Visual attention during L1 and L2 sounds perception: an eye-tracking study

Bianca Sisinni[1], Mirko Grimaldi[1], Elisa Tundo[1], Andrea Calabrese[2]
[1]Department of Philology, Linguistics and Literature, University of Salento, Italy
[2]Department of Linguistics, University of Connecticut, Massachusetts, USA

https://doi.org/10.36505/ExLing-2010/03/0043/000163

Abstract

Visual information affects speech perception as demonstrated by the McGurk effect (McGurk & McDonald, 1976): when audio /ba/ is dubbed with a visual /ga/, what is perceived is /da/. This study aims at observing how visual information, intended as articulatory orofacial movements, is processed by eye, i.e., if gaze is related to articulatory information processing. The results indicate that visual attentional resources seem to be higher during multisensory (AV) than unisensory (A; V) presentation. Probably, higher visual attentional resources are needed to integrate inputs coming from different sources. Moreover, audiovisual speech perception seems to be similar across languages (e.g., Chen & Massaro, 2004) and not language-specific (Ghazanfar et al., 2005).

Key words: Audiovisual speech, multisensory integration, native and non-native perception

Introduction

Visual information affects speech perception as proved by the McGurk effect (McGurk & McDonald, 1976) that shows that when audio /ba/ is dubbed with a face saying /ga/, what is perceived is /da/.

There are conflicting results on the audiovisual (AV) non-native perception. Navarra and Soto Faraco (2007) suggested that the addition of visual information about the articulatory gestures enhanced the ability to discriminate sounds in second language and that the integration of visual-gestural plus auditory information can produce a specific improvement in phonological processing. Hazan et alii (2002) found that non-native speech perception seemed not to be affected by the addition of visual information and concluded that the sensitivity to acoustic and visual cues in L2 acquisition process can be not so evident especially in the early stage of the process and that this sensitivity can largely vary across learners.

Our aim is to observe if gaze behaviour changes according (i) to the speech stimuli presented, i.e., native (L1) and non native (L2) stimuli, and (ii) to the modality of stimuli presentation, i.e., audio (A), video (V) and audio-video (AV), during an almost no-task experiment.

Method

Participants

Nine subjects (1 male) participated (mean age: 19). They all were native speakers of Salento Italian, a variety spoken in the south part of Apulia, and they all attended the first year at Salento University, Faculty of Foreign Language and Literature. In a questionnaire, they reported a formal knowledge of English (mean English education: 12 years) and also declared to have a normal or corrected to normal vision and normal hearing. They were informed about the experimental procedure, approved by the ASL/LE Ethical Committee, and gave their written consent.

Stimuli

Stimuli were produced by two female speakers differing in native language, i.e., Italian (L1) and American (L2). The Italian speaker produced /a/, /e/, /i/, /u/ and the American one produced /æ/, /ɜ˞/, /ɪ/ and /ʊ/ vowels. Speakers were filmed in a soundproof room by a the camera Camescope Canon MV 960 KIT and their productions were recorded by an SM58 microphone by means of Computer Speech Lab (sampling rate: 22.05 kHz). Video and audio stimuli have been then synchronised. Each stimulus presented the head and the higher part of speakers' shoulders and had a total duration equal to the vowel duration and 0,15 sec before and after the production of the vowel, with the speakers preserving a neutral facial expression.

The L1 and L2 stimuli were randomly presented 15 times in the A, V and AV modalities for a total of 3240 stimuli (8 stimuli x 15 repetitions x 3 modalities x 9 subjects). Subjects were simply asked to watch the computer screen where stimuli were presented. A question slide about what subjects heard (or saw) was randomly presented in order to be sure that they were actually performing the task. The audio of the stimuli was presented throughout loudspeakers at a comfortable level.

Procedure

The experiments were performed in a quiet dark room. Subjects were seated in front of the pc monitor Hp 1702 at a distance of 80 cm. They positioned their head in a headrest and chinrest in order to prevent movements. Right eye movements were monitored by means of ASL Eye-trac 6000, by an infrared camera Canon VC-C50i (sampling rate 120 Hz). A 9-point calibration was used. Spatial error between true gaze position and the computer measurement was less than 1 degree and the precision on a point was better than 0.5 degree. A region of interest (ROI) has been studied in this work, i.e., the mouth. In order to detect this ROI, single frames of the two speakers producing, respectively, the vowels /a/ and /æ/, have been overlapped by a MATLAB script and a ROI compatible with both the two

frames have been considered. This ROI "mouth" was delimited by an semi-horizontal visual angle of 1,9° and a semi-vertical visual angle of 1,5°.

