Object perception by ear and eye

**fMRI studies on multisensory processes
in human cerebral cortex**

Cover Design: Petra Janson

ISBN 10: 90-5278-593-7

ISBN 13: 978-90-5278-593-6

Printed by Datawyse / Universitaire Pers Maastricht

Copyright © 2006 by Marcus J. Naumer. All rights reserved.

Object perception by ear and eye

fMRI studies on multisensory processes in human cerebral cortex

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr. G.P.M.F. Mols
volgens het besluit van het College van Decanen, in het openbaar te verdedigen op
dinsdag 19 december 2006 om 12:00 uur

door

Marcus Johannes Naumer

Geboren op 1 september 1973 te Neustadt an der Weinstrasse

Promotor:

Prof. dr. R. Goebel

Copromotor:

Dr. L.F. Muckli (Max-Planck-Institut für Hirnforschung, Frankfurt am Main, Duitsland)

Beoordelingscommissie:

Prof. dr. J. van Heerden (voorzitter)

Prof. dr. F. Di Salle

Dr. E. Formisano

Prof. dr. D.E.J. Linden (University of Wales, Bangor, UK)

Prof. dr. B. Röder (Universität Hamburg, Duitsland)

“The integration of inputs from different sensory modalities not only transforms some of their individual characteristics, but does so in ways that can enhance the quality of life. Integrated sensory inputs produce far richer experiences than would be predicted from their simple coexistence or the linear sum of their individual products.”

Stein & Meredith, 1993

To my loved ones

Contents

CHAPTER 1	Functional imaging of human crossmodal identification and object recognition	9
<hr/>		
CHAPTER 2	Retinotopic effects during spatial audio-visual integration	43
<hr/>		
CHAPTER 3	Natural sounds activate object-related visual cortex	65
<hr/>		
CHAPTER 4	Audio-visual integration of artificial and natural object features in the human brain	79
<hr/>		
	Summary	105
<hr/>		
	Samenvatting	109
<hr/>		
	Acknowledgements	115
<hr/>		
	Curriculum Vitae	117
<hr/>		

CHAPTER 1

Functional imaging of human crossmodal identification and object recognition

Based on manuscript:

Amedi, A. *, von Kriegstein, K. *, van Atteveldt, N.M. *, Beauchamp, M.S. *, Naumer, M.J. * (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research* 166, 559-571.

*equal contribution

Abstract

The perception of objects is a cognitive function of prime importance. In everyday life, object perception benefits from the coordinated interplay of vision, audition, and touch. The different sensory modalities provide both complementary and redundant information about objects, which may improve recognition speed and accuracy under many circumstances. We review crossmodal studies of object recognition in humans that mainly employed functional magnetic resonance imaging (fMRI). These studies show that visual, tactile, and auditory information about objects can activate cortical association areas that were once believed to be modality-specific. Processing converges either in multisensory zones or via direct crossmodal interaction of modality-specific cortices without relay through multisensory regions. We integrate these findings with existing theories about semantic processing and propose a general mechanism for crossmodal object recognition: The recruitment and location of multisensory convergence zones varies depending on the information content and the dominant modality.

Introduction

We experience our environment via several sensory modalities at the same time. For example, in an opera show or a movie, we perceive visual and sound information in parallel. The information provided by these diverse sensory systems is synthesized in our brains to create the coherent and unified percepts we are so familiar with. The multisensory nature of our perceptions provides several behavioral advantages, for example, speeded response and improved recognition in noisy environments (e.g. Newell 2004). Until recently, the neural correlates of crossmodal integration and its development have been investigated mainly using (invasive) electrophysiology in a variety of animal species and brain structures (Stein and Meredith 1993; Wallace et al. 2004a). The advent of non-invasive functional neuroimaging techniques led to investigations of crossmodal processes important for human cognition, such as linguistic processing, person identification, and object categorization.

In this paper, we mainly review fMRI investigations of human crossmodal object recognition. We define objects as: “something material that may be perceived by the senses” (Merriam-Webster online dictionary). We employ this definition in a broader sense by including not only concrete objects such as vehicles, tools, and persons, but also more abstract objects such as letters or speech with its accompanying lip movements. As these objects have characteristic attributes in multiple sensory modalities, recognition might strongly benefit from the coordinated interplay of vision, audition, and touch. We will focus on studies of audio-visual and visuo-tactile object recognition that have recently produced a number of consistent findings. After a brief introduction to unimodal visual, auditory, and tactile object recognition, we will give an overview of studies including audio-visual and visuo-tactile processing. In the discussion, we will specifically focus on determinants of crossmodal convergence, on the role of imagery and furthermore, describe a model of distributed semantic processing.

Unimodal object recognition

Visual object recognition

Based on the vast amount of anatomical data in monkey visual cortex, the picture of a highly diverse and hierarchically structured system has emerged (Felleman and

Van Essen 1991). There is a widely accepted notion of two processing streams between early visual areas in the occipital cortex and higher-order processing centers in the temporal and parietal lobes, referred to as the dorsal and ventral streams, respectively (Ungerleider and Haxby 1994). As the dorsal stream is also involved in visuomotor transformations, Goodale and Milner proposed to distinguish between *vision for action* (dorsal stream) and *vision for perception* (ventral stream; Goodale et al. 1994). In this review, we will focus on object perception and thus discuss the ventral stream in more detail.

As the ventral stream contains structures devoted to the fine-detailed analysis of a visual scene such as form and color, it is also known as the “what” pathway. It consists of areas V1 to V4 in the occipital lobe and several regions that belong to the lateral and ventral temporal lobe (see blue regions in **Fig. 1**). Electrophysiological studies in non-human primates have identified organizing principles, such as retinotopy, selectivity for simple features like spatial orientation in V1, and selectivity for categories of complex stimuli such as faces or spatial layouts in inferior temporal (IT) cortex. These findings have been confirmed non-invasively in human and non-human primates using fMRI (Tootell et al. 2003; see also **Fig. 1**). An active debate is ongoing as to the actual organization of the ventral stream (Grill-Spector and Malach 2004; Tootell et al. 2003). One area in the lateral occipital cortex, the lateral occipital complex (LOC; Malach et al. 1995) has also been the subject of intense investigation. LOC responds strongly to pictures of intact objects compared to scrambled objects. In ventral temporal cortex, specialized areas for faces (Kanwisher et al. 1997), scenes (Epstein and Kanwisher 1998), human body parts (Downing et al. 2001; see also **Fig. 1**) letters (Gauthier et al. 2000; Polk et al. 2002) and visual word forms (McCandliss et al. 2003) have been described. However, developing a theoretical framework for understanding these specialized regions appears to be problematic, and the notion of widely distributed and overlapping cortical object representations remains a possible principle of organization (Haxby et al. 2000; Haxby et al. 2001). Effects of perceptual expertise for certain object categories (Gauthier et al. 2000) and, more recently, different category-related resolution needs (Hasson et al. 2003) have also been proposed to explain the organization of the human ventral stream.

Figure 1. Human cortical “what” pathways for vision, audition, and touch. Prominent cortical areas/regions that belong to the visual, auditory, and tactile object recognition pathways are shown in blue, yellow, and red, respectively. Their relative locations are indicated on inflated representations of the left (LH) and right (RH) cerebral hemispheres in ventral (LH) and lateral (RH) views, respectively. Abbreviations: V1, V2, VP, and V4: retinotopic visual areas; LOC: lateral occipital complex; FFA: fusiform face area; PPA: parahippocampal place area; VOT: ventral occipito-temporal cortex; EBA: extrastriate body area; A1: tonotopic primary auditory cortex; PT: planum temporale; aSTG: anterior superior temporal gyrus; aVR: anterior voice region; pSTG: posterior STG; SI-SII: somatotopic primary and secondary somatosensory cortices; IPS: intra-parietal sulcus.

Auditory object recognition

Auditory objects are complex spectro-temporal sounds that can be either associated with a visual object (e.g., a certain string sound with a cello or a specific voice with a specific face) or completely independent (e.g., a melody). In analogy to the visual system, it is thought that auditory “what” information is processed in a ventral as opposed to the dorsal “how/where” stream (Arnott et al. 2004; Belin and Zatorre 2000; Kaas and Hackett 1999; Rauschecker and Tian 2000; Romanski et al. 1999). These processing streams are thought to originate at least in part from the planum temporale (PT, **Fig. 1**) which responds to any kind of complex spectro-temporal sounds (Griffiths and Warren 2002). Auditory processing may be organized in a specialized fashion with regard to different categories of auditory stimuli. Evidence for such an organization comes from behavioral and lesion data including cases of

auditory agnosia specific to environmental sounds, music, words, and voices (Polster and Rose 1998). Recent functional imaging data show that anterior auditory regions (namely anterior superior temporal sulcus and gyrus (aSTS/STG); **Fig. 1**) respond during the discrimination and identification of complex auditory forms (Binder et al. 2004; Zatorre et al. 2004). Furthermore, a region in the aSTS (aVR; **Fig. 1**) is activated specifically by speaker identity (Belin and Zatorre 2003; von Kriegstein et al. 2003) and thus may be the auditory equivalent to the fusiform face area (FFA; Kanwisher et al. 1997) in visual face recognition. In contrast, stimulation with environmental sounds has been shown to activate posterior regions in the STS and the medial temporal gyrus (MTG; Lewis et al. 2004), which might reflect semantic processing and/or temporal integration (Thierry et al. 2003). Traditionally, the left posterior STG, and more specifically the PT, has been associated with speech perception (Wernicke 1874). Some investigations have suggested that this region is involved in the analysis of rapidly changing acoustic cues that is especially important, but not specific, for speech perception (Jäncke et al. 2002).

Tactile object recognition

Tactile information processing seems to be hierarchically organized demonstrating growing complexity as one moves from area 3b in the primary somatosensory cortex (SI; see also **Fig. 1**) towards areas located more posterior along the post-central gyrus and sulcus. In non-human primates, there is increasing receptive field complexity as one moves along the anterior-posterior axis of the post-central gyrus (from area 3b, via areas 1 and 2, to areas 5 and 7 that are located around the anterior intraparietal sulcus (IPS)). In humans, there is evidence for a similar functional hierarchy as one moves from areas 3b and 1 (activated by any tactile input) to area 2 (showing selectivity to the attributes of objects such as curvature) and to anterior IPS (showing preference to overall shape compared to shape primitives such as curvature; Bodegard et al. 2001). Activation in IPS was also found in a variety of tasks requiring analysis of object shape, length, and curvature using both simple geometrical shapes (Bodegard et al. 2001; O'Sullivan et al. 1994; Roland et al. 1998) and natural everyday objects (Amedi et al. 2001; Amedi et al. 2002; Deibert et al. 1999; Reed et al. 2004). Thus, although it has been less studied than the ventral visual stream, an analogous pathway for tactile object recognition might also exist in primates. It also suggests similarities with the hierarchical organization evident in vision and audition. However, in general, haptic object recognition can be

viewed as a more serial process than visual or even auditory object recognition (i.e. in active palpation one needs to integrate the tactile input over time, sampling the different parts of an object). Still, this functional hierarchy is not supported yet by as much anatomical evidence as in the visual system (Felleman and Van Essen 1991).

