NeuroImage 124 (2016) 876–886

Contents lists available at ScienceDirect

NeuroImage

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

How prior expectations shape multisensory perception

Remi Gau *, Uta Noppeney

Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK

A R T I C L E I N F O

Article history:
Received 7 August 2015
Accepted 20 September 2015
Available online 28 September 2015

Keywords:
fMRI
Multisensory perception
McGurk
Prefrontal cortex
Causal inference
Crossmodal integration
Congruency
Learning
Perceptual illusion

A B S T R A C T

The brain generates a representation of our environment by integrating signals from a common source, but segregating signals from different sources. This fMRI study investigated how the brain arbitrates between perceptual integration and segregation based on top-down congruency expectations and bottom-up stimulus-bound congruency cues.

Participants were presented audiovisual movies of phonologically congruent, incongruent or McGurk syllables that can be integrated into an illusory percept (e.g. "ti" percept for visual «ki» with auditory /pi/). They reported the syllable they perceived. Critically, we manipulated participants' top-down congruency expectations by presenting McGurk stimuli embedded in blocks of congruent or incongruent syllables.

Behaviorally, participants were more likely to fuse audiovisual signals into an illusory McGurk percept in congruent than incongruent contexts. At the neural level, the left inferior frontal sulcus (lIFS) showed increased activations for bottom-up incongruent relative to congruent inputs. Moreover, lIFS activations were increased for physically identical McGurk stimuli, when participants segregated the audiovisual signals and reported their auditory percept. Critically, this activation increase for perceptual segregation was amplified when participants expected audiovisually incongruent signals based on prior sensory experience.

Collectively, our results demonstrate that the lIFS combines top-down prior (in)congruency expectations with bottom-up (in)congruency cues to arbitrate between multisensory integration and segregation.

Introduction

In everyday life, our senses are bombarded by a plethora of visual and auditory information. Our brain seemingly effortlessly makes sense of this cacophony and merges the signals into a coherent percept of our environment (Ernst and Buelthoff, 2004). For instance, in speech processing, a video clip of mouth movements articulating «ga» can alter the percept of a concurrent auditory speech signal /ba/ into an integrated 'illusory' "da" percept — a phenomenon called the McGurk–MacDonald illusion (McGurk and MacDonald, 1976). Critically, the brain should integrate signals originating from a common source, but segregate those emanating from different sources. Human observers have been shown to arbitrate between information integration and segregation by combining bottom-up sensory evidence with prior common source expectations (i.e. the so-called common source prior) optimally in line with predictions from Bayesian Causal Inference (Körding et al., 2007; Magnotti et al., 2013; Rohe and Noppeney, 2015a, 2015b; Shams and Beierholm, 2010; van Wassenhove, 2013).

First, sensory signals are likely to come from a common source when they co-occur in time and space and are structurally similar with respect to phonology, semantics or other higher-order statistical relationships (e.g. gender of voice) (Körding et al., 2007; Lee and Noppeney, 2011; Magnotti et al., 2013; Munhall et al., 1996; Wassenhove et al., 2007). We refer to these spatiotemporal and structural correspondences as bottom-up sensory congruency cues, because they are stimulus-bound and need to be extracted from the sensory inputs. Previous research has demonstrated that human observers indeed use bottom-up congruency cues such as temporal coincidence and spatial colocalization to arbitrate between information integration and segregation. Thus, perceptual illusions such as spatial ventriloquism or the McGurk–MacDonald illusion break down when sensory signals are brought into large spatial or temporal conflict (Körding et al., 2007; Magnotti et al., 2013). At the neural level, the left inferior frontal sulcus (lIFS) may play a key role in controlling information integration and segregation based on bottom-up incongruency cues, as it shows increased activations for incongruent relative to congruent audiovisual inputs (Adam and Noppeney, 2010; Noppeney et al., 2008; Ojanen et al., 2005).

Second, top-down prior 'congruency' expectations, which are equivalent to the so-called common source prior in Bayesian Causal Inference (Körding et al., 2007; Magnotti et al., 2013), determine whether signals are integrated or segregated. For instance, in a conversational setting with a single speaker, we should be more inclined to integrate his/her facial movements with the syllables s/he is uttering for improved speech comprehension. By contrast, in a busy pub where we are bombarded with many conflicting auditory and visual speech signals, unconstrained information integration would be detrimental.

* Corresponding author at: Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK.
E-mail address: remi.gau@tuebingen.mpg.de (R. Gau).

http://dx.doi.org/10.1016/j.neuroimage.2015.09.045
1053-8119/

Indeed, recent behavioral evidence suggests that participants are influenced by contextual congruency expectations (Nahorna et al., 2015, 2012). Participants were more likely to integrate audiovisual signals into an illusory McGurk–MacDonald percept after exposure to audiovisually congruent than incongruent speech signals. While several previous fMRI studies have compared neural activations when participants integrate or segregate sensory signals into a coherent percept, the neural mechanisms by which prior top-down congruency expectations (i.e. common source prior) control these processes remain unexplored. Classical studies of cognitive control have shown increased activations in the lIFS after a series of incongruent trials in a Stroop paradigm, which reduced Stroop interference on subsequent trials (Kerns et al., 2004). This response profile suggests that the lIFS may not only be sensitive to bottom-up incongruency, but also reflect top-down congruency expectations that are constantly up dated based on past and current congruent or incongruent stimuli. Thus, one may ask whether lIFS plays a key role in controlling information integration and segregation in multisensory perception (Shenhav et al., 2013).

