Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus

Rebecca Watson [a,b,*], *Marianne Latinus* [b,c], *Ian Charest* [b,d], *Frances Crabbe* [b] and *Pascal Belin* [b,c,e]

[a] *Maastricht Brain Imaging Center, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands*
[b] *Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK*
[c] *Institut des Neurosciences de La Timone, UMR 7289, CNRS & Université Aix-Marseille, Marseille, France*
[d] *Medical Research Council-Cognition and Brain Sciences Unit (MRC-CBU), Cambridge, UK*
[e] *International Laboratories for Brain, Music and Sound (BRAMS), Université de Montréal & McGill University, Montreal, Canada*

ARTICLE INFO

Article history:
Received 12 December 2012
Reviewed 28 March 2013
Revised 27 June 2013
Accepted 25 July 2013
Action editor Jason Barton
Published online 2 August 2013

Keywords:
Audiovisual integration
Superior temporal sulcus
Face sensitivity
Voice sensitivity

ABSTRACT

The functional role of the superior temporal sulcus (STS) has been implicated in a number of studies, including those investigating face perception, voice perception, and face–voice integration. However, the nature of the STS preference for these 'social stimuli' remains unclear, as does the location within the STS for specific types of information processing. The aim of this study was to directly examine properties of the STS in terms of selective response to social stimuli. We used functional magnetic resonance imaging (fMRI) to scan participants whilst they were presented with auditory, visual, or audiovisual stimuli of people or objects, with the intention of localising areas preferring both faces *and* voices (i.e., 'people-selective' regions) and audiovisual regions designed to specifically integrate person-related information. Results highlighted a 'people-selective, heteromodal' region in the trunk of the right STS which was activated by both faces and voices, and a restricted portion of the right posterior STS (pSTS) with an integrative preference for information from people, as compared to objects. These results point towards the dedicated role of the STS as a 'social-information processing' centre.

1. Introduction

In the last decade, the human superior temporal sulcus (STS) and surrounding areas have been widely studied (see Hein & Knight, 2008 for a review). The STS is a major sulcal landmark in the temporal lobe, lying between cortices on the surface of the superior temporal gyrus (STG) and middle temporal gyrus (MTG). An extensive region, it can be divided into three distinct sections: the anterior, mid, and posterior STS (aSTS, mid-STS, pSTS). Furthermore, in most individuals, the pSTS divides into two spatially separable terminal ascending branches – the so-called anterior and posterior terminal ascending branches. Thus, the STS can also be anatomically separated into the branch, bifurcation (equivalent to pSTS) and trunk parts (equivalent to mid-STS, aSTS) (Ochiai et al., 2004). There is now a large body of evidence which suggests

* *Corresponding author*. Oxfordlaan 55, Maastricht 6229 EV, The Netherlands.
E-mail addresses: r.watson@psy.gla.ac.uk, rebecca.watson@maastrichtuniversity.nl (R. Watson).
0010-9452/$ – see front matter
http://dx.doi.org/10.1016/j.cortex.2013.07.011

the STS is a major player in social perception – particularly, the pSTS region. This evidence has been provided from two separate camps of research; the first which has investigated unimodal face and voice processing, and the second which has pointed to the role of the pSTS in multisensory integration of social signals (Allison, Puce, & McCarthy, 2000).

We rely greatly on information gathered from both facial and vocal information when engaging in social interaction. Along with the inferior occipital gyri (IOGs) and lateral fusiform gyrus (FG) [specifically, the fusiform face area (FFA) (Kanwisher, McDermott, & Chun, 1997)] the pSTS has been highlighted as a key component of the human neural system for face perception (Haxby, Hoffman, & Gobbini, 2000). It appears to be particularly involved in processing the more dynamic aspects of faces: when attending to these aspects the magnitude of the response to faces in the FFA is reduced and the response in the pSTS increases (Hoffman & Haxby, 2000). Although perhaps not as strong as for faces, evidence for voice-selective regions, particularly in the STS, is accumulating. Several fMRI studies (e.g., Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Ethofer, Van De Ville, Scherer, & Vuilleumier, 2009; Grandjean et al., 2005; Linden et al., 2011) have demonstrated the existence of voice-selective neuronal populations: these voice-selective regions of cortex [‘temporal voice areas’ (TVAs)] are organized in several clusters distributed antero-posteriorly along the STG and STS bilaterally, generally with a right-hemispheric preponderance (Belin et al., 2000; Kreifelts, Ethofer, Shiozawa, Grodd, & Wildgruber, 2009). The aSTS and pSTS in particular appear to play an important role in the paralinguistic processing of voices, such as voice identity (Andics et al., 2010; Belin & Zatorre, 2003; Latinus, Crabbe, & Belin, 2011). Thus parts of the pSTS appear to show greater response to social signals compared to non-social control stimuli in both the visual and auditory modalities, although the relative location of face- and voice-sensitive regions in pSTS remains unclear.