The possible role of SII in tactile object recognition and the possibility of alternative pathways (a ventrolateral pathway stretching from SI to inferior parietal regions including SII and the insula; Mishkin 1979; Reed et al. 2004) are still under debate. Early work suggested that there is a division of labor between processing of micro-geometry aspects of the stimulus, such as surface texture properties, in the parietal operculum (SII) and macro-geometry properties, such as shape attributes of objects, in the anterior IPS (O'Sullivan et al. 1994; Roland et al. 1998). This view was challenged recently both in humans (Reed et al. 2004: showing object preference in SII and the insula for tactile object recognition versus palpation of nonsense objects) and in non-human primates (for a review, see Iwamura et al. 1998). In addition, previous studies have shown that lesions to SII and related structures give rise to tactile agnosia (e.g. Reed and Casselli 1994).

Audio-visual object recognition

In everyday object recognition, the coordinated interplay of two or more sensory systems is the rule rather than the exception (as illustrated by the influence of lip reading in speech perception or by crossmodal priming in person recognition). However, relatively little is known about multisensory integration in human cerebral cortex. Given the distributed organization of the visual and auditory systems, how do these systems interact? We do not perceive the sight and the sound of a musical instrument as two different events. But what are the underlying neural mechanisms that accomplish this crossmodal sharing and union of information? In the following section, we will describe studies investigating possible neural mechanisms underlying the perception of different categories of audio-visual information such as speech, persons, and common objects.

Linguistic information

Mouth movements of a speaker form a natural visual component of speech perception. Movements and phonetic information are inherently linked by

characteristics such as temporal onset, duration and frequency-amplitude information (Calvert 2001; Munhall and Tohkura 1998). Correspondences between speech sounds and mouth movements (often referred to as ‘audio-visual (AV) speech’) are learned implicitly and early in development by exposure to heard speech together with the sight of the speaker (Kuhl and Meltzoff 1982). In contrast, the visual representation of spoken language by written language is an artificial cultural invention. That is to say, letters and speech sounds exhibit no naturally corresponding characteristics but are related in a completely arbitrary way. Therefore, associations between letters and speech sounds are not learned automatically, but require explicit instruction (Gleitman and Rozin 1977; Liberman 1992).

Figure 2. A summary of the reported superior temporal regions (mainly STG, STS, and MTG) involved in AV integration of linguistic information. The most important activations of the reviewed studies are projected on sagittal slices of the MNI template brain transformed into Talairach space. LH is shown at x=-50, RH at x=50. Circles reflect results from studies using speech and lip movement stimuli (i.e. naturally related AV stimuli), stars reflect results from studies using letters and speech sounds (i.e. artificially related AV stimuli).

Viewing the face of a speaker during listening enhances speech perception, especially under noisy conditions (Sumby and Pollack 1954). Lip reading can also change the speech percept, as shown by the McGurk illusion, in which the auditory

speech percept is altered by non-corresponding visually perceived speech information (McGurk and MacDonald 1976). A possible neural basis for these perceptual effects is the activation of auditory processing regions by silent lip reading, as is reported by several fMRI studies (Bernstein et al. 2002; Calvert et al. 1997; Olson et al. 2002; Paulesu et al. 2003; Sekiyama et al. 2003). Consistently reported auditory regions include the STS, STG, and auditory association regions, like the PT, that surround the primary auditory cortex (A1). The involvement of A1, as reported by Calvert et al. (1997), is a matter of debate (Bernstein et al. 2002; Paulesu et al. 2003). In addition to studies focusing on the visual component of AV speech, the integration of auditory and visual speech has also been addressed directly. Despite wide variations in methodology, a consistent finding from AV speech studies is the involvement of the left STG, STS, and MTG in the integration of heard and seen speech (Callan et al. 2003; Calvert et al. 2000; Macaluso et al. 2004; Sekiyama et al. 2003; Wright et al. 2003; see also **Fig. 2**). Olson et al. (2002) report that the claustrum (but not the STS) integrates AV speech information. Interestingly, crossmodal effects in the STS seem to be strongest when auditory information is degraded by noise (Callan et al. 2001; Callan et al. 2003; Sekiyama et al. 2003), and this is consistent with behavioral findings (Sumbly and Pollack 1954). A mechanism for AV speech perception has been proposed in which the auditory and visual speech signals are integrated in STS/STG, followed by modulation of the relevant unisensory cortices (auditory association cortex and MT/V5) via feedback projections (Calvert et al. 1999; Calvert et al. 2000).

The neural basis of the associations between speech sounds and their written representations (i.e. letters) is much less extensively investigated. In a recent fMRI study (van Atteveldt et al. 2004), several bilateral STS/STG regions were identified that responded stronger to bimodal congruent (i.e. matching) letter-sound pairs as compared to their respective unimodal components (see **Fig. 2**, and blue-colored regions in **Fig. 3**). Furthermore, the response to speech sounds in early unisensory auditory regions in Heschl's sulcus (HS) and PT was found to be modulated by the degree of congruency with a simultaneously presented letter. Because these early auditory regions did not respond to unimodal visual letters, the authors interpret this modulation as a feedback effect from integration sites in the STS/STG. An MEG study provides converging time course information for the processing and integration of letters and speech sounds. The estimated source locations also indicate the

involvement of the STS in the integration of letters and speech sounds (Raij et al. 2000). More evidence for this function of the STS/STG is provided by a recent learning experiment (Hashimoto and Sakai 2004). Learning of new letter-sound mappings involved a network of inferior temporal and parieto-occipital regions, while the STG/MTG was active during processing of already acquired matching letter-sound combinations.

Figure 3. A summary of regions found in studies for crossmodal object recognition. Active regions are shown in color on lateral and ventral views of an inflated cortical surface model. Yellow color: regions for VT object recognition (from Amedi et al. 2002; Amedi et al. 2001). Green color: regions for AV integration of complex object information (from Beauchamp et al. 2004). Blue color: regions for AV letter recognition (from van Atteveldt et al. 2004). Red color: During the task of speaker recognition on familiar speakers' voices the FFA interacts with the middle STS voice area (analysis of functional connectivity from von Kriegstein et al. in press). Purple color: regions for AV recognition of natural or common objects (from Naumer et al. 2004).

Persons: faces and voices

Auditory and visual person-related information can be associated either inherently (e.g. face and voice) or arbitrarily (e.g. person and name). Person identification in general is a special form of subordinate object classification which consists of binding concurrent information about face, voice, body, and name with a specific person. Studies on unimodal person perception have revealed cortical areas and activation patterns that are more responsive to faces than to other visually or haptically presented object categories (Kanwisher et al. 1997; Haxby et al. 2001; Pietrini et al. 2004) and the same holds true for the auditory system where regions in the right STS

showed a higher fMRI signal increase in response to voices than to any other category of auditory objects (Belin et al. 2000). These regions are anatomically segregated and the neural mechanisms underlying the combination of visual and auditory person identity information is unclear. An influential model of person recognition (Burton et al. 1990; Ellis et al. 1997) assumes that the combination of unimodal information takes place at a supramodal stage serving as a connecting node for voice, face, and name. However, it has been proposed recently that category-specific visual and auditory association cortices can also interact directly without a relay through supramodal regions (von Kriegstein et al. in press).

In contrast to arbitrary face-name associations, voice and face information is often encoded in parallel over an extended period of time. This close association between facial and vocal information does not only allow the ability to read speech from facial movements and emotion from face and voice (Dolan et al. 2001), but in addition, it allows to infer information from the identity of a voice to the identity of a face even if voice and face have never been encountered before (Kamachi et al. 2003). When people become familiar with a person, vocal and facial information influences the speed with which voices and faces are recognized (Ellis et al. 1997). In a recent study with familiar and non-familiar speakers' voices and faces, von Kriegstein et al. (in press) showed that if subjects recognize familiar speakers' voices, activity in face responsive regions is enhanced. According to the person recognition model, crossmodal activation in face responsive regions should be accomplished via a supramodal node. Possible person identity nodes (e.g. regions regularly responding to familiar faces, voices and names more than to non-familiar stimuli) are bilateral anterior temporal lobes, temporo-parietal junctions, and precuneus (Gorno-Tempini et al. 1998; Leveroni et al. 2000; Nakamura et al. 2000; Gorno-Tempini and Price 2001; Shah et al. 2001). Interestingly, functional connectivity analyses revealed that these supramodal areas were not functionally connected with the FFA during recognition of familiar persons' voices (von Kriegstein et al. in press). In contrast, connectivity was enhanced to the STS voice regions suggesting a direct interaction of these regions during familiar speaker recognition (red-colored regions in **Fig. 3**).

Common objects

Although common objects are often perceived audio-visually under natural conditions, only few crossmodal fMRI studies have employed common object stimuli. Among the cortical regions reported in these studies, three regions appear to be

most prominently involved in AV recognition of common objects: lateral temporal cortex (especially the STS), ventral temporal cortex, and frontal cortices.

Lateral temporal cortex

Two recent studies (Beauchamp et al. 2004; Naumer et al. 2004) showed that posterior STS responds most strongly to auditory and visual objects and shows an enhanced response when auditory and visual objects are presented together (see green- and purple-colored regions in **Fig. 3**). Beauchamp and colleagues (2004) demonstrated that this holds true for different categories of common objects (animals and manipulable tools), stimulus configurations (photograph vs. line drawing vs. moving video), and behavioral tasks (simple one-back, naming, and semantic judgments). As multisensory activity in individual subjects always appeared in posterior STS, they named this region the STS multisensory region (STS-MS). This is also consistent with a recent proposal (Wallace et al. 2004a) that crossmodal integration occurs in multisensory zones that arise at the borders between unisensory cortical regions. While STS-MS shows a preference for semantically consistent, compared to inconsistent, audio-visual stimuli, the exact nature of the computations performed in this region remains unclear.

Ventral temporal cortex

Amedi and colleagues recently demonstrated tactile object-related activation in dorsal parts of LOC, an object-related region of the ventral visual pathway (Amedi et al. 2001; yellow-colored regions in **Fig. 3**). As stimulation with sounds related to the same objects did not strongly activate this region (Amedi et al. 2002), they concluded that this lateral occipital tactile-visual (or LOtv) region is devoted to the processing of shape. However, other fMRI studies have reported crossmodal AV effects in the ventral visual pathway during subordinate categorization and the detection of semantic (in)consistencies between pictures and sounds (Adams and Janata 2002; Beauchamp et al. 2004; Laurienti et al. 2003; Naumer et al. 2004; Naumer et al. 2002a). In addition, two recent EEG studies (Molholm et al. 2004; Murray et al. 2004) provided converging time course information for the processing of common object-related pictures and sounds, and the estimated source locations in these studies also indicated an involvement of the ventral visual pathway in object-related AV integration. These findings can be integrated by taking into account the hierarchical organization of the ventral visual stream (Grill-Spector 2003). We assume that

regions in ventral occipito-temporal cortex (VOT; see **Fig. 1**) located ventral to LOC are processing more abstract stimulus properties (such as object category) that are also accessible via audition, thus abstracting from those properties that are only accessible via vision and touch (such as 3D structure). Preliminary fMRI data (Naumer et al. 2002b) provided the first empirical support for this assumption by indicating category-related AV convergence in VOT.

