/**
	2.5 /bv/	2.5 /d/	7.5 /d, dh, g/
/d/	**87.5 /d/**	**37.5 /d/, 22.5 /t/**	**47.5 /d/, 12.5 /t/**
	7.5 /l/, 5 /g/	15 /g/, 7.5 /th/	25 /b, p/, 15 /g, k/
		5/dh/, 5 /p/, 2.5 /f/	
/g/	**17.5 /g/**	**30 /g/, 30 /k/**	**50 /g/, 15 /k/**
	39 /d/, 14 /l/	20 /b, p/	25 /d/, 10 /t/
	22.5 /h/, 5 /dh/, 2 /b/	10 /d/, 5 /t/, 5 /n/	
/v/	**85 /v/, 12.5 /f/**	**40 /v/, 35 /f/**	**70 /f/**
	2.5 /g/	10/dh/ 10 /b/, 5 /g/	15 /b/, 5 /d/, 10 /g, k/
/dh/	**95 /dh/**	**40 /dh/, 20 /th/**	**10 /dh/, 10 /th/**
	5 /d/	10 /d,t/	50 /d/, 15 /b/, 15 /p/, 15 /g/

Table 2: Percentage of responses in the video-only condition for native English speakers (NES), advanced-intermediate EFL learners (AEL), and elementary EFL learners (EEL). Correct identifications are presented in boldface.

Regarding the congruent AV condition, as illustrated in Table 3, the NES group was highly successful at identifying the stimuli correctly. The non-native listeners were also able to identify the stop consonants correctly, albeit with a number of voiceless stop responses, as expected from their responses to the auditory-only responses. Again the voiced labiodental and the voiced dental fricatives posed the greatest challenge. As with the previous two conditions, the AEL group outperformed the EEL group regarding these two sounds. The AEL group reached 85% correct identification of /v/, as opposed to only 10% by EEL. Similarly, the AEL perceived /dh/ correctly 50% of the time, and responded /d/ 35% of the time, whereas the EEL's responses overwhelmingly involved /d/ (95%).

The results for the incongruent AV condition are presented separately for each group in Tables 4, 5 and 6 for the NES, EEL and AEL groups respectively. For the sake of brevity, we will mainly focus on the general trends observed for each group. Regarding the canonical example of the McGurk effect, that is, fused (i.e. /da/) responses for the visual /ga/-auditory /ba/ combination, the results of this study showed a moderate effect. Native speakers yielded a majority of fused responses (63%), followed by auditory responses (37%). The non-native groups showed a smaller amount

of fused responses (30-35%) and more auditory responses, perhaps due to differences across subjects in testing conditions.

Target sound	*NES*	*AEL*	*EEL*
/b/	**98 /b/**	**55/b/, 42.5 /p/**	**75 /b/, 20/p/**
	2 /d/	2.5 /m/	2,5 /t,d/
/d/	**98 /d/**	**70 /d/**	**80 /d/, 15 /t/**
	2 /l/	20 /th/, 10 /dh/	2.5 /g/, 2.5 /k/
/g/	**100 /g/**	**40 /g/, 60 /k/**	**40 /g/, 60 /k/**
/v/	**90 /v/**	**85 /v/**	**10 /v/,** 10 /f/
	5 /dh/, 2.5 /g/, 2.5 /dhv/	10 /b/, 5 /t/	70 /b/, 5 /d/, 5 /t/
/dh/	**97.5 /dh/**	**50 /dh/**	**90 /d/,** 5/t/
	2.5 /d/	35 /d/, 10 /th/, 5 /bh/	2.5 /b/ 2.5 /g/

Table 3: Percentage of responses in the congruent audio-visual condition for native English speakers (NES), advanced-intermediate EFL learners (AEL), and elementary EFL learners (EEL). Correct identifications are presented in boldface.

Comparing across groups, we can see again that the responses of the NES group differ the most from those of the EEL group, with the AES group falling somewhere in between. For the native speakers, cue weight was related to cue salience on most occasions. For instance, visual and fused responses were most frequent with labial, labiodental and dental gestures than with alveolar and velar articulations, denoting the less visible nature of the latter. In addition, a visual bias was also more frequent with fricative gestures than with stops. On the whole, the non-native listeners appeared to be more influenced by the relationship between the target stimuli and their L1 phonological system. In cases where the combination involved a native and a non-native sound, the non-native listeners tended to provide the response that corresponded to the native cue. For example, given the visual /ba/-auditory /va/ stimulus, the EEL listeners yielded 85% /ba/ responses, while in the visual /va/-auditory /ba/ stimulus, they responded /ba/ 80% of the time. The AEL group approximated the native speakers more closely, providing more non-L1 responses and more fricative responses with fricative gestures.

