The content of auditory feedback to human early visual cortex and its impact on visual perception.

Giusi Pollicina

Thesis submitted in fulfilment of the

degree of Doctor of Philosophy

Department of Psychology

Royal Holloway, University of London

May 2023

Declaration of Authorship

I, Giusi Pollicina, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed: Pollicina Giusi

Date: 14/02/2023

Abstract

The goal of the current research project was to investigate auditory feedback to cortical regions at the early stages of visual processing (i.e., V1, V2 and V3), while also looking at the consequences this flux of information has on visual perception itself. First, I conducted a functional magnetic resonance imaging (fMRI) study on blindfolded participants, which aimed to understand the semantic content of auditory information that is fed back to early visual cortex and its degree of specificity in the absence of visual stimulation. I presented different types of natural sounds, categorised in a hierarchical fashion, and tested whether early visual cortex received sufficient information to differentiate between semantic sound categories. This was done through a classifier algorithm in a multivariate technique called Multi-Voxel Pattern Analysis (MVPA). I found that sounds belonging to superordinate categories such as animate and inanimate, and to more specific categories such as humans, animals, vehicles and objects, could be decoded from neural activity patterns in early visual cortex. Additionally, human sounds seemed to be better decoded than other sound categories. I then conducted a binocular rivalry experiment with the scope of investigating whether sounds with varying degrees of semantic congruency can offer an advantage towards the perception of ambiguous human visual stimuli. These novel findings suggest that semantically congruent sounds can boost perception of congruent visual stimuli, but only when these are fully congruent rather than merely related (e.g., a clapping sound will boost perception of clapping hands, but not the image of a baby). This effect was only observed in the case of human sounds and human images. Taken together, the findings from these studies contribute to the evidence that suggests early visual cortex might belong to the network of multisensory processing, particularly when the stimuli are of human nature.

Table of contents

Table of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Acknowledgements

I would like to express my deepest gratitude to Professor Petra Vetter and Professor Polly Dalton for their invaluable help and support throughout this journey. These past years have been intense and unexpected in so many ways, but your guidance helped me keep my spirits high and overcome all the challenges I encountered.

I am also grateful to the people that helped me during data collection and analysis, I could not have undertaken this journey without the help of MRI technician Ari Lingeswaran, Dr Ayan Sengupta, and Dr Tibor Auer. Special thanks also go to my peers and colleagues, who offered help in recruiting participants, as well as moral support during long days. Additionally, I would like to mention my housemates, my friends and my family for their patience and encouragement, even when they were not entirely sure of what I was talking about.

Lastly, to my partner, Angelo. I had many difficult days and moments of discouragement throughout this journey, thank you for your endless emotional and practical support, but mostly thank you for believing in me, even when I did not. To my cats, Loki and Ivy, thank you for bringing a sprinkle of joy into my most intense days.

Chapter 1: Introduction

1.1: The processing of sensory information

Understanding how the cerebral cortex processes information is a major aim of neurobiology and cognitive neuroscience, with implications for a variety of disciplines (i.e., medicine, computer science and psychology in general). However, investigating information processing in the brain is not a simple task, as researchers need to account for a high number of confounding factors, individual differences, as well as challenges concerning how to analyse the data collected. In the past few decades, findings from neuropsychology, brain imaging, computational modelling and many other branches have been combined to create a plausible picture of the organization of cortical processes.

1.1.1: The feedforward model of sensory processing

According to the archetypical model, sensory information from each modality arrives at its specific primary sensory area within milliseconds of stimulus presentation. For example, data from experiments using electroencephalography (EEG) has shown event-related potentials evoked in the visual cortex after 50-100 milliseconds from stimulus presentation (Mangun, 1995). Information is then passed forward in a hierarchical fashion: the cortical areas that first receive the signal from the sensory organs seem to process basic, low-level features, while the regions with greater synaptic distance from the sensory organ deal with more and more complex features (Felleman and Van Essen, 1991), often combining more than one sensory input, and only after the information processing carried out in sensory-specific areas. Eventually, inputs from different sensory modalities converge (Shams and Beierholm, 2011) and are elaborated to create representations of the world around us, with the help of other cognitive functions such as memory.

Following this model in relation to vision, information from the eyes arrives to the primary visual cortex (V1), the first cortical processing stage, after passing through the lateral geniculate nucleus and other subcortical structures. V1 is located around the calcarine fissure and is concerned with processing different low-level visual features such as orientation, direction in motion (Hubel & Wiesel, 1959 and 1979), spatial frequency (De Valois and De Valois, 1990) and temporal frequency (Movshon et al., 1978a). Here, and in other regions low in the cortical hierarchy of visual processing, the visual input can be precisely mapped onto the cortex with a topographic organisation, meaning that the spatial distribution of visual information is preserved in early visual cortex. After V1, the stream of information bifurcates into a dorsal and a ventral stream (Milner and Goodale, 1995; Kravitz et al., 2013) and it is passed onto other regions of the visual cortex where more complex visual features are processed (Figure 1.1). Concerning the dorsal stream, area V3 stands at the beginning, and together with MT is involved in the perception of motion, with MT playing a major role also in the guidance of eye movements (Born and Bradley, 2005). Visual inputs are further transmitted up to regions of the posterior parietal cortex (i.e., superior parietal lobule and intraparietal sulcus), which is concerned with spatial reasoning, planning movements, and performing actions (Marshall and Fink, 2001; Marshall et al., 2002). As for the ventral stream, V2 seems to respond again to features like orientation and spatial frequency, but also to more complex features such as edges and contours (von der Heydt and Peterhans, 1989; Leventhal et al., 1998), and illusory contours (Anzai, Peng and Van Essen, 2007). The ventral portion of V3 responds to contours too, but it also contains populations of colour-selective neurons (Nasr, Polimeni and Tootell, 2016), while area V4 is specialised in responding to colour (Zeki, 1985) and simple geometric shapes (Kobatake and Tanaka, 1994). The ventral stream extends into the inferior temporal cortex, roughly divided into posterior, central and anterior, which has been extensively reported as being the locus of representation of complex categories such as faces (Kanwisher, McDermott and Chun, 1997; Tsao et al., 2006), bodies (Downing et al., 2001; Peelen and Downing, 2005), objects (Martin et al., 1996) and scenes (Epstein and Kanwisher, 1998; Aguirre, Zarhan and D’Esposito, 1998) (for a review on the organisation of the ventral stream, see Conway et al., 2018).

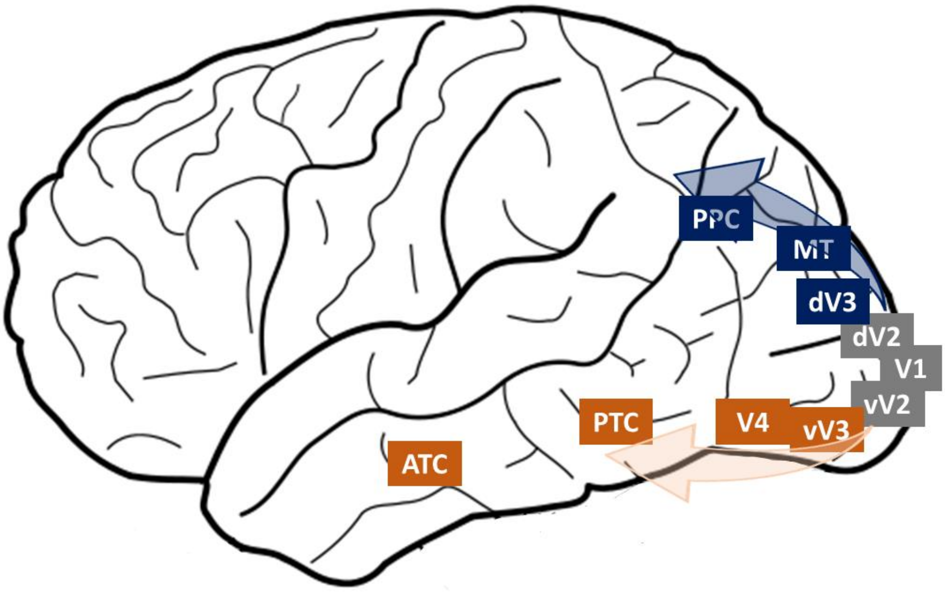


Figure 1.1: Visual representation of the dorsal (blue) and ventral (orange) visual streams (Milner and Goodale, 1995).

As they branch off to the temporal and parietal lobes, the two streams integrate visual information with other perceptual domains (Fiehler and Rosler, 2010, Conway et al., 2018).

This traditional account saw the information processed in the dorsal and ventral streams as segregated, with dorsal areas being concerned with the spatial, action-related feature of objects, and ventral areas with their recognition and categorisation. More recent evidence however, appears to be in support of a less distinct view of visual processing, one that represents information about identity and spatial properties – in some measure – along both streams (Freud, Plaut and Behrmann, 2016). For example, visual object properties related to its position, dimension and orientation, can be decoded from activity patterns in the ventral cortex (Zoccolan et al., 2007; Hong et al., 2016), as well as movement-related modulations (Astafiev et al., 2004; Gallivan et al., 2013). Moreover, humans with lesions to the ventral visual stream still represent 3D object representations (Freud et al., 2015) in parietal areas, suggesting that both streams produce dissociable representations that combine towards a stable visual percept.

The ventral and dorsal pathways, either as separate entities or as part of a whole account of the visual system, have been extensively studied to gain a better sense of how the brain processes and represents different types of visual stimuli, and while there are still open questions on what exactly plays a role in the category specificity we see in object representation, the literature seems to agree that a hierarchy of processing exists within the visual system (Bracci and Op de Beeck, 2023). An fMRI study by Bracci and Op de Beeck (2016) nicely illustrates this by presenting participants with a set of visual stimuli where it was possible to dissociate shape from category. They observed how the content represented in occipitotemporal regions of interest progressed in complexity, going from silhouette representations in early visual cortex regions, to regions representing both shapes and categories in regions higher up in the hierarchy (e.g., face-, body-, and object-selective loci along the posterior inferior temporal cortex), and ultimately to regions with a preference for object category and not shape (e.g., intraparietal sulcus). While objects with similar orientations were represented in a similar way in early visual regions, objects with elongated shapes were represented similarly in higher-level regions, regardless of their orientation.

Interestingly, higher-level regions like these seem to be involved in the categorisation of other sensory modalities too, which is consistent with what was said before about multiple senses being integrated at higher levels of cortical processing. Regions like the middle temporal gyrus and intraparietal cortex for example, have been implicated in the categorisation of animal versus tool sounds (Lewis et al., 2005), and a wider fronto-parietal network has been found to preferentially activate for non-living versus living sounds (Engel et al., 2009). This indicates the presence of categorical frameworks at these higher-level locations, which suggests that they not only play a role in the recognition of visual and auditory stimuli, but it also makes them optimal association areas where multisensory information can converge and combine.

1.1.2: The top-down model of sensory processing: feedback pathways

The idea that sensory processing is a unidirectional flow of feedforward operations of increasing complexity is juxtaposed to a different account, which sees the already-known bottom-up model accompanied by top-down processes (McMains & Kastner, 2011; Gilbert & Li, 2013).

Top-down influence refers to the influence that prior knowledge and higher-order representations have on the early stages of information processing, implemented via a large number of feedback pathways coming from regions traditionally thought to partake in later/other stages of processing, from association areas and from other sensory areas (Felleman and Van Essen, 1991; Singer, 2013; Markov et al., 2014). The feedback received by early sensory areas carries a sufficient amount of information to alter the way neurons respond to a stimulus, effectively influencing information processing carried out in those sensory areas.

In the context of the visual system, this implies that regions of the early visual cortex receive top-down influences from other brain areas that facilitate the recognition of a complex environment. Top-down influences can assume many forms: in terms of spatial attention, their effect can enhance responses to relevant stimuli and suppress those coming from outside the focus of attention (Motter, 1993; Desimone and Duncan, 1995; Ito, Westheimer and Gilbert, 1998). Attention can also highlight feature-based differences, allowing us to process entire scenes by distributing resources towards the recognition of a specific category, such as colour or direction of motion (Motter, 1994; Treue and Martinez Trujillo, 1999). An additional type of top-down control comes in the form of the perceptual task being performed. This can shape the way neurons react to stimuli by discarding influences that are not relevant to the current task and solely focusing on the task-relevant ones, so that even when the stimulation is identical, the neural responses change according to the task (Li, Piech and Gilbert, 2004). Memory can also influence sensory areas in a top-down fashion, an example of which is offered by fMRI studies in which oriented gratings held in working memory could be predicted by activity patterns in early visual cortex, in absence of real visual stimuli (Harrison and Tong, 2009).

Evidence of the presence of these top-down influences has been found at all stages of the visual hierarchy, from the lateral geniculate nucleus (O'Connor et al., 2002) to extrastriate cortices (Chelazzi et al., 1993; Treue and Maunsell, 1996; Reynolds and Desimone, 2003); a full review is presented by Gilbert and Li, (2013).

The visual processing that would result from this model is accomplished by an interaction of both feedforward and feedback processes, bringing together sensory inputs and influences from attention, memory and mental state, making the parsing of a visual scene an easier task. However, among the top-down influences on visual areas, I have not discussed those coming from other sensory domains. In the following sections, I will describe in more detail the literature available on the early visual cortex and its connections to other sensory regions.

1.2: Early visual cortex

Primates, in particular humans, have evolved to rely on vision for a variety of tasks and behaviours, leading to an expansion of the cortical and subcortical structures dedicated to processing visual information (Kaas, 2013). Due to the complexity of visual processing and to the wide range of ways in which it can be impaired while maintaining intact other functions, the visual system has been extensively studied for more than a century. The importance of the occipital lobe (the locus of the primary visual cortex) for the processing of visual information has been known since the early 20th century (Birch-Hirschfeld and Inouye, 1909; Lister and Holmes, 1916), but the whole process employs a complex network of areas including also temporal and parietal cortical areas.

I described in section 1.1.1 the hierarchy of visual processing in the brain, and how V1, V2 and V3 are the early stages of cortical processing, where low-level visual features are tackled. I will refer to the ensemble of these three as early visual cortex (EVC).

Neurons in the EVC are strongly tuned to very specific types of stimuli, to the point where it is possible to specifically locate neurons that respond, for example, to a certain orientation in a certain part of the visual field. These regions pass visual information forward and receive feedback from other visual areas which, later in time, makes the same neurons also responsive to more global representations of a scene (David, Vinje and Gallant, 2004; Sillito, Cudeiro and Jones, 2006). However, there is also a large number of feedback connections coming from other sensory modalities and many other brain areas, which might send information that potentially interacts with vision.

1.2.1: Feedback to early visual cortex

There is a substantial body of evidence coming from different species of primates and mammals of afferents going from visual and non-visual regions to EVC, both direct and passing through association cortices. Some examples coming from animal studies include evidence of projections from V2, V3, V4 and MT to V1 (Lyon and Kaas, 2002), from primary somatosensory sensory cortex and auditory cortex to V1 (Henschke et al., 2015), from frontal eye fields and intraparietal sulcus to human V1 (Griffis et al., 2015). While currently there are no directly comparable studies in humans, inferences for similar structures can be drawn from studies using Diffusion Tensor Imaging, an MRI technique that can provide maps of the axonal organisation of the brain and measures of structural connectivity. These have shown the presence of structural connectivity between low-level auditory regions and the early visual cortex in the form of direct white matter connections (Beer, Plank and Greenlee, 2011), both to foveal and peripheral areas in the calcarine sulcus, and connections from fronto-parietal regions to V1 (Griffis et al., 2015; Morimoto, Uchishiba and Saleem, 2021).

Despite being aware that EVC is at the receiving end of extra-visual cortical processes, Muckli and Petro (2013) discussed how little is known about top-down projections to V1 compared to bottom-up, and how this could be the key to providing a fuller understanding of sensory processing. Feedback connections to the early visual cortex may support some higher-level functions that have so far been neglected by research, perhaps due to early visual cortex's greater responsiveness to low-level features. As suggested by Deco and Lee (2004), due to the retinotopic fashion in which fine-grained information is processed in early visual cortex, V1 naturally provides a spatially registered common space for all the higher order perceptual inferences to come back together. Thus, the early visual cortex might be not only a place for the processing of low-level features, but also for the employment of higher-level features to understand and process visual stimuli. Moreover, the existence of feedback connections from other sensory regions and the fact that other sensory modalities can modulate activity in the early visual cortex, seem to suggest that some forms of multisensory interaction could be happening in early visual areas.

Beyond the anatomy, another way to identify, albeit indirectly, the presence of projections for other sensory regions to EVC is by looking for the information transferred to EVC via those connections. Since it has been shown anatomically and functionally that feedback from auditory cortex arrives to early visual cortex, Vetter, Smith and Muckli (2014) tried to understand what information this feedback might carry and what role it might have. During a series of functional magnetic resonance imaging (fMRI) experiments, they blindfolded participants and presented them with three different sounds: birds chirping, traffic noise, and people talking. They then analysed data using a multivariate method of analysis called Multivoxel pattern analysis (or MVPA, Haxby et al., 2001), in which an algorithm is used to identify whether activity patterns elicited by the same stimulus category can be distinguished from those produced by another category (more details on MVPA can be found in Chapter 2). They found that it was possible to significantly distinguish the activity patterns elicited by bird, people, and car sounds in early visual cortex. In other words, since the algorithm was able to decode which sounds subjects were listening to, solely based on the activity patterns in early visual cortex when participants listened to sounds, Vetter et al. (2014) suggested that the feedback from auditory cortex received by early visual cortex carries information about the meaning of those sounds. They also found out that spatial activity patterns generalised across sounds belonging to the same category (even sounds that were not encountered by the algorithm before), meaning that semantic categories might be what is represented in the information fed back to visual cortex. The information encoded in the activity patterns elicited by sounds in the early visual cortex allowed Vetter and colleagues to support the idea that categorical and semantic information about non-visual sensory domains is fed back all the way down to the early visual cortex, but the purpose of this flux of data is still up for debate. More literature on the multisensory interactions in EVC will be discussed in section 1.3.

1.2.2: Predictive coding theory

In light of the evidence discussed so far, one of the theories that gains support from all the numerous top-down influences across the brain is predictive coding. Predictive coding refers to a neural framework that sees the brain as constantly engaged in predictions about its future states, which are used to generate and update internal models of the world to anticipate and minimise prediction errors (Friston, 2010; Clark, 2013).

The theory has its origins in a concept described by von Helmholtz (1860), who proposed that the basis of how the brain perceives is the unconscious inferences it makes about the world, and it can be considered one of the most biologically plausible in terms of its implementation (Mumford, 1992; Rao and Ballard, 1999).

According to this model, as the brain perceives any kind of sensory stimulus, the predictions it formulates are organised in a hierarchical fashion, with lower levels representing simpler details (e.g., colours, contours) and higher levels dealing with more complex and abstract representations (e.g., object categories, emotions, decisions) coming from multiple lower-level modules. Predictions from higher levels are fed back to early stages, where they are compared with incoming sensory signals. The discrepancy between the predictions created by the model and the actual sensory inputs is referred to as prediction error (Clark, 2013). The prediction errors are then fed forward to higher processing levels, where they serve as inputs that are matched against predictions formed at those stages. These cycles of prediction and error correction repeat constantly along the processing hierarchy, with the goal of minimising prediction errors, so that the sensory system is able to accurately deal with the environment in a dynamic, adaptive way.

When we apply this model to a sensory processing system such as vision, the regions at the early stages of the cortical hierarchy (i.e., V1) send forward the input received from the eyes, which is compared in V2 with the prediction made by V2, and the error calculated is then fed back to V1, which in turn communicates the updated model to other regions along the hierarchy. This is then repeated at all stages until the model is a satisfactory prediction of the real sensory stimulus (Rao and Ballard, 1999).

Thus, a possible explanation for the presence of the numerous, intricate feedback connections between higher and lower-level sensory areas, might be the need the brain has to continuously adjust its models to know how to properly react to, and interact with, the outside world.

1.2.3: The role of mental imagery

Among the top-down processes that occur in EVC, it is crucial to mention visual mental imagery, which occurs when someone imagines visual stimuli but does not actually see them. It is a process that involves a rich network of areas in the frontal, parietal and temporal cortices, but notably, it activates V1 with some degree of variability, depending on the exact nature of the imagery (Winlove et al., 2018). Furthermore, new evidence coming from multivariate analyses suggests the content of mental imagery in EVC is sufficient to be decoded in areas V1 and V2 (Naselaris et al., 2015; Dijkstra, Bosch, & van Gerven, 2017). The feedback EVC receives from auditory cortex seems to also contain information about mental imagery, as shown by the above-mentioned study by Vetter et al. (2014). When they asked participants to imagine the sounds rather than actually listen to them, they found that imagined sounds elicited activity patterns similar to the actual sounds in EVC. The fact that this was not the case for decoding in auditory cortex (i.e., the patterns elicited by imagined sounds were not similar to those elicited by listening to real sounds), supports the idea that information from both real and imagined sound information is fed back to EVC.

1.3: Multisensory interaction and integration

Every second, the brain receives an immense amount of information from all sensory modalities (i.e., we are constantly touching, hearing or seeing something), which needs to be processed in order to make sense of what surrounds us. Research into multisensory interactions has been increasing exponentially in the last couple of decades, allowing us to better understand how layered and complex these processes can really be. In the past, it was understood that multisensory processes took place in certain subcortical structures in the earlier stages, and later in higher-order association cortices. The thalamus, for example, is one of those structures, whose neural activity is linked to audiovisual speech processing (Musacchia et al., 2006) and multisensory attention tasks (Vohn et al., 2007). Or the superior colliculus, which is involved in saccadic eye movements but has been found to respond also to somatosensory and auditory stimuli (Meredith, Wallace and Stein 1992).

Regarding the higher-order association cortices, the superior temporal sulcus (STS) is an example, reportedly involved in audio-visual integration (see Hein and Knight, 2008). The inferior frontal gyrus has also been extensively reported as a region sensitive to audiovisual decision-making and speech perception (Miller and D'Esposito, 2005; Noppeney et al., 2010; Tse et al., 2015).

However, subcortical structures and association cortices are not the only components of multisensory processing. As mentioned in previous sections, the idea that primary sensory regions such as V1 are strictly uni-sensory at the early stages and that multisensory interactions happen later in the hierarchy of processes, was one of the classical views of multisensory processing (Stein and Meredith, 1993). This model has now been challenged by numerous studies showing how low-level sensory cortices can store and respond to information from other sensory modalities (see Ghazanfar and Schroeder, 2006; Van Atteveldt et al., 2014a for reviews), thus suggesting that this convergence is a key part of primary sensory regions' function. A study by Martuzzi and colleagues (2007) using fMRI has shown that V1 can show increased blood-oxygen-level-dependent (BOLD) activation in response to auditory information, while A1 behaves similarly, showing increased BOLD activation when presented with visual information. In addition to that, a modulation of peak activity latency in primary sensory regions was found when comparing multisensory BOLD responses with either auditory or visual responses, suggesting multisensory interaction is happening at those sensory locations and is modulating their neural responses (Martuzzi et al., 2007). When using multivariate techniques of analyses, it is also possible to decode sounds from EVC (Vetter et al., 2014, 2020) and visual content from primary auditory cortex (Meyer et al., 2010), which further supports the theory that multisensory processing is not exclusively a feedforward operation. The information received by primary sensory regions is detailed enough that it can be distinguished from other modalities i.e., primary auditory cortex can distinguish visual from tactile stimulation (Liang et al., 2013).

Together, the evidence seems to support an account of multisensory processing that involves a network of structures which includes (but is not limited to) primary sensory cortices.

The fact that information from different sensory modalities arrives in the same brain area in close temporal proximity does not necessarily mean that multisensory integration is happening within that area. Multisensory integration is defined as the brain's ability to identify whether two or more sensory stimulations come from the same source, combining them into a single percept, and is accompanied by neural evidence in the form of response enhancement (Calvert et al., 2001). However, response enhancement is not the only metric that has been used in the literature, and the question of which statistical measure can better identify multisensory integration in the brain when using traditional fMRI experiments (Beauchamp, 2005; Laurienti et al., 2005; Stevenson and James, 2009), still has no definitive answer (Murray et al., 2012). What can be more easily identified is the behavioural consequences of multisensory integration, which I will discuss in the following sections of this Chapter.

It has been proposed by Meredith and Stein (1986, 1987) that integration is more likely to happen when stimuli occur close in time and close in location. Thus, through multisensory integration, we are also able to solve ambiguous situations and to focus attention when there are too many sources. For example, a study by Van der Burg and colleagues (2011) found that when presented with a visual target in a cluttered environment, participants had difficulties in finding them, while their performance was greatly improved when the visual target was presented simultaneously with a "beep" noise. This suggests that integrating multiple sources of information can help direct attention and inform the brain on how to optimally process external cues. Moreover, Stein and Meredith (1986, 1987) also suggests that multisensory integration results in a multiplicative, rather than additive change in neural activity, and that this is particularly striking in the event of unimodal stimuli that evoked weak, close to threshold signals; the weaker the individual stimuli, the greater it was the potential for response amplification due to multisensory integration. Their proposition is also supported by new evidence that sees multisensory integration being enhanced for weak unimodal stimuli, or stimuli with a poor signal-to-noise ratio (Holmes 2007, 2009; Leone and McCourt 2013), or when the sources of stimuli are farther from the participant, multisensory integration offers a more effective advantage compared to when the sources are nearer (Van der Stoep et al., 2016). This supports the idea that the brain benefits more from multisensory integration when stimuli are unclear or ambiguous, and is thus a valuable tool in our everyday exploration of the world, where plenty of stimuli are outside our immediate focus. However, whether this integration actually happens in primary sensory regions or exclusively in association cortices is still the subject of debate.

1.3.1: Multisensory processing in EVC

Studies on non-primate mammals have shown how V1 pyramidal neurons can integrate sounds, speaking in favour of the theory that EVC not only supports multimodal processing but also integration (Chanauria et al., 2019; McClure and Polack, 2019). Since the discovery that EVC receives feedback from other sensory modalities, a large library of experiments has been conducted to understand the purpose of this flux of information transferred to EVC.

For example, a study by Macaluso, Frith and Driver (2000) found that tactile stimulation of a hand would enhance activity in the early visual cortex when paired with a visual target on the same side of the visual field. It had already been found that touch can improve vision near the location of the touch (Butter, Buchtel and Santucci, 1989; Spence et al., 1998) and it was believed to be due only to multimodal regions working to focus spatial attention, but Macaluso’s experiment suggested that early visual cortex might also be involved in this multimodal integration process. It has also been shown that salient sounds evoke potentials in EVC in a spatially specific manner, activating neurons in the visual area corresponding to the apparent location of the sound, even when sounds are task-irrelevant (McDonald et al., 2013), highlighting once again the role of top-down influences on the visual cortex.

