NeuroImage 183 (2018) 25–36

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

# Free viewing of talking faces reveals mouth and eye preferring regions of the human superior temporal sulcus

Johannes Rennig, Michael S. Beauchamp *

*Department of Neurosurgery and Core for Advanced MRI, Baylor College of Medicine, Houston, TX, USA*

ARTICLE INFO

*Keywords:*
Audiovisual
Face
Multisensory
Speech
Eye tracking
fMRI

ABSTRACT

During face-to-face communication, the mouth of the talker is informative about speech content, while the eyes of the talker convey other information, such as gaze location. Viewers most often fixate either the mouth or the eyes of the talker's face, presumably allowing them to sample these different sources of information. To study the neural correlates of this process, healthy humans freely viewed talking faces while brain activity was measured with BOLD fMRI and eye movements were recorded with a video-based eye tracker. *Post hoc* trial sorting was used to divide the data into trials in which participants fixated the mouth of the talker and trials in which they fixated the eyes. Although the audiovisual stimulus was identical, the two trials types evoked differing responses in subregions of the posterior superior temporal sulcus (pSTS). The anterior pSTS preferred trials in which participants fixated the mouth of the talker while the posterior pSTS preferred fixations on the eye of the talker. A second fMRI experiment demonstrated that anterior pSTS responded more strongly to auditory and audiovisual speech than posterior pSTS eye-preferring regions. These results provide evidence for functional specialization within the pSTS under more realistic viewing and stimulus conditions than in previous neuroimaging studies.

## 1. Introduction

Conversing with another human face-to-face exposes us to an abundance of sensory input. In the auditory modality, the voice of the talker conveys speech content. In the visual modality, the movements of the talker's mouth also convey speech content (since different mouth movements produce different speech sounds), while the talker's eyes carry other types of information, such as the spatial location of the talker's gaze. A growing body of evidence suggests that the human posterior superior temporal sulcus (pSTS) contains distinct regions specialized for processing these multiple information sources.

Wernicke first observed that damage to pSTS and nearby regions of lateral temporal cortex impairs speech perception. Functional neuroimaging has increased our understanding of the neural computations performed by this important piece of cortex. In the auditory domain, pSTS contains regions that are highly selective for the human voice (Belin et al., 2000) as well as particular speech sounds (Chang et al., 2010). Functional subdivisions of the pSTS also exist in the visual domain. Pelphrey and colleagues (2005) presented silent videos of a computer-generated face, either opening and closing its mouth or moving its eyes. BOLD fMRI activations within the pSTS to mouth movements were located more anteriorly while activations to eye movements were located more posteriorly. This direct comparison of mouth and eye movements was consistent with studies finding more posterior pSTS activity for eye-gaze observations compared to scrambled images (Hoffman and Haxby, 2000); that anterior pSTS activity is observed for mouth movements contrasted against still mouths (Calvert and Campbell, 2003); and that visual mouth movements related to speech production activate regions of the STS that are more anterior than those activated by non-speech mouth movements, such as yawns (Bernstein et al., 2011). Zhu and Beauchamp (2017) replicated the findings of Pelphrey and colleagues using silent videos of real human faces making mouth or eye movements. Mouth-preferring regions compared to eye preferring regions were located more anterior in the pSTS and responded strongly to unisensory auditory stimuli, especially speech.

A better understanding of the relationship between auditory speech and visual face processing in the STS requires presenting both unisensory and multisensory stimuli. However, most previous studies presented only unisensory auditory or visual stimuli (Belin et al., 2000; Bernstein et al., 2011; Chang et al., 2010; Pelphrey et al., 2005b; Zhu and Beauchamp, 2017). The unisensory visual stimuli in these studies consisted of silent videos of faces making mouth or eye movements in isolation, while in

---

* Corresponding author. Baylor College of Medicine, Department of Neurosurgery, Core of Advanced Magnetic Resonance Imaging, 1 Baylor Plaza, Houston, TX, 77030, USA.

*E-mail address:* Michael.Beauchamp@bcm.edu (M.S. Beauchamp).

https://doi.org/10.1016/j.neuroimage.2018.08.008
Received 23 January 2018; Received in revised form 31 July 2018; Accepted 5 August 2018

real world conditions, humans are confronted with audiovisual talking faces making both eye and mouth movements.

Under natural viewing conditions, humans fixate either the mouth or the eyes of the talker for time intervals that can extend to a second or more (Gurler et al., 2015). A recent fMRI study capitalized on the existence of these extended fixations to search for the neural correlates of fixating the eyes of a dynamic talking face (Jiang et al., 2016). The authors describe the existence of an "eye contact network" that includes portions of the pSTS, the temporo-parietal junction and other brain areas.

In the present study, we searched for brain areas that were more active when participants fixated the mouth of the talker. Our hypothesis was that this "mouth contact network" should include regions important for speech perception, especially areas responsive to visual mouth movements and auditory speech in anterior portions of the pSTS, and that these regions should demonstrate multisensory integration. To test this hypothesis, we performed two independent fMRI experiments. In the first, participants freely viewed dynamic talking faces while their eye movements were monitored in order to identify mouth and eye-selective regions. In the second, participants viewed blocks of auditory, visual, and audiovisual speech in order to determine functional specialization and multisensory integration of mouth and eye-selective regions.

## 2. Methods

34 healthy right-handed participants (16 females, mean age 26.5, range 18–45) with normal or corrected to normal vision and normal hearing provided written informed consent under an experimental protocol approved by the Committee for the Protection of Human Subjects of the Baylor College of Medicine, Houston, TX. 29 of 34 participants were native English speakers (2 German speakers, 3 Mandarin speakers).