The gaze behaviour has been quantified in terms of (i) the percentage of time spent gazing in the ROI, (ii) the number of fixations (i.e., a fixation last 100ms at least) in the ROI, and (iii) the average duration of fixations in the ROI. These dependent variables have been analysed in series of ANOVA investigating separately the effect of the Condition of stimuli presentation (A, V, AV) and of Language (L1 and L2). Only the significant results ($p<0,05$) are reported.

Results

When comparing each of the three dependent variables on the basis on the Condition, the results showed that, in both languages, there was a significant difference [$F(2,1619) = 21,320$ to $43,584$ $p<0,05$]. The post hoc test showed that the dependent variables were higher in AV than in V than in A, as visible in Table 1.

Table 1. The percentage of time spent gazing (% time), the number of fixations (N° fix) and the average duration of fixations (A.d.(s)) in the "mouth" ROI, in each condition (A, V, AV) for each language (L1, L2). Standard deviations are in parenthesis.

	A L1	A L2	V L1	V L2	AV L1	AV L2
% time	2,48 (5,95)	3,56 (8,43)	5,64 (11,27)	5,45 (11,60)	8,02 (12,57)	8,29 (13,15)
N° fix	0,41 (0,97)	0,48 (1,12)	0,84 (1,56)	0,66 (1,34)	1,13 (1,75)	1,01 (1,51)
A.d.(s)	0,037 (0,07)	0,043 (0,08)	0,05 (0,09)	0,05 (0,09)	0,09 (0,11)	0,081 (0,10)

In the comparison of the dependent variables between each language in each single condition, there was a significant difference for the percentage of time fixation in A [$F(1,1079) = 5,898$ $p < 0,05$], since this variable was higher in L2 than in L1, and a significant difference in V [$F(1,1079) = 4,012$ $p < 0,05$] for the number of fixations, higher in L1 than in L2 (Table 1). On the whole, there were no other relevant differences among the dependent variables in the two languages.

Discussion and conclusion

The first most significant finding in our results is the concentration of the gaze on the mouth in the AV condition. The fact that visual attention—i.e., the variables related to gaze behaviour—is higher in AV condition than in V or in A alone, indicate that visual attentional resources are higher during

multisensory presentation than during unisensory presentation. Our results can be considered in line with the results of Tiippana et al. (2004) which found that visual attention is required to combine visual and auditory speech features since when it is disrupted, audiovisual integration is less efficient.

The other significant finding is that gaze behaviour was not different when processing L1 and L2 sounds. Namely, visual information during L2 vowel speech perception seems not to be processed in a different way than the L1 vowel speech perception. Navarra and Soto Faraco (2007) found that vowel discrimination was not different between L1 and L2 speakers when visual information was added. They argued that in order to interpret linguistic speech gestures, the listener does not need to be familiar being the interpretation of visual gestures not language specific. Accordingly, our data seem to provide further evidence for this hypothesis. The lack of differences in visual processing during perception of L1 and L2 sounds could be accounted for by assuming that the underlying process of audiovisual speech perception is similar across languages (e.g., Chen & Massaro, 2004), so that no special visual processing is needed to perceive foreign sounds. It follows that the higher visual attentional resources that are observed during multisensory presentation—another finding of our study—are needed to integrate sensory inputs from different sources as part of a perceptual process, which is not language-specific and is based on a general process of perception also found in non-human primates (Ghazanfar et al., 2005).

References

Chen, T., Massaro, D.W. 2004. Mandarin speech perception by ear and eye follows a universal principle. Perception and Psychophysics**.** 66**,** 820–836.

Ghazanfar, A.A., Maier, J.X, Hoffman, K.L., Logothetis, N.K, 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. The Journal of Neuroscience. 25(29), 5004-5012

Hazan, V., Sennema, A., Faulkner, A. 2002. Audiovisual perception in L2 learners. In Proc. of the Intern. Conference for Spoken Language Processing, 1685-1688.

McGurk, H., McDonald, J. 1976. Hearing lips and seeing voices. Nature, 265, 746-748.

Navarra, J., Soto Faraco, S. 2007. Hearing lips in a second language. Psychological Reasearch, 71, 4-12.

Tiippana, K., Andersen, T.S., Sams, M. 2004. Visual attention modulate audiovisual speech perception. European J. of Cognitive Psychology. 16, 457-462.