BRAIN RESEARCH 1385 (2011) 229–245

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Crossmodal interaction of facial and vocal person identity information: An event-related potential study

Julia Föcker[*,1], *Cordula Hölig, Anna Best, Brigitte Röder*

University of Hamburg, Biological Psychology and Neuropsychology, Von-Melle-Park 11, D-20146 Hamburg, Germany

ARTICLE INFO

Article history:
Accepted 9 February 2011
Available online 15 February 2011

Keywords:
Priming
Crossmodal
Multisensory
Event-related potentials
Faces
Voices

ABSTRACT

Hearing a voice and seeing a face are essential parts of person identification and social interaction. It has been suggested that both types of information do not only interact at late processing stages but rather interact at the level of perceptual encoding (<200 ms). The present study analysed when visual and auditory representations of person identity modulate the processing of voices. In unimodal trials, two successive voices (S1–S2) of the same or of two different speakers were presented. In the crossmodal condition, the S1 consisted of the face of the same or a different person with respect to the following voice stimulus. Participants had to decide whether the voice probe (S2) was from an elderly or a young person. Reaction times to the S2 were shorter when these stimuli were person-congruent, both in the uni- and crossmodal conditions. ERPs recorded to the person-incongruent as compared to the person-congruent trials (S2) were enhanced at early (100–140 ms) and later processing stages (270–530 ms) in the crossmodal condition. A similar later negative ERP effect (270–530 ms) was found in the unimodal condition as well. These results suggest that identity information conveyed by a face is capable to modulate the sensory processing of voice stimuli.

1. Introduction

Both human faces and voices provide information about the identity, gender, age and emotional state of other people. The recognition of a person's identity is based on the analysis of vocal and facial *invariants* in the acoustic and in the visual input (Belin et al., 2004; Haxby et al., 2000): characteristic features of a voice such as the mean fundamental frequency (pitch) and spectral formant frequencies (timbre) (Relander and Rämä, 2009; von Kriegstein et al., 2010) and invariant facial features such as the configurational arrangement of facial features (Haxby et al., 2000; Maurer et al., 2007) provide unique information about a person's identity.

Partially distinct and partially overlapping brain systems have been suggested to be involved in the processing of faces and voices (Belin et al., 2000, 2002; Belin and Zatorre, 2003; Haxby et al., 2000; Hoffman and Haxby, 2000; von Kriegstein et al., 2003). Representations of unvarying facial features and thus person identity are assumed to exist in the right fusiform face area (FFA) (Haxby et al., 2000). By contrast, the perception

[*] *Corresponding author*. University of Rochester, Brain and Cognitive Sciences, 404 Meliora Hall, Rochester, NY 14627, USA. Fax: +1 585 442 9216.

E-mail addresses: jfoecker@bcs.rochester.edu (J. Föcker), cordula.hoelig@uni-hamburg.de (C. Hölig), anna.best@uni-hamburg.de (A. Best), brigitte.roeder@uni-hamburg.de (B. Röder).

URL: http://www.bcs.rochester.edu (J. Föcker).

[1] *Present address:* University of Rochester, Brain and Cognitive Sciences, 404 Meliora Hall, Rochester NY 14627, USA.

0006-8993/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2011.02.021

of dynamic facial features such as eye gaze and lip movements have been associated with the superior temporal sulcus (STS) (Hoffman and Haxby, 2000; Puce et al., 1998). This model has been supported in other species as well as by brain lesion data in humans (Campbell et al., 1990; Heywood and Cowey, 1992; Tranel et al., 1988; Young et al., 1996).

Similar to the perception of dynamic facial features, voices have also been found to elicit activation in the STS (Belin et al., 2002, 2000; Belin and Zatorre, 2003; von Kriegstein et al., 2003): recent fMRI studies have suggested a higher activation in the STS for vocal sounds compared to environmental sounds (Belin et al., 2000) and a modulation of the STS activation by a speaker's identity but not by the recognition of the verbal content of an oral message (Belin and Zatorre, 2003; von Kriegstein et al., 2003; von Kriegstein and Giraud, 2004). In particular, the right anterior STS (Belin and Zatorre, 2003; von Kriegstein et al., 2003) and the right precuneus (von Kriegstein et al., 2003) have been found to be relevant in person identification based on human voices.

ERP studies have provided insights into the time course of identity recognition of human faces and voices (Beauchemin et al., 2006; Campanella et al., 2000; Münte et al., 1998; Schweinberger, 2001; Spreckelmeyer et al., 2009; see Campanella and Belin, 2007 for a review). For example, the N170 has been proposed to indicate the structural encoding of faces (Eimer, 2000). Moreover, its amplitude can be modulated by facial identity (Campanella et al., 2000). Face identity effects have also been observed between 200 and 300 ms after stimulus onset (N250r, *r* = repetition, Schweinberger et al., 1995, 2002) or even later (350 ms; Münte et al., 1998).