Studies using Transcranial Magnetic Stimulation (TMS) have also investigated auditory modulations in EVC (Ramos-Estebanez et al., 2007; Romei et al., 2009). TMS is a brain stimulation method in which electric currents run through a figure of 8 coil, inducing a magnetic field perpendicular to this coil. The magnetic pulse travels through the scalp and induces an electric field in the brain tissue underneath, causing temporary disruption or enhancement of a specific region. It can be directed at fairly specific areas of the brain, with no harm to subjects. When used on the primary visual cortex, subjects can experience spots of light in their visual field called phosphenes. The experimenters calculated the minimum intensity of the TMS pulse able to produce phosphenes in each participant's visual field, and then send pulses to V1 right below that level. They found that when paired with a sound, the TMS pulse would cause participants to see a phosphene in significantly more trials compared to when the TMS pulse was presented alone (Ramos-Estebanez et al., 2007), giving further evidence that EVC activity is responsive to information coming from other senses, and these can modulate the excitability of EVC.

When comparing congruent and incongruent audio-visual stimulation, it was found that EVC seems to preferentially respond to incongruent stimuli, as opposed to regions higher in the visual hierarchy, which prefer congruent stimulation (Meienbrock et al., 2007). This activity was interpreted as a correlate of the recurrent feedback reaching EVC during mismatch processing, which could mean that when an auditory-visual mismatch is detected in multisensory regions, low-level sensory regions need more processing power to make sense of the incongruent information. Thus, it is plausible that the function of multisensory information in EVC is of disambiguation or clarification in the event of unclear visual stimulation.

When humans view naturalistic videos paired with semantically incongruent sounds however, decoding of the visual stimuli's identity from V2 and V3 is significantly impaired, compared to when the sound is congruent (de Haas et al., 2013). These findings can be explained by incongruent stimuli introducing noise in the activity patterns in EVC, and are to be interpreted as complementary to Meienbrock et al. (2007). While we know that incongruent audiovisual stimuli activate EVC more than congruent ones, the activity patterns they induce in EVC are more difficult to distinguish. Taken together, these results suggest that when EVC receives incongruent feedback from other sensory modalities, it might engage in reprocessing or in the re-evaluation of visual stimuli.

1.4: Audio-visual integration

Over the course of the last 60 years, audio-visual integration has been thoroughly studied, informing us of the importance that low-level spatiotemporal factors have for integration to happen. As mentioned before, Stein and Meredith (1993) described multisensory integration as more likely to happen when unisensory percepts are close in space, in time, and when the multisensory stimuli evoke stronger responses compared to when stimuli are presented in isolation. When multimodal stimulations meet these criteria, the source of these is more likely to be attributed to the same object.

When discussing the criteria the brain uses to make object identity decisions, Bedford (2001) mentions how conscious awareness is not necessary, suggesting that multiple modalities can be integrated to recognise a unique source even in the absence of awareness. This allows us to study audiovisual integration under an umbrella of different techniques, which I will describe in the following sections.

1.4.1: Audio-visual integration in the presence and absence of consciousness

One way to study audiovisual integration is to use ambiguous stimuli, since looking at how multimodal stimulation facilitates disambiguation or perception is an effective measure of integration. Phenomena like the McGurk effect (McGurk and MacDonald, 1976), in which speech sounds are miscategorised when paired with incongruent visual cues from the speaker’s lips, are examples of integration that happen when participants are fully aware but with ambiguous stimuli. Another example is the sound-induced flash illusion (Shams, Kamitani and Shimojo, 2000), in which the number of auditory cues alters the perception of the number of visual flashes that are being simultaneously presented, so that hearing two auditory cues will induce the illusion of two visual flashes appearing even if only one was actually presented. Neural activity also reflects the illusion, as the illusory visual flash is associated with enhanced activity in V1 (Watkins et al., 2006). These are only a few examples, but the versatility of these phenomena makes them useful tools to identify the conditions that make integration possible.

An interesting technique that allows to make ambiguous any type of visual stimulus is binocular rivalry (Crick, 1996). By showing two different images to the eyes, it is possible to induce a fluctuation in perception where the stimuli are perceived one at a time rather than superimposed. This is thought to happen due to the brain receiving two inputs that hold equal salience and that compete for control. Combining this phenomenon with potentially disambiguating information from other sensory modalities, it is possible to see whether multisensory integration has any effect on which image is perceived.

Conrad et al. (2010) presented different random-dot kinematograms to each participant's eye paired with sounds that had various degrees of motion cues (congruent and incongruent movement, non-motion, no sound). They calculated the dominance times for each visual percept and found that congruent directional motion sounds influenced more the temporal dynamics of binocular rivalry (i.e., congruent motion sounds increased dominance times of congruent visual stimuli), compared to incongruent and non-motion sounds cues, highlighting the importance of congruence for multisensory integration. Semantic congruence is also important, as shown by the work of Chen, Yeh and Spence (2011), who presented participants with line drawings of objects to each eye (i.e., a drawing of a car and a bird) and paired them with either a sound congruent to one of the images or completely incongruent. They found that in the first case, perception of the image non-congruent to the sound was shortened in favour of the congruent one, an effect that was not visible in the case of incongruent sounds.

This top-down influence does not seem to be intrinsic to semantically congruent stimuli, but can also be a previously learned association. Einhäuser, Methfessel, and Bendixen (2017) investigated whether sound stimuli they had previously trained participants with would facilitate switches in rivalrous perception to favour the image associated with each specific sound. They found that even rapid association of sound-visual stimuli can aid visual perception, which seems to suggest that as long as the brain recognizes a link between two multimodal stimuli, this can influence sensory perception in a top-down way.

Another way to investigate audiovisual integration is in the absence of visual awareness: hiding or suppressing the presence of certain stimuli and testing whether presenting a stimulus from a different modality would facilitate perception. A technique used to study unconscious visual perception is continuous flash suppression (CFS), developed by Tsuchiya and Koch (2004; 2005), in which a static image is presented to one eye, while the other sees a group of rich, rapidly changing stimuli. This creates a conflict similar to binocular rivalry, but more powerful since it effectively erases the static image from conscious perception. It can be used to investigate the preferential processing of certain stimuli (particularly when leading to conscious perception), comparing the time they take to break through from suppression, but it can be equally useful in the case of audiovisual integration. Aller et al. (2015) for example, found that when a visual flash is masked with CFS, its conscious perception is facilitated by a concurrent, co-localised beep sound. CFS has also been used to demonstrate how sounds that are semantically congruent to a hidden image (i.e., a Chinese character and the word associated with it) can prompt a faster release from suppression, compared to semantically incongruent ones, offering further evidence that multisensory integration exists at the semantic level (Yang and Yeh, 2011). This is also true for more complex visual scenes: where congruent auditory stimuli are presented in conjunction with the masked visual scene, the scene breaks into awareness faster (Tan and Yeh, 2015).

1.5: Summary of the current thesis, research questions and hypotheses

To summarise the current state of the field, there is sufficient evidence that points to a widespread network of areas that processes multisensory information, and indicates that primary sensory areas not only receive inputs from other sensory modalities, but also represent these inputs to some extent. However, in the context of EVC and auditory feedback specifically, the exact content and function of this flux of information are yet to be determined.

Stemming from Vetter's results (Vetter et al., 2014), the current project aims to understand the nature of the auditory information fed back to the early visual cortex and, possibly, shed light onto the purpose of auditory information representation in early visual cortex. The first experiment aims to understand what kind of semantic sound information, and to which level of detail, is fed back to early visual cortex. Even though it has been shown that it is possible to distinguish between general categories of sounds (e.g., animate versus inanimate sounds, Vetter et al., 2014), it is still unclear to what extent different semantic categories of auditory stimuli can be discerned in EVC, e.g., animal sounds versus people sounds. Determining which semantic categories of sounds are represented in early visual cortex would give significant insight into the content of information that is communicated between auditory and visual cortex. This, in turn, could help us understand a) how this information is used by the visual cortex, particularly EVC, during actual visual perception.

The first study described in this thesis is an fMRI study similar to Vetter et al. (2014) but using a wider range of sounds organised in a hierarchical categorical structure. More specifically, 'animal' and 'human' were used as sub-categories of the 'animate' category, and 'vehicle' and 'object' as sub-categories of the 'inanimate' category (Figure 1.2). The results were analysed using a multivariate approach in order to see if it was possible to identify the category of a sound on the basis of neural activity patterns in EVC, and to reveal how fine-grained this categorical distinction can be (i.e. going from the highest tier of the hierarchy of categories to the lowest).

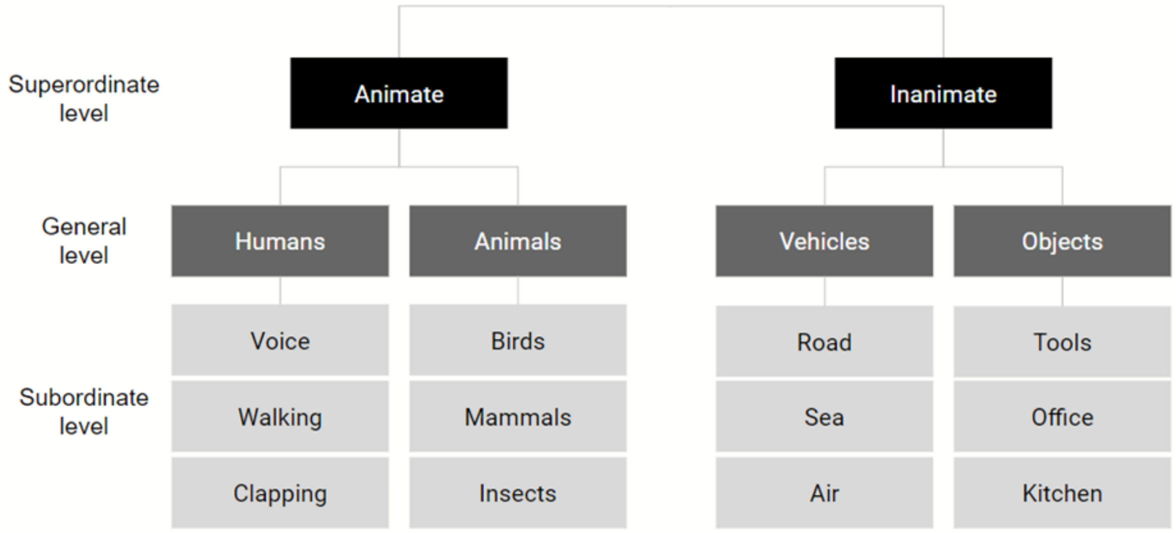


Figure 1.2: Hierarchy of the sound categories investigated in the first experiment. Within the subordinate level, there are individual sounds which will be later referred to as 'exemplars'.

The decision to compare human, animal, vehicle and object sounds, rather than other semantic categories, is supported by previous research showing that several brain regions can make the distinction between these categories in different sensory domains. For example, regions in the ventral visual pathway preferentially activate for animals versus tools and vice versa (Chao and Martin, 2000; Beauchamp et al., 2002), as well as being able to distinguish objects according to animacy (Blumenthal et al., 2018). This was found for sound stimuli as well, since Engels and colleagues (2009) reported that sounds coming from different categories of animate and inanimate sources could activate distinct cortical networks in auditory cortex. The hierarchical organisation of this study's semantic categories was also inspired by the results of Kriegeskorte, Mur and Bandettini (2008), in which they were able to find similarities between neural representations of images belonging to the same semantic category in both humans and monkeys. In higher regions of the visual cortex, the distinction between animate and inanimate and between human and animal was represented in neural activity patterns. Moreover, as in Bracci and Op de Beeck (2016), the categories used in the current studies were more generic (i.e., vehicles rather than cars), in order to avoid low-level visual properties similarities in the visual mental images potentially evoked by sounds.

As the main goal of this experiment was to determine the information content that is communicated from auditory cortex to EVC, I hypothesise two possible outcomes: 1) the early visual cortex is able to distinguish sounds only to some extent, e.g. only at the 'general' or superordinate level; 2) the level of information communicated down to EVC is sufficient to allow the distinction between all of the sounds and their classification into categories. If there is evidence suggesting that early visual cortex is able to distinguish sounds at a more specific level, it would further challenge the idea that primary visual areas only represent visual information and support the idea that primary sensory areas might be more multisensory in nature (Ghazanfar and Schroeder, 2006; Murray et al., 2016).

The focus of Experiment 1 was on the content of the flux of information that goes from auditory cortex to early visual cortex, but did not provide insight into its purpose or how it is employed by the early visual cortex. To make inferences about that, we would need a way to link behavioural performance with the auditory feedback in early visual cortex, and as a first step towards that, I chose to test whether auditory information could influence visual perception in a behavioural experiment when visual information is ambiguous. The second part of the current thesis aims to investigate the behavioural repercussions of auditory feedback on the computations happening in EVC by indirectly assessing a task that has neural bases in EVC (Tong and Engel, 2001; Kamitani & Tong, 2005; Haynes & Rees, 2005): binocular rivalry. The second experiment employed a selection of sounds used in Experiment 1 accompanied by a set of corresponding visual stimuli, in a task where subjects were instructed to indicate which visual stimulus they were perceiving at any given time point. This provided us with a timeline of the fluctuation in visual perception between the two eyes, depending on congruent or incongruent sound stimulation. I aimed to see 1) whether congruent sound stimuli would have an effect on the time spent perceiving each image; 2) whether sounds belonging to the same macro-category as the image but not completely congruent would facilitate perception, too; and 3) whether the congruency effect of sounds on images differed between semantic categories. The categories used were once again humans, animals and vehicles; I removed objects and selected a smaller sample of sounds for each category for practical reasons (explained in Chapter 4). Based on the results from Experiment 1, I also decided to focus our analyses on the human category. I hypothesised that: 1) congruent human sounds would facilitate perception of human images, and 2) partially congruent sounds (i.e., sounds belonging to the same broad semantic category of human, referred to as "relevant") might, to some extent, also facilitate perception of human images.

The experiments are presented in Chapters 3 and 4 in the form of complete scientific articles.

Chapter 2: Methods

After describing the theoretical background behind our research questions and hypotheses, I will discuss in this chapter the methods and techniques employed during this research project. First, I will go over the experimental design and methods used to collect behavioural and brain imaging data for Experiment 1, then describe and discuss the techniques of data analyses. The second section will be structured in the same way to cover Experiment 2, albeit shorter given the simpler design.

2.1: Auditory stimulus development

As mentioned in the introductory chapter, Experiment 1 involved presenting a large sample of sounds to participants inside an MRI scanner and comparing the neural activity patterns generated in response to them, in a way that allowed us to understand if it was possible to decode the semantic category of sounds. When designing the experiment, the first objective was to find a good set of stimuli that we could reliably use to address this question. To do this, we conducted two different pilot studies on a selection of sounds sampled from different online databases and platforms (e.g., www.youtube.com, www.freesound.org, www.soundbible.com).

The initial stimulus categories were identified during a focus group-like session comprising of six people (including the author). While we did not explicitly check for familiarity, we aimed at selecting sounds that the majority of people in our pool of subjects would be able to identify (i.e., animate or inanimate sounds they were likely to come in contact often either via personal experience or television). The criteria we used to narrow down the list of potential sounds were: a) sounds should have a high percentage of recognizability and distinctiveness, meaning they had to be as unambiguous as possible when presented to human ears; b) sounds should adequately represent the semantic category they were selected for (i.e. be a good example of what a dog sounds like if the sounds belonged to the category ‘dog’); c) animate sounds should not have any emotional or linguistic content immediately detectable in order to avoid confounds with emotional valence or speech perception, as these have high salience and can draw more attention (e.g. Gerdes, Wieser and Alpers, 2014). According to these criteria, we did not use samples of laughter, crying or angry speech, we excluded meaningful speech in any language as well as sounds that were considered easily mistakable for something else (i.e., cats' and foxes' vocalisations can often sound like human voices, bacon sizzling in a pan can sound like rain, etc.). Moreover, all the sounds we selected involved movement in some measure, and while we cannot be certain there are no differences between the vividness of the impressions they create, there should not be a sound that attracts more attention than others due to a more dynamic mental image. Once we had a large pool of suitable sounds, we began the piloting phase.

The aim of pilot study 1 was to identify, for each of our semantic categories, the most recognizable audio recordings. For each of the categories we wanted to include in the study, three different samples were selected; they were sampled from different audio recordings of the same category (i.e., different recordings of a dog barking). The sounds were 3 seconds long and were presented to 50 participants through an online Qualtrics questionnaire, in which they were asked to listen to each sound and write what they thought was producing the sound. The stimuli were presented with fMRI noise added on top of them, to ensure that participants would still be able to recognise the target sounds while in the scanner. The results showed that despite coming from the same category of stimuli, some of the samples were recognised with higher accuracy compared to others (Figure 2.1). For example, one of the car recordings was often mistaken for a growling animal, one of the recordings for the jet sound was miscategorised as a racing car, or the sound made by a fly was mistaken for indistinct buzzing. The trials selected for the following steps were those with recognition rates higher than 80%, with some categories being used twice (i.e., we used both walking on gravel and walking on heels from the walking category).

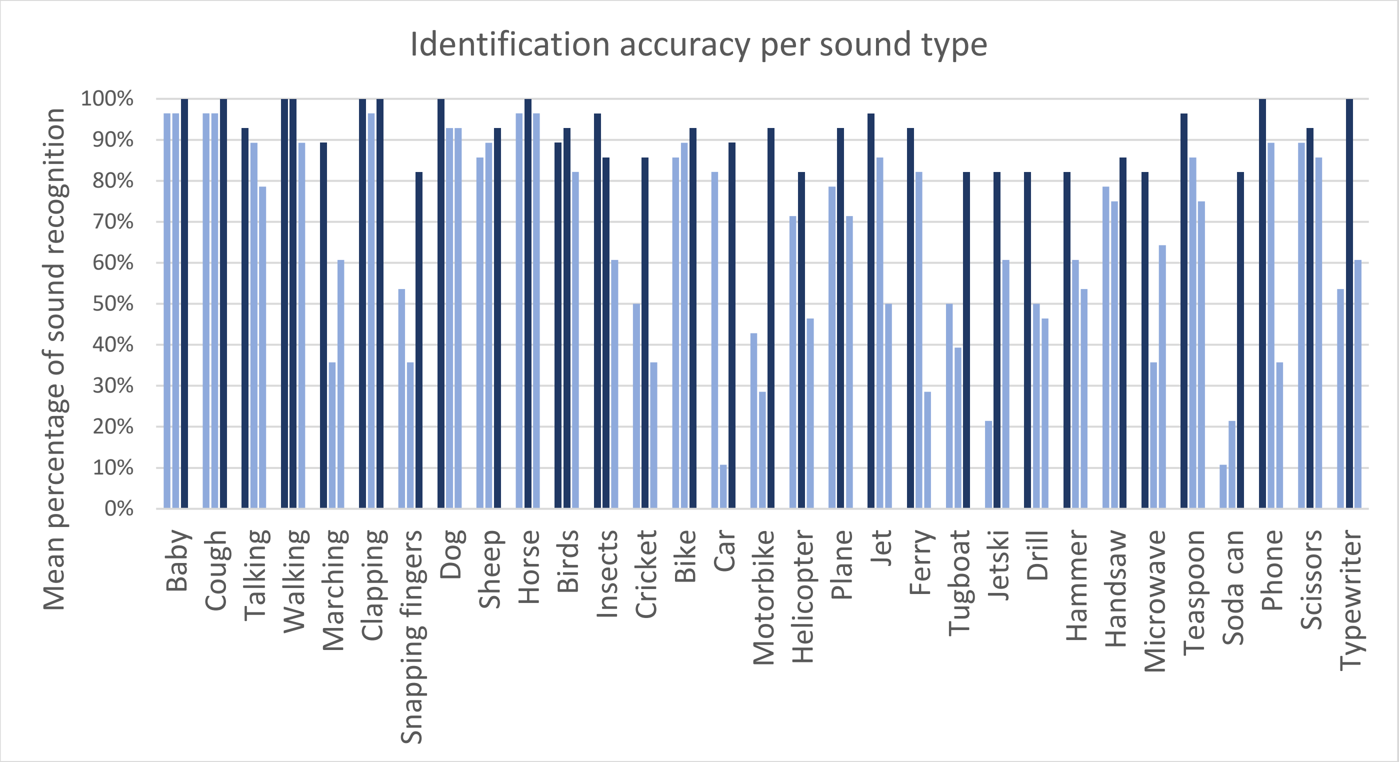


Figure 2.1: Accuracy ratings for each sound piloted in Pilot Study 1, showing the mean percentage of recognizability of all sounds, with sounds selected for the next part of piloting marked in dark blue.

As we also wanted to optimise the time inside the scanner, the aim of pilot study 2 was to present shorter versions of the same sounds and see whether participants would still recognise the sound with enough confidence. Another online Qualtrics questionnaire was presented to 15 volunteers, in which they listened to 2-second versions of the previously piloted sounds and had to answer with their guess for what was producing the sound and their confidence level. We saw that participants were able to recognise the sounds after 2 seconds of exposure, with accuracy above 70% (Figure 2.2), confirming that a shorter amount of time was still sufficient to prompt semantic recognition.

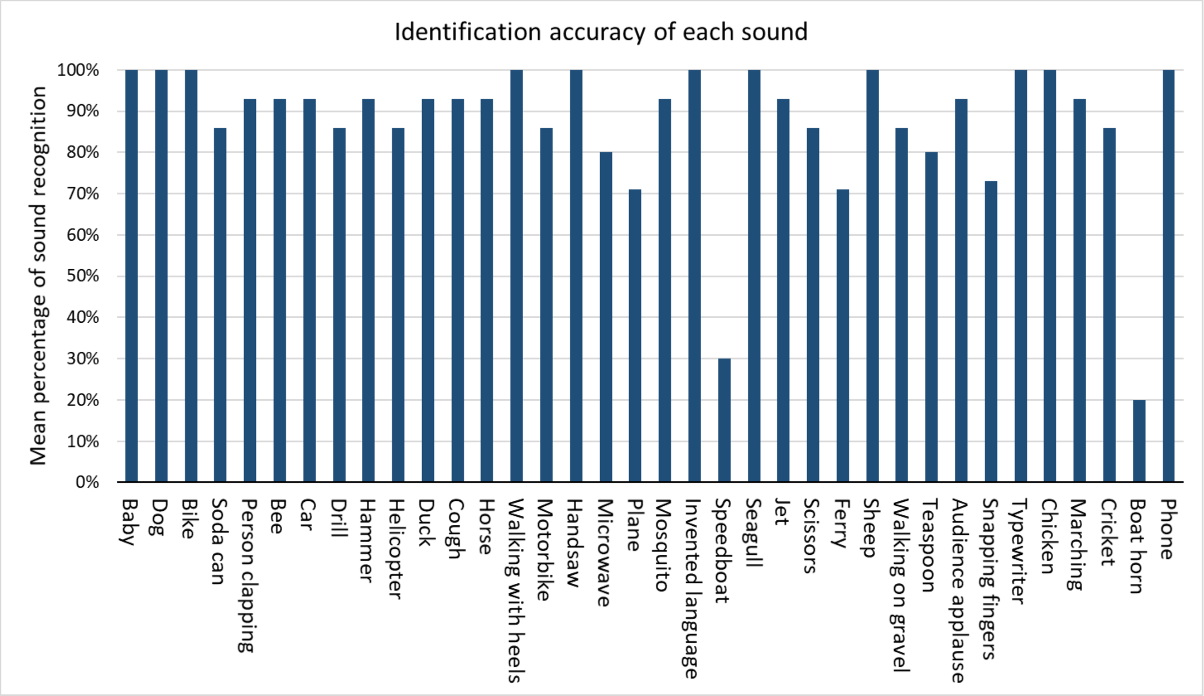
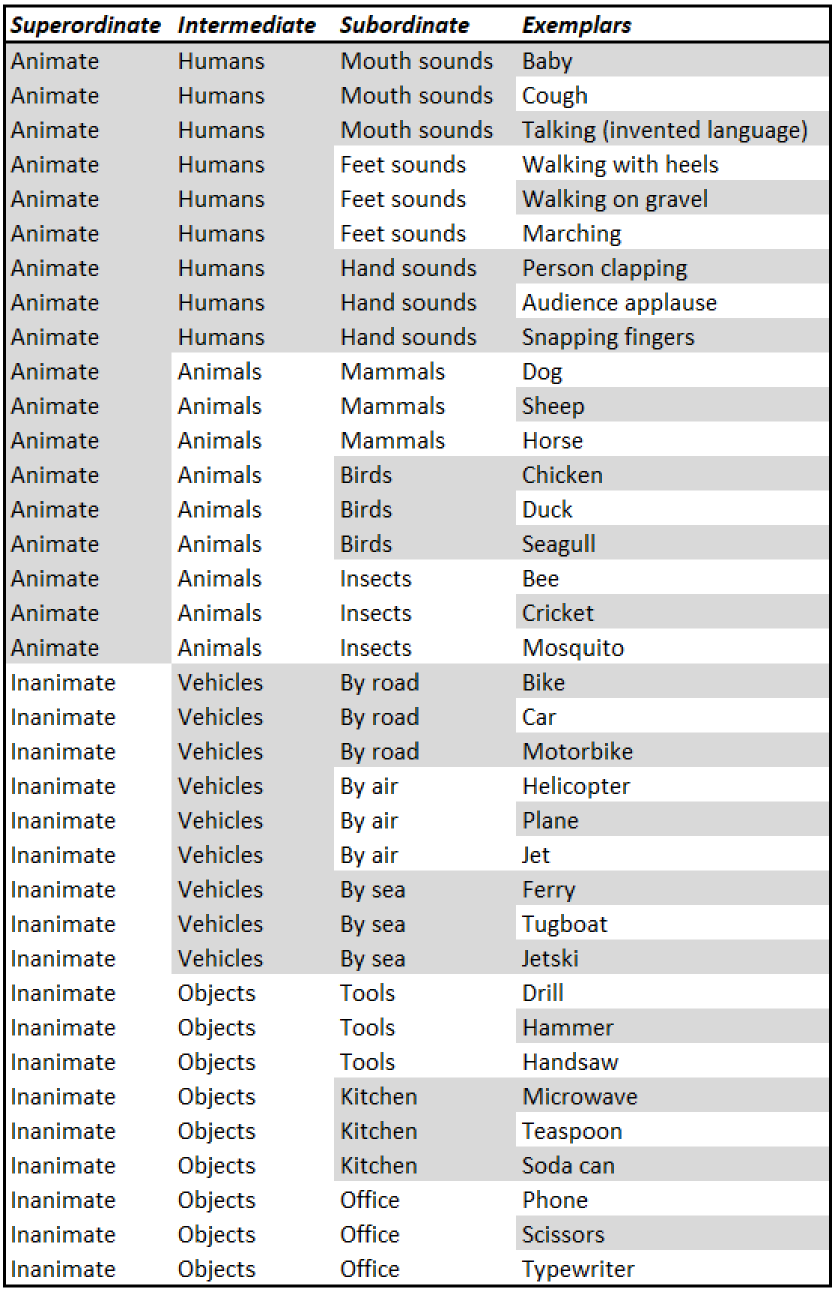


Figure 2.2: Accuracy ratings for each sound piloted in Pilot Study 2, showing the mean percentage of recognizability of all sounds.