Each participant was scanned using a 3T Siemens Trio MRI scanner equipped with a 32-channel head coil at Baylor College of Medicine's Core for Advanced MRI (CAMRI). During a single scanning session, participants performed two different fMRI experiments. The two experiments were analyzed independently to eliminate bias. Stimuli were presented using Matlab (The Mathworks, Inc., Natick, MA, USA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and viewed on an MR compatible screen (BOLDscreen32, Cambridge Research Systems, Rochester, UK) placed behind the bore of the scanner. Auditory speech was presented using high-fidelity MR compatible headphones (Sensimetrics, Malden, MA, USA). Behavioral responses were collected using a fiber-optic button response pad (Current Designs, Haverford, PA, USA) and eye movements were recorded during scanning using the Eye Link 1000 (SR Research Ltd., Ottawa, Ontario, Canada) in MR-compatible mode with a sampling rate of 500 Hz.

In the first fMRI experiment, eye-tracking was performed in the MR scanner while participants viewed audiovisual movies presented in an event-related design. Each 2-s movie consisted of a talker saying a single syllable. Each participant viewed 240 movies, equally distributed across six different types: three syllables (AbaVba, AgaVga, AbaVga) × two talkers (one male and one female). Following each movie, participants identified the presented syllable with a button press.

In the second fMRI experiment, participants viewed long blocks (20 s) of auditory, visual or auditory-visual speech, with a single female talker reading Aesop's fables (Nath and Beauchamp, 2012). The eye image from the eye tracker was monitored to ensure the participant's alertness but no eye tracking was performed and there was no task.

### 2.1. Eye tracking data analysis

Fig. 1A shows frames from a stimulus movie. To simulate natural viewing conditions, each face movie was preceded with a gray screen containing a fixation crosshairs presented in a random location distant from the spatial position where the face would later appear (Gurler et al.,

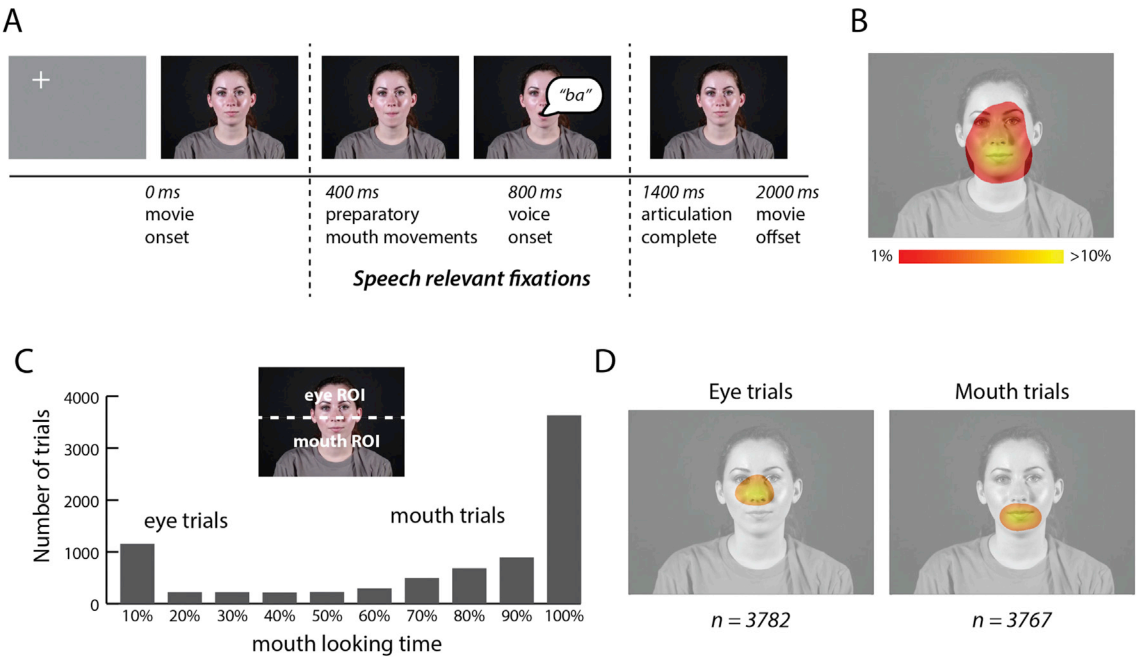

**Fig. 1.** Stimulus and eye movement analysis. **A.** Within each trial, participants viewed a 2-s duration audiovisual movie of a talker speaking a single syllable (still frames from single movie shown for illustration). Preparatory mouth movements began ~400 ms after stimulus onset, voice onset occurred at ~800 ms and articulation was complete by 1400 ms. Only the speech relevant fixations between 400 ms and 1400 ms were included in the analysis. **B.** Fixations for 34 participants viewing the audiovisual movies. Color scale indicates percent fixation time for each image location. **C.** Each movie was divided into an upper region, corresponding to the eye region of the face, and a lower region, corresponding to the mouth region of the face (dashed white line, not present in actual stimulus). For each trial, the percent of time fixating the eye and mouth regions of the face was calculated. The histogram shows the number of trials in each bin, with bins sorted by increasing amounts of time fixating the mouth. Within each participant, each trial was classified as an eye or a mouth trial, based on that participant's median fixation time. **D.** Average fixation locations across 34 participants for all eye and all mouth trials (*n* shows number of trials used for the fMRI analysis).