Frontal cortices

AV fMRI studies also revealed contributions of frontal cortices to crossmodal cognitive functions such as object categorization (Adams and Janata 2002) and the detection of semantic inconsistencies (Laurienti et al. 2003; Naumer et al. 2004; Naumer et al. 2002a). Adams and Janata (2002) instructed subjects to match either pictures or environmental sounds with simultaneously presented words. During visual as well as auditory judgments, neural activity in (predominantly left) inferior frontal gyrus (IFG) appeared to reflect the level of categorization (subordinate versus basic), thus indicating a role of the IFG in integrating multisensory object representations with concepts in semantic memory. Varying the degree of AV semantic consistency, Laurienti and colleagues (2003) found significantly greater activation in anterior cingulate gyrus and medial prefrontal cortex for consistent as compared to inconsistent AV objects, whereas the inverse activation profile was found in a fronto-parietal network of regions (bilaterally including the precentral sulci and inferior parietal lobules) in recent studies by Naumer and colleagues (2004; 2002a).

Visuo-tactile object recognition

The visual and tactile hierarchies

The visual and tactile systems share many similarities. Both systems are arranged in a hierarchical order of increasing complexity. Neurons in V1 and area 3b within SI have relatively small receptive fields of simple features and contralateral specificity (hand or visual field), while the receptive field size and complexity (or functional selectivity) increase along the pathways as was detailed above. The two systems also share the principle of topographical organization, that is, adjacent parts of the world (whether in the visual field or on the body) are represented in adjacent parts within retinotopic and somatotopic cortical maps, respectively. In both systems, the strict topographical representation of the contralateral sensory field at early

processing stages decreases as one moves up along the hierarchy. For instance, LOC responds to visual stimuli in both hemifields and areas 1 and 2 have bilateral tactile receptive fields (Iwamura et al. 1994, Iwamura 1998). Finally, the visual system includes parallel pathways to analyze different aspects of the sensory world (Ungerleider and Mishkin 1982) and a similar pathway devoted to tactile form processing might also exist in the somatosensory system (Bodegard et al. 2001; Reed et al. 2004).

Evidence for visuo-tactile integration

The crossmodal delayed match to sample (DMS) task is a widely used approach to study crossmodal interactions. In this task, the cue is presented in one sensory modality and the target in the other. Human PET and fMRI studies employing DMS tasks on ellipsoids with varying curvature (Hadjikhani and Roland 1998), 2D patterns (Saito et al. 2003), or simple arbitrary 3D shapes created from wooden spheres (Grefkes et al. 2002) have all found greater activation following crossmodal versus intramodal matching in the insula/claustrum (Banati et al. 2000; Hadjikhani and Roland 1998), anterior IPS (Grefkes et al. 2002), and posterior IPS (Saito et al. 2003). These results suggest that the insula and IPS play a crucial role in binding visual and tactile information. It was proposed that the insula plays a role in crossmodal matching by serving as a mediating area allowing unisensory areas to communicate and exchange information. This was based on crossmodal matching experiments (Hadjikhani and Roland 1998) and the notion that the insula is highly connected to the various sensory areas and thus might be an ideal candidate to accomplish this function. On the other hand, the IPS was hypothesized to be a multisensory convergence site for visual, tactile, and possibly also auditory information. To study the potential role of visual areas in human tactile processing, an early study investigated tactile discrimination of grating orientations (Sathian et al. 1997). They reported robust tactile activation in the parieto-occipital cortex, a region previously considered part of the dorsal visual stream. In a later study, these same investigators demonstrated an interference in task performance using transcranial magnetic stimulation (TMS) to disrupt activity in this area (Zangaladze et al. 1999).

Figure 4. Putative VT object/form integration sites in humans. Summary of the sites showing multisensory convergence from the different articles reviewed here (see also Table 1). The Talairach coordinates reported in each of the studies were projected on an inflated cerebral representation of a Talairach normalized brain. The approximate locations are denoted by their respective symbol. The relevant putative multisensory areas are found in the bilateral occipito-temporal cortex (centered on LOTv), parietal cortex (centered on the IPS), and the right insula. Please note that in the study of Stoesz et al. another more ventral cluster in the parahippocampal gyrus was found, but is hidden from view here. The study of Deibert et al. includes several occipital areas but the exact Talairach coordinates are not available and thus are not marked on the figure.

Integration of visual and tactile object processing in the ventral visual pathway

Visuo-tactile convergence of object-related information occurs in LOTv (Amedi et al. 2001), which is a subregion within the human LOC. The defining feature of this region is that it is robustly activated during both visual and tactile object recognition. It shows a preference for objects compared to textures and scrambled objects in both modalities and is only weakly activated by the motor, naming and visual imagery components of object recognition. The category-related specialization for faces, objects, and scenes observed in ventral temporal cortex for visually presented objects is also found when objects are recognized by touch (Amedi et al. 2002; Pietrini et al. 2004). As shown by Pietrini and colleagues (2004), these category-related responses are correlated across touch and vision, suggesting that a common

representation of 3D objects is activated by both these modalities. Stoesz et al. (2003) and Prather et al. (2004) showed that both IPS and LOtv are preferentially activated by macrospatial shape recognition (of imprinted symbol identity) but not during a microspatial task (gap detection or orientation judgment). This demonstrates that 2D shape stimuli activate LOtv and that this activation is maintained without active exploration, thus excluding a contribution from the motor system. James and colleagues showed fMRI activation in occipital areas during haptic exploration of novel abstract objects. These objects were employed in order to reduce the potential influence of naming and visual imagery (James et al. 2002). By demonstrating that the magnitude of tactile-to-visual priming was similar to the magnitude of visual-to-visual priming, their findings supported the idea that vision and touch share common representations (see also Easton et al. 1997; Reales and Ballesteros 1999). Reed and coworkers (Reed et al. 2004) used abstract and nonsense 3D shapes combined with familiar and meaningful 3D objects. Lateralized activation was found, suggesting that right LOtv is more involved in recovering the shape of objects while the left LOC is more focused on the association of tactile input with semantic information and possibly also with familiarity knowledge. Taking into account the differences in stimuli, tasks, contrasts, and analysis techniques, LOtv seems to be highly consistent in location (average Talairach coordinates (x, y, z): -47, -61, -5 for the left and 51, -56, -7 for the right hemispheres, respectively).

Normally, individuals can easily recognize faces by touch. Interestingly, Kilgour and Lederman have recently reported evidence of an individual with haptic prosopagnosia (i.e. a deficit in recognizing familiar faces by touch), in addition to visual prosopagnosia following lesions to occipital, temporal and prefrontal regions (Kilgour et al. 2004). Some evidence suggests that patients with visual agnosia also suffer from tactile agnosia (Feinberg et al. 1986; Morin et al. 1984; Ontake et al. 2001; for a short review see Amedi et al. 2002). In an experimental setting, TMS can be used to create “virtual lesions” in normal individuals (Pascual-Leone et al. 2000). Zangaladze et al. (1999) used this technique to show interference with a tactile orientation task during TMS of area PO, while Merabet et al. (2004) showed that inhibitory repetitive TMS (at 1 HZ) to the occipital cortex reduced performance in tactile distance judgments but not in roughness judgments, suggesting an involvement of the occipital cortex in tactile distance processing. Both studies support the view that the occipital cortex is engaged in tactile tasks requiring fine spatial

discriminations. To conclude, three cortical regions seem to be important for visuo-tactile form integration (see **Fig. 4**): LOTv and adjacent areas, parietal regions (especially IPS), and the claustrum/insular cortex.

Discussion

Multisensory regions of temporal, parietal, frontal, and insular cortices are involved in crossmodal binding of AV and VT object-related information (Calvert and Lewis 2004; Newell 2004). The fMRI studies reviewed here demonstrate that these regions are not equally responsive to all types of unimodal stimulation and their combinations, but rather show a degree of specialization. After a short summary of the most prominent neural systems involved in AV and VT integration, we will further discuss findings that provide more insight in to what determines the location of crossmodal convergence and the role of imagery in crossmodal object processing. Finally, we will conclude by proposing a model for crossmodal object recognition by distributed semantic processing.

The posterior STS and STG play an important role in the integration of naturally as well as artificially related AV “what” information (i.e. speech with lip movements or letters, and sounds with pictures of common objects). This is in accordance with the suggested general role of the STS in integration of object identity information (Beauchamp et al 2004; Calvert 2001). The functional role of the STS is also supported by the recent finding that activity in this area is affected by the temporal synchrony of AV speech information, but is insensitive to the spatial displacement of the visual and auditory stimuli (Macaluso et al. 2004). However, the commonalities between vision and audition that elicit STS activation are still elusive. A variety of studies have emphasized the involvement of STS in dynamic aspects of visual object processing (e.g. eye gaze (Hoffman and Haxby 2000), visual speech (Calvert and Campbell 2003)), as well as auditory object processing (von Kriegstein and Giraud 2004; Thierry et al. 2003).

On the other hand, in VT object recognition, the LOTv, IPS, and right insular cortex appear to be prominently involved. It was recently demonstrated that tactile responses in LOTv during object identification are bilateral irrespective of the palpating hand in both sighted and blind subjects (Amedi et al. 2004). This further supports the view that LOC is involved in high-order tactile recognition. Such

crossmodal convergence might reflect the processing of features that are common to both vision and touch. Besides their unimodal features (e.g., color and consistency), both contribute to the evaluation of surface characteristics of objects (e.g., geometrical shape). In general, it seems that LOtv and IPS are more involved in object shape analysis and recognition, while the claustrum/insula region is more involved in the transfer and/or binding of unimodal information. The possible division of labor between these regions and their functional interplay are far from being understood and thus remain to be addressed in future studies. In contrast to vision and touch, audition contributes much less to the evaluation of surface characteristics, but instead, presents the dynamics or energetics of a situation (e.g. whether two objects pass each other or collide; Bregman 1990).

What determines a point of crossmodal convergence?

One determinant of where crossmodal convergence of object information takes place is which modality provides the most accurate information about the object (Welch & Warren, 1986; Ernst & Bulthoff, 2004). As vision dominates touch and audition in human object perception, crossmodal convergence might take place in the visual “what” pathway. However, it has been demonstrated in several experimental settings that the dominance of vision is not exclusive. If for example, the tactile input provides richer information, the tactile rather than the visual input will dominate. In this situation, multisensory convergence might occur in tactile rather than visual cortices. Indeed, some of the studies reviewed here showed VT effects in the parietal cortex (especially the IPS; Banati et al. 2000; Grefkes et al. 2002; Saito et al. 2003) and the right insula (Banati et al. 2000; Hadjikhani and Roland 1998; Prather et al. 2004). Multisensory responses in the parietal cortex were also reported in non-human primates (though their homology in humans is debated. For reviews, see Andersen et al. 1997; Colby and Goldberg 1999).

Another factor that could determine the recruitment of a certain brain region for processing of diverse sensory input might be the task demands of certain computations. This was suggested recently by Pascual-Leone and Hamilton (2001) in their “metamodal theory”. This would mean that the crucial aspect of processing is not the kind of sensory input the brain receives, but the contribution that a certain area makes and how the available information from the relevant senses is extracted for successful task performance.

It has recently been proposed that multisensory convergence zones are mainly located at the borders of modality-specific cortical areas (Wallace et al. 2004a). The location of the involved modality-specific regions is therefore another important determinant for multisensory convergence zones. The STS/STG is located anatomically between the visual and auditory “what” systems, which makes this region a suitable candidate for integrating auditory and visual information about objects.