Two sounds were not recognised with sufficient accuracy (speedboat and boat horn), so they were excluded from the experiment and substituted with two other sounds (tugboat and jet ski), giving us the final sample of 36 sounds employed in Experiment 1 (Figure 2.3).



Superordinate	*Intermediate*	*Subordinate*	*Exemplars*
Animate	Humans	Mouth sounds	Baby
Animate	Humans	Mouth sounds	Cough
Animate	Humans	Mouth sounds	Talking (invented language)
Animate	Humans	Feet sounds	Walking with heels
Animate	Humans	Feet sounds	Walking on gravel
Animate	Humans	Feet sounds	Marching
Animate	Humans	Hand sounds	Person clapping
Animate	Humans	Hand sounds	Audience applause
Animate	Humans	Hand sounds	Snapping fingers
Animate	Animals	Mammals	Dog
Animate	Animals	Mammals	Sheep
Animate	Animals	Mammals	Horse
Animate	Animals	Birds	Chicken
Animate	Animals	Birds	Duck
Animate	Animals	Birds	Seagull
Animate	Animals	Insects	Bee
Animate	Animals	Insects	Cricket
Animate	Animals	Insects	Mosquito
Inanimate	Vehicles	By road	Bike
Inanimate	Vehicles	By road	Car
Inanimate	Vehicles	By road	Motorbike
Inanimate	Vehicles	By air	Helicopter
Inanimate	Vehicles	By air	Plane
Inanimate	Vehicles	By air	Jet
Inanimate	Vehicles	By sea	Ferry
Inanimate	Vehicles	By sea	Tugboat
Inanimate	Vehicles	By sea	Jetski
Inanimate	Objects	Tools	Drill
Inanimate	Objects	Tools	Hammer
Inanimate	Objects	Tools	Handsaw
Inanimate	Objects	Kitchen	Microwave
Inanimate	Objects	Kitchen	Teaspoon
Inanimate	Objects	Kitchen	Soda can
Inanimate	Objects	Office	Phone
Inanimate	Objects	Office	Scissors
Inanimate	Objects	Office	Typewriter

Figure 2.3: Full list of sound stimuli as used in the fMRI Experiment, divided by category.

While recognisability of the sounds was important to consider during sound selection, it was also a useful measure to include during the actual experiment. After the fMRI session, participants were presented with a similar questionnaire to the one used for Pilot 2, in which we presented the 36 sounds once again, and asked them not only to identify the sound and give a confidence rating to their guess, but also to confirm whether they recognised the sound as such when they heard it inside the scanner. If participants cannot correctly identify the majority of the sounds, their data might reflect how those semantic categories of sounds are represented in the brain.

2.2: FMRI methods

Functional magnetic resonance imaging (fMRI) is a neuroimaging technique that measures brain activity indirectly by relying on the relationship between cerebral blood flow and neural activity. When neurons are active, they require more energy, so the brain supplies oxygenated blood to those regions (hemodynamic response). These changing concentrations of oxygenated and deoxygenated blood create local magnetic field distortions, as oxygenated and deoxygenated blood have different magnetic properties (Thulborn et al., 1982), and can be detected and imaged with magnetic resonance imaging (MRI). A higher ratio of oxygenated over deoxygenated blood corresponds to an increased blood-oxygen-level-dependent (BOLD) signal, which is associated with higher local neural activity in a specific region.

The spatial resolution of an MRI scan is determined by the size of the imaging voxel, which is the three-dimensional equivalent of a pixel. A smaller voxel size allows for higher spatial resolution, meaning that the imaging technique can distinguish smaller structures within the brain, but higher resolution results in longer scans and larger datasets. In terms of temporal resolution of MRI, this is determined by the repetition time (TR): the time between successive acquisitions of a single volume of data. A shorter TR allows for higher temporal resolution, meaning that the imaging technique can distinguish signals that occur closer together in time (Huettel, Song and McCarthy, 2014).

A typical fMRI experiment would include a high-resolution structural scan and functional data acquired through an Echo Planar Imaging (EPI) sequence (Mansfield, 1977), which uses a particular radiofrequency pulse that allows for the acquisition of a single image (or "slice" of the brain) in a few milliseconds. Generally, data collected using this sequence are in the form of ~ 32 two-dimensional slices sampled with a TR of 2 seconds per volume, with a spatial resolution of around 3x3x3 mm voxel size (Glover, 2011). These parameters can be tweaked to fit the specific purposes of the individual experiments.

When analysing functional data, the BOLD signal can be modelled with a mathematical function called a Gamma distribution, which is then referred to as the Hemodynamic

Response Function, or HRF. This peaks around 6 seconds after the onset of the stimulus that triggers its response, its width is of around 4 seconds, and it is followed by an average decrease in BOLD signal below the pre-stimulus baseline, called post-stimulus undershoot. Altogether, the HRF lasts around 20 seconds, and sampling with a TR of 2 seconds while jittering the onset of the stimuli ensures an accurate modelling of its shape.

For an efficient fMRI experimental design, it is important to keep participants as busy as possible inside the scanner (Henson, 2007), minimising the time they spend not engaged in a task. This was particularly relevant to consider for Experiment 1, due to the repetitive, passive nature of our task. Participants would be laying in the scanner in the dark and with blindfolds on, with the only instruction to listen to some natural sounds. Thus, the risk of them falling asleep or losing focus in favour of mind-wandering was high. It is common practice in experiments like these to implement an additional task as an attention check, which we did in the form of adding "beep" noises on top of 10% of trials in random positions and asking participants to press a button whenever they heard the beep. This was also done to ensure that participants paid attention to each sound. We verified the accuracy of detecting beep noises post-hoc to ensure each participant paid attention and did not fall asleep.

2.2.1: Univariate Analysis

During an fMRI experiment, sets of stimuli or tasks are typically employed to induce different cognitive states in subjects, which in turn generate hemodynamic responses. The BOLD signal recorded in a voxel across an entire experimental run is called a time-series, and there are several ways to interpret and analyse these responses, the most traditional approach perhaps being the standard univariate analysis, generally done through a General Linear Model (GLM). The GLM is used to model the BOLD signal in terms of one or more explanatory variables, on a voxel-by-voxel basis. Usually, it is composed of predictors of interest, which are derived from the experimental task, as well as some other nuisance regressors (i.e., breathing, head movements, etc.). The goal is to construct a model which can explain or predict most of the variance recorded across each voxel, expressed in beta estimates, and return small residuals, or the differences between the fMRI signal predicted by the GLM and the actual measured fMRI signal.

Beta estimates from the same predictor (typically an experimental condition) are pooled together and their mean is calculated for each voxel, which is then compared against baseline or another condition. The resulting statistical maps are projected onto the brain and show the areas where the activation during the two conditions compared is significantly different (Huettel, Song and McCarthy, 2014).

This approach can be applied to the voxels in the whole brain or can be circumscribed to specific regions of interest (ROIs), which are useful when the research question focuses on specific areas, or when predictions can be made on the basis of existing literature. As the number of voxels in an ROI is greatly reduced compared to the whole brain, it offers fewer chances of finding false positives. However, when conducting a GLM analysis at the whole-brain level, it is important to correct for multiple comparisons, which can be done through several techniques (e.g., cluster-level thresholding, family-wise error correction and false discovery rate correction, to mention some of the more widely used).

2.2.2: Multivariate Pattern Analysis

Univariate analyses are useful for answering questions regarding the involvement of a brain region in a certain cognitive process but are limited when it comes to tackling the stimulus-dependent relationships between regions, or understanding the different ways in which a region can respond to different stimuli. Multivariate analyses offer alternatives to this approach, allowing us to investigate neural responses in the form of activity patterns (Haxby et al., 2014).

As we mentioned earlier, neural activity is acquired across space and time through recordings of multiple voxels at different time points. The concept at the basis of multivariate analyses is representing each voxel at each time point as a separate dimension in a high-dimensional representational space. In other words, we are able to look at the neural response to certain stimuli as distributed patterns and apply different types of analyses. There are various types of encoding and decoding analyses that can be performed on activity patterns, but for the scope of this thesis, I will focus on multivariate pattern analysis (MVPA), a technique that allows us to establish how distinct a brain state is from another (Haxby et al., 2014). Neural responses are vectorised within neural representational spaces, and a machine learning algorithm is used to classify patterns and associate them with an experimental condition.

MVPA can be broken down into two stages: training and testing. In the training phase, a machine learning algorithm, i.e., a classifier, is trained with a portion of fMRI data recorded during the experiment, to which we assign labels that identify their category. In our case, fMRI data recorded when listening to a horse might be given the label "animal", whereas data recorded when listening to a typewriter is given the label "object". It is important that the test data are not used in the training of the classifier and totally independent from it (Kriegeskorte et al. 2009). If we consider our neural data as vectors, the classifier is trained to assign areas of the vector space to the labels provided with the training set. Once the classifier is sufficiently trained to distinguish activity patterns associated with one label from data associated with another label, its efficiency of distinguishing activity patterns is tested on a portion of the data previously not encountered. Each activity pattern is represented in the vector space and the classifier accuracy is calculated as the percentage of test vectors that are assigned the correct label (Pereira, Mitchell and Botvinick, 2009; Haxby et al., 2014).

If the activity patterns contain enough information to make each stimulus class sufficiently distinct from the others, the classifier should be able to distinguish the activity patterns in the test data with an above-chance level of accuracy. For example, if the classifier has been trained to learn to distinguish activity patterns elicited by human sounds and animal sounds, looking at a pattern corresponding to a person talking should produce the label "human" as result. Assuming a good quality of data, the bigger the dataset with which the classifier is trained, the better its performance will be (Pereira, Mitchell and Botvinick, 2009).

For the testing phase, a common way of assessing classifiers' performance is through cross-validation (Hastie et al., 2009). Rather than training and testing the classifier only once, the process is done multiple times with different samples used for training and testing. There are several ways in which a dataset can be partitioned for cross-validation, but the current work uses the leave-one-out method.

For this approach, the dataset is split into subsets, and the classifier is trained with all subsets but one, and its performance is tested on the remaining one. The process is then repeated, cycling through all groups, with the advantage of a) always having a big sample of data to train the classifier with, making it more accurate and b) obtaining a performance more generalizable to other, unseen data. The fMRI data we obtained from Experiment 1 was naturally divided into 4 subsets offered by the division into 4 experimental runs, which meant that the classifier was always trained on 3 runs and tested on one, and the results of this cross-validation were then pooled across all four iterations and averaged into a single accuracy score.

Although cross-validation is an excellent way of testing a classifier's performance (Hastie et al., 2009), it is not a way to determine the statistical significance of accuracy scores. To do that, we performed permutation analyses, in which the classifier is trained and tested repeatedly (in our case 1000 iterations) on the dataset with randomly-assigned labels to create a null distribution (Golland and Fischl, 2003; Stelzer, Chen and Turner, 2013). P-values are then derived from the probability of obtaining, in the randomised distribution, a value as large or higher as the one obtained in the observed performance. This returns the smallest possible p-value of 0.001. This was done within each subject and then, for the group level, p-values were derived from the mean randomisation distribution and the mean real label performance.

Another aspect to consider with MVPA is feature selection. In order to make a classifier faster and more efficient, it is possible to reduce the input data by eliminating non-relevant features which, in the case of fMRI data, are non-relevant voxels (Pereira, Mitchell and Botvinick, 2009). Our main hypothesis focused on early visual cortex regions, thus we restricted the brain areas on which we conducted our MVPA to three regions of interest (ROIs): V1, V2, and V3. We also included auditory cortex as positive control region as we expected the classifier to distinguish the activity patterns for different sounds very well in auditory cortex.

MVPA can also be conducted at the whole-brain level, without the a priori identification of ROIs. Searchlight analysis (Kriegeskorte and Bandettini, 2007a; Kriegeskorte et al., 2006) is an MVPA method in which a sphere of a predetermined radius is created around each voxel in the brain, and within that, classification and cross-validation are performed. The result of the analysis is a map projected onto a brain surface that indicates all the areas where the classifier performance was higher than chance level, corrected for multiple comparisons with a cluster level threshold test. Findings from searchlight analyses can give insight into which regions contain enough information to allow the classifier to distinguish semantic categories of sounds. Understanding which and how many brain areas show this, could be important when proposing which brain networks might be involved in sound categorisation and information flow from auditory to visual cortex, which is why a searchlight analysis was also performed for Experiment 1.

As already mentioned previously, in the realm of fMRI analysis, univariate and multivariate approaches are used to draw different types of conclusions, which is why a certain type of approach might be better suited for one experiment but not another. For the purpose of answering questions regarding which sound categories are represented in early visual cortex, multivariate analyses are best suited. Nonetheless, looking at voxel-wise neural activity with a standard univariate approach is also a necessary control. So far, our understanding is that univariate analysis is not sensitive enough to pick up the small difference in spatial activity patterns in early visual cortex (Vetter et al., 2014, 2020), but this might not necessarily be the case for the experimental conditions we are interested in. If a difference between sound categories were found in early visual cortex averaging across a region of interest, i.e., on the univariate level, that would mean that the MVPA classifier results could have been driven by overall activity differences between sounds, independent of the spatial distribution of activity across a region of interest. For example, different categories of sounds attracting different levels of attention could affect overall activity in early visual cortex (Kastner et al., 1999). For this reason, differential activity levels on the univariate level need to be investigated also within each region of interest.

2.2.3: Retinotopic mapping

When choosing what method to use to define our ROIs, there were several options to consider: a) using a probabilistic atlas of function on which to base our assumptions of where the region boundaries are; b) using structural landmarks; or c) drawing the region boundaries around areas active during a functional localiser. To identify the auditory ROI, we used functional localisation by identifying the region of interest in the temporal cortex that showed peak activation for the univariate contrast of sound stimulation (independent of semantic category) compared to baseline. For the visual ROIs, we used retinotopic mapping (Sereno et al., 1995; Wandell, Dumoulin, and Brewer, 2007).

Retinotopic mapping generally refers to the spatially specific mapping of visual information coming from the retina directly to specific neurons of the visual system, and to the concept that each visual region represents a full or partial map of the visual field which can be reliably identified. Subjects fixate at the centre of their visual field while stimuli travel through their entire visual field, giving us a way to reliably associate neural activation with the stimulation of a certain part of the visual field. Retinotopic mapping is usually along two dimensions: eccentricity and polar angle, where in order to map different eccentricities the stimuli used are expanding or constricting circles, and for polar angle maps (i.e., mapping different spatial locations in the upper, middle and lower visual field) there is one rotating wedge revolving around the centre of the visual field. By applying linear correlation analyses to fMRI data in relation to the precise timing and position of the visual stimuli, it is possible to determine the boundaries of each visual region with great accuracy. In the interest of keeping the total time of the fMRI experiment as short as possible, we presented participants with stimuli mapping the polar angle only. Each participant thus had maps for V1, V2 and V3 drawn by hand for both hemispheres, which gave us higher confidence when making claims on the location of neural activity we analysed.

2.3: Binocular rivalry

As described in the previous chapter, section 1.4.1, the range of analyses employed for Experiment 1 allowed us to make inferences regarding the flux of information that goes from auditory cortex to early visual cortex and particularly its content, but they did not inform us about the function it might have, or how the auditory information is employed by early visual cortex. A way to address this would be adding a behavioural component to our neuroimaging experiment and then finding a way of linking behavioural performance with the auditory feedback arriving at early visual cortex. We opted to test whether auditory information could influence visual perception in a behavioural experiment when visual information is ambiguous.

When designing the experiment, we evaluated a few different techniques that could be used to induce visual ambiguity, such as the previously mentioned continuous flash suppression or visual masking. For reasons that will be discussed throughout this section, we selected binocular rivalry as the most appropriate technique. As briefly explained in the introductory chapter, binocular rivalry is a visual phenomenon that arises when two different stimuli are presented to each eye. Rather than fusing the two into a unique visual experience, the conscious perception fluctuates between the two images irregularly, creating for a few moments the perception of one image which then switches to the perception of the other image. Binocular rivalry can be achieved through the use of a stereoscope, a device that separates the visual field through the use of mirrors and makes the left side of the visual field be displayed in the left eye, and the right side of visual field displayed to the right eye.

Subjects look through the stereoscope and, usually, their task is to record the periods of time in which they experience one of the two percepts, shown on a computer screen (Carmel et al., 2010).

Binocular rivalry relies on stimuli being in equal measure perceivable (Blake and Logothetis, 2002), it is a very easy effect to induce, and the data allow for straightforward analyses. It can be achieved using inexpensive red-blue goggles or a mirror stereoscope, and it has been shown to work well with complex stimuli (Conrad et al., 2013). However, with binocular rivalry, the suppression of one of the rivalling stimuli is rarely a complete suppression. Pieces of the suppressed image can break through and result in a conscious percept that has elements of both images. In extreme cases, this phenomenon is called piecemeal rivalry (Kovacs et al., 1996): unless one of the images is clearly more visible than the other, participants might find trouble in determining dominance during these percepts. Smaller images tend to cause less piecemeal rivalry, which is why they are more indicated for binocular rivalry experiments (Carmel et al., 2010).

Due to the common occurrence of incomplete suppression, binocular rivalry is not a technique indicated when asking questions about unconscious processing, unless additional measures of verification are employed (Carmel et al., 210). For that, a technique inducing stronger rivalry like continuous flash suppression (CFS) would be more appropriate.

Among the visual parameters that can impact binocular rivalry, the most prominent one is stimulus strength (Levelt, 1965). This has been originally conceptualised as the combination of density, luminance contrast and blur of stimulus contours, all features associated with stimuli located away from the viewer (Brascamp, Klink and Levelt, 2015). Factors such as previous exposure (Goryo, 1969), familiarity (Engel, 1956) and cognitive salience (LoSciuto and Hartley, 1963) can also modulate dominance, and these are all elements to take into consideration when selecting stimuli for a binocular rivalry task. Moreover, binocular rivalry does not happen in the same way for every participant, as some see the switch in perception happen quite fast, while it might be slower for others (Blake and Logothetis, 2002).

Our main goal was to draw conclusions on multisensory integration and provide evidence of how auditory feedback might induce switches in visual perception, as well as being interested in showing images of equal strength to each eye. For these reasons, we concluded that the potential cons of binocular rivalry would not be weighing on our study, and its simplicity would offer a strong basis on which to build our audio-visual design.

In our experiment, subjects were instructed that the images appearing on the screen would belong to the human category on one side, and either to the vehicle or animal category on the other side. While they viewed these images in pairs, they listened to different samples of natural sounds that would either be congruent to the human picture (e.g. the sound of footsteps with a picture of feet walking), ‘relevant’ to the human picture (this is the term that I will use for sounds that are incongruent with the picture but still belonging to the human category, e.g. the sound of footsteps with a picture of a woman coughing), or ‘incongruent’, meaning that they belonged to the other category (either vehicle or animal). The main measure we used in the analyses was the percentage of time spent by each subject seeing the human stimuli or the ‘other’ stimuli over the course of a 45s trial. We based these elements on the experimental design of a similar study (i.e., Chen, Yeh and Spence, 2011).

In order to potentially make claims regarding the use that early visual cortex makes of auditory feedback, we decided to employ for Experiment 2 the same stimulus categories as in Experiment 1. However, binocular rivalry trials are typically longer than other psychophysics experiments, with an average length of around 60 seconds (e.g., Tong et al., 1998; Zhang et al, 2011, Chen, Yeh and Spence, 2011). This is to ensure that, during each trial, participants can switch perception regularly at least a few times. Given that our proposed trial time was 45 seconds and the difference in trial length between Experiment 1 and 2, it was not feasible to use all categories in Experiment 2, as that meant pairing 36 sounds with different combinations of images multiple times. So, we opted for a selection of ten, which we used for both sounds and images.

For the sound stimuli, we selected ten of the sounds used in Experiment 1 and either cut a longer sample from the original sound files, or artificially lengthened them by repeating certain parts. In this way, we ended up with ten 45-second sounds that we already knew participants would be able to identify. The image stimuli were new, as we did not have any visual stimulation in Experiment 1, and were selected from online databases to match the sound selected for that category. However, as mentioned earlier, binocular rivalry relies on images being in equal measure perceivable (Blake and Logothetis, 2002), thus we needed to normalise luminance and contrast across the whole set. The images were converted into grayscale first, then the background was manually modified to remove any standout features that could possibly drive visual perception. We then used the SHINE toolbox for Matlab (Willenbockel et al., 2010) to equalise luminance, and we manually adjusted them to have low contrast.

Before proceeding with data collection, the experiment was piloted on three subjects to verify that the images and the setup were correctly inducing binocular rivalry. Subjects reported an inability to fully switch percepts in certain trials, thus we adjusted the contrast of these images such that regular switches happened, as piloted in a new batch of four subjects.

Chapter 3: Experiment 1

Title: Early visual cortex represents semantic sound categories with a preference for human sounds.

3.1: Abstract

Auditory feedback to early visual cortex elicits distinguishable neural activity in early visual cortex when listening to different natural sounds in the absence of visual stimulation or sight (Vetter, Smith & Muckli, 2014, Current Biology; Vetter et al., 2020, Current Biology). However, it is still unclear what kind of information is contained in this feedback, and to what degree of specificity sound information is represented in early visual cortex. To address this question, we presented a large sample of 36 natural sounds, hierarchically organised into semantic categories, to blindfolded male and female human participants while acquiring functional MRI data. We analysed the fMRI activity patterns elicited by these sounds in each early visual region V1, V2 and V3, as determined with individual retinotopic mapping, using Multivoxel Pattern Analysis (MVPA). In early visual cortex, the MVPA classifier successfully distinguished between animate and inanimate sounds, as well as between human, animal, vehicle and object sounds. Pairwise classification of the different sound categories demonstrated that sounds produced by humans were generally better distinguished than other semantic categories. Whole-brain searchlight analyses showed that sound decoding also worked in regions of higher-level visual and multisensory processing. Thus, auditory feedback relays categorical and semantic information about natural sounds, particularly human sounds, to brain areas once believed to be exclusively specialised for vision. We conclude that early visual cortex function is not restricted to the processing of low-level visual features but includes representation of semantic and categorical sound information, potentially to be used to predict visual information.

3.2: Introduction

Early visual cortex can receive different types of sensory inputs: visual information coming from the retina, passing through domain-specific afferent pathways, as well as information coming from other sensory modalities, e.g., audition or touch (e.g., Petro, Paton and Muckli, 2017; Sathian, 2016). In the absence of visual stimulation from the retina, i.e., during eyes-closed conditions or during congenital blindness, natural sounds such as birds chirping, traffic noise and people talking can be decoded from activity patterns in early visual cortex (Vetter et al., 2014; Vetter et al., 2020). Thus, content-specific and categorical information from audition is fed back to early visual cortex and is represented there, possibly for the purpose to predict incoming visual information. However, it remains unclear which type of auditory information content is transferred all the way down to early visual cortex: is it only broad semantic sound categories, such as animate and inanimate sounds (Vetter et al., 2014) or are more fine-grained semantic sound categories (e.g., human, animal, vehicle and object sounds) also represented in early visual cortex?

Much previous audio-visual interaction research has focussed on how signals from audition and vision are integrated across space and time (e.g., Koelewijn et al., 2010; Doehrmann & Naumer, 2008), usually using simple stimuli such as beeps and flashes or visual gratings (e.g., Doehrmann & Naumer, 2008; Chen et al., 2011; Caclin et al., 2011, Chanauria et al., 2019). However, beeps and gratings do not contain ecologically valid information content, and the question of the semantic information content that is communicated between audition and vision has hardly been addressed. In ecologically valid environments we encounter in everyday life, the sounds that accompany our vision carry substantial information content with far-reaching consequences for our interactions with the world. For example, when we walk down a busy road and hear the sound of an approaching car from behind, our brain prepares to see a car a moment later (e.g., Bar, 2007; Friston, 2010). When we aim to cross the street, it is critical for our survival that our brain identified the sound behind us as that of an approaching car rather than a stationary car and prepared our visual sense accordingly.

The current study aimed to determine which semantic categories of sounds are fed down to and represented in early visual cortex, and to what degree of categorical specificity. To this aim, we presented a large sample of natural sounds, hierarchically organised into semantic categories, to blindfolded sighted participants in a functional magnetic resonance imaging (fMRI) experiment and decoded different sound categories from early visual cortex activity using multivariate pattern analyses (MVPA).

Regions higher in the hierarchy of visual processing have been extensively reported to be sensitive to different semantic categories of visual objects, with studies suggesting that the ventrotemporal cortex is structured to differentiate between animate and inanimate stimuli (e.g. Grill-Spector et al., 2014; Bracci & op de Beeck, 2016; Martin et al., 2018; Blumenthal et al., 2018) and to preferentially activate for animals versus tools and vice versa with a certain degree of spatial distinction (Chao and Martin, 2000; Beauchamp et al., 2002, Bola et al., 2022;). Moreover, visual object stimuli belonging to the same semantic category are represented with similar patterns of neural activation in IT cortex (Kriegeskorte et al., 2008; Bracci & op de Beeck, 2016). We drew from this body of evidence to select the categories of our auditory stimuli: animate, subdivided into human and animal sounds and inanimate, subdivided into vehicle and object sounds (see Fig.1). We tested which of those semantic categories, shown to be distinguished in high level visual cortex when presented visually, could be distinguished in early visual cortex when presented auditorily, and in the absence of visual input. This allowed us to determine the semantic information content contained in auditory feedback to early visual cortex.

3.3: Materials and methods

3.3.1: Participants

21 healthy adult volunteers were recruited for the study, all signed informed consent and were paid for their participation. The data of 3 participants were excluded due to poor sound stimuli recognition or head movements in the scanner, leaving us with a sample size of 18 participants (12 females, mean age 24). The study was approved by the ethics committee of Royal Holloway, University of London.