Turning away from unimodal face and voice processing, another vital skill for effective social communication is the ability to combine information we receive from multiple sensory modalities into one percept. Converging results point to the role of the pSTS in multisensory integration, particularly in audiovisual processing. The logic of fMRI experiments on audiovisual integration has been to search for brain regions which are significantly involved in the processing of unimodal visual and auditory stimuli, but show an even stronger activation if these inputs are presented together—the so-called ‘supra-additive response’, where the response to the bimodal stimuli is larger than the sum of the unimodal responses. Integration of speech (Calvert, Campbell, & Brammer, 2000; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003), affective (Ethofer et al., 2006; Kreifelts et al., 2009; Pourtois, de Gelder, Bol, & Crommelinck, 2005), and identity (Blank, Anwander, & von Kriegstein, 2011) information from faces and voices have all been found in the pSTS. However, it should also be noted that integration of ‘non-social’ information – such as tools and their corresponding sounds (Beauchamp, Lee, Argall, & Martin, 2004) and letters and speech sounds (van Atteveldt, Formisano, Goebel, & Blomert, 2004) – has also been observed in the pSTS, and thus it is unclear whether this region performs a more ‘general’ integrative role, or shows preferences for particular stimulus categories.

Here we brought together these distinct lines of research by examining properties of the STS in terms of selective response to social stimuli. Normal adult volunteers participated in an ‘audiovisual localiser’ scan during which they were stimulated with auditory, visual, or audiovisual stimuli of people or objects. We proposed, given that face-selective, voice-selective and integrative regions are found within the STS, that in addition to areas preferring both faces *and* voices (i.e., ‘people-selective’ regions) there could also be audiovisual regions that are more sensitive to social stimuli, as compared to information from non-social categories, such as objects.

We found that a restricted portion of the right pSTS was characterised by a conjunction of (1) an ‘integrative’ response, i.e., stronger response to audiovisual stimuli compared to visual and compared to auditory stimuli and (2) ‘people-selectivity’, i.e., preference for social stimuli irrespective of the modality (voice > objects; face > objects). Furthermore, a large region further extending down the trunk of the right STS was observed to be heteromodal: that is, this region was activated by both faces and voices, but did not necessarily show integrative properties.

2. Materials and methods

2.1. Participants

Forty English-speaking participants (15 males and 25 females; mean age: 25 years ± 5 years) took part in the scan. All had self-reported normal or corrected vision and hearing. The ethical committee from the University of Glasgow approved the study. All volunteers provided informed written consent before, and received payment for, participation.

2.2. Stimuli

24 people (12 males and 12 females) were video-recorded producing a variety of vocal expressions, both speech and non-speech (e.g., saying the word ‘had’, humming, yawning). Recordings took place in the television studio at the Learning and Teaching Centre, Glasgow University, and participants were paid at the rate of £6 per hour. The participants were filmed under standard studio lighting conditions (standard tungsten light), and sat directly facing the camera, at a distance so that the whole face was in frame. Videos were recorded with 25 frames per second (40 msec per frame) using a Panasonic DVC Pro AJD 610 camera, fitted with a Fujiform A17 × 7.8 BERM-M28 lens, and transferred and edited using Adobe Premier Elements. Within the video recording, vocalisations were recorded with 16-bit resolution at a sampling frequency of 44,100 Hz. Under the same conditions, 24 moving objects producing sound were also filmed (e.g., a moving toy car, a ball bouncing, a violin being played). The objects were filmed with the intention of recording the canonical view. Videos were edited so that every production of a vocal sound by a participant formed a separate clip, with the clips lasting