A possible role of crossmodally triggered visual imagery

The literature reviewed here demonstrates that crossmodal AV effects often depend on previous associations between unimodal stimulus components. This suggests that crossmodal auditory activation of the ventral visual stream might be different from visual activation of these same regions. The responsiveness of visual cortices to auditory events or tactile objects might correspond to crossmodally triggered visual imagery (i.e. perceptual information accessed from memory; Kosslyn et al. 2001). Along these lines, Sathian et al. (1997) suggested that visual imagery is important for successful performance in tactile tasks.

However, the question arises whether this association of unimodal information is an epiphenomenon mediated by top-down processes, or is actually helpful in the recognition process. The latter would only be observed if unisensory cortices could interact directly before object recognition in unisensory cortices takes place. It has recently been shown (Falchier et al. 2002; Rockland and Ojima in press) that in the macaque areas V1 and V2 receive direct input from core and parabelt auditory areas. As connectivity is eccentricity-dependent (peripheral but not central areas receive these inputs), the function may be related to a foveation mechanism designed to reduce behavioral orientation response time to peripheral stimuli. In humans, functional connectivity analyses in a study not explicitly requiring imagery suggest that crossmodal AV coupling of highly familiar stimuli (familiar faces and voices) is subserved by a direct association of unimodal information (von Kriegstein et al. in press). This low-level association leaves open the possibility that even before recognition of the object is accomplished in one modality, information is combined. This effect might explain the behavioral interaction of multisensory information in recognition of unisensory stimuli (Ellis et al. 1997). This is further supported by recent studies in the tactile domain that showed a similar level of LOTv activation during

tactile object recognition, even in subjects that have had no visual experience before (i.e. congenitally blind subjects; Pietrini et al. 2004; Amedi et al. 2004).

A distributed model of semantic processing

Independent of their computational function, multisensory regions presumably interact with unisensory cortices via feedback connections and thus subserve a cortical network in which semantic knowledge about objects is represented in a distributed fashion (Martin and Chao 2001). According to this model, brain regions that process a specific type of incoming sensory information also serve as a knowledge store about the attributes of particular objects. For instance, a hammer has a number of salient attributes (such as a characteristic color, shape, sound). Ventral temporal cortex processes incoming information about visual form and color. If the incoming form and color match the stored hammer template, ventral temporal cortex identifies the object as a hammer. Lateral temporal cortex processes information about object motion. If the incoming motion information matches the stored template (e.g. up, down, impact) then the object is identified as a hammer. Frontal and parietal regions store information about the hand orientation and arm posture that would be required to grasp the hammer. The representations in these different regions are linked, so that simply seeing a photograph of a static hammer also activates the respective visual motion and motor representations (Beauchamp et al. 2002). Thus, networks of neurons in auditory association cortex that respond to an auditory object are also linked to visual, tactile, and motor information about the same object represented in temporal, parietal, and frontal regions.

The operation of multisensory regions during crossmodal object recognition remains to be determined. Mesulam (Mesulam 1998) suggests that associative areas contain "road maps" for binding distributed information in different modalities. In this view, associative areas do not contain multimodal representations but merely link the unimodal representations to make them accessible to each other. Contrary to this view, recent studies of crossmodal speaker recognition revealed enhanced functional connectivity between face responsive regions in extrastriate visual cortex and voice regions in anterior/middle STS without involvement of supramodal areas (von Kriegstein and Giraud 2004a; von Kriegstein et al. in press). Thus a direct interaction of these regions during familiar speaker recognition might rely on direct anatomical connections between lateral and ventral temporal regions of human cortex (Catani et al. 2003).

A possible explanation for the diverse findings regarding the involvement of multisensory associative regions may be that it is dependent on the type of information that has to be bound together. This view is supported by findings of different patterns of anatomical connectivity (feedforward vs. feedback) between unisensory and multisensory areas in primates (Falchier et al. 2002; Schroeder et al. 2003). Direct sharing of information between modalities, without recruitment of multisensory regions, might be only expected for information that is important to identify an object quickly and automatically. Examples would be alerting, orienting, and localizing responses, motion perception, and processing of socially important information (e.g. face/voice identity). On the contrary, more complex or arbitrarily related information (e.g. audio-visual speech, letters, and common objects) might be integrated through associative nodes (e.g. STS-MS) that are able to form more complex and flexible mappings between information from different modalities.

After placing the reviewed neuroimaging findings on unimodal and crossmodal object processing in the broader context of crossmodal and semantic processing theories, we propose the following tentative model for AV and VT object processing: Modality-specific representations in unisensory association cortices are linked either directly or through a heteromodal binding site depending on the information type and the goal or the requirements of the crossmodal combination. In case of a mediating heteromodal binding site, this probably reflects an associative node rather than a complex heteromodal object representation (Mesulam 1998). The location of crossmodal convergence, either within a unisensory system or in heteromodal association cortex, is determined by the dominance of one of the modalities, the relevant computational capabilities of a region, and the anatomical location of the uni- and multisensory cortices relative to each other.

To better understand the rapid computations performed during object perception, future studies should combine high-resolution fMRI with methods that allow examination of the temporal sequence of information processing (such as MEG/EEG or TMS). Such an approach should deepen our understanding of crossmodal object recognition and also shed light on unisensory mechanisms of object processing. The employment of more natural stimulation conditions (Bartels and Zeki 2004a; Bartels and Zeki 2004b; Hasson et al. 2004) should substantially contribute to human object recognition theories with increased ecological validity (de Gelder and Bertelson 2003).

Acknowledgements

This research was funded by a Horowitz Foundation fellowship (A.A.), the Bundesministerium für Bildung und Forschung (BMBF; K.v.K., M.J.N.), the Volkswagenstiftung (K.v.K.), and the Max Planck Society (M.J.N.). The authors thank Nikolas Francis, Axel Kohler (for help with the figures), Lotfi Merabet, Wolf Singer, Lars Muckli, and three anonymous reviewers (for their helpful comments on earlier versions of this paper).

References

- Adams RB, Janata P (2002) A comparison of neural circuits underlying auditory and visual object categorization. *Neuroimage* 16:361-377
- Amedi A, Jacobson G, Hendler T, Malach R, Zohary E (2002) Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cereb Cortex* 12:1202-1212
- Amedi A, Malach R, Hendler T, Peled S, Zohary E (2001) Visuo-haptic object-related activation in the ventral visual pathway. *Nat Neurosci* 4:324-330
- Amedi A (2004) Multisensory object-related processing in the visual cortex of sighted and its reversed hierarchical organization in blind humans. Presented at the 5th International Multisensory Research Forum in Sitges, Spain, Abstract #149
- Andersen RA, Snyder LH, Bradley DC, Xing J (1997) Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annu Rev Neurosci* 20:303-330
- Arnott SR, Binns MA, Grady CL, Alain C (2004) Assessing the auditory dual-pathway model in humans. *Neuroimage* 22:401-408
- Banati RB, Goerres GW, Tjoa C, Aggleton JP, Grasby P (2000) The functional anatomy of visual-tactile integration in man: a study using positron emission tomography. *Neuropsychologia* 38:115-124
- Bartels A, Zeki S (2004a) Functional brain mapping during free viewing of natural scenes. *Hum Brain Mapp* 21:75-85
- Bartels A, Zeki S (2004b) The chronoarchitecture of the human brain - natural viewing conditions reveal a time-based anatomy of the brain. *Neuroimage* 22:419-433
- Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41:809-823

- Beauchamp MS, Lee KE, Haxby JV, Martin A (2002) Parallel visual motion processing streams for manipulable objects and human movements. *Neuron* 34:149-159
- Belin P, Zatorre RJ (2000) 'What', 'where', and 'how' in auditory cortex. *Nat Neurosci* 3:965-966
- Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14:2105-2109
- Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. *Nature* 403:309-312
- Bernstein LE, Auer ET, Jr., Moore JK, Ponton CW, Don M, Singh M (2002) Visual speech perception without primary auditory cortex activation. *Neuroreport* 13:311-315
- Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD (2004) Neural correlates of sensory and decision processes in auditory object identification. *Nat Neurosci* 7:295-301
- Binkofski F, Buccino G, Posse S, Seitz RJ, Rizzolatti G, Freund H (1999) A fronto-parietal circuit for object manipulation in man: evidence from an fMRI-study. *Eur J Neurosci* 11:3276-3286
- Bodegard A, Geyer S, Grefkes C, Zilles K, Roland PE (2001) Hierarchical processing of tactile shape in the human brain. *Neuron* 31:317-328
- Bregman AS (1990) *Auditory scene analysis*. MIT Press, Cambridge, MA
- Burton AM, Bruce V, Johnston RA (1990) Understanding face recognition with an interactive activation model. *Br J Psychol* 81(Pt3):361-380
- Callan DE, Callan AM, Kroos C, Vatikiotis-Bateson E (2001) Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. *Brain Res Cogn Brain Res* 10:349-353

- Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E (2003) Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport* 14:2213-2218
- Calvert GA (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb Cortex* 11:1110-1123
- Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS (1999) Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10:2619-2623
- Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. *Science* 276:593-596
- Calvert GA, Campbell R (2003) Reading speech from still and moving faces: the neural substrates of visible speech. *J Cogn Neurosci* 15:57-70
- Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol* 10:649-657
- Calvert GA, Lewis JW (2004) Hemodynamic studies of audiovisual interactions. In: Calvert G, Spence C, Stein BE (Eds) *The handbook of multisensory processes*. MIT Press, Cambridge, MA, pp 483-502
- Catani M, Jones DK, Donato R, ffytche DH (2003) Occipito-temporal connections in the human brain. *Brain* 126:2093-2107
- Colby CL, Goldberg ME (1999) Space and attention in parietal cortex. *Annu Rev Neurosci* 22:319-49
- De Gelder B, Bertelson P (2003) Multisensory integration, perception and ecological validity. *Trends Cogn Sci* 7:460-467
- Deibert E, Kraut M, Kremen S, Hart J, Jr. (1999) Neural pathways in tactile object recognition. *Neurology* 52:1413-1417

- Dolan RJ, Morris JS, de Gelder B (2001) Crossmodal binding of fear in voice and face. *Proc Natl Acad Sci USA* 98:10006-10010
- Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. *Science* 293:2470-2473
- Easton RD, Srinivas K, Greene AJ (1997) Do vision and haptics share common representations? Implicit and explicit memory within and between modalities. *J Exp Psychol Learn Mem Cogn* 23:153-163
- Ellis HD, Jones DM, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. *Br J Psychol* 88:143-156
- Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. *Nature* 392:598-601
- Ernst M, Bühlhoff H (2004) Merging the senses into a robust percept. *Trends Cogn Sci* 8:162-169
- Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. *J Neurosci* 22:5749-5759
- Feinberg TE, Rothi LJ, Heilman KM (1986) Multimodal agnosia after unilateral left hemisphere lesion. *Neurology* 36:864-867
- Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1:1-47
- Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci* 3:191-197
- Gauthier I, Tarr MJ, Moylan J, Skudlarski P, Gore JC, Anderson AW (2000) The fusiform "face area" is part of a network that processes faces at the individual level. *J Cogn Neurosci* 12:495-504
- Gleitman LR, Rozin P (1977) The structure and acquisition of reading I: relations between orthographies and the structure of language. In: Reber A, Scarborough D (Eds) *Towards a psychology of reading: the proceedings of the CUNY conferences*. Lawrence Erlbaum Associates, Hillsdale, NJ