3.3.2: Sound stimuli:

We selected the natural sound stimuli based on behavioural piloting, sampling the sounds from different online databases (e.g., Youtube, FreeSound, SoundBible). We selected them to fit the following criteria: a) be as recognisable and unambiguous as possible and b) not conveying any emotion or containing meaningful speech. Thus, samples of laughter and crying were excluded, as well as sounds that could easily be mistaken for something else (i.e., bacon sizzling in a pan was often mistaken for rain, so it was not optimal for this experiment). Based on the behavioural piloting (n = 15), we selected sounds that were recognised with a minimum accuracy of 75%. The final set of stimuli used in the experiment consisted of 36 natural sounds with lengths ranging from 2-3 seconds (full list and length of each sound in the supplemental information at the end of this chapter). These were categorised in a hierarchical fashion: inanimate and animate sounds at the highest tier; animals, humans, objects and vehicles at the intermediate level, and further labelled with their specific subcategory in the last tier (See Figure 3.1). Three sound exemplars were used for each subcategory (i.e., ‘seagull’, ‘duck’ and ‘chicken’ for the ‘birds’ category). All sounds were normalized for amplitude, equalized with custom digital equalization filters (Sensimetrics) and presented binaurally and mono. As the sound length was not consistent for all our samples, it is possible that those differences might have influenced the results. However, the vast majority of sounds was 2-second long, with the exception of 3 samples, distributed in the animal, vehicle and object category, it is unlikely that the differences in length had an impact on the resulting neural activity or the analyses performed.

3.3.3: Data acquisition and experimental procedure

We acquired blood oxygen level dependent (BOLD) signals in a 3 T Siemens Tim Trio MRI scanner with a 32-channel phased array head coil, bandwidth 1628 Hz (TR = 2.0 s, TE = 30.6 ms, resolution 2.0 x 2.0 x 2.0 mm, 48 slices, flip angle 78°), using a multiband EPI sequence. A 6 min long T1 weighted structural MRI scan was also acquired. Participants were placed in the MRI scanner first for retinotopic mapping and then for the functional runs with sound stimulation. For retinotopic polar mapping, participants were instructed to fixate on a red cross at the centre of their visual field while a black and white checkered wedge (22.5 deg wide) rotated anticlockwise across their visual field. To ensure participants kept central fixation, they were asked to press a button with their index finger every time the cross at the centre changed colour. The run lasted about 8 min, comprising 12 cycles of wedge rotation, in line with standard retinotopic polar mapping procedures (e.g., Wandell et al., 2007; Schira et al., 2009; Muckli, Naumer & Singer, 2009). After retinotopic mapping, participants were taken out of the scanner, had a short break, and were given blindfolds and in-ear headphones to wear during the functional runs with sound stimulation. The use of blindfolds is justified by the fact that visual stimulation can create noise in the activity patterns, making the decoding of sounds more difficult (Vetter, Smith and Muckli, 2014; Vetter et al., 2020). They were instructed to keep their eyes closed at all times and to listen carefully to a series of sounds, which were presented individually and followed by an inter-stimulus interval of 4 seconds (jittered at 0.5 s). To ensure participants paid attention to all sounds, they were asked to press a button with their index finger every time they heard a “beep” noise on top of one of the stimuli. The control noises were either high pitch (800 Hz) or low pitch (400 Hz) pure tones, present in 10% of the trials. For an additional 10% of trials, we presented them with null events to allow brain activity to return to baseline. We used 3 sound exemplars in each of the 12 subcategories (see Fig. 1), adding up to 36 sounds stimuli in total. Each sound was presented 12 times, in a pseudo-randomised order, totalling 432 trials, divided into 4 functional runs. At the end of the scanning sessions, participants listened to the 36 sounds once more on the lab PC and reported the identity of the sounds, how confident they felt in their choice on a scale 1-10, and whether their perception matched that inside the scanner. The purpose of this was to ensure that participants identified the sounds accurately. Participants also completed a Vividness of Visual Imagery Questionnaire (VVIQ; Marks, 1973), which included picturing items or scenes in their minds first with eyes closed and then with eyes open and rating the vividness of their imagination on a scale of 1-5 (full list of items in supplementary information).

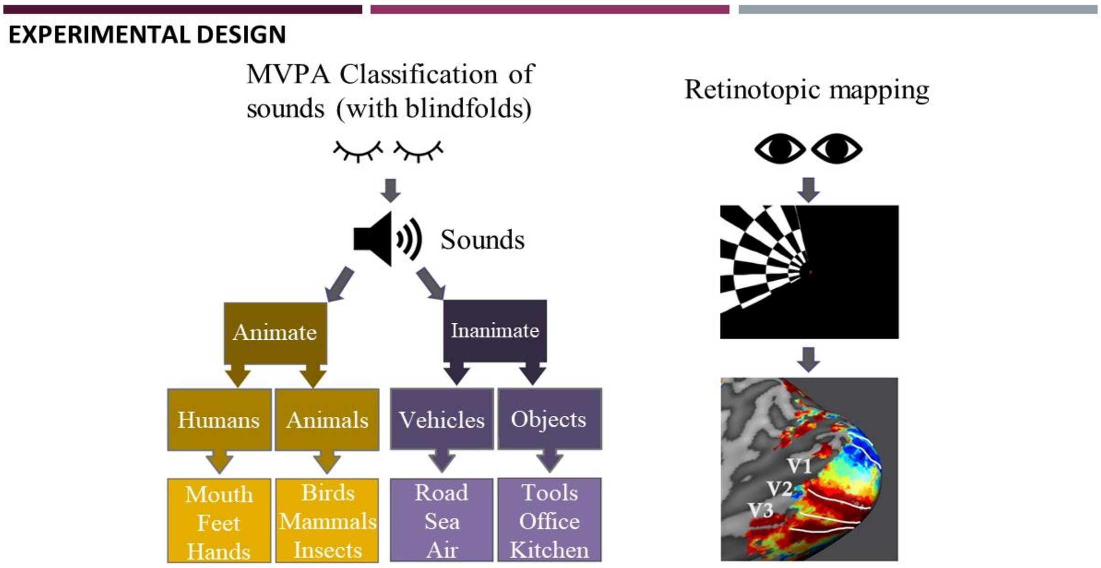


Figure 3.1: fMRI experimental procedure. Participants were blindfolded and attentively listened to 36 different natural sounds, subdivided into several semantic categories, while detecting an occasional target tone present in 10% of the trials. For retinotopic polar mapping of early visual cortex regions, participants watched a flickering rotating wedge and detected a colour change at the centre.

3.3.4: Data analyses

Functional MRI data were pre-processed with BrainVoyager 22.2 (BrainInnovation) with standard preprocessing steps (i.e., slice scan time correction, temporal high-pass filter, 3D rigid body motion correction), but without spatial smoothing in the sound stimulation runs to retain the small activity differences across voxels critical for MVPA. Regions of interest (ROIs) for V1, V2 and V3 were defined on individual reconstructed cortical surfaces using the activation gradients from the retinotopic polar mapping. V1, V2 and V3 were collated to create an additional ROI, comprising the whole early visual cortex (EVC). ROIs for auditory cortex were defined using the contrast all sounds > no sound stimulation (background MRI noise was still present). After pre-processing, two participants were excluded from analyses due to head movements and one participant for not recognising the majority of the sound stimuli.

Univariate whole-brain and ROI analyses were run on BrainVoyager 22.2 (BrainInnovation). Functional runs were combined and data were entered into a series of General Linear Models (GLM), where each sound category was modelled as a separate predictor (two for the animate-inanimate division and four for the intermediate 4-category division). A z-transform was performed on the resulting beta estimates.

MVPA analyses were run with custom-written MATLAB (version r2015b) scripts (adapted from Smith & Muckli, 2009; Vetter et al., 2014; Vetter et al., 2020, see https://github.com/Muckli-lab/MVP-analysis-tool-box). BrainVoyager data was handled through the NeuroElf toolbox (v 1.1) in MATLAB. We estimated beta weights for each sound event in all vertices of each ROI. These were fed into a linear support vector machine classification algorithm (LIBSVM toolbox, http://www.csie.ntu.edu.tw/~cjlin/libsvm). ROIs were combined across hemispheres to obtain higher statistical power. For each classification, the algorithm was trained on 3 runs and tested on 1, in a leave-one-out cross-validation procedure; the results were then averaged across the four instances. A mean classification accuracy score was computed for each individual participant and then results for each ROI were averaged across subjects. To assess statistical significance, we used a permutation analysis, as it is more robust than a t-test (Stelzer, Chen & Turner, 2013). For each subject and each ROI, the classifier was trained and tested 1000 times with randomised labels for our trials, thus we obtained distributions with 1000 classifier performance scores. Then, the p-value was derived from the classifier performance with the correct labels and the probability of getting a value as large as the correct labels' performance in our randomised labels distribution. To calculate the group p-value, the same process was applied, computing the group means of the randomised distribution and the group means of the correct label performance. With this method, we computed one group p-value per ROI, for each classification performed.

Whole-brain searchlight analyses were performed on the voxel level with the SearchMight toolbox (Pereira and Botvinick, 2011) in MATLAB, using a linear SVM (3-voxel radius). We assessed statistical significance by testing whether the mean accuracy across participants was significantly higher than chance at each voxel (i.e., 1/2 for the animate-inanimate categories and 1/4 for the four intermediate categories). The resulting maps were created using a voxelwise threshold of $p <.001$ and corrected for multiple comparisons with a cluster threshold correction ($p <.01$) calculated through the BrainVoyager Cluster Threshold Plugin tool (Goebel, Esposito and Formisano, 2006).

3.4: Results

We first tested whether natural sounds of the super-ordinate categorical level “animate” versus “inanimate” could be distinguished from activity patterns in early visual cortex. We successfully decoded all animate versus all inanimate sounds significantly above chance in V1 and V2 (Figure 3.2a; $p < .05$, as determined with permutation analyses).

Next, we decoded sounds belonging to the intermediate categories “humans”, “animals”, “vehicles” and “objects”. We found that the classifier was able to accurately predict intermediate sound categories in V2 and V3 (Figure 3.2b; $p < .050$, $p < .010$). Classification across all twelve subordinate categories was not successful in early visual cortex ($p \geq .115$), potentially due to classification being performed on the basis of 1/12 of the data given the 12 subcategories (Figure S2 in supplemental information).

Given the successful classification of sounds in early visual areas along the animate-inanimate division, as well as along the division of human – animal – vehicle – object, we also ran a pairwise MVPA analysis on each pair of the 4 categories separately (i.e., excluding data from the other pairs from the dataset, resulting in 50% of trials and statistical power compared to the animate-inanimate classification). The results showed significant above-chance classification accuracy for the human-animal pair in V1 ($p < .05$), the human-object pair in V2 and V3 ($p < .01$) and the human-vehicle pair in V2 and V3 ($p < .05$; Figure 3.2c).

In auditory cortex, we performed the same analyses resulting in successful classification accuracy across all sound categories and their pairwise combinations (p = .001, Figure 3.2c).

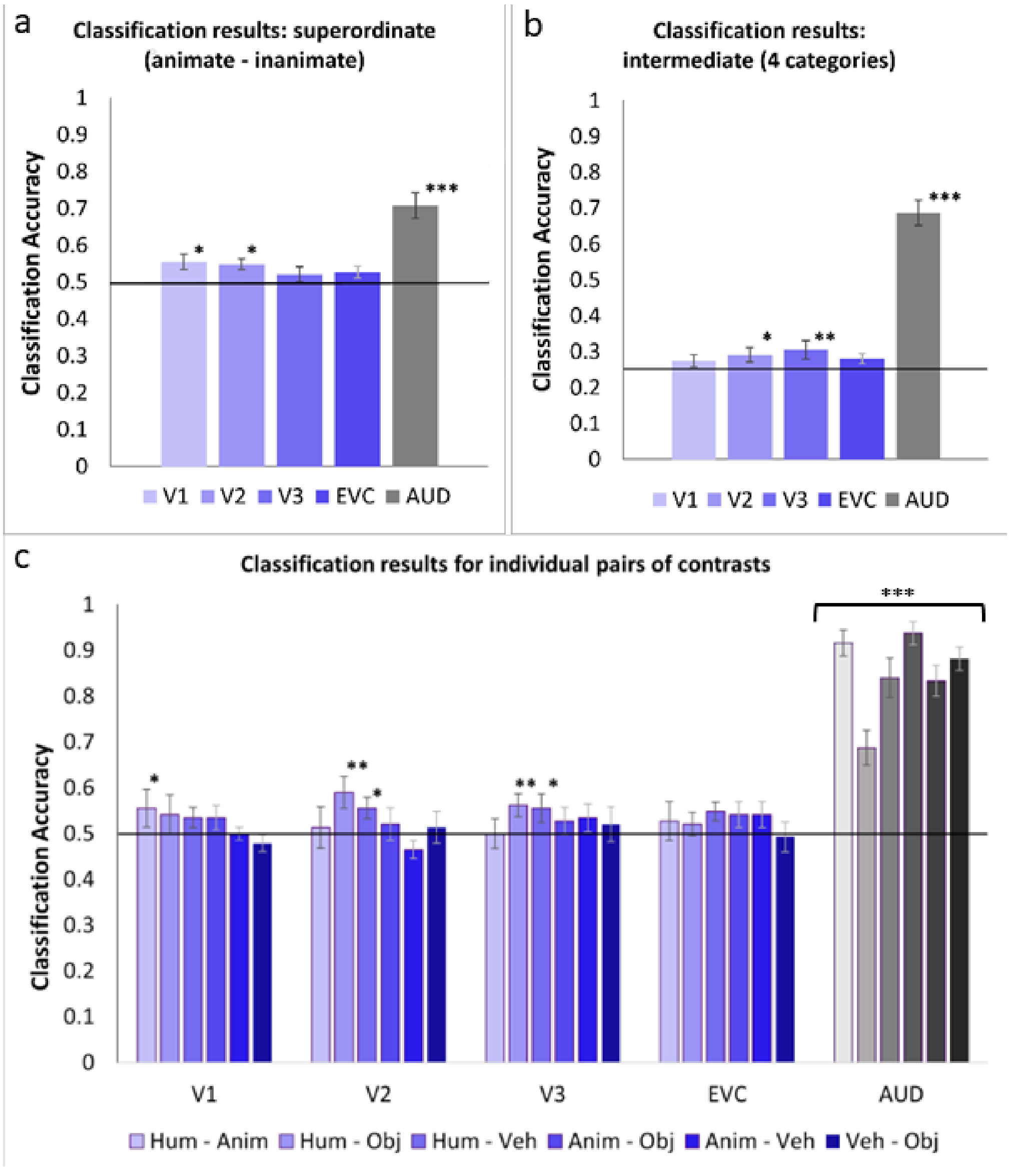


*Figure 3.2: classification results for a) animate and inanimate sounds; b) humans, animals, vehicles and objects sounds; c) individual pairs of contrasts. All p values testing above chance classification were derived from permutation analyses. The bracket indicates a main effect of classification accuracy across all classification pairs. Error bars indicate SEM. * p< .05; ** =p<.01; *** p= .001.*

Classification accuracy for all pairs of comparisons did not differ across early visual regions (repeated-measures ANOVA F(5, 85)= .874, p ≥.463), but differed in auditory cortex (F(1,17) = 1482.98, p < .001, Figure 3.2c). Post-hoc tests with Bonferroni correction revealed that the human-object classification accuracy was significantly lower than human-animal (p<.001), animal-object (p<.001) and vehicle-object (p =.034) classification.

To investigate whether other brain regions could discriminate the semantic sound categories, whole-brain searchlight analyses were performed on the voxel level using a linear SVM, implemented with the SearchMight toolbox (Pereira and Botvinick, 2011). This analysis revealed above-chance classification for animate versus inanimate sounds in the auditory cortex bilaterally, as well as in the middle temporal gyrus, right middle frontal gyrus, left fusiform gyrus and bilateral parahippocampal gyrus (Figure 3.3a). As for the four-way classification, auditory cortex predictably showed above-chance decoding bilaterally once again, as well as bilateral middle temporal gyrus (MTG) and left parahippocampal gyrus. Additionally, bilateral inferior frontal gyrus (IFG), bilateral posterior superior temporal sulcus (pSTS), right precuneus, left lateral occipital complex (LOC) and left transverse occipital sulcus (TOS) were also identified as regions with above-chance classification (Figure 3.3b).

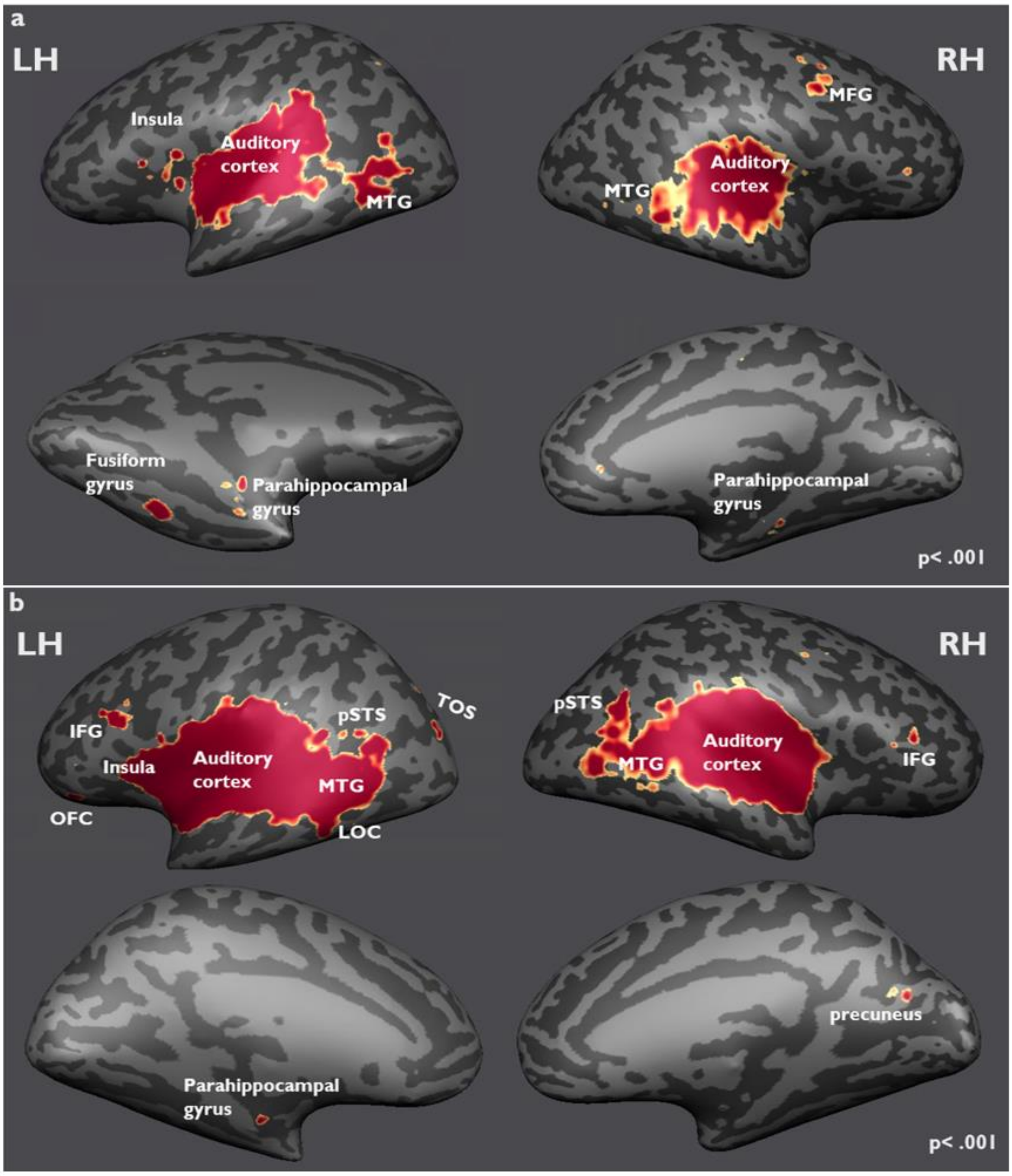


Figure 3.3: Results of the whole-brain searchlight analysis for a) the animate-inanimate classification and b) the human-animal-vehicle-object classification, voxelwise, cluster threshold corrected. Displayed on an inflated MNI template cortical surface reconstruction.

Univariate analyses:

We also performed a series of general linear model (GLM) analyses. In line with previous findings (Vetter et al., 2014, 2020), a whole-brain GLM with the contrasts inanimate>animate revealed that each semantic category of sound activated distinct regions. Animate sounds activated more lateral regions of the auditory cortex, whereas inanimate sounds activated regions more medial of the auditory cortex and a more widespread network including the left TOS, IFG, MFG and LOC (Figure 3.4).

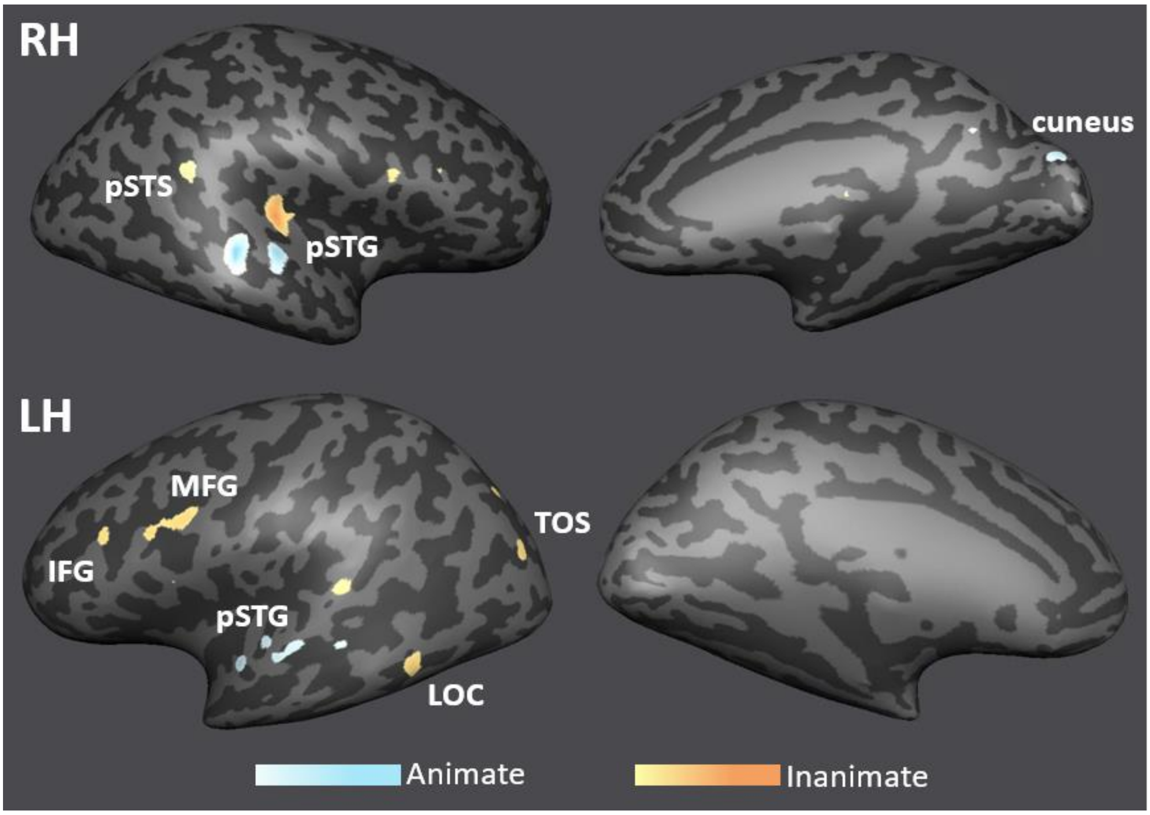


Figure 3.4: GLM analysis of inanimate>animate contrast projected onto an inflated MNI template cortical surface reconstruction with p=.01 (FDR corrected), where warm colours represent inanimate sounds and cold colours the animate category.

The contrasts comparing the categories “human”, “animal”, “vehicle” and “object” against baseline highlighted a pattern of activation across the brain quite similar for all categories, including the temporal regions around auditory cortex and fronto-parietal clusters (Figure 3.5b). Comparing activation of each individual category against the other three did not produce significant differences in early visual cortex, but revealed a distinct pattern across the temporal brain for each category, particularly noting the preference of left pSTG and right LOC for animals and a larger network of regions including bilateral IFG and MTG, and left MFG for objects (Figure 3.5a).

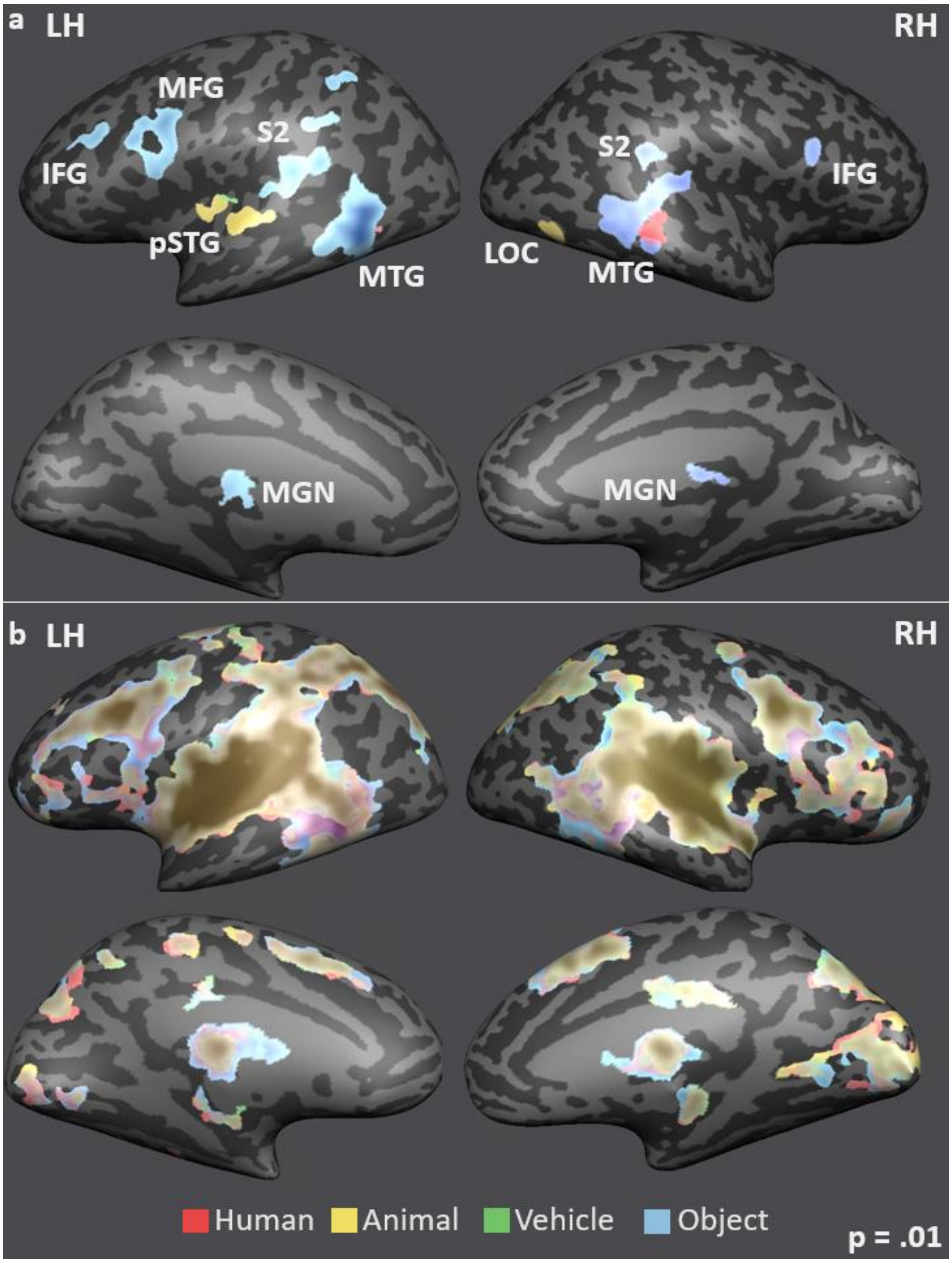


Figure 3.5: GLM analyses of a) each of the four categories (human, animal, vehicle and object) against the other three (colour-coded as per legend), and b) the same four categories against baseline. Both contrasts are projected onto an inflated MNI template cortical surface reconstruction with p=.01 (FDR corrected).

