BRAIN RESEARCH 1407 (2011) 13–26

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Person identification through faces and voices: An ERP study

Ileana Quiñones González[a,*], María Antonieta Bobes León[a], Pascal Belin[b], Yaiselene Martínez-Quintana[a], Lidice Galán García[a], Manuel Sánchez Castillo[c]

[a]Cognitive Neuroscience Department, Cuban Neuroscience Center, Cuba
[b]Voice Neurocognition Level, Psychology Department, University of Glasgow, UK
[c]Electronics Department, Cuban Neuroscience Center, Cuba

ARTICLE INFO

Article history:
Accepted 11 March 2011
Available online 17 March 2011

Keywords:
Person recognition
Cross-modal interaction
Faces
Voices

ABSTRACT

Different models have been proposed to explain how identity is extracted from faces and voices and how these two sensory systems interact. The neural loci of audio–visual interactions have been studied using neuroimaging techniques; however, the time course of these interactions is not well established. Here, we use event related potentials (ERPs) to study the temporal dynamics of the interaction of face and voice processing modules. We presented to the subjects either faces alone (F), voices alone (V) or faces and voices together (FV) in a familiarity detection task. Responses obtained for FV were compared with the sum of the responses obtained for F plus responses for V, for familiar and unfamiliar stimuli. This comparison shows differences in amplitude for different latencies, indicating cross-modal interactions. For unfamiliar stimuli, this interaction began very early (around 200 ms) and was restricted to the time window corresponding to the face N170 component. For familiar stimuli, the interaction was longer, began earlier and remained until after the N170 component. These results indicate that the interaction between faces and voices occurs from the initial stages of processing and continues as the person identification process goes on. This study is the first electrophysiological evidence of cross-modal interaction of faces and voices during the recognition of acquaintances' identities. It suggests that the assessment of person familiarity can result in direct information sharing between voice and face sensory modules from the early processing stages, before access to the person identity nodes.

1. Introduction

The recognition of people is a multimodal process essential in social life. This process is carried out using different cues such as faces and voices. Different models have been proposed to explain how identity recognition is extracted from faces and voices and how these two sensory systems, visual and auditory, interact. Ellis et al. (1997) postulated a cognitive model that consists of two processing modules working in parallel, corresponding to the auditory and visual modalities. Each includes homologous stages: a step for "structural encoding" (auditory and visual) followed by modality-specific "recognition units" [face recognition units (FRU) and voice recognition units (VRU)]. These two pathways converge in a "person identity node" (PIN), which associates semantic information. In this model, the cross-modal interaction only

* *Corresponding author*. Cuban Neuroscience Center, 25th Ave and 158, #15802, Cubanacán, P.O. Box 6414, Playa, Havana City 10600, Cuba. Fax: +53 7 208 6707.
E-mail address: fucacu@yahoo.com (I.Q. González).

0006-8993/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2011.03.029

occurs at the PIN level. For their part, Belin et al. (2004) proposed an anatomical model which predicts functional dissociations analogous to those proposed for faces in Bruce and Young's model (1986). This model proposes that different processing modules interact across the homologous stages in the processing architecture, besides the cross-modal interaction at the PIN level (Belin et al., 2004; Campanella and Belin, 2007).

Both proposals attempt to explain the integration of the two sensory systems during person recognition processing; however, this phenomenon is relatively less well researched and the evidence for supporting the models is insufficient. However, face and voice integration has also been studied in the context of speech perception (McGurk and MacDonald, 1976; Sams et al., 1991; Van Wassenhove et al., 2005) and affective information processing (Foxe and Schroeder, 2005; Pourtois et al., 2005; Kreifelts et al., 2007). One example of face and voice speech integration is provided by the robust illusion known as the McGurk effect (McGurk and MacDonald, 1976; Sams et al., 1991; Van Wassenhove et al., 2005), whereby incongruent facial and vocal phonetic information results in an intermediate percept. In this case, perception may represent a merger of face and voice processing. Most neuroimaging (Munhall et al., 2009; Stevenson and James, 2009; von Kriegstein et al., 2010), ERP (Sams et al., 1991; van Wassenhove et al., 2005) and behavioral (Arnal et al., 2009) studies have tried to characterize the functional architecture involved in the integration process, especially as it relates to speech perception. These studies suggest that the integration of speech, identity and affective information from faces and voices involves different cortical regions (Campanella and Belin, 2007). For this reason, in this paper we focus on the evidence related to the integration process during identity processing.