- Goodale MA, Meenan JP, Bulthoff HH, Nicolle DA, Murphy KJ, Racicot CI (1994) Separate neural pathways for the visual analysis of object shape in perception and prehension. *Curr Biol* 4:604-610
- Gorno-Tempini ML, Price CJ (2001) Identification of famous faces and buildings: a functional neuroimaging study of semantically unique items. *Brain* 124:2087-2097
- Gorno-Tempini ML, Price CJ, Josephs O, Vandenberghe R, Cappa SF, Kapur N, Frackowiak RS, Tempini ML (1998) The neural systems sustaining face and proper-name processing. *Brain* 121(Pt11):2103-2118
- Grefkes C, Weiss PH, Zilles K, Fink GR (2002) Crossmodal processing of object features in human anterior intraparietal cortex: an fMRI study implies equivalencies between humans and monkeys. *Neuron* 35:173-184
- Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. *Trends Neurosci* 25:348-353
- Grill-Spector K (2003) The neural basis of object perception. *Curr Opin Neurobiol* 13:159-166
- Grill-Spector K, Malach R (2004) The human visual cortex. *Annu Rev Neurosci* 27:649-677
- Hadjikhani N, Roland PE (1998) Cross-modal transfer of information between the tactile and the visual representations in the human brain: a positron emission tomographic study. *J Neurosci* 18:1072-1084
- Hashimoto R, Sakai KL (2004) Learning letters in adulthood: direct visualization of cortical plasticity for forming a new link between orthography and phonology. *Neuron* 42:311-322
- Hasson U, Harel M, Levy I, Malach R (2003) Large-scale mirror-symmetry organization of human occipito-temporal object areas. *Neuron* 37:1027-1041
- Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. *Science* 303:1634-1640

- Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425-2430
- Hoffman EA, Haxby JV (2000) Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat Neurosci* 3:80-84
- Iwamura Y (1998) Hierarchical somatosensory processing. *Curr Opin Neurobiol* 8:522-528
- James TW, Humphrey GK, Gati JS, Servos P, Menon RS, Goodale MA (2002) Haptic study of three-dimensional objects activates extrastriate visual areas. *Neuropsychologia* 40:1706-1714
- Jäncke L, Wüstenberg T, Scheich H, Heinze HJ (2002) Phonetic perception and the temporal cortex. *Neuroimage* 15:733-746
- Kaas JH, Hackett TA (1999) 'What' and 'where' processing in auditory cortex. *Nat Neurosci* 2:1045-1047
- Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E (2003) "Putting the face to the voice": matching identity across modality. *Curr Biol* 13:1709-1714
- Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* 17:4302-4311
- Kilgour AR, de Gelder B, Lederman SJ (2004) Haptic face recognition and prosopagnosia. *Neuropsychologia* 42:707-712
- Kilgour AR, Lederman SJ (2002) Face recognition by hand. *Percept Psychophys* 64:339-352
- Kosslyn SM, Ganis G, Thompson WL (2001) Neural foundations of imagery. *Nat Rev Neurosci* 2:635-642
- Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. *Science* 218:1138-1141

- Laurienti PJ, Wallace MT, Maldjian JA, Susi CM, Stein BE, Burdette JH (2003) Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Hum Brain Mapp* 19:213-223
- Leveroni CL, Seidenberg M, Mayer AR, Mead LA, Binder JR, Rao SM (2000) Neural systems underlying the recognition of familiar and newly learned faces. *J Neurosci* 20:878-886
- Lewis JW, Wightman FL, Brefczynski JA, Phinney RE, Binder JR, DeYoe EA (2004) Human brain regions involved in recognizing environmental sounds. *Cereb Cortex* AoP:May 27th
- Liberman AM (1992) The relation of speech to reading and writing. In: Frost R, Katz L (Eds) *Orthography, phonology, morphology and meaning*. Elsevier Science Publishers B.V., Amsterdam
- Macaluso E, George N, Dolan R, Spence C, Driver J (2004) Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21:725-732
- Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, Tootell RB (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc Natl Acad Sci USA* 92:8135-8139
- Martin A, Chao LL (2001) Semantic memory and the brain: structure and processes. *Curr Opin Neurobiol* 11:194-201
- McCandliss BD, Cohen L, Dehaene S (2003) The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn Sci* 7:293-299
- McGurk H, MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264:746-748
- Merabet L, Thut G, Murray B, Andrews J, Hsiao S, Pascual-Leone A (2004) Feeling by sight or seeing by touch? *Neuron* 42:173-179
- Mesulam MM (1998) From sensation to cognition. *Brain* 121(Pt6):1013-1052

- Mishkin M (1979) Analogous neural models for tactual and visual learning. *Neuropsychologia* 17(2):139-151
- Molholm S, Ritter W, Javitt DC, Foxe JJ (2004) Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. *Cereb Cortex* 14:452-465
- Morin P, Rivrain Y, Eustache F, Lambert J, Courtheoux P (1984) Visual and tactile agnosia. *Rev Neurol (Paris)* 140:271-277
- Munhall KG, Tohkura Y (1998) Audiovisual gating and the time course of speech perception. *J Acoust Soc Am* 104:530-539
- Murray MM, Michel CM, Grave de Peralta R, Ortigue S, Brunet D, Gonzalez AS, Schnider A (2004) Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage* 21:125-135
- Nakamura K, Kawashima R, Sato N, Nakamura A, Sugiura M, Kato T, Hatano K, Ito K, Fukuda H, Schormann T, Zilles K (2000) Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. *Brain* 123:1903-1912
- Naumer MJ, Petkova V, Havenith MN, Kohler A, Singer W, Muckli L (2004) Paying attention to multisensory objects. OHBM abstract #TH99
- Naumer MJ, Singer W, Muckli L (2002a) Audio-visual perception of natural objects. OHBM abstract #15600
- Naumer MJ, Wibrall M, Singer W, Muckli L (2002b) FMRI-studies of category-specific audio-visual processing – visual cortex. IMRF abstract #25
- Newell FN (2004) Cross-modal object recognition. In: Calvert G, Spence C, Stein BE (Eds) *The handbook of multisensory processes*. MIT Press, Cambridge, MA, pp 123-139
- Ohtake H, Fujii T, Yamadori A, Fujimori M, Hayakawa Y, Suzuki K (2001) The influence of misnaming on object recognition: a case of multimodal agnosia. *Cortex* 37:175-186

- Olson IR, Gatenby JC, Gore JC (2002) A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Res Cogn Brain Res* 14:129-138
- O'Sullivan BT, Roland PE, Kawashima R (1994) A PET study of somatosensory discrimination in man. Microgeometry versus macrogeometry. *Eur J Neurosci* 6:137-148
- Pascual-Leone A, Hamilton R (2001) The metamodal organization of the brain. *Prog Brain Res* 134:427-445
- Pascual-Leone A, Walsh V, Rothwell J (2000) Transcranial magnetic stimulation in cognitive neuroscience - virtual lesion, chronometry, and functional connectivity. *Curr Opin Neurobiol* 10:232-237
- Paulesu E, Perani D, Blasi V, Silani G, Borghese NA, De Giovanni U, Sensolo S, Fazio F (2003) A functional-anatomical model for lipreading. *J Neurophysiol* 90:2005-2013
- Pietrini P, Furey ML, Ricciardi E, Gobbini MI, Wu WH, Cohen L, Guazzelli M, Haxby JV (2004) Beyond sensory images: object-based representation in the human ventral pathway. *Proc Natl Acad Sci USA* 101:5658-5663
- Polk TA, Stallcup M, Aguirre GK, Alsop DC, D'Esposito M, Detre JA, Farah MJ (2002) Neural specialization for letter recognition. *J Cogn Neurosci* 14:145-159
- Polster MR, Rose SB (1998) Disorders of auditory processing: evidence for modularity in audition. *Cortex* 34:47-65
- Pons TP, Garraghty PE, Friedman DP, Mishkin M (1987) Physiological evidence for serial processing in somatosensory cortex. *Science* 237:417-420
- Prather SC, Votaw JR, Sathian K (2004) Task-specific recruitment of dorsal and ventral visual areas during tactile perception. *Neuropsychologia* 42:1079-1087
- Raij T, Uutela K, Hari R (2000) Audiovisual integration of letters in the human brain. *Neuron* 28:617-625

- Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc Natl Acad Sci USA* 97:11800-11806
- Reales JM, Ballesteros S (1999) Implicit and explicit memory for visual and haptic objects: cross-modal priming depends on structural descriptions. *J Exp Psychol Learn Mem Cog* 25:644-663
- Reed CL, Caselli RJ (1994) The nature of tactile agnosia: a case study. *Neuropsychologia* 32:527-539
- Reed CL, Shoham S, Halgren E (2004) Neural substrates of tactile object recognition: an fMRI study. *Hum Brain Mapp* 21:236-246
- Rockland KS, Ojima H (2003) Multisensory convergence in calcarine visual areas in macaque monkey. *Int J Psychophysiol* 50(1-2):19-26
- Roland PE, O'Sullivan B, Kawashima R (1998) Shape and roughness activate different somatosensory areas in the human brain. *Proc Natl Acad Sci USA* 95:3295-3300
- Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat Neurosci* 2:1131-1136
- Saito DN, Okada T, Morita Y, Yonekura Y, Sadato N (2003) Tactile-visual cross-modal shape matching: a functional MRI study. *Brain Res Cogn Brain Res* 17:14-25
- Sathian K, Zangaladze A, Hoffman JM, Grafton ST (1997) Feeling with the mind's eye. *Neuroreport* 8:3877-3881
- Schroeder CE, Smiley J, Fu KG, McGinnis T, O'Connell MN, Hackett TA (2003) Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *Int J Psychophysiol* 50:5-17
- Sekiyama K, Kanno I, Miura S, Sugita Y (2003) Auditory-visual speech perception examined by fMRI and PET. *Neurosci Res* 47:277-287

- Shah NJ, Marshall JC, Zafiris O, Schwab A, Zilles K, Markowitsch HJ, Fink GR (2001) The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. *Brain* 124:804-815
- Stein BE, Meredith MA (1993) *The merging of the senses*. MIT Press, Cambridge, MA
- Sumbly WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 26:212-215
- Thierry G, Giraud AL, Price C (2003) Hemispheric dissociation in access to the human semantic system. *Neuron* 38:499-506
- Tootell RB, Tsao D, Vanduffel W (2003) Neuroimaging weighs in: humans meet macaques in "primate" visual cortex. *J Neurosci* 23:3981-3989
- Ungerleider LG, Haxby JV (1994) 'What' and 'where' in the human brain. *Curr Opin Neurobiol* 4:157-165
- Ungerleider LG, Mishkin M (1982) Two cortical visual streams. In: Ingle DJ, Goodale MA, Mansfield RJW (Eds) *Analysis of visual behavior*. MIT Press, Cambridge, MA
- van Atteveldt N, Formisano E, Goebel R, Blomert L (2004) Integration of letters and speech sounds in the human brain. *Neuron* 43:271-282
- von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Res* 17:48-55
- von Kriegstein K, Giraud AL (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* 22:948-955
- von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud AL (in press) Interaction of face and voice areas during speaker recognition. *J Cog Neurosci*.
- Wallace MT, Ramachandran R, Stein BE (2004a) A revised view of sensory cortical parcellation. *Proc Natl Acad Sci USA* 101:2167-2172

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004b) Unifying multisensory signals across time and space. *Exp Brain Res* epub ahead of print: April 27th