ROI analysis:

A GLM analysis conducted on our ROIs V1, V2 and V3 allowed us to specifically look at differences in univariate levels of activation in early visual cortex in response to the different sound categories. Beta values for each ROI were averaged across hemispheres for each of our sound category distinctions and then across participants (Figure 3.6).

Paired-samples t-tests on the two-way categorisation results revealed no significant difference between activity levels for animate and inanimate sounds, in any of our visual ROIs (t(17) <. 1.67, p >.113). When using the four-way categorisation as predictors, a repeated-measures ANOVA found a main effect of sound category in V3 (F(3,51)= 3.30, p=.027), which post hoc comparisons revealed being driven by the mean activation for animals being significantly higher than vehicles (p =.024, Bonferroni corrected). However, a one-sample t-test against zero showed that mean activation for animal sounds was not significantly different from baseline (t(17)= 1.93, p =.070). ANOVAs on V1 and V2 did not result in any significant differences between the means of our four categories (F(3,51) < 2.62, p > .061).

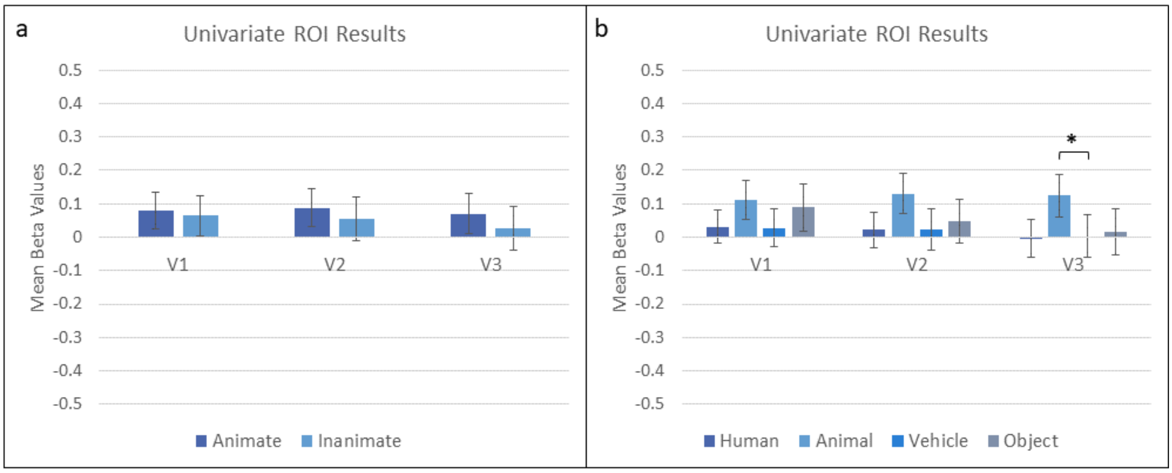


*Figure 3.6: Univariate ROI activity levels for the two-way (a) and four-way categorisation. Mean beta values are depicted for each sound condition in V1, V2 and V3, relative to baseline. Error bars indicate SEM. * indicates p < .05.*

Correlation analyses with Vividness of Imagery:

Scores from the VVIQ questionnaires were correlated with classification accuracy scores for all ROIs in early visual cortex and auditory cortex using Pearson's correlation coefficient. Higher VVIQ scores, corresponding to less vivid visual imagery, positively correlated with 4-categories classification accuracy in V3 (r(17) = .658, p = .003; FDR-corrected p =.045). This demonstrated that participants with less vivid visual imagery displayed more distinguishable activity patterns in V3 than participants with highly vivid visual imagery (Figure 3.7).

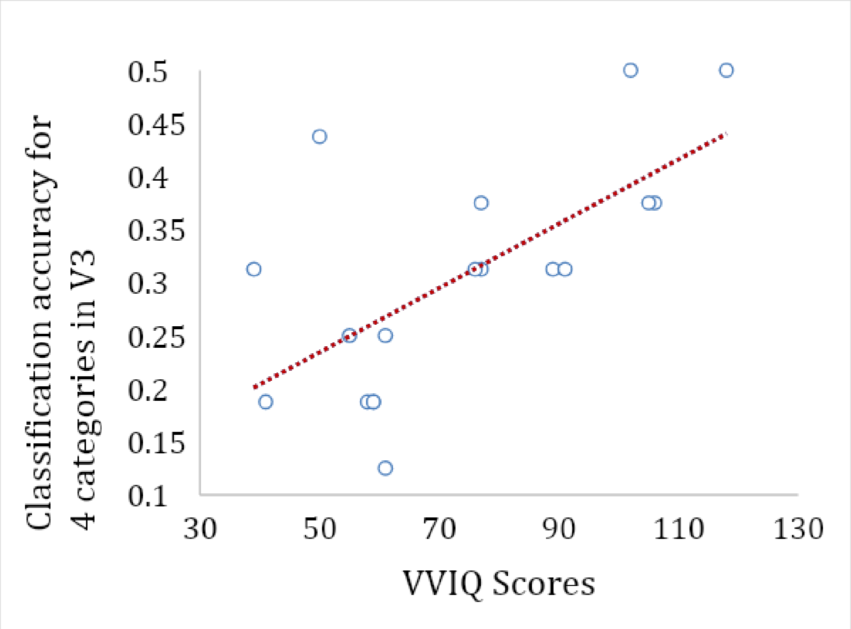


Figure 3.7: Scores from the Vividness of Visual Imagery Questionnaire plotted as a function of sound classification accuracy for the 4-category classification in V3. Note that high VVIQ scores denote less vivid visual imagery.

3.5: Discussion

The current study investigated which types of auditory semantic information content are represented in the neural activity patterns of early visual cortex. We found that natural sounds belonging to several semantic categories can be decoded from neural activity patterns in early visual cortex in blindfolded participants. Sounds belonging to superordinate semantic categories like ‘animate’ and ‘inanimate’ could be decoded in early visual cortex, replicating and extending our previous results (Vetter et al., 2014; Vetter et al., 2020) with a much larger variety of sound stimuli. In addition, we show here that sounds belonging to more specific categories such as humans, animals, vehicles and objects can also be decoded from activity patterns in early visual cortex. This latter novel finding suggests that the information transferred from auditory cortex to early visual cortex carries semantic content that goes beyond the animacy/inanimacy distinction, and in fact also follows the more fine-grained categorical distinction of human/animal/vehicle/object. To our knowledge, this is first evidence for fine-grained categorical sound distinction in early visual cortex in sighted participants, and it is nicely mirrored by sound decoding results in blind participants, both in early visual cortex and ventral temporal cortex (van der Hurk et al., 2017; Mattioni et al., 2020).

Importantly, our classifier was able to reliably distinguish between individual pairs of sound categories when the pair included the ‘human’ category. These results show that human sounds are represented in the visual brain, even down to early visual cortex, in a way that makes them distinguishable from others. This suggests that the brain treats stimuli from the human category more distinctly than other semantic categories, even when sensory information comes from audition and is represented at the earliest stages of visual processing. Again, our results from sighted people mirror findings in congenitally blind individuals showing that early visual cortex distinguishes meaningful from non-meaningful human speech stimuli (Musz et al., 2022).

As expected, the classifier performance in auditory cortex was well above chance. This was expected given that animate and inanimate sounds have been reported to activate distinct areas of auditory cortex even using less sensitive univariate analyses (Engel et al., 2009). However, the classifier performance for the human-object classification was significantly worse than for the other pairs. A plausible explanation could be that the samples we chose for object sounds mostly came from human-produced actions (e.g., spoon stirring in a teacup, hammering, etc.), which could have led to similar representations in auditory cortex as those elicited by the other human sounds (e.g., hand clapping, footsteps), resulting in less distinct representations and decreased classification accuracy.

Why are human-made stimuli more distinctly represented in the brain than other categories? It is possible that the brain devotes a lot of computational power to interpreting and representing human stimuli due to their high relevance for social interaction. Examples come from literature on face and speech perception which indicate how, for humans, conspecifics are better and preferentially detected than other living beings (Pascalis, De haan & Nelson, 2002; Vouloumanos et al., 2010). Humans are able to extract a variety of information during exposure to human faces and voices, ranging from identity to personality, to emotion and mental state (Oruc, Balas & Landy, 2019; Maguinness, Roswandowitz & von Kriegstein, 2018; McAleer, Todorov & Belin, 2014). For these reasons, the current findings on the ability of early visual cortex to distinguish human-made sounds fit nicely with the literature that supports this tuning for conspecifics in the human brain. One potential explanation for our results is that the brain prioritises the information about other humans in the environment and sends this information via feedback to several brain areas, including early visual cortex, such that it is optimally prepared to precisely identify and predict human stimuli most relevant for social interaction.

The distinctiveness of neural representation of human sounds in early visual cortex is unlikely to be driven by a similarity of the acoustic features of our human sound samples, since the human category contained several sound exemplars of walking, clapping, coughing and talking that contain a wide range of acoustic frequency patterns. Likewise, the sounds from the other categories were equally diverse in terms of acoustic features (see Figure S1 of supplemental information for the full list of sounds employed). Thus, our classification results are more likely to have been driven by the information content that is shared between different sound exemplars of one category, and by the information content that is distinct from that of the other categories. For the same reason, our results are unlikely to be driven by neural activity potentially evoked by sound-induced visual imagery – the potential visual images evoked by the sounds would differ substantially in visual features across exemplars within a category (e.g., finger snapping and footsteps; horse galloping and seagull). Again, even if sounds elicited visual imagery, the shared information content of all stimuli within one category must have driven the distinction of activity patterns from those evoked by the other categories across all their stimulus exemplars. The univariate whole-brain results revealed patterns of activation consistent with the activity found in other studies investigating differences in the brain areas recruited for the processing of different sound categories. Particularly the widespread activation we found across the left hemisphere in response to the object category, comprising the MFG, IFG and MTG, is similar to the 'mirror network' encountered by Lewis and colleagues and active for tools over animal sounds (Lewis et al., 2005). Although the univariate contrasts sound stimulation > baseline did not reveal any significant positive (or negative) activation difference in early visual cortex, ROI analyses revealed very weak positive activation for animal sounds to be higher than vehicles in V3. While this might have partly contributed to the significant classifier performance in the four-way classification, it cannot fully explain our results, particularly those in the pairwise comparisons not involving animals, or those in other ROIs. These results make it overall unlikely that sounds evoked differential visual activation through visual imagery or attention. Furthermore, we demonstrated in a previous study that sounds along the categories human-animal-inanimate can be distinguished in early visual cortex in participants who are blind from birth and lack visual imagery (Vetter et al., 2020). The correlation of VVIQ scores with classifier performance for the four-way classification fits in this framework revealing that participants with less vivid visual imagery had higher classification accuracy in V3, which suggests that vivid imagery during sound presentation might have interfered with sound classification in early visual cortex, or at least did not boost classifier performance.

Our findings from the whole-brain searchlight analysis revealed that a network of regions beyond auditory cortex is sensitive to the categorical distinction between animate and inanimate sounds and human, animal, vehicle and object-made sounds. Above chance performance for the four-way classification in regions implicated in multisensory integration such as the precuneus (Park and Kayser, 2019, Ripp et al., 2018), and particularly audiovisual integration, such as the pSTS (Naumer et al., 2011, Beauchamp et al., 2004), suggests that category-specific information about sounds is also represented in multisensory regions. These might serve as mediators between primary sensory areas, relaying information from auditory to visual cortex. Both searchlight analyses revealed above chance classification in the insula which is often active during auditory stimulation (Uddin et al. 2018), but more importantly, was found to be active during experience of perceptually related audiovisual stimuli (Naghavi et al., 2007).

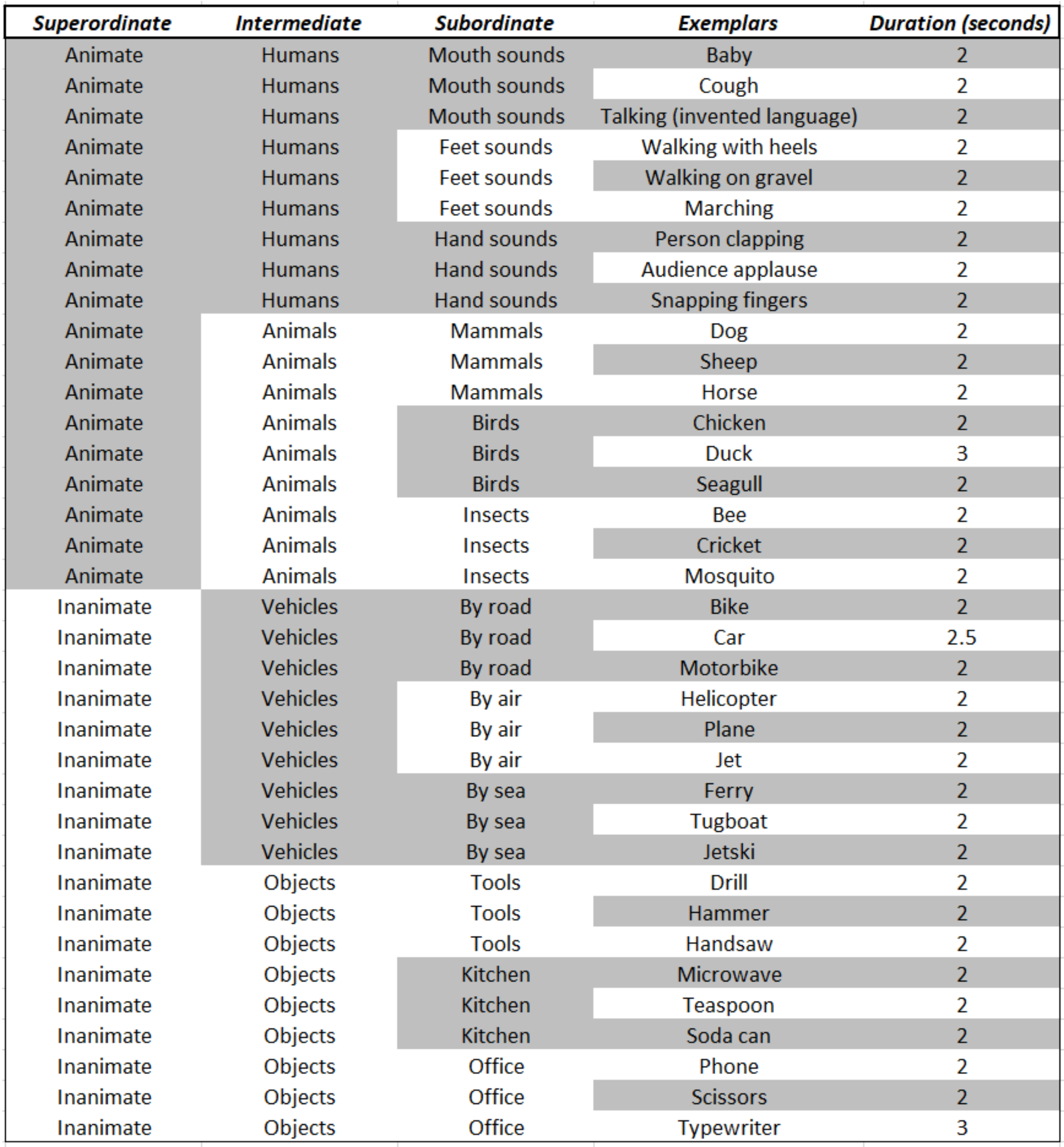
Animate and inanimate sound decoding was also successful in fusiform gyrus and parahippocampal gyrus, regions of the ventrotemporal cortex involved in higher-level visual processing. Fusiform gyrus plays a crucial role in face recognition (Kanwisher, McDermott, & Chun, 1997; Ganel et al., 2005), while the parahippocampal gyrus is active during scene recognition (Epstein and Kanwisher, 1998). Their domain selectivity has been reported to exist beyond visual stimulation, as the FFA and PPA of congenitally blind people has been shown to robustly respond to haptic exploration of 3D faces (Ratan Murty et al., 2020) and scenes (Wolbers et al., 2011). This, coupled with our findings from audition, suggest that these regions may integrate information from different sensory modalities for the recognition of human stimuli.

Interestingly, LOC was also able to significantly distinguish between sound categories in our four-way classification. LOC is involved in object recognition (Grill-Spector et al., 1999) and has been found to be implicated in audiovisual integration of objects (Giovanelli et al., 2016) and to be active when objects are recognised not only visually, but haptically (Amedi et al., 2001) and during visual-to-auditory sensory substitution (Amedi et al., 2007). While these earlier studies using standard univariate analyses failed to show LOC activity with just auditory stimulation, we show here with more sensitive MVPA that LOC can in fact distinguish between auditory objects and categories.

What is the reason for auditory cortex feeding category-specific information back to early visual cortex? Both touch and sound have been shown to enhance activity in early visual cortex (Macaluso et al., 2000; Ramos-Estebanez et al., 2007; Martuzzi et al., 2007), so a potential way in which the early visual cortex employs auditory feedback is to predict visual information and improve visual perception. This is in line with theories of predictive coding (Friston 2010, Clark, 2013), which suggests that the purpose of feedback information arriving in EVC could be to provide categorical expectations that help the visual system be prepared for the incoming visual stimuli. This in turn might facilitate their recognition, prompt faster reactions and faster binding with other sensory information. Particularly, auditory information could be used to resolve ambiguities in vision (e.g., Chen & Spence, 2011), and when multisensory integration is critical to producing stable perceptual experiences (Spence and Squire, 2003).

To conclude, our results provide further evidence that early visual cortex function is not restricted to the processing of low-level visual features from retinal input, but also for the representation and potential employment of higher-level features from other sensory modalities. Early visual cortex might be more multisensory in nature than previously thought, which calls for more studies exploring this concept for other primary sensory regions as well.

3.6: Supplemental information:



Superordinate	Intermediate	Subordinate	Exemplars	Duration (seconds)
Animate	Humans	Mouth sounds	Baby	2
Animate	Humans	Mouth sounds	Cough	2
Animate	Humans	Mouth sounds	Talking (invented language)	2
Animate	Humans	Feet sounds	Walking with heels	2
Animate	Humans	Feet sounds	Walking on gravel	2
Animate	Humans	Feet sounds	Marching	2
Animate	Humans	Hand sounds	Person clapping	2
Animate	Humans	Hand sounds	Audience applause	2
Animate	Humans	Hand sounds	Snapping fingers	2
Animate	Animals	Mammals	Dog	2
Animate	Animals	Mammals	Sheep	2
Animate	Animals	Mammals	Horse	2
Animate	Animals	Birds	Chicken	2
Animate	Animals	Birds	Duck	3
Animate	Animals	Birds	Seagull	2
Animate	Animals	Insects	Bee	2
Animate	Animals	Insects	Cricket	2
Animate	Animals	Insects	Mosquito	2
Inanimate	Vehicles	By road	Bike	2
Inanimate	Vehicles	By road	Car	2.5
Inanimate	Vehicles	By road	Motorbike	2
Inanimate	Vehicles	By air	Helicopter	2
Inanimate	Vehicles	By air	Plane	2
Inanimate	Vehicles	By air	Jet	2
Inanimate	Vehicles	By sea	Ferry	2
Inanimate	Vehicles	By sea	Tugboat	2
Inanimate	Vehicles	By sea	Jetski	2
Inanimate	Objects	Tools	Drill	2
Inanimate	Objects	Tools	Hammer	2
Inanimate	Objects	Tools	Handsaw	2
Inanimate	Objects	Kitchen	Microwave	2
Inanimate	Objects	Kitchen	Teaspoon	2
Inanimate	Objects	Kitchen	Soda can	2
Inanimate	Objects	Office	Phone	2
Inanimate	Objects	Office	Scissors	2
Inanimate	Objects	Office	Typewriter	3

Figure S1: Full list of sound stimuli divided by category, including their duration in seconds.

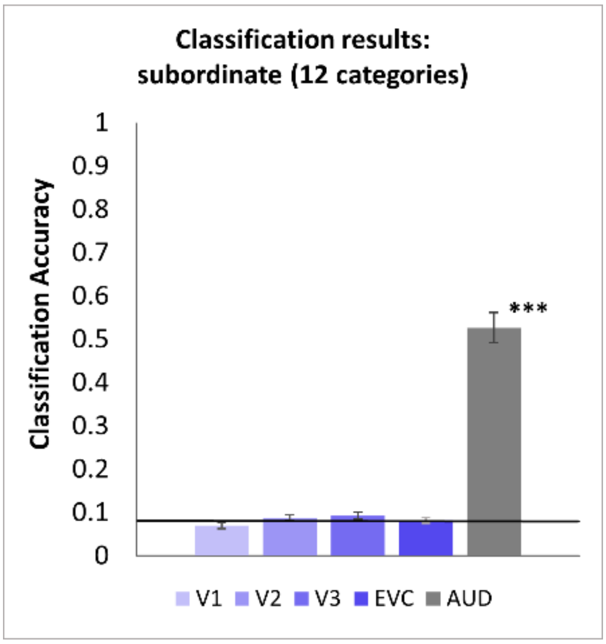


*Figure S2: classification results for the subordinate groups of sounds (mouth, hands, feet, mammals, birds, insects, road vehicles, air vehicles, sea vehicles, kitchen utensils, office objects, tools.). Error bars indicate SEM. * p< .05; ** =p<.01; *** p= .001.*

VVIQ questionnaire:

For each scenario try to form a mental picture of the people, objects, or setting. Consider carefully the vividness of your visual imagery experience. Does some type of image come to mind? Rate how vivid the image is using the 5-point scale. The rating scale is as follows:

1. Perfectly realistic, as vivid as real seeing
2. Realistic and reasonably vivid
3. Moderately realistic and vivid
4. Dim and vague image
5. No image at all, I only “know” I am thinking of the object

The questions were answered two times, one with eyes open and the other with eyes closed.

For items 1-4, think of some relative or friend whom you frequently see (but who is not with you at present) and consider carefully the picture that comes before your mind’s eye.

1. The exact contour of face, head, shoulders and body ________________

2. Characteristic poses of head, attitudes of body etc. ________________

3. The precise carriage, length of step etc., in walking ________________

4. The different colours worn in some familiar clothes ________________

Visualise a rising sun. Consider carefully the picture that comes before your mind's eye.

5. The sun rising above the horizon into a hazy sky ________________

6. The sky clears and surrounds the sun with blueness ________________

7. Clouds. A storm blows up with flashes of lightning ________________

8. A rainbow appears ________________

Think of the front of a shop which you often go to. Consider the picture that comes before your mind's eye.

9. The overall appearance of the shop from the opposite side
of the road ________________

10. A window display including colours, shapes and details
Of individual items for sale ________________

11. You are near the entrance. The colour, shape and
details of the door. ________________

12. You enter the shop and go to the counter. The counter
assistant serves you. Money changes hands ________________

Finally think of a country scene which involves trees, mountains and a lake. Consider the picture that comes before your mind's eye. ________________

13. The contours of the landscape ________________

14. The colour and shape of the trees ________________

15. the colour and shape of the lake ________________

16. A strong wind blows on the trees and on the lake causing
waves in the water. ________________

Chapter 4: Experiment 2

Title: Semantically congruent sounds increase perceptual dominance times of human images during binocular rivalry

4.1: Abstract

During binocular rivalry, two different images are presented to each eye at the same spatial location, and observers perceive a fluctuation between the two. It has been shown that binocular rivalry can be modulated by stimuli from other sensory modalities, and particularly that sounds can facilitate perception of semantically congruent images (Chen, Yeh and Spence, 2011). Neural activity in early visual cortex is correlated with conscious perception of binocular rivalry (Lee and Blake, 2002; Haynes et al., 2005), and our previous fMRI experiment (Chapter 3) revealed that early visual cortex receives information about human sounds, representing these better than other sounds. This prompted us to focus on the audio-visual integration of human-related stimuli. To investigate the degree of relatedness necessary to promote conscious perception during binocular rivalry, we presented participants with human images (e.g., people, feet or hands) paired with either vehicle or animal images, while they listened to congruent sounds, 'relevant' sounds (which were human-related but not congruent with the image), incongruent sounds, or silence. Our results show that human sounds increased the mean perceptual dominance times of human visual stimuli, but only when these were specifically congruent to the image, as hearing relevant (but not entirely congruent) sounds of a human nature did not influence perceptual dominance of human images. This effect was not found for sounds congruent with animal or vehicle images, whose dominance times did not differ from the no sound condition. These results support the idea that semantically congruent sounds can promote disambiguation of ambiguous visual percepts, particularly when these are human and strictly congruent.

4.2: Introduction

When two different images of equal salience are displayed to each eye, visual perception fluctuates between the two images in a phenomenon called binocular rivalry (BR) (Crick and Koch, 1990; Crick, 1996). This can be a useful technique when investigating visual awareness and the mechanisms that modulate visual perception, because the brain receives full visual information from both eyes, but only one image at a time dominates visual awareness.