Stimuli		Responses			
Visual	Audio	Visual	Audio	Fused	Other
ba	va	42.5	42.5		10 /bv/, 5 /dhv, bdhv/
ba	dha	22.5	42.5	27.5 /v/	2.5 /dhb/, 5 /bv, vdh/
ba	da	5	77.5		12.5 /bd/, 2.5 /bdhv/, 2.5 /v/
ba	ga		85.5		7.5 /bg/, 5 /v, w/
va	ba	65	32.5		2.5 /g/
va	dha	49	46		5 /vdh/,
va	da		97.5		2.5 /dhd/
dha	ba	67.5	12.5	20 /d/	
dha	va	92.5	5		2.5 /d/
dha	da	5	82.5		2.5 /dhd/, 2.5 /dhg/
da	ba	36 /d/ 32 /dh/	27.5		2.5 /dhv/, 2.5 /g/
da	va		56.3	36.3 /dh/	5 /vdh/, 2.5 /g/
da	dha		97.5		2.5 /v/
da	ga		100		
ga	ba		37	26 /d/, 37 /dh/	
ga	da	2.5	95		2.5 /b/

Table 4: Percentage of responses in the incongruent audio-visual condition for native English speakers

Stimuli		Responses			
Visual	Audio	Visual	Audio	Fused	Other
ba	va	85	10		5 /dh/
ba	dha	70	30		
ba	da	22.5	70		5 /dh/, /2.5 /g/
ba	ga	15	75	10 /d/	
va	ba	10 /f/	80	10 /bh/	
va	dha	20 /f/			45 /b/, 35 /d/
va	da	5 /f/	90	5 /dh/	
dha	ba	5	30	65 /d, t/	
dha	va	10	5 /f/		30 /b/, 5 /g/, 50 /d/
da	ba	37.5	52.5		5 /g/, 5 /th, dh/
da	va	55	5	25 /b/	5 /g/, 10 /dh/
da	dha	85	10		5 /g/
da	ga		100		
ga	ba	5	65	30 /d/	
ga	da		95		5 /b/

Table 5: Percentage of responses in the incongruent audio-visual condition for the EEL group

Stimuli		Responses			
Visual	Audio	Visual	Audio	Fused	Other
ba	va	40	55		5 /t/
ba	dha	25	20	35 /v/, 15 /d, t/	5 /g/
ba	da	25	67.5		7.5 /dh/
ba	ga	10	90		
va	ba	20	75		5 /d/
va	dha	45	10		40 /d/, 5 /b/
va	da		95	5 /dh/	
dha	ba	25	35	15 /v/, 25 /d/	
dha	va	65	10		15 /d/, 5 /b/
da	ba	30	60		5 /v/, 2.5 /dh/, 2.5 /l/
da	va	20	25	10 /b/, 40 /dh/	5/g/
da	dha	40	60		
da	ga		100 /g, k/		
ga	ba	10	55	35 /d/	
ga	da		100		

Table 6: Percentage of responses in the incongruent audio-visual condition for the AEL group

4. Discussion and conclusions

This study set out to explore the contribution of auditory and visual cues in the perception of English consonants and to evaluate the effect of cue salience and L1-L2 differences in phonemic status on non-native perception. The effect of the native language system on the audiovisual perception of non-native consonants was clear in a number ways. First, cross-linguistic differences in stop VOT and fricative voicing accounted for the fact that the non-natives, but not the natives, often perceived voiceless consonants. The influence of the L1 inventory was also particularly clear in the non-natives' perception of /v/ and /dh/. The voiced labiodental was perceived as /b/ when only the sound was present and as /f/ when the only visual cue was present, conforming to the closest L1 categories. In the AV congruent condition, the non-natives again tended to perceive the target sounds /v/ and /dh/ as /b/ and /d/, respectively. With

respect to the relative weight of auditory vs. visual cues, the results for incongruent AV condition indicate a greater effect of visual cues for native speakers and a strong influence of L1-L2 phonemic inventory differences for non-native speakers.

An effect of FL learning was evident in that the AEL group outperformed the EEL group in all conditions. The EEL group showed a greater influence of the native language inventory, as illustrated by their /f/ and /d/ responses for the visual /v/ and /dh/ stimuli, respectively, as well as the number of /b/ and /d/ responses for the auditory and audiovisual /v/ and /dh/ targets. The AEL group, on the other hand, showed evidence of L2 learning; even though as a whole the AEL group's results were still numerically lower than the NES's, the advanced learners were able to distinguish /ba/ from /va/ and /da/ from /dha/ both visually and auditorily.