Welch RB, Warren DH (1986) Intersensory interactions. In: Boff KR, Kaufman L, Thomas J (Eds) *Handbook of perception and human performance*. Wiley, New York

Wernicke C (1874) *Der aphasische Symptomenkomplex, eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert, Breslau

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb Cortex* 13:1034-1043

Zangaladze A, Epstein CM, Grafton ST, Sathian K (1999) Involvement of visual cortex in tactile discrimination of orientation. *Nature* 401:587-590

Zatorre RJ, Bouffard M, Belin P (2004) Sensitivity to auditory object features in human temporal neocortex. *J Neurosci* 24:3637-3642

Zeki SM (1978) Functional specialization in the visual cortex of the rhesus monkey. *Nature* 274:423-428

CHAPTER 2

Retinotopic effects during spatial audio-visual integration

Based on manuscript:

Meienbrock, A.* , Naumer, M.J.* , Doehrmann, O., Singer, W., Muckli, L. (2006).
Retinotopic effects during spatial audio-visual integration. *Neuropsychologia* [Epub
ahead of print]

*equal contribution

Abstract

For the successful integration of visual and auditory information it is important to identify whether visual and auditory signals originate from corresponding places in the external world. Here, we report crossmodal effects of spatially congruent and incongruent audio-visual (AV) stimulation. Visual and auditory stimuli were matched to one of four horizontal locations in external space. Seven healthy human subjects had to assess the spatial fit of a visual stimulus (i.e. a gray-scaled picture of a comic dog) and a simultaneously presented auditory stimulus (i.e. a barking sound). Employing functional magnetic resonance imaging (fMRI), we were able to reveal two distinct networks of cortical regions one preferentially processing spatially congruent the other preferentially processing spatial incongruent AV stimulation. Whereas earlier visual areas responded preferentially to incongruent AV stimulation higher visual areas of the temporal and parietal cortex and prefrontal regions responded preferentially to congruent AV stimulation (left inferior temporal gyrus [ITG], right posterior superior temporal gyrus/sulcus [pSTG/STS], left intra-parietal sulcus [IPS], left pre-central gyrus [PreCG], and left dorsolateral pre-frontal cortex [DLPFC]). A position-resolved analysis revealed three robust cortical representations for each of our four visual stimulus locations in retinotopic visual regions corresponding to the representation of the horizontal meridian in area V1 and at the dorsal and ventral borders between areas V2 and V3. While these regions of interest (ROIs) did not show any significant effect of spatial congruency, we found subregions within ROIs in the right hemisphere that showed an incongruency effect (i.e. an increased fMRI-signal during spatially incongruent as compared to congruent AV stimulation). We interpret this finding as a correlate of spatially disperse recurrent feedback during mismatch processing: whenever a spatial mismatch is detected in multisensory regions (such as the IPS) processing resources are re-directed to low-level visual areas.

Introduction

Perceiving an object can involve information from different modalities, thereby engaging several streams of processing. Although these streams operate at least in early stages largely independently, visual, auditory, and tactile cues are eventually integrated to form a unified percept of the object. From the large set of possible parameters, which might facilitate this kind of multisensory integration, research has focused on temporal synchronicity, as well as spatial and semantic congruency as the crucial dimensions (see e.g. Calvert and Thesen 2004). Unlike ‘natural’ objects, which are usually characterized by a high coincidence of these factors, experimental setups aim to investigate the individual impact of these stimulus dimensions. Here, we study the variation of spatial audio-visual (AV) congruency.

The importance of “temporal” and “spatial” congruency was investigated in animal studies using invasive electrophysiology (for an early review see Stein and Meredith, 1993). These pioneering experiments revealed several features of particular neurons - so-called *multisensory integrative* (MSI) *cells* - whose activity is assumed to play a key role for multisensory integration. The most striking among their features is the super-additive response property, i.e. a response to (coincident and/or congruent) bimodal stimulation that clearly exceeds the linear sum of the neuron’s responses to the respective unimodal stimuli. MSI cells have been found in a number of species and in both cortical (Wallace et al., 2004) as well as sub-cortical brain structures as early as the superior colliculus (Meredith and Stein, 1996).

In humans, spatial AV integration has been investigated predominantly with psychophysical (e.g. McDonald et al. 2000) and electrophysiological methods such as EEG and MEG (for a review see Eimer and Driver 2001). Recent ERP findings (Teder-Sälejärvi et al. 2005) point towards different cortical activation patterns for spatially congruent versus incongruent AV stimuli. Due to the limited spatial resolution of EEG/MEG, it has remained unsettled yet, which cortical sites are involved in processing these stimuli. We used functional magnetic resonance imaging (fMRI), which offers the means to overcome these difficulties.

Several imaging studies have identified different general networks of cortical regions that are involved in semantic (Amedi et al. 2005) and spatial crossmodal integration (Macaluso and Driver 2005). Till now, most imaging studies of spatial crossmodal integration have employed visuo-tactile stimuli, while similar

investigations in the AV domain are relatively rare. In a recent positron emission tomography (PET) study Macaluso et al. (2004) varied both the spatial congruency and temporal synchronicity of auditory (spoken word lists) and visual stimuli (faces mouthing these same words). Spatially incongruent blocks were associated with activity in more dorsal areas while ventral occipital areas and the superior temporal sulcus (STS) were unaffected by relative location. In addition, lateral and dorsal occipital areas were selectively activated for synchronous bimodal stimulation at the same external location. The spatial resolution of PET might have been insufficient to detect differential effects in retinotopically organized visual areas. The focus of our study was the investigation of spatial congruency and incongruency effects on retinotopic visual areas by presenting auditory and visual stimuli at two different locations within each external hemi-space.

Methods

Subjects

Seven subjects (four female, one left-handed) participated in the fMRI experiment; their mean age was 26.0 years (range 23 to 33 years). All subjects had normal or corrected-to-normal vision (four subjects). All participants received information on MRI and a questionnaire to check for potential health risks and contraindications. Volunteers gave their informed consent after having been introduced to the procedure in accordance with the declaration of Helsinki.

Stimuli

The stimulus image signal was generated by a notebook (DELL 7500, 750 MHz) at a frame rate of 60 Hz. The image was projected onto a vertical screen positioned inside the scanner with an LCD projector (Sony, VPL PX 20) equipped with a custom-made lens. Subjects viewed the screen through a mirror. Mirror and projection screen were fixed onto the head coil. The subjects' field of view was 50 degrees visual angle (maximum horizontal distance).

Figure 1. Experimental design. Audio-visual (AV) stimulation consisted of gray-scaled pictures of a comic dog and barking sounds. Auditory and visual stimuli were presented synchronously, but could be either spatially congruent (**A**) or incongruent (**B**), depending on whether the unimodal stimulus components were presented at the same or at two different locations. In the incongruent conditions, the auditory stimulus was displaced from the visual stimulus by 20 degrees visual angle towards the contralateral hemi-space. (**C**) Illustrates the four locations in external space where stimuli could occur. We employed a block design with eight experimental conditions (i.e. two degrees of spatial congruency for each of four locations). Size of laterally and medially presented comic dogs differed, in order to compensate for the cortical magnification factor (Sereno et al. 2005). Information on sizes and positions of the visual stimuli is provided in **Table 1**.

The visual stimulation (**Fig. 1**) consisted of four gray-scaled pictures of comic dogs, one located peripherally and one centrally in each visual hemifield. They were adjusted in size to compensate for the cortical magnification factor (Sereno et al. 1995). All dogs were positioned along the horizontal meridian. Information on sizes and positions of the visual stimuli is provided in **Table 1**. In the center of the screen, a fixation cross was positioned, which was present during the whole experiment. The screen was of blue color to prevent subjects from being glared. Employing a classical block design, only one dog was present at a time during stimulation blocks. The dogs were painted in black and white and inverted their contrast every 100 ms (“checkerdog”) in order to obtain maximal visual stimulation. Each of eight experimental conditions (lasting 12 s per stimulation block) was repeated four times

during the experiment in a pseudo-randomized order, alternating with fixation periods (20 s). The experiment thus consisted of 32 stimulation periods alternating with 33 fixation periods, and lasted 17min and 20s.

Table 1. Sizes and positions of the visual stimuli

	L	LM	RM	R
length	11°	3°	3°	11°
height	7.3°	2°	2°	7.3°
horizontal position	-16.5°	-3.5°	+3.5°	+16.5°

Abbreviations: left [L], left medial [LM], right medial [RM], and right [R] locations

Auditory stimuli were generated on the same notebook and passed through an amplifier (JVC AX R5BK). Subjects received these sounds through MR-compatible headphones, which were fitted into an earmuff (Bilsom 727, sound insulation of 33 dB at 1000 Hz). We built these headphones by removing the permanent magnets of standard stereo headphones (impedance of 32 ohm, frequency ranging from 20 till 20000 Hz) and using the external magnetic field of the MR tomograph. Sound quality was in general superior to conventional MR-compatible stimulation systems working with sound transmission based on tubes.

The auditory stimulation consisted of four spatially different barking sounds, mapped on the four visual stimulus locations by varying the amplitude of the sounds presented to the left and the right ear, respectively. The spatial fit to the four locations was assessed in a separate behavioral experiment prior to the fMRI experiment. Here, three of our subjects judged 20 different sounds with respect to their spatial fit to the four visual stimulus locations. These 20 sounds were created by using a wave-file of a barking dog (with equal volume for both channels) and adjusting the volume of the right and left channels systematically in steps of 5% in opposite directions.

This resulted in intracranial percepts ranging from the very left to the very right (Zimmer et al. 2004). Interaural time difference remained unchanged. The four sounds rated to be located closest to each of the stimulated visual locations were used for the fMRI experiment. Stimulation during each of our experimental conditions consisted of the respective barking sound (length ~1s) that was repeated twelve times per block. During the experiment, each of the four auditory stimuli was presented during eight blocks of the same experimental run. In half of these cases resulting AV stimulation was spatially congruent (**Fig. 1A**). During spatially

incongruent stimulation (**Fig. 1B**), the distance between the respective auditory and visual stimulus locations remained constant and amounted to 20 degrees of visual angle. The auditory stimuli were displaced towards the respective contralateral hemi-space. Auditory and visual stimulation onsets were synchronous.

Procedure

Subjects were instructed to pay attention to the experimental stimuli while maintaining central fixation in order to minimize eye-movements. The subjects' task was to indicate via button-press, whether AV stimulation was perceived to be spatially congruent or incongruent.

Behavioral training

In a separate training session outside the tomograph, subjects were seated in a chair, wearing conventional stereo headphones (Philips SBC HS 400/00), and head position was restrained by a chin rest. Distance between the subject's eyes and the screen of the notebook (DELL 7500, 750 MHz) was 33 cm, resulting in the same visual angle of the visual stimulation as inside the tomograph. Subjects had to decide about the spatial congruence of the AV stimulation and were trained until individual performance was well above chance level.

Control of eye movements

To control for confounding stimulus-related eye movements, we used an infrared eye tracker to monitor eye movements in three of our subjects during the behavioral training session outside the tomograph. We found no detectable pattern of eye movements correlated with the locations of visual stimulation.