Similar as the N170 has been proposed to indicate face specific processes, voice specific responses were observed with latencies ranging between 150 and 250 ms (Beauchemin et al., 2006; Charest et al., 2009; Schweinberger, 2001; Spreckelmeyer et al., 2009; Zäske et al., 2009) and between 250 and 500 ms (Beauchemin et al., 2006; Levy et al., 2001; Spreckelmeyer et al., 2009) depending on the paradigm and task. Amplitude modulations of ERPs to human voices compared to other environmental sounds were observed at 164 ms with a fronto-temporal and occipital distribution (Charest et al., 2009) whereas sung voices elicited a positive potential at a latency of 320 ms with an anterior distribution (Levy et al., 2001).

However, only a few studies have investigated the time course of person identity recognition in human voices (Beauchemin et al., 2006; Schweinberger, 2001). ERPs to familiar and unfamiliar voices have been reported to differ starting at about 200 ms (Beauchemin et al., 2006). Moreover, ERP voice priming effects were observed for the P200 for both familiar and unfamiliar human voices (Schweinberger, 2001). Since these early repetition priming effects were elicited with backwards played voice probes, Schweinberger (2001) concluded that this repetition priming effect mainly indicates a perceptual priming based on the frequency spectrum because phonetic and articulatory information that depends on temporal information is eliminated in backward speech. Similar to identity matching effects observed for facial stimuli (Münte et al., 1998) vocal identity matching effects were observed in the time range between 300 and 1000 ms (Toivonen and Rämä, 2009: 350–700 ms, Spreckelmeyer et al. 2009: 300–1000 ms).

A separate processing hierarchy for vocal and facial features has been proposed (Ellis et al., 1997): in this model, voices and faces are first separately encoded on a basic level and subsequently examined for familiarity in voice and face recognition units. Voice and face recognition units are assumed to project to person identity nodes (supramodal nodes) which provide biographical information of a person such as profession or name and can be assessed from either facial or vocal information or both (Ellis et al., 1997; Shah et al., 2001). Prosopagnosic patients, who were impaired in recognizing faces but who had spared voice recognition substantiated the assumption of independent processing hierarchies for faces and voices. Moreover, Neuner and Schweinberger (2000) described the so-called phonagnostic patients who are characterised by the reversed impairment pattern.

The proposed initially independent processing streams for faces and voices bear some similarity with initial views of multisensory interaction which assumed that integration of different sensory inputs takes place in multisensory convergence zones (such as posterior STS, middle temporal gyrus (MTG), perirhinal cortex, intraparietal sulcus, posterior cingulated cortex) which are known to receive input from sensory-specific processing streams (Beauchamp et al., 2004; Benevento et al., 1977; Calvert et al., 2001; Shah et al., 2001; see Kayser and Logothetis, 2007 for a review).

However, more recent fMRI studies have provided evidence for the idea that predominantly auditory and visual sensory areas may be directly involved in multisensory interactions and that this holds for person recognition as well (for a review, see Kayser and Logothetis, 2007): for example, there are results demonstrating that auditory cortex activity is modulated by visual input, e.g., auditory cortex activity can be elicited by seeing lip movements of a speaker (Besle et al., 2008; Calvert et al., 1997; Pekkola et al., 2005; Sams et al., 1991; see Besle et al., 2009 for a review). Similarly, familiar human voices have been observed to activate the "visual" fusiform face area (FFA) (von Kriegstein et al., 2005).

Currently, little is known about the time course of audiovisual interactions during person recognition (see Campanella and Belin, 2007 for a review). Recent animal and human studies have revealed evidence for early face–voice interactions (Ghazanfar et al., 2005, 2008; Joassin et al. 2004). In rhesus monkeys, local field potentials were recorded with a mean peak latency of 84 ms to both voices and faces in the auditory core region of rhesus monkeys (Ghazanfar et al. 2005). In an EEG study in humans (Joassin et al., 2004), participants were asked to indicate whether an initially presented name (S1) matched with a subsequently presented face, voice or with a bimodal voice–face stimulus (S2). Multisensory interactions were analysed by subtracting the sum of the ERPs to unimodal stimuli from the ERP elicited by bimodal face–voice stimuli. The authors reported the first multisensory interactions to occur between 90 and 130 ms at central electrodes. Dipol analyses indicated sources in the fusiform gyrus which were taken as evidence for auditory influences in brain regions associated with face processing (see also von Kriegstein et al., 2005).

A different approach was used in the present study. A S1–S2 paradigm was employed. There were two different conditions: either both S1 and S2 were vocal stimuli (unimodal