A large body of evidence supports the idea that BR can be affected by cues from other sensory modalities such as audition (Kang and Blake, 2005), olfaction (Zhou et al., 2010), touch (Holcombe and Seizova-Cajic, 2008; Lunghi, Binda and Morrone, 2010; Lunghi and Morrone, 2013), as well as combinations of multiple senses (Lunghi, Morrone & Alais, 2014). In the case of audio-visual effects, modulation of BR is more likely when the stimuli are congruent, e.g., when they represent the same direction of movement (Conrad et al., 2010) or similar emotions (Jertberg, Levitan and Sherman, 2018). When the ambiguous visual stimulus is musical notation, the playing of music facilitates perception of the congruent notes, but only if participants are musically literate (Lee et al., 2015). The association between audio-visual stimuli does not have to be well established in order to have effects on BR, as even rapidly learned associations between a sound and a visual object can effectively modulate dominance during BR, i.e., associations between gratings and a low or high-pitch tone (Einhäuser, Methfessel and Bendixen, 2017; Piazza, Denison and Silver, 2018). However, semantically congruent sounds and visual objects implicitly have that association, reinforced by everyday experiences, and the question of under which degree of relatedness sounds can influence vision in a BR task is one that has yet to be answered. Chen, Yeh and Spence (2011) looked at whether hearing the sound of a car, a bird or an irrelevant sound (tableware recording at a restaurant) could modulate perception of car and bird line drawings during BR. They found that incongruent sounds to a target image (but congruent to the other image) inhibit its perception, indirectly concluding that congruent sounds can increase perception time of the congruent image. The irrelevant sound had no effect, suggesting semantic context is important and that a general naturalistic sound is not enough to modulate perception during BR. In a series of follow-up studies using a visual masking task, Chen and Spence (2018) demonstrated how presenting sounds related to a visual stimulus (i.e. belonging to the same subordinate category) before visual stimulation, can speed up image discrimination more than an irrelevant sound but not as much as a fully congruent sound. However, the categorical specificity of the relation between auditory and visual stimuli has not been fully investigated, especially in the context of ambiguous stimuli.

The current study used natural sounds and images in a binocular rivalry set-up to investigate the specificity of auditory influence on the visual perception of ambiguous stimuli.

Specifically, we investigated to what extent sounds that are related in semantic meaning to a visual stimulus can affect its dominance time in a BR task. We presented 25 healthy human participants with pairs of images viewed through a stereoscope to induce binocular rivalry. One of the images always depicted human stimuli (e.g., a baby, hands clapping, etc.), the other image contained either animals or vehicles, and we simultaneously played natural sounds that varied with respect to their congruency with the visual images. The decision to explore audiovisual interactions on these three semantic categories and to focus our analyses around the human stimuli was driven by results from Experiment 1. In it, we showed how early visual cortex receives feedback from auditory cortex containing information about the semantic nature of sounds, making it possible to decode whether participants were listening to humans, animals, vehicles or objects from the activity pattern in the early visual cortex alone (Experiment 1 in Chapter 2). Particularly, we found human sounds to be better decoded, compared to the other categories, which led us to hypothesise that the influence of auditory feedback on vision might be more pronounced for human-related stimuli.

During the experiment, each sound was either congruent to the human image (e.g., the sound of a baby gurgling paired with the image of a baby), ‘related’, by which we mean belonging to the human category but not congruent with the image (i.e., the sound of a woman coughing paired with the image of a baby), or fully incongruent (i.e., either animal or vehicle sounds paired with a human image). We compared how the total length of conscious image perception was modulated by the different semantic categories of sounds by measuring the percentage of time participants spent perceiving each stimulus, as well as measuring whether sounds increased congruent image perception when compared with a no-sound condition. We hypothesised that human sounds would facilitate the perception of congruent human images, but we were also interested in seeing whether human sounds related but not fully congruent to human images would modulate perception. Stimuli that are more naturalistic in composition or more complex, seem to produce stronger modulations of BR (Conrad et al., 2013), which is also supported by experiments done with other paradigms (i.e., continuous flash suppression, Tan and Yeh, 2015). For this reason, we decided to use photographs rather than simple drawings for the visual stimuli.

4.3: Methods

4.3.1: Participants

Twenty-five subjects (seven males, mean age 24.57) participated in the experiment, all with normal or corrected-to-normal vision and hearing. They all gave informed consent and were paid for their time. Due to the exclusion criteria outlined in the following sections, four subjects were excluded from the analyses, leaving a total of twenty-one participants. The study was approved by the ethics committee of Royal Holloway, University of London.

4.3.2: Apparatus and Stimuli

The experiment was conducted in a dark and quiet sound-attenuated booth. Visual stimuli were sourced from online databases of stock images (www.unsplash.com; www.gettyimages.co.uk), while the audio samples were adapted from Experiment 1. The experiment was presented using PsychoPy (Pierce et al., 2019) on a 22-inch LCD monitor (Samsung syncmaster 2243nw) at a resolution of 1680x1050 pixels with 60 Hz refresh rate. Participants looked at the screen through a mirror-based stereoscope at a distance of 65 cm, resting their heads on a chin rest, and listened to the auditory stimuli through headphones (Sennheiser Hd201). Their responses were provided and recorded via keys pressed on a keyboard.

For the auditory stimuli, we selected 10 sounds from Experiment 1, two for each of the following categories: mouth, feet, hands (all human, 2 of each sub-category), animals (sheep and chicken) and vehicles (car and helicopter). They were modified to be 45 seconds long to cover the duration of an entire trial, either by using longer recordings or repeating shorter sound samples. As for the visual stimuli, we selected from online databases 10 images semantically congruent with the sounds (Figure 4.1). The background of each image was modified to remove distracting features and the final images were converted into grayscale and normalised for luminance and contrast using the SHINE toolbox (Willenbockel et al., 2010). After a first run of piloting, subjects reported difficulties in seeing the perceptual switch from one image to the other with certain images (e.g., the car and the baby), so the images’ contrast was further reduced manually and consistently for the problematic stimuli.

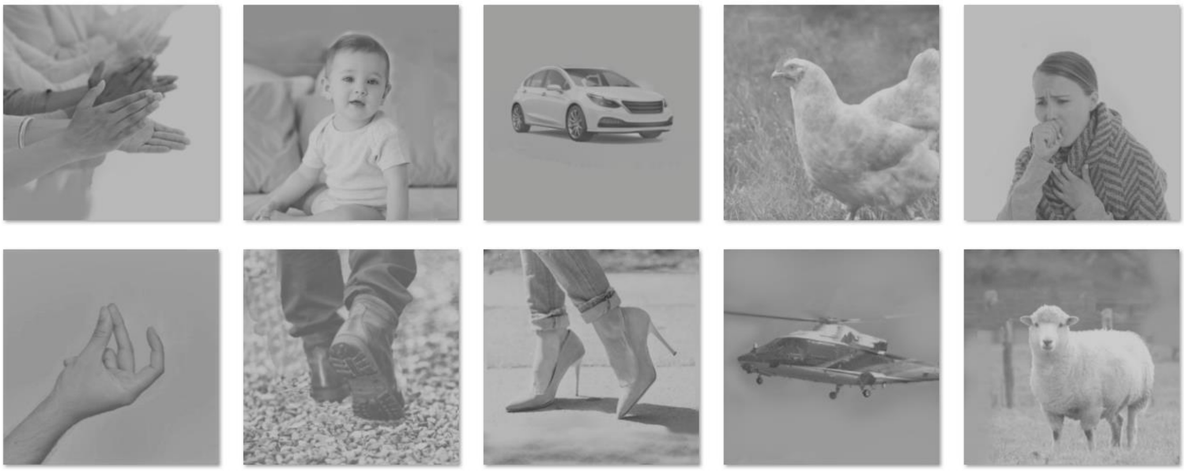
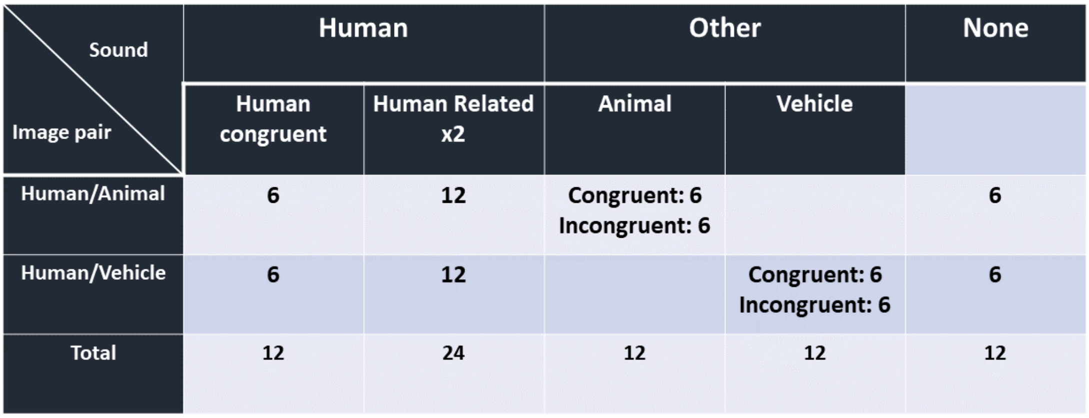


Figure 4.1: The set of visual stimuli used in the experiment.

4.3.3: Experimental Design

We created a total of 72 trials by combining a human image, a non-human image, and a sound, keeping the number of times each image appeared on the left or right-hand side of the screen the same (Figure 4.2). Trials had to be long enough to give participants time to switch perception several times (the trial length generally used in BR experiments is between 30 and 90 seconds e.g., Tong et al., 1998; Zhang et al, 2011, Chen, Yeh and Spence, 2011). We opted for trials 45 seconds long due to the large number of stimuli in our experimental design. Human images were repeated 11-13 times, to maintain equal numbers of trials in each category, but note that the non-human images were repeated more often, as they were 4 in total to accompany the 6 human images. As we were particularly interested in investigating the influence of sounds on human images, we assigned each trial a sound that was either congruent to the human image, related to the human image (still belonging to the human category but not entirely congruent), or completely incongruent (i.e., animal or vehicle). We also included trials where pairs of images were accompanied by silence, to use as a baseline and to investigate whether some images dominate perception due to their visual features only, independent of sound. Each sound was repeated 6 times, with the exception of the baby' sound (5) and the 'heels' sound (7), to maintain the balance of trials while keeping into account how often each image would appear on either side of the screen. The human sounds were divided into 2 repetitions for congruent trials and 4 repetitions for relevant trials. This resulted in 8 blocks of 9 trials each, which we divided over two separate sessions, so as not to pose too much strain on participants' eyes. Note that, for the 'related to human' trials, the total number of trials is double the amount of the human congruent trials. This is due to the human stimuli being three different types instead of two (e.g., mouth, feet and hands), so the pool of related sounds had more stimuli. During the experiment, the order of blocks was counterbalanced across participants as well as which button was assigned to which category, and the order of trials randomised within each block.



Sound / Image pair	Human		Other		None
	Human congruent	Human Related x2	Animal	Vehicle	
Human/Animal	6	12	Congruent: 6 Incongruent: 6		6
Human/Vehicle	6	12		Congruent: 6 Incongruent: 6	6
Total	12	24	12	12	12

Figure 4.2: Table depicting the distribution of trials according to which auditory and visual stimuli were presented.

4.3.4: Procedure

Before each session, participants became acquainted with the image set (Figure 1 was used to familiarise them with the stimuli) and with the binocular rivalry paradigm: they were told that they would be seeing pairs of images while listening to sounds and that their perception would fluctuate from one image to the other. They were asked to press two buttons on the keyboard depending on what they were seeing at the moment (Figure 4.3), or to press the spacebar if they saw a mixture of percepts (which we term 'piecemeal') rather than a clear fluctuation. Each session started with the adjustment of the mirror stereoscope and chin rest. Participants had then the opportunity to get acquainted with the stimuli, ask questions if the task was unclear, and have a practice trial followed by a break where they could either confirm they understood or practise more if needed. After the experiment, participants were all debriefed at the end of the second session.

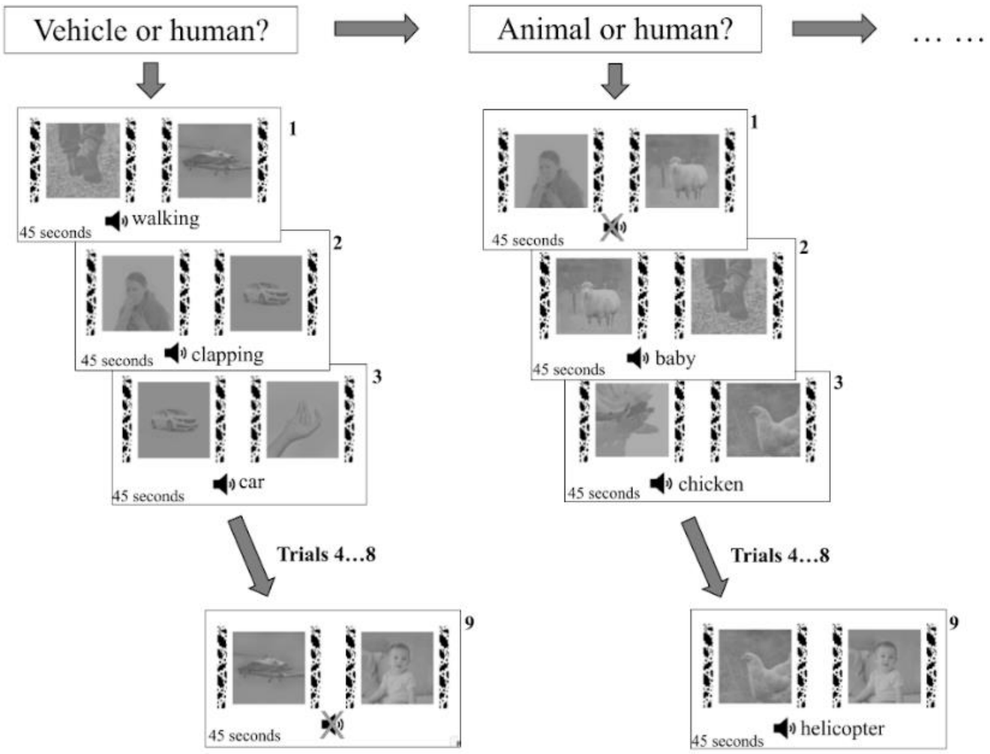


Figure 4.3: Experimental procedure. It consisted of four experimental conditions: sounds congruent to one of the human images, sounds related but not entirely congruent to the human image, incongruent sounds, and silent trials without any sound. Each trial lasted 45 seconds.

4.4: Results

Out of 25 participants tested, four were removed from analyses. Three could not see the perceptual switch in more than 50% of the trials, while the fourth had more than 50% piecemeal perception in more than half of the trials (i.e., a mixture of the two images shown on the screen). In the remaining participants, trials with more than 50% of piecemeal perception were discarded (7 trials across all participants), as well as trials where one of the two visual objects was perceived 0% of the time (15 trials across all participants).

For each subject, we obtained a sequence of button presses for each of the 72 trials they completed. Raw data of the length of each button press was transformed into proportions of time by taking the sum of their durations and dividing it by the total length of the trial. For each trial we obtained a) the proportion of time the left stimulus was dominant, b) the proportion of time the right stimulus was dominant and c) the proportion of time participants perceived a combination of both images.

For the initial analyses, we coded our 72 trials according to whether the sound was congruent to the human image, relevant to the human image (i.e., incongruent but still belonging to the human category), incongruent (i.e., belonging to either animal or vehicle category), or whether no sound was played.

The mean proportion of time either the human or the other picture was dominant across the four experimental conditions is plotted in Figure 4.4. A repeated-measures ANOVA revealed a main effect of experimental condition over the dominance times of human images (F(1.59,60) = 4.07, p = .035, Greenhouse-Geisser corrected), which post-hoc paired sample t-tests showed to be due to the human image being perceived for significantly longer during the human congruent sounds, compared to the other three categories (p = .031 for relevant sounds, p = .041 for incongruent sounds, p = .040 for no sound). The mean proportion of time the other image (animal or vehicle) was dominant was not significantly different across the four sound categories (F(1.70,60) = 3.20, p = .061). These results suggest that congruent human sounds can lengthen the conscious visual perception of human stimuli.

Additionally, we wanted to test whether there was a difference between the amount of time participants perceived the *human* and the *other* images across all experimental conditions. Paired-sample t-tests conducted between the averaged dominance times for each experimental condition revealed that there was no significant difference between the dominance times of *human* and the *other* images (t(20) = 0.04, p = .970 for relevant human sounds, t(20) = -0.32, p = .750 for incongruent ‘other’ sounds, t(20) = -0.03, p = .974 for no sound condition), except for the congruent trials, in which the human images were perceived for significantly longer than the other images when a congruent sound was played (t(20) = 2.30, p = .033, Figure 4.4).

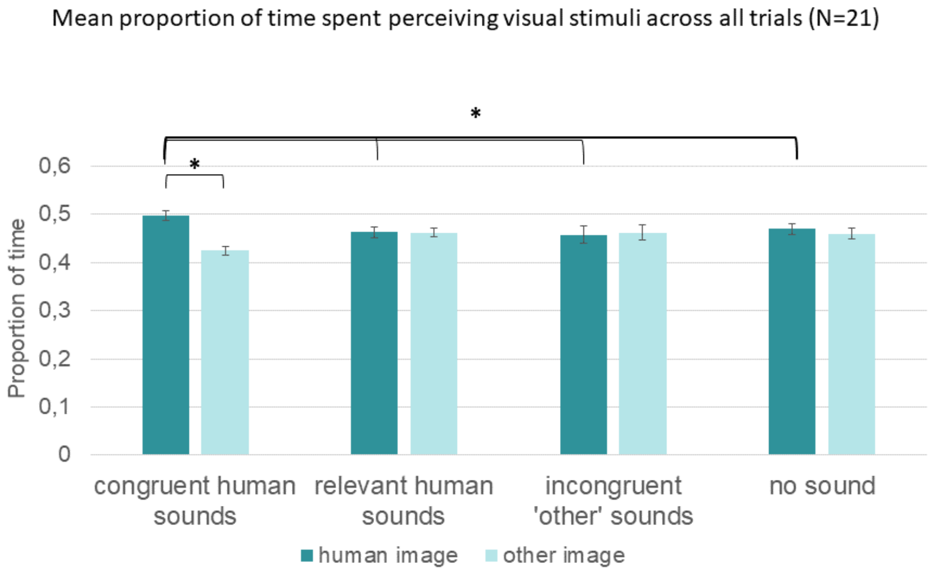


Figure 4.4: Bar plot showing the mean proportions of time spent perceiving the 'human' and 'other' image, grouped by experimental condition: congruent sounds (12 trials per participant), relevant human sounds (2x12 = 24 trials), incongruent other sounds (24), and no sound (12). P-values were derived from t-tests and repeated measures ANOVA and indicate significant differences at p < .050. Error bars indicate SEM.

However, it is possible that participants varied in the amount of time they spent perceiving each picture independent of which sound was displayed. To account for these differences both across individual visual images and across individual participants, we used the proportion of time each participant saw *human* and *other* during the no sound condition as baseline, and subtracted it from the respective proportions of time they perceived the same image in the other three experimental conditions that included sounds. We refer to this measure as absolute difference in dominance time (sound condition - no sound condition). This computation normalised dominance times for any possible dominance differences that are due to the visual stimulus alone, for each participant and each image combination separately. One-sample t-tests comparing the absolute differences against 0 (where 0 is the baseline, no sound condition) revealed a significant difference only in the case of human stimuli during the perception of congruent human sounds (t(20) = 2.20, p = .040, uncorrected), supporting the results previously reported in the previous section (Figure 4.5). All other experimental conditions did not show any significant deviations from 0 (t(20) < .61, P > .087).

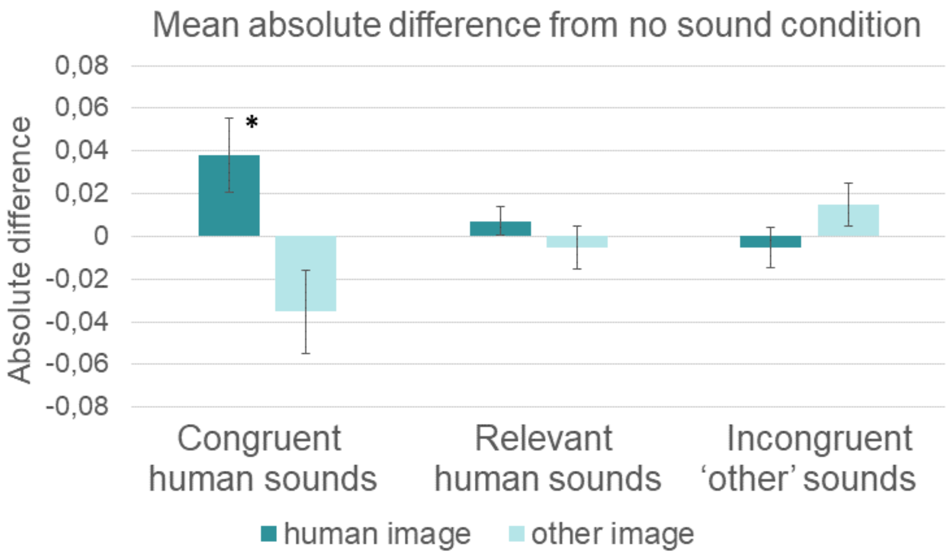


Figure 4.5: Bar plot showing the mean absolute difference in dominance times of human and other images, calculated by subtracting proportions of time of the no sound trials from trials with congruent sound, relevant human sounds and incongruent other sounds. Dark bars represent the human images' score while the lighter ones represent the other images (vehicle or animal). P-values were derived from t-tests and indicate significant differences from 0 at p < .050. Error bars indicate SEM.

In order to examine whether the effect was unique to the human category, we re-coded all the trials that had sounds congruent to either the vehicle or the animal picture as *congruent animal* (six trials per participant) and *congruent vehicle* (six trials per participant). Please note that while these are small trial counts, each trial contains several instances of perception switching from one image to the other over 45 seconds. For the purpose of this analysis, we also split the *other* images in the human congruent trials into animals and vehicles.

First, we performed paired-sample t-tests on the congruent animal and congruent vehicle trials, which revealed that vehicle images were dominant significantly longer than human images in trials with vehicle sounds (t(20) = 3.95, p < .001), while trials with animal sounds showed the opposite pattern, with human images being dominant for longer during animal sounds presentation (t(20) = -3.59, p = .002, Figure 4.6).

Then, we wanted to see whether splitting the *other* images of the human congruent trials into animals and vehicles showed any differences compared to human congruent images (depicted in the first group of columns in Figure 4.6). Paired-sample t-tests revealed that the proportion of time human images are dominant is not significantly different from that of the vehicle images (t(20) = 1.70, p = .105), but it is different from the animal ones (t(20) = 3.18, p = .005) when human sounds are displayed.

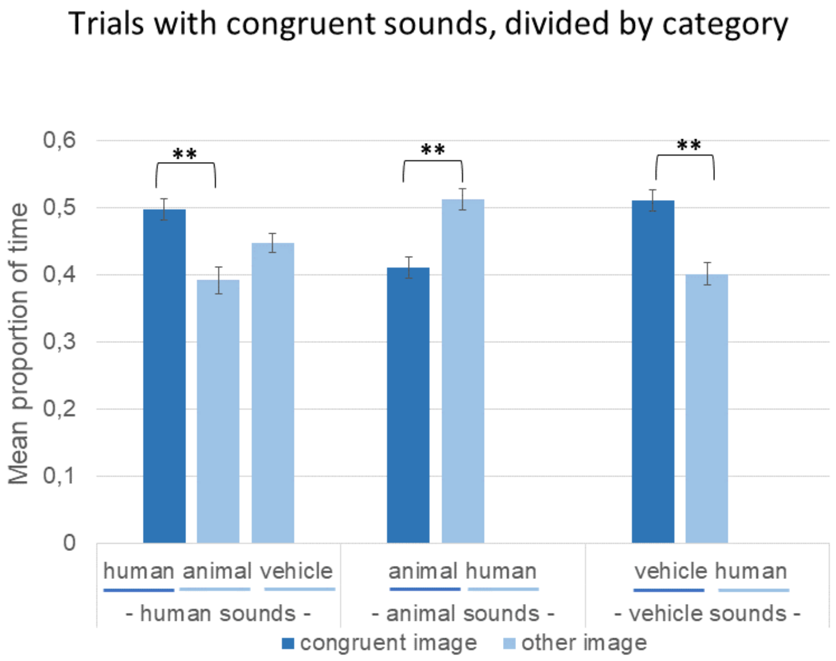


Figure 4.6: Bar plot showing the proportions of time spent perceiving the 'congruent' and 'other' image, averaged across subjects and trials, comparing categories with congruent sounds: human (12 trials per participant), animal (6 trials per participant) and vehicle (6 trials per participant). P-values were derived from t-tests and indicate significant differences (= p < .050, ** = p < .010). Error bars indicate SEM.*

We then looked at the same categories using the absolute difference (Figure 4.7), and one-sample t-tests against 0 revealed that the perception of vehicle images during congruent human sounds was significantly diminished (t(20) = -2.94, p = .008). Please note that Figure 4.7 is also showing human images during congruent human sounds to be significantly different from 0, this is the same result as already reported in Figure 4.5. There were no noticeable effects arising from congruent animal or congruent vehicle sounds (figure 4.7).

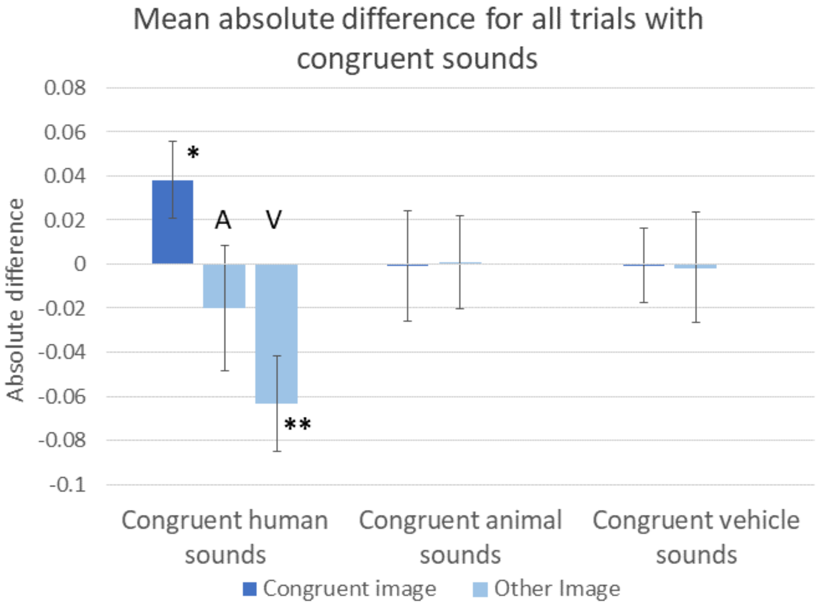


Figure 4.7: Bar plot showing the mean absolute difference of congruent and other images, comparing the trials with congruent sounds divided into three categories: human (12), animal (6) and vehicle (6). Dark bars represent the congruent images' score while the lighter ones represent the others. P-values were derived from t-tests and indicate significant differences from 0 (= p < .050, ** = p < .010). Error bars indicate SEM.*

However, while the t-tests we conducted between the *human* and *other* dominance times in the no sound condition did not reveal any significant difference, after looking at the results shown in Figures 4.6 and 4.7, we decided to divide the data of the *other* images of the no sound condition into animal and vehicle as well. Note that the number of trials for each of the two groups was again 6 and 6. A repeated measures ANOVA on the now three proportions of time in the no sound condition, showed a main effect of image category on length of perception (F(2,40) = 11.27, p < .001) with vehicle images being perceived for significantly longer than both human (p = .044) and animal (p = .001) when no sound was displayed. This suggested that vehicle images dominated BR even without any sounds, possibly due to differences in image contrast/energy or other image characteristics. This result could explain the overall higher dominance times of vehicle images compared to animal images, reinforced by congruent vehicle sounds (Fig. 6).