Imaging

fMRI data acquisition was performed using a 1.5 Tesla Siemens Magnetom Vision tomograph (Siemens, Erlangen, Germany) at the Institute of Neuroradiology, University Hospital, Johann Wolfgang Goethe-University, Frankfurt am Main. A gradient-recalled echo-planar-imaging (EPI) sequence was used with the following parameters: 16 slices, oriented approximately in parallel to the AC-PC plane (AC, anterior commissure; PC, posterior commissure); TR, 2081 ms; TE, 69 ms; FA, 90°; FOV, 200 mm; in-plane resolution, 3.13 x 3.13 mm²; slice thickness, 5 mm; gap thickness, 1 mm. In addition, a detailed T1-weighted anatomical scan was acquired for all subjects using a Siemens fast low-angle-shot (FLASH) sequence (isotropic voxel size 1 mm³). For each subject, a magnetization-prepared rapid-acquisition

gradient-echo (MP-RAGE) sequence was used (TR = 9.7 ms, TE = 4 ms, FA = 12°, matrix = 256 x 256, voxel size 2.0 x 1.0 x 1.0 mm³) for realignment with the detailed anatomical scan that had been measured in a previous session.

Data Analysis

Data were analyzed using the BrainVoyager™ 2000 (version 4.9) and BrainVoyager™ QX (version 1.6) software packages (Rainer Goebel, Brain Innovation, Maastricht, The Netherlands, <http://www.brainvoyager.com>). The first four volumes of each scan were discarded to preclude T1 saturation effects. Preprocessing of the functional data included the following steps: (i) three-dimensional motion correction, (ii) linear-trend removal and temporal high-pass filtering at 0.01 Hz, (iii) slice-scan-time correction with sinc interpolation, and (iv) the alignment of individual cortices to one another using an algorithm accounting for an optimal fit of the main gyrification with minimal distortion between the individual cortices (see below). Cortex-based statistical analysis was performed using multiple linear regression. For every cortical surface vertex, the time course was regressed on a set of dummy-coded predictors representing our eight experimental conditions. To account for the shape and delay of the hemodynamic response (Boynton et al. 1996), the predictor time courses (box-car functions) were convolved with a gamma function. A contrast-based analysis was performed, using a t-statistics. We thereby defined sets of predictors (experimental conditions) for each of our subjects individually (so-called “separate subject predictors”), and employed all of them during multiple linear regression to obtain our group-statistics.

In a first step, we revealed the general cortical network involved in spatial AV integration. Then, applying a position-resolved analysis, we further searched for modulatory effects in retinotopically organized low-level visual cortices. During the initial location-spanning analysis, we compared all four spatially congruent to all four incongruent AV conditions (resulting contrasts: *congruent* > *incongruent* and *incongruent* > *congruent*, respectively). Vertices were included into the statistical maps if the obtained p-value was < 0.01 (uncorrected). We performed separate analyses based on either all trials (“physically” congruent vs. incongruent) or correct trials only (“corrected” analyses). In a second step, we performed a balanced contrast of one spatial location with the other three locations. Vertices were included into the statistical map if the obtained p-value was < 0.0001 (corrected). Finally, we contrasted the spatially congruent and incongruent conditions for each of the four

locations, separately. Vertices were included into the respective statistical map if the obtained p-value was < 0.001 (uncorrected).

During analysis, statistical maps of group data were superimposed on inflated hemispheres of one subject (DL). The response profiles for each of our regions-of-interest (ROIs) were visualized by plotting either the event-related BOLD-signal averages or the GLM beta weights for the different experimental conditions.

Retinotopic Mapping

Phase-encoded retinotopic mapping was assessed in each subject and included mapping of eccentricity and polar angle (Engel et al. 1994; Sereno et al. 1995; Goebel et al. 1998; Muckli et al. 2005). In the eccentricity mapping experiment, black and white checkerboard patterns were presented in a ring-shaped configuration and were flickered at a rate of 4 Hz. The ring started with a radius of 1° and slowly expanded to a radius of 12° visual angle within 64 s. In the polar-angle mapping experiment, the checkerboard pattern consisted of a ray-shaped disk segment subtending 22.5° of polar angle. The ray started at the right horizontal meridian and slowly rotated clockwise for a full cycle of 360° (within 64 s). Each mapping experiment consisted of seven repetitions of a full expansion (each cycle lasting for 64 s) or ten repetitions of rotation, respectively.

The analysis of the retinotopic-mapping experiment was conducted by the use of a cross-correlation analysis. We used the predicted hemodynamic signal time course for the first 1/8 of a stimulation cycle (corresponding to 45° visual angle in the polar mapping experiment) and shifted this reference function successively in time (time steps correspond to the recording time for one volume, TR; see also Muckli et al. 2005). Sites activated at particular eccentricities and polar angles were identified through selecting the lag value that resulted in the highest cross-correlation value for a particular voxel. The obtained lag values at particular voxels were encoded in pseudocolor on corresponding surface patches (triangles) of the reconstructed cortical sheet. Based on the polar-angle mapping experiment, the boundaries of retinotopic cortical areas V1, V2, V3, V3A, and V4v were estimated manually on the inflated cortical surface.

Cortical-Surface Reconstruction and Visualization

The high-resolution T1-weighted 3D recordings were used for surface reconstruction of both cerebral hemispheres of each subject (Kriegeskorte and Goebel 2001). The

white/gray-matter border was segmented with a region-growing method preceded by inhomogeneity correction of signal intensity across space. The borders of the two resulting segmented subvolumes were tessellated to produce a surface reconstruction of each hemisphere.

Cortex-based inter-subject alignment

To improve the spatial correspondence between subjects' brains beyond Talairach space matching, the reconstructed hemispheres were aligned using curvature information reflecting the gyral/sulcal folding pattern (see van Atteveldt et al. 2004 for details). As demonstrated recently for a similar analysis tool (Argall et al., 2006), such a surface-based approach can help to improve t-statistics in the average activation map as compared to the volume average.

Folded cortical representations of each subject and hemisphere were morphed into a spherical representation that provided a parameterizable surface for nonrigid alignment across subjects. The curvature information of the folded representation was preserved as a curvature map on the spherical representation. This folding pattern was smoothed along the sphere surface to provide spatially extended gradient information driving intersubject alignment. Following a coarse-to-fine matching strategy, the alignment started with highly smoothed curvature maps and progressed to only slightly smoothed representations. While the alignment of major gyri and sulci was achieved reliably using this method, smaller structures, reflecting idiosyncratic differences between the subjects' brains, were not completely aligned.

Cortex-based inter-subject alignment enabled us to align the time courses for multisubject GLM data analysis. Group-averaged functional data were then projected on inflated representations of the left and right cerebral hemispheres of a single subject (DL; see Fig. 2A). As a morphed surface always possesses a link to the folded reference mesh, functional data can be shown at the correct location of folded as well as inflated representations. This link was also used to keep geometric distortions to a minimum during inflation through inclusion of a morphing force that keeps the distances between vertices and the area of each triangle of the morphed surface as close as possible to the respective values of the folded reference mesh.

Results

At the behavioral level, subjects on average had classified 85.5 % of all trials correctly, but incongruent trials (92.2 % correct) were detected with a substantially higher accuracy than congruent trials (57.5 % correct). As subjects' performance was found to be markedly worse in the congruent condition, we conducted sensitivity analyses (d' analyses) for each of our subjects. **Table 2** lists these individual d' -scores that ranged between 0.68 and 3.13 (with a mean d' [$n=6$] of 1.83). Three subjects (BV, GS, and MN) performed rather weak and reached only low sensitivity values ($d' < 1.3$). Moreover it seems that these subjects were employing a conservative response criterion that prevented a positive (i.e. 'congruent') response except in cases of high certainty. In order to use perceptually pure conditions we only used certified trials in later steps of data analysis ('corrected').

Table 2. Behavioral sensitivity data (individual d' analyses)

subject	d'
AF	2.45
BV	1.05
CM	3.13
DL	2.45
GS	1.22
MN	0.68
n=6	mean=1.83

Presentation of our AV stimuli activated a distributed network of cortical regions in the occipital, temporal, parietal, and frontal lobes (**Fig. 2**). In a location-spanning analysis, we could reveal two independent networks whose activation reflected the degree of spatial AV congruency and incongruency in a complementary manner. Statistical maps for the contrast *congruent* > *incongruent* AV stimulation were based on multiple regression of group-averaged data, likewise for the contrast *incongruent* > *congruent*. Hereby, separate analyses were based either on all trials (7 subjects; **Fig. 2A**) or on correct trials only (6 subjects; **Fig. 2B**) and resulted in largely similar cortical activation maps ($P < 0.01$, uncorrected).

Figure 2. Cortical regions sensitive to spatial congruency. **(A)** Analysis based on physical stimulation. Inflated representations of one subject's cerebral hemispheres in lateral (lateral panels) and posterior views (medial panels). Statistical maps for the contrasts (physically) congruent > incongruent (in orange) and (physically) incongruent > congruent AV stimulation (in light blue) were based on multiple regression of group-averaged ($n = 7$) data. **(B)** Analysis based on correct trials only. Inflated representations of one subject's cerebral hemispheres in lateral (lateral panels) and posterior views (medial panels). Statistical maps for the contrasts congruent > incongruent (in orange) and incongruent > congruent AV stimulation (in light blue) were based on multiple regression of group-averaged ($n = 6$) data. **(C)** Shows the mean BOLD signal intensity changes during perception of spatially congruent (in orange) and incongruent AV stimulation (in light blue) for eight selected cortical regions from **(B)**. All maps were significant at $P < 0.01$ (uncorrected).

Regions showing a preference for spatial AV congruency (see orange-colored regions in **Fig. 2**) included left inferior temporal gyrus (ITG), intra-parietal sulcus (IPS), pre-central gyrus (PreCG), and dorsolateral pre-frontal cortex (DLPFC), and right posterior superior temporal gyrus/sulcus (pSTG/STS). On the other hand,

retinotopic low-level visual regions (including V1) in bilateral occipital cortex showed a significantly higher BOLD-signal during spatially incongruent as compared to congruent AV stimulation (see light blue-colored regions in **Fig. 2**).

In a second step, our goal was to find separable foci of BOLD activation that corresponded to the four different locations along the horizontal meridian of the visual field. In order to carry out an analysis of single-subject data, we performed a retinotopic mapping procedure (see Methods for details) and defined the borders of retinotopic low-level visual areas for each subject individually. The horizontal meridian (i.e. where we presented our experimental stimuli) is represented at three distinct locations within human retinotopic visual areas V1-V3: in the middle of V1, at the border between dorsal V2 and V3, and between ventral V2 and V3. Using a surface-based analysis of the individual data sets, we identified those regions showing a BOLD-signal increase in response to our visual stimuli and overlapping with the respective cortical representations of the horizontal midline. We were able to separate foveal from peripheral activation by contrasting the response to each of our four visual stimulus locations [Left (L), Left-middle (LM), Right-middle (RM), Right (R)] and pooling the respective congruent and incongruent conditions for each of the locations. The mapping results are shown based on group-averaged data ($n = 7$) in **Fig. 3**. For each of our visual stimulus locations, cortex-based group analysis revealed three distinct representations in the contralateral hemisphere. One of these representations consisted of surface patches extending from the border between dorsal V2 and V3 to the border between ventral V2 and V3, including parts of V1. The two other representations were located more laterally in LO (tentatively labeled *lateral occipital*) and dorsally around the IPS (see lower panels in **Fig. 3** for the respective response profiles). While all these ROIs showed a highly significant location preference ($P < 0.0001$, corrected), none of them showed a significant preference for spatially congruent or incongruent AV stimulation at the respective location of the visual field.