These outcomes will need to be contrasted with follow-up studies exploring the effect of EFL proficiency on audio-visual perception in more controlled testing conditions. Still, despite the preliminary nature of the results of this study, the difference between the two non-native groups points to the fact that learning new (L2) auditory categories goes hand in hand with the learning of the visual or gestural characteristics of those sounds.

Notes

1. Work supported by grant FFI2010-19206/FILO from the Spanish Ministry of Science and Innovation and a *José Castillejos* grant (JC2010-0246) from the Spanish Ministry of Education to the first author.

2. Throughout the paper, "dh" and "th" stand for the voiced and voiceless dental fricatives, IPA symbols /ð/ and /θ/, respectively.

Works Cited

Aliaga-García, Cristina 2010: "Measuring Perceptual Cue Weighting after Training: a Comparison of Auditory vs. Articulatory Training Methods". A K. Dziubalska-Kołaczyk, M. Wrembel and M. Kul, ed. *New Sounds. 2010 Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech*. Berne, etc: Peter Lang. 77-82.

García Lecumberri, María Luisa and Michael P. Cooke 2006: "Effect of Masker Type on Native and Non-native Consonant Perception in Noise". *Journal of the Acoustical Society of America* 119.4: 2445-54.

Hardison, Debra M. 1996: "Bimodal Speech Perception by Native and Nonnative Speakers of English: Factors Influencing the McGurk Effect". *Language Learning* 46.1: 3-73.

_____ 2003: "Acquisition of Second-Language Speech: Effects of Visual Cues, Context, and Talker Variability". *Applied Psycholinguistics* 24: 495-522.

_____ 2010: "Visual and Auditory Input in Second-Language Speech Processing". *Language Teaching* 43.1: 84-95.

Hayashi, Yasuko and Sekiyama, Kaoru 1998: "Native-Foreign Language Effect in the McGurk Effect: a Test with Chinese and Japanese". Eric Vatikiotis-Bateson, Denis Burnham and Jordi Robert-Ribes, eds. *Proceedings of AVSP 1998*. Sydney?: s.n. 61-66.

Hazan, Valerie, Anke Sennema, Midori Iba and Andrew Faulkner 2005: "Effect of Audiovisual Perceptual Training on the Perception and Production of Consonants by Japanese Learners of English". *Speech Communication* 47: 360-78.

Hazan, Valerie, Anke Sennema, Andrew Faulkner, Marta Ortega-Llebaria, Midori Iba and Hyunsong Chung 2006: "The Use of Visual Cues in the Perception of Nonnative Consonant Contrasts". *Journal of the Acoustic Society of America* 119: 1740-51.

Hazan, Valerie, Jeesum Kim and Yuchun Chen 2010: "Audiovisual Perception in Adverse Conditions: Language, Speaker and Listener Effects". *Speech Communication* 52: 996-1009.

Massaro, Dominic W., Minoru Tsuzaki, Michael Cohen, Antoinette Gesi and Roberto Heredia 1993: "Bimodal Speech Perception: An Examination across Language". *Journal of Phonetics* 21: 445-78.

McGurk, Harry and John MacDonald 1976: "Hearing Lips and Seeing Voices". *Nature* 264: 746-48.

Ortega-Llebaria, Marta, Andrew Faulkner and Valerie Hazan 2001: "Auditory-visual L2 Speech Perception: Effects of Visual Cues and Acoustic Phonetic Context for Spanish Learners of English". *Proceedings of AVSP*: 49-154.

Sekiyama, Kaoru and Yoh'ichi Tohkura 1993: "Inter-Language Differences in the Influence of Visual Cues in Speech Perception". *Journal of Phonetics* 21: 427–44.

Wang, Yue, Dawn M. Behne and Haisheng Jiang 2009: "Influence of Native Language Phonetic System on Audio-Visual Speech Perception". *Journal of Phonetics* 37: 344–56.

THIS TEXT IS PART OF THE VOLUME:

Martín Alegre, Sara (coord. and ed.), Melissa Moyer (ed.), Elisabet Pladevall (ed.) & Susagna Tubau (ed.). *At a Time of Crisis: English and American Studies in Spain*. Departament de Filologia Anglesa i de Germanística, Universitat Autònoma de Barcelona/AEDEAN, 2012. ISBN-10: 84-695-4273-7, ISBN-13: 978-84-695-4273-6. Available from www.aedean.org