In the current psychophysics-fMRI study, we presented participants with McGurk–MacDonald syllables (e.g. "ti" percept for visual «ki» with auditory /pi/) that were interspersed within a sequence of phonologically congruent (e.g. visual «ki» with auditory /ki/) or incongruent (e.g. visual «ki» with auditory /pa/) syllables. On each trial, participants reported their percept. Based on models of Bayesian Causal Inference (Körding et al., 2007; Magnotti et al., 2013; Shams and Beierholm, 2010), we hypothesized that participants would form stronger congruency expectations (i.e. a higher common source prior) in congruent than incongruent contexts. As a consequence of this higher congruency expectation (i.e. common source prior), they should be more likely to fuse signals into an illusory McGurk–MacDonald percept in congruent than incongruent contexts. At the neural level, we first identified brain areas sensitive to bottom-up stimulus-bound (in)congruency cues by directly comparing incongruent and congruent audiovisual Non-McGurk–MacDonald speech signals. Second, we investigated how the brain controls whether audiovisual McGurk–MacDonald signals are integrated into an illusory percept and how these integration processes are modulated by top-down (in)congruency expectations (i.e. the interaction between fused/segregated McGurk–MacDonald percept and (in)congruency context). We predicted that the lIFS as our region of interest (ROI) plays a critical role in controlling information integration and segregation in multisensory perception based on top-down prior (in)congruency expectations.

Material & methods

The study was approved by the joint human research review committee of the Max Planck Society and the University of Tübingen.

Participants

Sixteen healthy right-handed German native speakers (6 females; 10 males; age: mean: 30.1, standard deviation: 7.1, range: 22–45 years) gave written informed consent to participate in this fMRI study. Participants reported no history of psychiatric or neurological disorders, and no current use of any psychoactive medications. All had normal or corrected to normal vision and reported normal hearing.

Experimental design

In the main experiment, participants were presented with audiovisually congruent, incongruent or McGurk–MacDonald stimuli. Critically, the congruent and incongruent stimuli were presented in blocks. Further, the McGurk–MacDonald trials were interspersed either in the incongruent or congruent blocks. This enabled us to characterize the neural processes of perceptual integration vs. segregation and their contextual modulation by analyzing the BOLD-responses to McGurk–MacDonald stimuli in a 2×2 factorial design manipulating (i) top-down prior (in)congruency expectations (i.e. McGurk–MacDonald stimuli in incongruent vs. congruent blocks) and (ii) participants' percept: fused McGurk–MacDonald percept vs. segregated auditory percept (Fig. 1C). Furthermore, we identified 'bottom-up incongruency' by comparing audiovisually incongruent and congruent trials. We acknowledge that formally speaking the bottom-up incongruency effects cannot be fully dissociated from contextual effects of (in)congruency, as incongruent trials were only presented in incongruent blocks (and vice versa). Yet, given the extensive body of research demonstrating audiovisual incongruency effects in lIFS under randomized presentation (Adam and Noppeney, 2010; Noppeney et al., 2010, 2008; Ojanen et al., 2005), we decided to optimize the design efficiency for our main contrasts of interest (i.e. the contrasts involving the McGurk–MacDonald trials). Including incongruent trials in congruent blocks and vice versa would have attenuated the contextual modulation on the McGurk–MacDonald trials.

Stimuli

Stimulus material was taken from close-up audiovisual recordings of a female actress' face on a dark background looking straight into the camera and uttering the following syllables: ba, be, bi, da, ga, ge, gi, gu, ka, ke, ki, ko, ku, pa, pe, pi, pu, tu. Audio and video were recorded with a camcorder (HVX 200 P; Panasonic). The video was acquired at 25 frames per second phase alternation line (PAL $= 768 \times 567$ pixels); audio was acquired at 48 kHz (two channels).

The recorded videos were edited (using PiTiVi 0.15.2) into 1520 ms long segments (38 frames). Each video started and finished with a neutral closed lip view of the speaker's face. To facilitate the cross-dubbing between the different movies we ensured that the acoustic burst onset of each syllable started 680 ms after the beginning of the movie. The first articulatory movement started on average 334 ms after the beginning of the movie (standard deviation: 172 ms).

We used the movies of 8 different syllables as congruent stimuli. Eight incongruent stimuli were generated by cross-dubbing the video and the audio-track from the different stimuli (e.g. the auditory component of the ba syllable was dubbed on the visual component of the gu syllable and vice versa). Critically, the re-combination of the congruent into incongruent stimuli was performed in a pairwise fashion, so that the auditory and visual components of the set of congruent and incongruent stimuli were identical (Fig. 1A and Table 1). In other words, congruent and incongruent conditions were matched in terms of the individual auditory and visual inputs and differed only in how auditory and visual signals were combined within an audiovisual stimulus. We also ensured that none of the incongruent stimuli elicited a McGurk–MacDonald percept (e.g. see Green and Gerdeman, 1995 that had previously reported illusory McGurk–MacDonald percepts despite vocalic change).

Next, we created five McGurk–MacDonald stimuli by dubbing the audio-tracks from the movies of the be, bi, pa, pe, and pi syllables onto the video-tracks of the ge, gi, ka, ke, and ki stimuli respectively (Fig. 1A and Table 1). The auditory and visual components of the McGurk–MacDonald stimuli as well as the resulting illusory McGurk–MacDonald percepts were distinct from the set of congruent and incongruent stimuli to avoid perceptual and response priming confounds within a congruent or incongruent block during the fMRI study. In this way we ensured that changes in the percentage of illusory McGurk–MacDonald percept could not be attributed to perceptual or response priming induced by the surrounding congruent or incongruent stimuli within a block.

Finally, for all classes of stimuli, audio-track intensities were normalized and the stimuli were digitized into MPEG-4 (H.264) format files.

Adapting stimuli to each participant

Fusion of the auditory and visual signal components into an illusory McGurk–MacDonald percept relies on the relative auditory and visual