Since the results of the human category showed that congruent sounds could modulate dominance times, we wanted to see if this was the case also for congruent sounds and animal and vehicle images, by comparing them with the no-sound condition. A significant difference between the conditions with congruent sounds and the no-sound conditions would suggest sounds are modulating BR also when non-human stimuli are used. Interestingly, the mean proportions of times for both the competitor and the human image in the trials with congruent sounds and the trials with no sound are not significantly different (p > .957), suggesting congruent sounds did not have any effect on categories other than human (Figure 4.8). Thus, the dominance of vehicle images over the other categories seems to be due to the images themselves, and is present regardless of which sound is played.

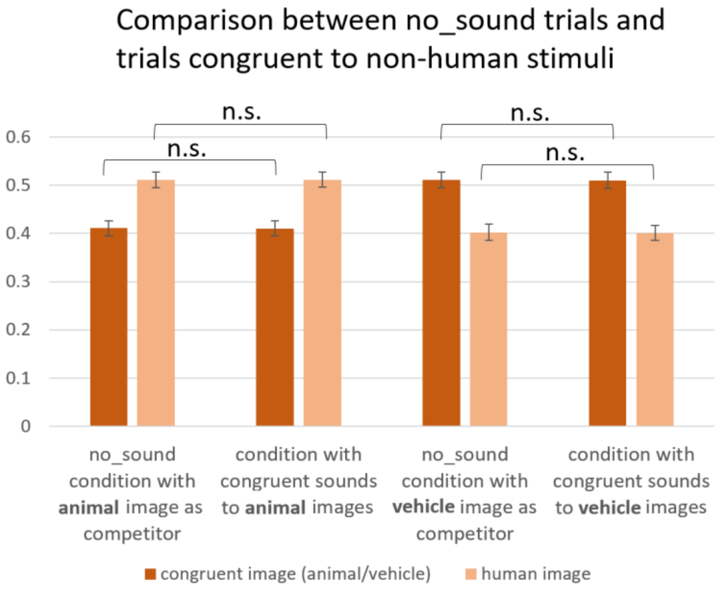


Figure 4.8: Bar plot showing the mean proportions of time spent perceiving the 'congruent' and human ('other') image, comparing no sound conditions with congruent sounds conditions (non-significant); animal (6 trials per participant) and vehicle (6 trials per participant) are compared against human (12 trials per participant). Error bars indicate SEM.

4.5: Discussion

To investigate how sounds of different semantic congruency modulated the dominance of human visual stimuli during binocular rivalry, we presented human images paired with animal or object images in the presence of congruent sounds, 'relevant' sounds (human but not entirely congruent, e.g., a marching sound with an image of a woman coughing), incongruent sounds, or silence. Our results showed that human sounds increased perceptual dominance times of human visual stimuli, but only when these were specifically congruent to the image. Hearing relevant but not entirely congruent sounds of other human nature had no effect over perceptual dominance of human images. This is consistent with previous studies that highlighted how semantically relevant sounds are facilitators of ambiguous visual percepts (Chen, Yeh and Spence, 2011; Tan and Yeh, 2015; Chen and Spence, 2018,). Semantically congruent sounds can even alter and speed up visual perception by allowing humans to extrapolate visual information more quickly (Williams and Stormer, 2019; Williams et al., 2022).

Our findings not only replicate the results shown by Chen, Yeh and Spence (2011), but extend them including a larger variety of stimuli. However, they also highlight a difference in how different categories of images interact with sounds during BR, which was not found in the aforementioned study. The boost of perceptual dominance was only found in trials with fully congruent human sounds, but when we looked at trials with congruent vehicle or animal sounds, these had no effect on the respective congruent vehicle or animal images. The vehicle images were perceived significantly more than the human images during the presentation of a congruent vehicle sound, but this was likely due to an effect of the vehicle images' properties, as vehicle images dominated BR to the same extent in the no sound condition. Interestingly, Chen, Yeh and Spence (2011) also reported a difference in the general dominance of their vehicle image versus the animal one, but whether this is due to the vehicle images being more complex or to vehicles driving more attention cannot be concluded with the findings from the present study.

Since we reported no differences between the no sound condition and the trials with congruent animal and vehicle sounds, it is also possible that audiovisual integration failed for those events. This could have happened because the images we selected were not optimal for BR. The animal images had a noisier background than most human images (i.e., the grass and the landscape) which may have made the animals less prominent than the human stimuli. This offers an explanation for the fact that animal images were dominant for less time than human and vehicle ones already in the no sound condition. On the other hand, the vehicle images had more horizontal shapes compared to the human ones, which might have also affected the way they were conflicting with the human ones. In the absence of sound, the vehicle images dominated perception in BR more, compared to the other categories, which perhaps means they already had a perceptual advantage due to visual features. Additional congruent sounds did not boost the perceptual advantage of vehicle images further.

As our experiment was designed to investigate various types of human sounds, the conclusions on the other two categories were drawn on the basis of unbalanced groups (i.e., congruent human trials were 12 trials per participants while congruent animal/vehicle were 6). To verify whether animal and vehicle sounds can in fact integrate with congruent pictures and boost their perception, a follow-up experiment should be conducted with an even (and larger) group of trials.

It should also be mentioned that, to our knowledge, there are no other BR studies using audio-visual stimulation that compared human stimuli with other categories. Human stimuli seem to be preferentially processed by the brain (Bracci and Op de Beeck, 2023), with pathways and regions dedicated to the processing of certain human features (e.g., body parts and faces). This is particularly relevant for multisensory processing, as efficient integration of human sounds and human images might be prioritised due to the importance of social interactions, which are by nature complicated and a prominent part of our lives.

Regardless of the differences between semantic categories we reported, our findings still support the facilitating role of sounds for the disambiguation of ambiguous visual percepts. But what are the mechanisms that allow semantically congruent auditory stimuli to influence visual perception? Sounds are very reliable cues for visual scenes, since they can usually be associated with their source (i.e., the voice of someone we know would prompt us to expect to see that person, the noise of an engine would make us expect a car, etc.), and vice versa for visual cues for auditory perception. It is possible that, due to the predictive relationship that ties audiovisual events together, perception in one modality may predict the relevant features in another modality, and as a consequence suppresses the irrelevant features. In the case of our experiment, hearing the sound of a woman coughing might suppress the irrelevant image of e.g., a sheep, and enhance the congruent image of a woman coughing, resulting in perceptual dominance during BR. The idea that sensory regions use categorical information coming from higher-level areas to predict signals and frame their recognition has been widely explored by predictive coding frameworks (Friston, 2010; Clark, 2013). The visual system might use information from the auditory domain to create predictions that favour the perception of the congruent stimulus.

However, some previous experiments have shown this semantic congruency to facilitate visual tasks even when the sound only shared the general category of the visual input (Chen and Spence, 2018, Williams and Stormer, in prep). In the latter, naturalistic sounds roughly related to the image sped up visual discrimination of a noisy visual stimulus, e.g. hearing a natural forest sound facilitated the perception of bird images, allowing participants to extract visual details more quickly than irrelevant sounds. Our study did not reflect that the general semantic category of sounds boosts perceptual dominance time of images. Our findings showed that human sounds relevant but not entirely congruent to human images (e.g. marching sound with a coughing woman image) did not facilitate image perception. Instead, only sounds fully congruent with the image (e.g., coughing sound with coughing woman image) boosted perceptual dominance in BR. Why was that the case? In the above-mentioned studies (Chen and Spence, 2018, Williams and Stormer, in prep), the sounds were not presented simultaneously with the visual stimuli, but always before, while the current study aimed to promote multisensory integration by presenting auditory and visual components together. Additionally, Chen and Spence (2018) and Williams and Stormer (in prep) did not present conflicting visual stimuli to participants, as opposed to our BR experiment. This could suggest that two different mechanisms have been tapped into: previous research investigated the predictive advantage multimodal stimulation can offer, which does not necessarily require auditory and visual events to be integrated, while our findings might instead reflect the need for integration of sound stimuli to have an effect on BR. In other words, when the brain receives ambiguous stimuli from two eyes, it integrates information from multiple modalities to make one visual stimulus prevail over the other, and this can only happen with fully congruent auditory-visual events. It should be noted that it is also possible that the ineffectiveness of the related sounds compared to the congruent has to do with the imbalance of trials in each category. The fact that trials in the related category were twice the amount of the trials used to make inferences about the congruent sounds, might have contributed to the difference reported. Future studies should take this into consideration and consider versions of this experiment where the number of trials is consistent or where subsamples of the larger group are used.

To conclude, our experiment supports the idea that semantically congruent, simultaneous sounds can boost conscious visual perception, including the disambiguation of ambiguous visual percepts during BR. This seems to happen only in the case of full semantic congruency, and only for human stimuli.

Chapter 5: General discussion

The goal of the current research project was to investigate auditory feedback to cortical regions at the early stages of visual processing (i.e., V1, V2 and V3), while also looking at the repercussions of this flux of information on visual perception itself.

First, I conducted an fMRI study on blindfolded participants, aimed at understanding the semantic information content of auditory feedback to early visual cortex and its degree of specificity in the absence of visual stimulation. I presented various types of sounds, categorised in a hierarchical fashion, and tested whether early visual cortex received enough information to be able to differentiate between semantic sound categories, as tested using MVPA. The findings revealed that sounds belonging to superordinate categories such as animate and inanimate, and to more specific categories such as humans, animals, vehicles and objects, can be decoded from neural activity patterns in early visual cortex (EVC). Further inspection also revealed that human sounds seem to be better decoded than other sound categories.

Second, I devised a binocular rivalry experiment with the scope of investigating the advantage sounds with varying degrees of semantic congruency can offer toward the perception of ambiguous human visual stimuli. This experiment provided novel findings that suggest that semantically congruent sounds can boost perception of congruent visual stimuli, but only when they are fully congruent (rather than merely related). This was only demonstrated in the case of human sounds and human images.

How do these findings fit together? Experiment 1 provides information about the type of auditory semantic content that is represented in EVC, but does not tell us much about the function that auditory feedback might have for vision or why it arrives in EVC in the first place. Experiment 2, on the other hand, addressed the question of what type of interactions auditory and visual stimuli have during ambiguous visual stimulation, although it lacks a neural component. Together, the results of both experiments complement each other and offer both neural and behavioural evidence to develop theories on auditory influence on both vision and EVC.

Previous research has extensively looked at neural correlates of binocular rivalry. It has been shown that primary visual cortex activity strongly correlates with perception during binocular rivalry (Polonsky et al. 2000; Lee and Blake, 2002; Haynes et al., 2005), as does activity in the lateral geniculate nucleus of the thalamus (Wunderlich, Schneider and Kastner, 2005), with more recent studies linking anatomical components of EVC (i.e., surface area) to the speed of perceptual transition across the visual field (Genç et al., 2015). The neural correlates of binocular rivalry do not stop at the early stages of vision. Activity linked to perceived stimuli is likely fed forward to inferior temporal regions of the visual system, which also reflect the alternation of perception. This is illustrated by rivalry studies in which perception switching from faces to houses respectively activates the fusiform face area and the parahippocampal place area (Tong et al., 1998). However, EVC reflects rivalry even in the event of unconscious perception (i.e., with stimuli made invisible), but the effect does not produce activation in fronto-parietal areas (Zou, He and Zhang, 2016).

These findings have been reconciled in support of theories that see rivalry being the result of processes not at just one location, but at different stages of the cortical hierarchy, depending on the specific type of stimuli and their disparity (Blake and Logothetis, 2002; Wilson, 2003). Ultimately, EVC seems to lie at the basis of this hierarchy.

As already mentioned in previous chapters, binocular rivalry has been found to be impacted by past knowledge and familiarity with the stimuli (Engel, 1956; LoSciuto and Hartley, 1963; Goryo, 1969, Lee et al., 2015) as well as information coming from other sensory modalities (Kang and Blake, 2005; Holcombe and Seizova-Cajic, 2008; Lunghi, Binda and Morrone, 2010; Zhou et al., 2010; Lunghi, Morrone & Alais, 2014). Interestingly, neuroimaging data found perceptual switches in binocular rivalry to be modulated by regions in the parietal cortex. A few transcranial magnetic stimulation (TMS) studies suggested that disrupting regions around the inferior parietal sulcus can act as a stabiliser of binocular rivalry switches, directing attention away from the bistable stimuli and thus lengthening each percept (Kanai, Bahrami and Rees, 2010; Zaretskaya et al., 2010).

Taken together, these findings could mean that the top-down influences arriving to EVC from other visual and non-visual regions might play a role in modulating binocular rivalry. Examining neuroimaging and behavioural results in light of this, I suggest that auditory information fed back to EVC (likely via temporoparietal multisensory areas i.e., posterior superior temporal sulcus, or pSTS) can modulate and influence perception during ambiguous stimulations such as binocular rivalry. When perception is unclear or unstable, feedback from other sensory modalities can act as a facilitator, providing new information that allows EVC to prioritise the relevant stimulus e.g., the one that is semantically relevant to the information received from other senses.

Why is pSTS a likely candidate as mediator for the communications between auditory and visual cortices? This is not only supported by the whole-brain MVPA searchlight findings in Experiment 1, but also by previous studies that reported decoding of auditory information in pSTS (Vetter, Smith and Muckli, 2014; Vetter et al., 2020). These are joined by structural and connectivity studies that found connections between pSTS and primary auditory and visual cortices (Rockland and Ojima, 2003; Beer, Plank and Greenlee, 2011). Notably, a recent study by Maguinnes and von Kriegstein (2021) found a correlation between the so-called 'face-benefit' of voice recognition and connectivity between pSTS and the fusiform face area (FFA). Voices are better identified when associated with faces, and in high-noise environments, the researchers found this facilitator effect to be positively correlated with increased functional connectivity between pSTS and FFA. Environments with low levels of noise instead, were associated with increased responses in FFA, which suggests pSTS might be employed in case of ambiguous/noisy environments to access information from other sensory modalities. Moreover, the fact that decoding of auditory information works better in areas of EVC representing the peripheral visual field (Vetter et al., 2014 & 2020) and that feedback connections from nonvisual areas seem more present in these peripheral areas too (Falchier et al., 2002), might further support the claim that EVC uses feedback from audition when visual information is not sufficiently clear (e.g., outside of the foveal visual field).

My interpretation of the current results fits with the theories that see EVC as an integral part of multisensory processing (Ghazanfar and Schroeder, 2006; Murray et al., 2016), speaking specifically for its role in audio-visual interactions. The neuroimaging study I conducted revealed that sounds are represented in EVC down to the categories of humans, animals, vehicles and objects. The binocular rivalry results, on the other hand, showed that only exactly congruent human sounds result in successful multisensory integration with corresponding visual images of humans, seemingly suggesting that detailed information about the semantic category is necessary. This could mean that, while EVC itself only represents certain information about sounds, to achieve audiovisual integration it needs to process multimodal events at different steps of the cortical hierarchy. So, after receiving input from the retina and auditory feedback, EVC might need to communicate with higher-level regions to make sense of the input from both modalities.

It is also possible that, in the interest of economising on resources, EVC receives a fine-grained level of feedback only when it is necessary i.e., when the visual stimulus is ambiguous. In the case of my fMRI study, since the sounds were presented in the absence of vision, it is plausible that the auditory information was transferred with reduced semantic specificity, because integration was not necessary.

As for binocular rivalry, while I do not have neuroimaging data on EVC's ability to represent sounds during binocular rivalry, I can theorise that feedback from auditory cortex mediated by multisensory regions and feedback from e.g., parietal regions involved in binocular rivalry, come together within EVC to disambiguate visual perception.

Alternatively, it is also possible that EVC does receive fine-grained semantic information about sounds. Although the 12-way classification I performed failed to show significant results, I cannot exclude the possibility that an experiment with more statistical power might reveal that more specific sounds can also be decoded in EVC.

In both experiments, I found evidence that supports the idea of human stimuli being important for the brain, or even preferentially treated. When studying how the brain represents sensory information, it is important to consider the processing of external stimuli in the context of the human experience. Humans relate to the world through actions and language, and most of our lives are filled with interactions with other humans. Our brains are able to extract an incredible amount of information from facial features and voices (Maguinness, Roswandowitz & von Kriegstein, 2018; Young, Frühholz and Schweinberger, 2020), so it would make sense for the brain's processes to reflect this significance.

In a recent review, Bracci and Op de Beeck (2023) proposed that the way the occipitotemporal cortex represents objects might be intrinsically related to the use we make of these representations in support of behaviour. Bodies facing each other and engaged in interactions are recognised more accurately than bodies not facing each other, an effect not found in the case of objects like chairs (Papeo, Stein and Soto-Faraco, 2017). Also, regions along the occipitotemporal cortex seem to represent animals according to how animate they are and how similar or dissimilar they are to objects (Sha et al., 2015). Notably, recently it has been shown how the variance in the representation of visual objects in the ventral temporal cortex can be explained through dimensions like their mobility, their similarity to animals and their agency (Jozkiw et al., 2022). These findings suggest that our interactions with the world might shape how the visual system treats stimuli, and given the consistent presence of human stimuli in our lives, there might be mechanisms that allow us to process them more efficiently.

Although the evidence described above speaks mostly for regions outside of EVC, this preference for processing human stimuli might exist there too. In that regard, the pSTS is well known for its representation of biological motion in faces and bodies (Beauchamp et al. 2002, Grossmann & Blake 2002). It also plays a role in audiovisual integration (Beauchamp et al., 2004) and, as already mentioned, has been found to feed information back to both primary auditory and visual areas (Naumer et al., 2011). Since I earlier suggested that auditory feedback might reach EVC through indirect routes i.e., passing through pSTS, it is plausible that this works particularly well for human auditory stimuli, given the area's predisposition to recognise human movements.

Results from both experiments could also speak in favour of the predictive coding theory I mentioned in Chapter 1 (Friston, 2010; Clark, 2013). As the brain uses newly available sensory information to update current predictions it makes of the world, it is possible that the activity patterns picked up in the EVC are the framework the brain is providing to EVC to better understand what visual stimuli to expect. The auditory stimulation received triggers the predictive model that, once arrived to EVC, produces content-specific activity patterns that are picked up by the classifier. Similarly, the advantage that congruent sounds give to perceptual dominance of images during binocular rivalry could also come from the fact that EVC has created a mental image of the stimuli that might help the brain perceive and recognise the visual input. If the brain expects to be in an environment where a certain category of stimuli is present (i.e., animals), the predictions that it is going to make are going to facilitate perception of what is congruent and produce errors if the sensory signal does not match.

Future directions

As my two experiments tackled two separate questions, one regarding the neural response of EVC to sounds, and one regarding the behavioural consequences of sounds on visual perception, it is important to find a way to combine these to propose a mechanism incorporating both.

Experiment 1 indicates that auditory cortex sends feedback information to EVC that is represented possibly according to specific semantic categories. However, while I can make inferences regarding the path taken by this flux of information, I cannot be sure of it as of yet. The current results and the evidence previously discussed seem to point towards the presence of an indirect path, one that might pass through pSTS and relay auditory feedback to EVC.

Future studies could employ techniques such as TMS to disrupt activity in specific regions to see if that prevents auditory information from reaching EVC. Additional to being a good candidate in terms of functionality and connectivity, pSTS has been reportedly used as locus for successful TMS experiments (e.g., Grossman, Battelli and Pascual-Leone, 2005, van Kemenade et al., 2012), which is important to consider when designing brain stimulation studies. If disrupting pSTS with TMS right before an fMRI procedure similar to the one employed in Experiment 1 has an impact on sound classification in EVC, this would provide a causal clue as to the path auditory information takes to reach EVC.

With these results, it would be interesting to design further experiments that implicate the auditory feedback to EVC in tasks such as binocular rivalry. If we disrupt pSTS with TMS, are participants still able to integrate congruent sounds and images in a way that advantages visual stimuli during binocular rivalry? Would dominance times of images congruent to sounds be affected by pSTS being disrupted? And if so, what is the temporal window of disruption that renders audiovisual integration ineffective?

Since I described evidence in favour of the theory that sees feedback information arriving to EVC when visual stimulation is unavailable or unclear, future studies could try to replicate my findings employing visual stimuli that are masked, or weak in terms of the activity they produce. As Stein and Meredith (1986, 1987) suggested that such stimuli are the ones demonstrating stronger multisensory integration at the neural level, the use of close-to-threshold stimuli to investigate EVC could potentially contribute to the hypothesis that EVC is not only representing information from other sensory modalities, but also integrating them.

A possible experiment to test whether auditory information is only represented in EVC when the visual information is ambiguous, would benefit from using these types of stimuli. Participants would be tested with three different types of trials in the scanner: a) audio-visual trials, where they see incomplete objects, perhaps with large parts of the visual field occluded, paired with congruent sounds; b) audio-visual trials with clear, intact visual stimuli; and c) auditory trials with blindfold, identical to those I used in the current fMRI experiment. If the theory that information is only represented in EVC when it is needed is correct, then we should also see a difference in the classifier accuracy between conditions.

As mentioned earlier, in Experiment 1 we could not significantly decode semantic sound categories down to the subordinate level in EVC. However, this might be the case if we include more trials within each category to classify. As the time inside the MRI scanner is limited, one possible way of doing this would be to reduce the number of categories i.e. investigating only humans and animals, or humans and objects, but keeping three subcategories with three sounds within each. In this way, the total number of categories is lower, and we can afford to include more repetitions of the same sound.

In addition to the ideas I have discussed so far, another possibility would be to conduct different types of analysis on the existing fMRI data I collected. Among the multivariate techniques, representational similarity analysis (RSA, Kriegeskorte, Mur and Bandettini, 2008) could offer some interesting insight that would complement what I reported with MVPA. Rather than looking for differences between activity patterns, RSA is used to plot how similar neural patterns are, which would allow me to see whether all sounds in a certain category are clustered together in their representation, or whether they can be distinguished into separate groups.

Another aspect that has not been covered by my current analyses and that would provide a different perspective from which to look at the fMRI data is functional connectivity (Friston, 1994). Looking at the connections between EVC and auditory cortex (passing through pSTS), could offer evidence towards the theory that pSTS is in fact mediating the communication between these primary sensory regions. Moreover, I proposed that the brain might be biased towards the perception of human stimuli; human faces are located faster and more accurately than other primates or animals (Simpson et al. 2014), and both faces and speech are preferentially attended by infants (Pascalis, De haan & Nelson, 2002, Vouloumanos et al., 2010). If this bias exists in terms of auditory feedback arriving to EVC, it could be reflected by increased functional connectivity between pSTS and EVC when participants are listening to human sounds, as opposed to other categories.

As for Experiment 2, I discussed in Chapter 4 how the design of the experiment was primarily focused on human audio-visual stimuli and how fewer trials were used to make inferences about the animal and vehicle stimuli. It has been shown that the facilitating effect of congruent sounds on the perception of ambiguous images exists for vehicles and animals, too (Chen, Yeh and Spence, 2011), so a future experiment might focus on replicating that. By designing a study with equal amounts of congruent trials across humans, vehicles and objects, it would be possible to verify whether human stimuli somehow overshadow the other categories or whether audiovisual integration still happens.

Conclusion

The content of the current thesis is divided into a neuroimaging and a behavioural component, but the conclusions that can be made from it fit nicely together in a theory that sees early visual cortex (EVC) as a focal point for non-visual sensory information that can help disentangle an ambiguous visual input. Results from Experiment 1 informed us that early visual cortex can represent information about the semantic category of sound down to the specific categories of humans, animals, vehicles and objects, and that human sounds are generally better decoded. With Experiment 2, I showed how listening to congruent human sounds can lengthen dominance times of congruent human images in a binocular rivalry setting, while ‘relevant’ sounds (incongruent but still human-related) have no such effect. Taken together, these results contribute to the growing body of evidence that implicates EVC in the network of multisensory processing, especially when the stimuli processed are of human nature. The mechanism I proposed involves EVC receiving feedback from auditory cortex mediated by posterior superior temporal cortex and possibly parietal cortices in the case of binocular rivalry. Future studies should focus on better understanding the conditions under which EVC receives auditory feedback and causally tie it to perceptual phenomena like binocular rivalry.

Bibliography

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage*, 8(4), 360-369.

Aller, M., Giani, A., Conrad, V., Watanabe, M., & Noppeney, U. (2015). A spatially collocated sound thrusts a flash into awareness. *Frontiers in Integrative Neuroscience*, 9, 16.

Amedi, A., Malach, R., Hendler, T., Peled, S., & Zohary, E. (2001). Visuo-haptic object-related activation in the ventral visual pathway. *Nature neuroscience*, 4(3), 324–330. https://doi.org/10.1038/85201

Amedi, A., Stern, W. M., Camprodon, J. A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P., & Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nature neuroscience*, 10(6), 687–689. https://doi.org/10.1038/nn1912

Anzai, A., Peng, X., & Van Essen, D. C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature neuroscience*, 10(10), 1313-1321.

Astafiev, S. V., Stanley, C. M., Shulman, G. L., & Corbetta, M. (2004). Extrastriate body area in human occipital cortex responds to the performance of motor actions. *Nature neuroscience*, 7(5), 542-548.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7), 280-289.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature neuroscience*, 7(11), 1190-1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41(5), 809-823.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron*, 34(1), 149-159.

Beauchamp, M. S. (2005). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics*, 3, 93-113.

Bedford, F. (2001). Towards a general law of numerical/object identity. *Current Psychology* *of Cognition*, 20(3/4), 113-176.

Beer, A. L., Plank, T., & Greenlee, M. W. (2011). Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. *Experimental brain research*, 213(2), 299-308.

Birch-Hirschfeld, A., & Inouye, T. (1909). Experimentelle und histologische Untersuchungen über Netzhautabhebung. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 70(3), 486-538.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience*, 3(1), 13-21.

Blumenthal, A., Stojanoski, B., Martin, C. B., Cusack, R., & Köhler, S. (2018). Animacy and real-world size shape object representations in the human medial temporal lobes. *Human brain mapping*, 39(9), 3779-3792.