One of the strongest pieces of evidence for the domain's specificity and analogy in the processing of faces and voices is provided by neuroimaging studies. In normal human subjects, two bilateral areas have been found to respond more to pictures of faces than to pictures of objects, especially in the right hemisphere (Kanwisher et al., 1997; Haxby et al., 1999; Gauthier et al., 2000; Rossion et al., 2003; Minnebuscha et al., 2009). These regions include the fusiform face area (FFA) and the occipital face area (OFA) and they are considered a module for face perception (Haxby et al., 2000). More recently, Belin et al. (2000) found a discrete region in the auditory cortex that exhibited a greater response to vocal sounds as compared to non-vocal sounds. These voice-sensitive cortical areas were located along the superior bank of the superior temporal sulcus (STS). This region might be equivalent to the FFA in visual processing.

Functional magnetic resonance imaging (fMRI) studies have provided evidence of cross-modal neural activations of face-specific regions in the context of speaker identification (von Kriegstein et al., 2005a; 2005b; von Kriegstein and Giraud, 2006). These studies, which measured brain activity during identification tasks where subjects focused on either the speaker's voice or the verbal content of sentences, found evidence that familiar persons' voices activated the FFA when the identification task was to focus on the speaker's identity. Additionally, functional connectivity between FFA and STS during familiar speaker recognition was obtained, leading to the conclusion that interactions between the person familiarity assessments can result from direct information sharing between auditory voice and visual face regions and do not necessarily engage supramodal cortical substrates. Thus, the authors suggest that interaction between face and voice regions can occur during person recognition without the participation of the supramodal cortical locus underlying person identity information (PIN) (von Kriegstein et al., 2005a; 2005b).

An alternative explanation to the findings discussed above, which cannot be discarded due to the poor temporal resolution of fMRI, is that the activation found in FFA is caused by feedback connections from other neural regions located after FFA in the feed-forward pathway, including the PIN. If the locus of the interaction between face and voice modules during person recognition is relayed onto specific sensory modules, as suggested by other authors (von Kriegstein et al., 2005a; 2005b; von Kriegstein and Giraud, 2006), it might occur earlier in time, at the level of the structural encoding stage. Given their high temporal resolution, ERP recordings may be an appropriate methodology for establishing the stage at which cross-modal interactions occur during person recognition through faces and voices.

Some ERP components have been associated with face and voice processing, for example, the N170 for faces and the "voice-specific response" (VSR) for voices. The N170 was first described by Bentin et al. (1996). They found a negative-going component that responds more to face stimuli than to other visual object categories, with topographic distribution restricted to temporo-parietal regions and latencies between 140 and 190 ms post-stimulus onset. Several lines of evidence indicate that the N170 is related to the structural encoding stage of face processing (Bentin and Deouell, 2000; Eimer, 2000a; 2000b; Sagiv and Bentin, 2001). The VSR was described by Levy et al. (2001), who compared the responses evoked by singing voices and tones played by different musical instruments. The VSR is a positivity peaking around 320 ms after stimulus onset with a right central topography. Although less studied, it has been suggested that this component reflects the identification and distinction of human voice stimuli indexed by earlier ERP differences (Levy et al., 2001). More recently, a fronto-temporal positivity to voices (FTPV) was found (Charest et al., 2009). This component appears at about 164 ms and exhibits a greater response to voices in comparison to non-voice sounds such as birdsong and environmental sounds. These authors suggest that the FTPV is comparable to the well-known face preferential N170.

Face familiarity processing has also been studied using ERPs (Schweinberger et al., 1995; Bentin and Deouell, 2000; Schweinberger et al., 2002; De Haan et al., 2003; Schweinberger et al., 2004; Neumann and Schweinberger, 2009). A specific ERP modulation resulting from the immediate repetition of faces in the time range immediately following the N170 has been identified. This component (N250r) is observed between 230 and 300 ms and peaks at approximately 250 ms over inferior temporal scalp regions. The N250r has a more anterior and ventral distribution than the N170, is lateralized to the right hemisphere and is larger for familiar than for unfamiliar faces, suggesting a relationship with recognition of individuals. It