Bola, Ł., Yang, H., Caramazza, A., & Bi, Y. (2022). Preference for animate domain sounds in the fusiform gyrus of blind individuals is modulated by shape–action mapping. *Cerebral Cortex*, 32(21), 4913-4933.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annu. Rev. Neurosci.*, 28, 157-189.

Bracci, S., & de Beeck, H. O. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience*, 36(2), 432-444.

Bracci, S., & Op de Beeck, H. P. (2023). Understanding Human Object Vision: A Picture is Worth a Thousand Representations. *Annual Review of Psychology*, 74, 113-135.

Brascamp, J. W., Klink, P. C., & Levelt, W. (2015). The ‘laws’ of binocular rivalry: 50 years of Levelt’s propositions. *Vision research*, 109, 20-37.

Butter, C. M., Buchtel, H. A., & Santucci, R. (1989). Spatial attentional shifts: Further evidence for the role of polysensory mechanisms using visual and tactile stimuli. *Neuropsychologia*, 27(10), 1231-1240.

Caclin, A., Bouchet, P., Djoulah, F., Pirat, E., Pernier, J., & Giard, M. H. (2011). Auditory enhancement of visual perception at threshold depends on visual abilities. *Brain research*, 1396, 35-44.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral cortex*, 11(12), 1110-1123.

Carmel, D., Arcaro, M., Kastner, S., & Hasson, U. (2010). How to create and use binocular rivalry. *JoVE (Journal of Visualized Experiments)*, (45), e2030.

Chanauria, N., Bharmauria, V., Bachatene, L., Cattan, S., Rouat, J., & Molotchnikoff, S. (2019). Sound induces change in orientation preference of V1 neurons: Audio-visual cross-influence. *Neuroscience*, 404, 48-61.

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, 12(4), 478-484.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427), 345-347.

Chen, Y. C., & Spence, C. (2018). Dissociating the time courses of the cross-modal semantic priming effects elicited by naturalistic sounds and spoken words. *Psychonomic bulletin & review*, 25(3), 1138-1146.

Chen, Y.-C., and Spence, C. (2011). When hearing the bark helps to identify the dog: semantically congruent sounds modulate the identification of masked pictures. *Cognition*, 114, 389–404.

Chen, Y. C., Huang, P. C., Yeh, S. L., & Spence, C. (2011). Synchronous sounds enhance visual sensitivity without reducing target uncertainty. *Seeing and perceiving*, 24(6), 623.

Chen, Y. C., Yeh, S. L., & Spence, C. (2011). Crossmodal constraints on human perceptual awareness: auditory semantic modulation of binocular rivalry. *Frontiers in psychology*, 2, 212.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci*. 2013;36:181–204.

Conrad, V., Bartels, A., Kleiner, M., & Noppeney, U. (2010). Audiovisual interactions in binocular rivalry. *Journal of Vision*, 10(10), 27-27.

Conrad, V., Kleiner, M., Bartels, A., Hartcher O'Brien, J., Bülthoff, H. H., & Noppeney, U. (2013). Naturalistic stimulus structure determines the integration of audiovisual looming signals in binocular rivalry. *PLoS One*, 8(8), e70710.

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual review of vision science*, 4, 381.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *In Seminars in the Neurosciences* (Vol. 2, pp. 263-275). Saunders Scientific Publications.

Crick, F. (1996). Visual perception: rivalry and consciousness. *Nature*.

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, 24(31), 6991-7006.

Deco, G., & Lee, T. S. (2004). The role of early visual cortex in visual integration: a neural model of recurrent interaction. *European Journal of Neuroscience*, 20(4), 1089-1100.

de Haas, B., Schwarzkopf, D. S., Urner, M., & Rees, G. (2013). Auditory modulation of visual stimulus encoding in human retinotopic cortex. *Neuroimage*, 70, 258-267.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193-222.

DeValois, R. L., & DeValois, K. K. (1990). *Spatial vision* (No. 14). Oxford University Press on Demand.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2017). Vividness of visual imagery depends on the neural overlap with perception in visual areas. *Journal of Neuroscience*, 37(5), 1367-1373.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain research*, 1242, 136-150.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470-2473.

Einhäuser, W., Methfessel, P., & Bendixen, A. (2017). Newly acquired audio-visual associations bias perception in binocular rivalry. *Vision Research*, 133, 121-129.

Engel, E. (1956). The role of content in binocular resolution. *The American Journal of Psychology*, 87-91.

Engel, L. R., Frum, C., Puce, A., Walker, N. A., & Lewis, J. W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. *Neuroimage*, 47(4), 1778-1791.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598-601.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, 22(13), 5749-5759.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex* (New York, NY: 1991), 1(1), 1-47.

Fiehler, K., & Rösler, F. (2010). Plasticity of multisensory dorsal stream functions: evidence from congenitally blind and sighted adults. *Restorative Neurology and Neuroscience*, 28(2), 193-205.

Freud, E., Macdonald, S. N., Chen, J., Quinlan, D. J., Goodale, M. A., & Culham, J. C.

(2018). Getting a grip on reality: Grasping movements directed to real objects and images rely on dissociable neural representations. *Cortex*, 98, 34-48.

Freud, E., Plaut, D. C., & Behrmann, M. (2016). ‘What’is happening in the dorsal visual pathway. *Trends in Cognitive Sciences*, 20(10), 773-784.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1–2), 56– 78.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, 11(2), 127-138.

Gallivan, J. P., McLean, D. A., Valyear, K. F., & Culham, J. C. (2013). Decoding the neural mechanisms of human tool use. *elife*, 2, e00425.

Ganel, T., Valyear, K. F., Goshen-Gottstein, Y., & Goodale, M. A. (2005). The involvement of the "fusiform face area" in processing facial expression. *Neuropsychologia*, 43(11), 1645–1654. https://doi.org/10.1016/j.neuropsychologia.2005.01.012

Genç, E., Bergmann, J., Singer, W., & Kohler, A. (2015). Surface area of early visual cortex predicts individual speed of traveling waves during binocular rivalry. *Cerebral cortex*, 25(6), 1499-1508.

Gerdes, A. B., Wieser, M. J., & Alpers, G. W. (2014). Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. *Frontiers in psychology*, 5, 1351.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory?. *Trends in cognitive sciences*, 10(6), 278-285.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5), 350-363.

Giovannelli, F., Giganti, F., Righi, S., Peru, A., Borgheresi, A., Zaccara, G., Viggiano, M. P., & Cincotta, M. (2016). Audio-visual integration effect in lateral occipital cortex during an object recognition task: An interference pilot study. *Brain stimulation*, 9(4), 574–576. https://doi.org/10.1016/j.brs.2016.02.009

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2), 133-139.

Goebel, R., Esposito, F. & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with Brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping*, 27, 392-401

Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. *In Information Processing in Medical Imaging: 18th International Conference, IPMI 2003, Ambleside, UK, July 20-25, 2003. Proceedings 18 (pp. 330-341).* Springer Berlin Heidelberg.

Goryo, K. (1969). The effect of past experience upon the binocular rivalry. *Japanese Psychological Research*, 11(2), 46-53.

Griffis, J. C., Elkhetali, A. S., Burge, W. K., Chen, R. H., & Visscher, K. M. (2015). Retinotopic patterns of background connectivity between V1 and fronto-parietal cortex are modulated by task demands. *Frontiers in human neuroscience*, 9, 338.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, 24(1), 187-203.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, 35(6), 1167-1175.

Grossman, E. D., Battelli, L., & Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision research*, 45(22), 2847-2853.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37(1), 435-456.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430.

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5), 686-691.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—it's my area: or is it?. *Journal of cognitive neuroscience*, 20(12), 2125-2136.

Helmholtz, H. V. (1860). *Theorie der Luftschwingungen in Röhren mit offenen Enden*.

Henschke, J. U., Noesselt, T., Scheich, H., & Budinger, E. (2015). Possible anatomical pathways for short-latency multisensory integration processes in primary sensory cortices. *Brain Structure and Function*, 220(2), 955-977.

Henson, R. (2007). Efficient experimental design for fMRI. *Statistical parametric mapping:*

The analysis of functional brain images, 193-210.

Holcombe, A. O., & Seizova-Cajic, T. (2008). Illusory motion reversals from unambiguous motion with visual, proprioceptive, and tactile stimuli. *Vision research*, 48(17), 1743-1757.

Holmes, N. P. (2007). The law of inverse effectiveness in neurons and behaviour: multisensory integration versus normal variability. *Neuropsychologia*, 45(14), 3340-3345.

Holmes, N. P. (2009). The principle of inverse effectiveness in multisensory integration: some statistical considerations. *Brain topography*, 21, 168-176.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nature neuroscience, 19(4), 613-622.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3), 574.

Hubel, D. H., & Wiesel, T. N. (1979). Brain mechanisms of vision. *Scientific American*, 241(3), 150-163.

Huettel, S. A., Song, A. W., & McCarthy, G. (2014). *Functional Magnetic Resonance Imaging*. Sinauer.

Ito, M., Westheimer, G., & Gilbert, C. D. (1998). Attention and perceptual learning modulate contextual influences on visual perception. *Neuron*, 20(6), 1191-1197.

Jertberg, R., Levitan, C. A., & Sherman, A. (2019). Multisensory processing of facial expressions in binocular rivalry. *Emotion*, 19(7), 1214.

Jozwik, K. M., Najarro, E., Van Den Bosch, J. J., Charest, I., Cichy, R. M., & Kriegeskorte, N. (2022). Disentangling five dimensions of animacy in human brain and behaviour. *Communications Biology*, 5(1), 1247.

Kaas, J. H. (2013). The evolution of brains from early mammals to humans. Wiley Interdisciplinary Reviews*: Cognitive Science*, 4(1), 33-45.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679-685.

Kanai, R., Bahrami, B., & Rees, G. (2010). Human parietal cortex structure predicts individual differences in perceptual rivalry. *Current biology*, 20(18), 1626-1630.

Kang, M. S., & Blake, R. (2005). Perceptual synergy between seeing and hearing revealed during binocular rivalry. *Psichologija*.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11), 4302-4311.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22(4), 751-761.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3), 856-867.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta psychologica*, 134(3), 372-384.

Kovács, I., Papathomas, T. V., Yang, M., & Fehér, Á. (1996). When the brain changes its mind: Interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences*, 93(26), 15508-15511.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, 17(1), 26-49.

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage*, 38(4), 649-662.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10), 3863-3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5), 535-540.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., & Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, 166, 289–297.

Lee, M., Blake, R., Kim, S., & Kim, C. Y. (2015). Melodic sound enhances visual awareness of congruent musical notes, but only if you can read music. *Proceedings of the National Academy of Sciences*, 112(27), 8493-8498.

Lee, S. H., & Blake, R. (2002). V1 activity is reduced during binocular rivalry. *Journal of* *vision*, 2(9), 4-4.

Leone, L. M., & McCourt, M. E. (2013). The roles of physical and physiological simultaneity in audiovisual multisensory facilitation. *i-Perception*, 4(4), 213-228.

Levelt, W. J. (1965). *On binocular rivalry* (Doctoral dissertation, Van Gorcum Assen).

Leventhal, A. G., Wang, Y., Schmolesky, M. T., and Zhou, Y. (1998). Neural correlates of boundary perception. *Vis. Neurosci.* 15, 1107–1118.

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., & DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. Journal of Neuroscience, 25(21), 5148-5158.

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience,* 7(6), 651-657.

Liang, M., Mouraux, A., Hu, L., & Iannetti, G. D. (2013). Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nature communications,* 4(1), 1-10.

Lister, W. T., & Holmes, G. (1916). Disturbances of vision from cerebral lesions, with special reference to the cortical representation of the macula. *Proceedings of the Royal Society of Medicine,* 9(Sect_Ophthalmol), 57-96.

LoSciuto, L.A. and Hartley, E.L., 1963: Religious affiliation and open-mindedness in binocular resolution, *Percept. Mot. Skills*, 17, 427–430.

Lunghi, C., & Morrone, M. C. (2013). Early interaction between vision and touch during binocular rivalry. *Multisensory Research*, 26(3), 291-306.

Lunghi, C., Binda, P., & Morrone, M. C. (2010). Touch disambiguates rivalrous perception at early stages of visual analysis. *Current Biology*, 20(4), R143-R144.

Lunghi, C., Morrone, M. C., & Alais, D. (2014). Auditory and tactile signals combine to influence vision during binocular rivalry. *Journal of Neuroscience,* 34(3), 784-792.

Lyon, D. C., & Kaas, J. H. (2002). Connectional evidence for dorsal and ventral V3, and other extrastriate areas in the prosimian primate, Galago garnetti. *Brain, Behavior and Evolution,* 59(3), 114-129.

Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science,* 289(5482), 1206-1208.

Maguinness, C., & von Kriegstein, K. (2021). Visual mechanisms for voice-identity recognition flexibly adjust to auditory noise level. *Human Brain Mapping*, 42(12), 3963-3982.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, 116, 179-193.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, 32(1), 4-18.

Mansfield, P. (1977). Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics*, 10(3), L55.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., ... & Kennedy, H. (2014). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1), 225-259.

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British journal of Psychology*, 64(1), 17-24.

Marshall, J. C., & Fink, G. R. (2001). Spatial cognition: where we were and where we are. *Neuroimage*, 14(1), S2-S7.

Marshall, J. C., Fink, G. R., Halligan, P. W., & Vallar, G. (2002). Spatial awareness: a function of the posterior parietal lobe?. *Cortex*, 38(2), 253-257.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, 379(6566), 649-652.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *elife*, 7, e31873.

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J. P., Maeder, P. P., Clarke, S., & Meuli, R. A. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cerebral Cortex*, 17(7), 1672-1679.

Mattioni, S., Rezk, M., Battal, C., Bottini, R., Cuculiza Mendoza, K. E., Oosterhof, N. N., & Collignon, O. (2020). Categorical representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *elife*, 9, e50732.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say ‘Hello’? Personality impressions from brief novel voices. *PloS one*, 9(3), e90779.

McClure Jr, J. P., & Polack, P. O. (2019). Pure tones modulate the representation of orientation and direction in the primary visual cortex. *Journal of neurophysiology*, 121(6), 2202-2214.

McDonald, J. J., Störmer, V. S., Martinez, A., Feng, W., & Hillyard, S. A. (2013). Salient sounds activate human visual cortex automatically. *Journal of Neuroscience*, 33(21), 9194-9201.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.

McMains, S., & Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *Journal of Neuroscience*, 31(2), 587-597.

Meienbrock, A., Naumer, M. J., Doehrmann, O., Singer, W., & Muckli, L. (2007). Retinotopic effects during spatial audio-visual integration. *Neuropsychologia,* 45(3), 531-539.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology,* 56(3), 640-662.

Meredith, M. A., Wallace, M. T., & Stein, B. E. (1992). Visual, auditory and somatosensory convergence in output neurons of the cat superior colliculus: multisensory properties of the tecto-reticulo-spinal projection. *Experimental Brain Research*, 88(1), 181-186.

Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., & Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature neuroscience*, 13(6), 667-668.

Miller, B. T., & D'Esposito, M. (2005). Searching for “the top” in top-down control. *Neuron*, 48(4), 535-538.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action.* Oxford, England: Oxford University Press.

Morimoto, M. M., Uchishiba, E., & Saleem, A. B. (2021). Organization of feedback projections to mouse primary visual cortex. *IScience*, 24(5), 102450.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of neurophysiology*, 70(3), 909-919.

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, 14(4), 2178-2189.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial and temporal contrast sensitivity of neurons in areas 17 and 18 of the cat's visual cortex. *The Journal of physiology*, 283(1), 101-120.

Muckli, L., & Petro, L. S. (2013). Network interactions: Non-geniculate input to V1. *Current opinion in neurobiology*, 23(2), 195-201.

Muckli, L., Naumer, M. J., & Singer, W. (2009). Bilateral visual field maps in a patient with only one hemisphere. *Proceedings of the National Academy of Sciences*, 106(31), 13034-13039.

Mumford, D. (1992). On the computational architecture of the neocortex: II The role of cortico-cortical loops. *Biological cybernetics*, 66(3), 241-251.

Murray, M. M., Cappe, C., Romei, V., Martuzzi, R., Thut, G., & Stein, B. E. (2012). Auditory-visual multisensory interactions in human primary cortices: Synthesis and controversies. *The new handbook of multisensory processes*, 223-238.

Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2016). The multisensory function of the human primary visual cortex. *Neuropsychologia*, 83, 161-169.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, 168(1), 1-10.

Musz, E., Loiotile, R., Chen, J., Cusack, R., & Bedny, M. (2022). Naturalistic stimuli reveal a sensitive period in cross modal responses of visual cortex: Evidence from adult-onset blindness. *Neuropsychologia*, 108277.

Naghavi, H. R., Eriksson, J., Larsson, A., & Nyberg, L. (2007). The claustrum/insula region integrates conceptually related sounds and pictures. *Neuroscience letters*, 422(1), 77–80. https://doi.org/10.1016/j.neulet.2007.06.009

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage*, 105, 215-228.

Nasr, S., Polimeni, J. R., & Tootell, R. B. (2016). Interdigitated color-and disparity-selective columns within human visual cortical areas V2 and V3. *Journal of Neuroscience*, 36(6), 1841-1857.

Naumer, M. J., van den Bosch, J. J., Wibral, M., Kohler, A., Singer, W., Kaiser, J., ... & Muckli, L. (2011). Investigating human audio-visual object perception with a combination of hypothesis-generating and hypothesis-testing fMRI analysis tools. *Experimental brain research*, 213, 309-320.

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience*, 30(21), 7434-7446.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature neuroscience*, 5(11), 1203-1209.

Olivares, E. I., Iglesias, J., Saavedra, C., Trujillo-Barreto, N. J., & Valdés-Sosa, M. (2015). Brain signals of face processing as revealed by event-related potentials. *Behavioural neurology,* 2015.

Oruc, I., Balas, B., & Landy, M. S. (2019). Face perception: A brief journey through recent discoveries and current directions. *Vision research*, 157, 1-9.

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The two-body inversion effect. *Psychological science*, 28(3), 369-379.

Park, H., & Kayser, C. (2019). Shared neural underpinnings of multisensory integration and trial-by-trial perceptual recalibration in humans. *Elife*, 8, e47001.

Pascalis, O., De Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life?. *Science*, 296(5571), 1321-1323.

Peelen, M. V., & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, 93(1), 603-608.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, 51(1), 195-203.

Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. *Neuroimage*, 56(2), 476-496.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1), S199-S209.

Petro, L.S., Paton, A.T., and Muckli, L. (2017). Contextual modulation of primary visual cortex by auditory signals. *Philos. Trans. R. Soc. B Biol. Sci.* 372.

Piazza, E. A., Denison, R. N., & Silver, M. A. (2018). Recent cross-modal statistical learning influences visual perceptual selection. *Journal of vision,* 18(3), 1-1.

Polonsky, A., Blake, R., Braun, J., & Heeger, D. J. (2000). Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature neuroscience,* 3(11), 1153-1159.

Ramos-Estebanez, C., Merabet, L. B., Machii, K., Fregni, F., Thut, G., Wagner, T. A., ... & Pascual-Leone, A. (2007). Visual phosphene perception modulated by subthreshold crossmodal sensory stimulation. *Journal of Neuroscience,* 27(15), 4178-4181.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1), 79-87.

Ratan Murty, N. A., Teng, S., Beeler, D., Mynick, A., Oliva, A., & Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, 117(37), 23011-23020.

Reynolds, J. H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron*, 37(5), 853-863.

Ripp, I., zur Nieden, A. N., Blankenagel, S., Franzmeier, N., Lundström, J. N., & Freiherr, J. (2018). Multisensory integration processing during olfactory-visual stimulation—An fMRI graph theoretical network analysis. *Human brain mapping*, 39(9), 3713-3727.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, 50(1-2), 19-26.

Romei, V., Murray, M. M., Cappe, C., & Thut, G. (2009). Preperceptual and stimulus-selective enhancement of low-level human visual cortex excitability by sounds. *Current biology*, 19(21), 1799-1805.

Sathian, K. (2016). Analysis of haptic information in the cerebral cortex. *Journal of neurophysiology*, 116(4), 1795-1806.

Schira, M. M., Tyler, C. W., Breakspear, M., & Spehar, B. (2009). The foveal confluence in human visual cortex. *J. Neurosci*, 29(28), 9050–9058.

Schlack, A., & Albright, T. D. (2007). Remembering visual motion: neural correlates of associative plasticity and motion recall in cortical area MT. *Neuron*, 53(6), 881-890.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., ... & Tootell, R. B. H. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889-893.

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The animacy continuum in the human ventral vision pathway. *Journal of cognitive neuroscience*, 27(4), 665-678.

Shams, L., & Beierholm, U. (2011). Humans' multisensory perception, from integration to segregation, follows Bayesian inference. *Sensory cue integration*, 251-262.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, 408(6814), 788-788.

Sillito, A. M., Cudeiro, J., & Jones, H. E. (2006). Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in neurosciences*, 29(6), 307-316.

Simpson, E. A., Buchin, Z., Werner, K., Worrell, R., & Jakobsen, K. V. (2014). Finding faces among faces: Human faces are located more quickly and accurately than other primate and mammal faces. Attention, Perception, & Psychophysics, 76, 2175-2183.

Singer, W. (2013). Cortical dynamics revisited. *Trends in cognitive sciences*, 17(12), 616-626.

Spence, C., & Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Current Biology*, 13(13), R519-R521.

Spence, C., Nicholls, M. E., Gillespie, N., & Driver, J. (1998). Cross-modal links in exogenous covert spatial orienting between touch, audition, and vision. *Perception & psychophysics*, 60(4), 544-557.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. The MIT Press.

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage*, 65, 69-82.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage*, 44(3), 1210-1223.

Tan, J. S., & Yeh, S. L. (2015). Audiovisual integration facilitates unconscious visual scene processing. *Journal of experimental psychology: human perception and performance*, 41(5), 1325.

Thulborn, K. R., Waterton, J. C., Matthews, P. M., & Radda, G. K. (1982). Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 714(2), 265-270.

Tong, F., & Engel, S. A. (2001). Interocular rivalry revealed in the human cortical blind-spot representation. *Nature*, 411(6834), 195-199.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron*, 21(4), 753-759.

Treue, S., & Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, 382(6591), 539-541.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736), 575-579.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science, 311*(5761), 670-674.

Tse, C. Y., Gratton, G., Garnsey, S. M., Novak, M. A., & Fabiani, M. (2015). Read my lips:

Brain dynamics associated with audiovisual integration and deviance detection. *Journal of cognitive neuroscience*, 27(9), 1723-1737.

Tsuchiya, N., & Koch, C. (2004). Continuous flash suppression. *Journal of Vision*, 4(8), 61-61.

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature neuroscience,* 8(8), 1096-1101.

Uddin, L. Q., Nomi, J. S., Hébert-Seropian, B., Ghaziri, J., & Boucher, O. (2017). Structure and function of the human insula. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, 34(4), 300.

Van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron*, 81(6), 1240-1253.

van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22), E4501-E4510.

Van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *Neuroimage,* 55(3), 1208-1218.

Van der Stoep, N., Van der Stigchel, S., Nijboer, T. C. W., & Van der Smagt, M. J. (2016). Audiovisual integration in near and far space: effects of changes in distance and stimulus effectiveness. *Experimental brain research*, 234, 1175-1188.

van Kemenade, B. M., Muggleton, N., Walsh, V., & Saygin, A. P. (2012). Effects of TMS over premotor and superior temporal cortices on biological motion perception. *Journal of Cognitive Neuroscience*, 24(4), 896-904.

Vetter, P., Bola, Ł., Reich, L., Bennett, M., Muckli, L., & Amedi, A. (2020). Decoding natural sounds in early "visual" cortex of congenitally blind individuals. *Current Biology*, 30(15), 3039-3044.

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology,* 24(11), 1256-1262.

Vohn, R., Fimm, B., Weber, J., Schnitker, R., Thron, A., Spijkers, W., ... & Sturm, W. (2007). Management of attentional resources in within-modal and cross-modal divided attention tasks: An fMRI study. *Human brain mapping*, 28(12), 1267-1275.

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The tuning of human neonates' preference for speech. *Child development,* 81(2), 517-527.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, 56(2), 366–383.

Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage*, 31(3), 1247-1256.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior research methods*, 42(3), 671-684.

Williams, J. R., & Störmer, V. S. (2019). Auditory information facilitates sensory evidence accumulation during visual object recognition. *Journal of Vision*, 19(10), 20c-20c.

Williams, J. R., Markov, Y. A., Tiurina, N. A., & Störmer, V. S. (2022). What You See Is What You Hear: Sounds Alter the Contents of Visual Perception. *Psychological science*, 33(12), 2109-2122.

Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*, 17(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, H. R. (2003). Computational evidence for a rivalry hierarchy in vision. *Proceedings of the National Academy of Sciences*, 100(24), 14499-14503.

Winlove, C. I., Milton, F., Ranson, J., Fulford, J., MacKisack, M., Macpherson, F., & Zeman, A. (2018). The neural correlates of visual imagery: A co-ordinate-based meta-analysis. *Cortex*, 105, 4-25.

Wolbers, T., Klatzky, R. L., Loomis, J. M., Wutte, M. G., & Giudice, N. A. (2011). Modality-independent coding of spatial layout in the human brain. *Current Biology*, 21(11), 984-989.

Wunderlich, K., Schneider, K. A., & Kastner, S. (2005). Neural correlates of binocular rivalry in the human lateral geniculate nucleus. *Nature neuroscience*, 8(11), 1595-1602.

Yang, Y. H., & Yeh, S. L. (2011). Accessing the meaning of invisible words. *Consciousness and cognition*, 20(2), 223-233.

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in cognitive sciences*, 24(5), 398-410.

Zaretskaya, N., Thielscher, A., Logothetis, N. K., & Bartels, A. (2010). Disrupting parietal function prolongs dominance durations in binocular rivalry. *Current biology*, 20(23), 2106-2111.

Zeki, S. (1983). The distribution of wavelength and orientation selective cells in different areas of monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 217(1209), 449-470.

Zhang, P., Jamison, K., Engel, S., He, B., & He, S. (2011). Binocular rivalry requires visual attention. *Neuron,* 71(2), 362-369.

Zhou, W., Jiang, Y., He, S., & Chen, D. (2010). Olfaction modulates visual perception in binocular rivalry. *Current Biology*, 20(15), 1356-1358.

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. Journal of Neuroscience, 27(45), 12292-12307.

Zou, J., He, S., & Zhang, P. (2016). Binocular rivalry from invisible patterns. *Proceedings of the National Academy of Sciences*, 113(30), 8408-8413.