In a final step, we thus directly contrasted the spatially congruent and incongruent experimental conditions for each of our four visual stimulus locations, separately. Only in the right hemisphere, we were able to reveal subdivisions within the above reported visual location ROIs that showed a significant BOLD-signal increase in response to spatially incongruent as compared to congruent AV stimulation ($P < 0.001$, uncorrected; **Fig. 4**).

Figure 3. Mean cortical representations of the four visual stimulus locations. Group-averaged ($n = 7$) location maps were projected on inflated representations of one subject's cerebral hemispheres (DL; shown in upper panels; view from posterior). Colors code for the different locations as follows: yellow, right (R); red, right medial (RM); blue, left medial (LM); green, left (L). For each of these locations, analysis revealed three distinct representations in the contralateral hemisphere. One of these representations consisted of surface patches extending from the border between dorsal V2 and V3 to the border between ventral V2 and V3, including parts of V1. The two other representations were located more laterally in LO (tentative label for *lateral occipital*) and dorsally around the posterior intraparietal sulcus (IPS). For these ROIs, the mean BOLD signal intensity changes (beta weights) are shown in response to each of our eight experimental conditions (lower panels). While all ROIs showed a highly significant location preference ($P < 0.0001$, corrected), none of them showed a significant preference for spatially congruent or incongruent AV stimulation.

Figure 4. Retinotopic effects of spatial AV incongruency (right hemisphere only). **(A)** Group-averaged ($n = 7$) location maps ($P < 0.05$, corrected) were projected on inflated representations of one subject's right cerebral hemisphere (as viewed from posterior). Colors code for the two different locations within the contralateral visual hemi-field as follows: blue, left medial (LM); green, left (L). In addition, group-averaged ($n = 6$) incongruency maps (as shown in lighter shades of the same colors; contrast: spatially *incongruent* > *congruent*; $P < 0.001$, uncorrected) have been computed for each of the two locations, separately. These incongruency maps (based on analyses of correct trials) partially overlapped with the respective location maps. **(B)** For each of the overlapping regions from **(A)** (circles) mean BOLD signal intensity changes are shown. Color code for the graphs: light blue: LM incongruent; dark blue: LM congruent; light green: L incongruent; dark green: L congruent.

Discussion

Employing BOLD-fMRI, we were able to reveal two distinct networks of cortical regions preferentially processing either spatially congruent (left ITG, IPS, PreCG, DLPFC, and right pSTG/STS) or spatially incongruent AV stimulation (bilateral

retinotopic low-level visual areas). A position-resolved analysis revealed robust cortical representations for each of our four visual stimulus positions in contralateral visual areas V1, V2d/V3d, V2v/V3v, LO, and around posterior IPS. While these ROIs did not show any significant effect of spatial congruency, we found subregions within right hemisphere ROIs that showed a spatial AV incongruency effect (i.e. a higher BOLD-signal during spatially incongruent as compared to congruent AV stimulation).

Effects of spatial AV congruency

Our findings concerning spatial congruency involve two networks of cortical regions. At a lower level, we found activation in left IPS and right pSTG/STS, which is consistent with a number of other neuroimaging findings (see Macaluso and Driver 2005; Amedi et al. 2005; Beauchamp 2005 for recent reviews). As these areas are considered to be important heteromodal (or multisensory) integration sites, they are likely to detect and gain from spatial congruency between semantically related auditory and visual stimuli.

The second network, operating at a higher processing level is constituted by frontal regions. At first glance, activation here could be due to a higher effort of the subjects during detection of *spatial congruency*, which might have resulted in a higher amount of task-directed attention resources (Corbetta and Shulman 2002). This would be in line with our subjects' behavioral performance differing significantly between the spatially congruent and incongruent conditions.

Spatial AV incongruency effects

A main result of our study was the observed effect of spatially incongruent AV stimuli in low-level visual regions. Although these regions might not be involved in detecting the spatial mismatch of AV stimulation, they might be affected as a consequence of this detected mismatch. Higher responses to spatially incongruent AV stimuli can not be explained by intensity differences, since the compared experimental conditions consist of nearly identical visual and auditory stimuli. The auditory stimuli were counterbalanced with respect to the different locations so each auditory stimulus that occurs as congruent for one position occurs also as an incongruent stimulus for another position. While the activation of the aforementioned network of frontal regions (during spatially congruent AV stimulation) might indicate an employment of task-directed attention resources (Corbetta and Shulman 2002), low-level visual areas show an AV incongruency effect. If it were the case that both networks

participate in the same task-related operations, we should have found rather a congruency effect in low-level visual areas – which we have not.

In addition, our analyses revealed that subjects exhibited neural response patterns in low-level visual cortices that differed between the left and right hemispheres. While we found a rather robust retinotopic AV incongruency effect in the right hemisphere (as shown in **Fig. 4**), we were not able to reveal a similar incongruency pattern in the left hemisphere (at least when applying the same statistical thresholds). We assume that this mainly reflects the fact that subjects made substantially more errors during physically congruent AV stimulation within the left (58.3% incorrect) as opposed to the right hemi-space (29.2% incorrect). In other words: subjects had a stronger tendency to perceive our AV stimuli as being spatially incongruent during stimulation within the left hemi-space.

In sum, we conclude that the activation of retinotopic low-level visual areas can neither be explained by stimulus energy differences nor by the specific behavioral task requirements. Thus, the variation of spatial AV congruency is the most likely factor to cause this finding.

Which mechanism might actually be involved in the processing of spatially incongruent multisensory stimuli? We favor the hypothesis that the activation of low-level visual areas reflects rather automatic and rapid processes to resolve a spatial mismatch that has been detected at the level of heteromodal (or multisensory) regions (such as IPS and pSTG/STS). It is likely that these processes are controlled via feedback (Mesulam 1998) from the aforementioned two networks – directly by the heteromodal regions and rather indirectly by the supramodal frontal network. This interpretation is supported by findings of studies investigating the timing of crossmodal spatial integration. McDonald et al. (2003) found a modulation of visual processing by a spatially varying auditory cue after 120-140 ms in the superior temporal cortex and 15-25 ms later in ventral occipital cortex. The authors conclude that this temporal pattern might reflect an attentional modulation of visual regions by recurrent feedback originating in heteromodal lateral temporal regions, which is in line with our interpretation. What is more, spatial AV incongruency could be interpreted as a salient stimulus, which automatically attracts attention resources to the visual domain (Corbetta and Shulman 2002; Kastner 1999). To resolve the detected incongruency with high speed, the involvement of low- to mid-level stages

of processing appears to be advantageous. This level could be constituted by the reported network involving low-level visual areas and left IPS and right pSTG/STS.

However, it remains the question why it is the visual cortex, which seems to have such a prominent role in the processing of spatial incongruency. With regard to modality dominance, research has generally found that vision is dominant over audition for spatial perception (*information reliability hypothesis*; Schwartz et al. 1998). In addition, *modality appropriateness* to the respective task (Warren 1979; Welch and Warren 1980) determines the direction of the crossmodal effect. In our case of spatial AV mismatch, the superior spatial resolution of low-level visual areas offers the most reliable information for successful conflict resolution.

Thus, our favored interpretation can be summarized as follows: Feed-forward processing involves unisensory auditory and visual cortices and is reduced whenever a spatial mismatch between visual and auditory stimuli has been detected at the level of multisensory regions (such as the IPS and pSTG/STS). In such a case the superior spatial resolution of the visual system is fully utilized via a feedback loop (down to V1) to re-direct attention resources to resolve the detected incongruency. This resulting interaction between low- and mid-level regions can be further modulated by frontal regions (e.g. DLPFC and PreCG), which enable the adaptation to the requirements of a specific task.

However, as 7 subjects is a rather low number, one has to be cautious with a very strong generalization of our findings as reported above. The revealed mechanism is at first only specific for our particular experimental conditions. There are several explanations possible for our data. Whether our explanation holds in different experimental conditions and whether the cortical networks described here are typical for different experimental variations remains to be shown in future experiments.

Acknowledgements

This research was funded by the German Federal Ministry for Education and Research (BMBF) and the Max Planck Society (MPS). We thank Lutz Fiesser (for helping to develop the auditory stimulation setup), Axel Kohler, Christian Altmann, and two anonymous reviewers (for helpful comments on earlier versions of this manuscript).

References

- Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, 166, 559–571.
- Argall, B. D., Saad, Z. S., & Beauchamp, M. S. (2006). Simplified intersubject averaging on the cortical surface using SUMA. *Hum Brain Mapp*, 27(1), 14-27.
- Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, 15(2), 145-153.
- Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13), 4207-4221.
- Calvert, G. A., & Thesen, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal Physiology Paris*, 98(1-3), 191-205.
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201-215.
- Eimer, M., & Driver, J. (2001). Crossmodal links in endogenous and exogenous spatial attention: Evidence from event-related brain potential studies. *Neuroscience and Biobehavioral Reviews*, 25(6), 497-511.
- Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., et al. (1994). FMRI of human visual cortex. *Nature*, 369(6481), 525.
- Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience*, 10(5), 1563-1573.
- 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.

- Kriegeskorte, N., & Goebel, R. (2001). An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes. *Neuroimage*, 14(2), 329-346.
- Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology*, 75(5), 1843-57.
- Muckli, L., Kohler, A., Kriegeskorte, N. Singer, W. (2005). Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS Biology*, 3(8), e265.
- Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neuroscience*, 28(5), 264-271.
- Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: A PET study. *Neuroimage*, 21(2), 725-732.
- McDonald, J. J., Teder-Sälejärvi, W. A., Di Russo, F., & Hillyard, S. A. (2003). Neural substrates of perceptual enhancement by cross-modal spatial attention. *Journal of Cognitive Neuroscience*, 15(1), 10-19.
- McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, 407(6806), 906-908.
- Mesulam, M. M. (1998). From sensation to cognition. *Brain*, 121 (Pt 6), 1013-1052.
- Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell (Ed.), *Hearing by eye: The psychology of lipreading* (pp. 3-51). Hove, U.K.: Lawrence Erlbaum Associates.
- Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., et al. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889-893.
- Stein, B. E., & Meredith, M. A. (1993). The merging of the senses. MIT Press,

Cambridge, MA.

- Teder-Sälejärvi, W.A., Di Russo, F., McDonald, J.J. & Hillyard, S.A. (2005). Effects of spatial congruity on audio-visual multimodal integration. *Journal of Cognitive Neuroscience*, 17, 1396-1409.
- van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, 43(2), 271-282.
- Wallace, M. T., Ramachandran, R., & Stein, B. E. (2004). A revised view of sensory cortical parcellation. *Proc Natl Acad Sci U S A*, 101(7), 2167-72.
- Warren, D. H. (1979). Spatial localization under conflict conditions: Is there a single explanation? *Perception*, 8(3), 323-337.
- Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88(3), 638-667.
- Zimmer, U., Lewald, J., Erb, M., Grodd, W., & Karnath, H. O. (2004). Is there a role of visual cortex in spatial hearing? *European Journal of Neuroscience*, 20(11), 3148-3156.

CHAPTER 3

Natural sounds activate object-related visual cortex

Based on manuscript:

Naumer, M.J., Wibral, M., Singer, W., Muckli, L. (2006). Natural sounds activate object-related visual cortex. *Submitted*.

Abstract

We investigated audio-visual object recognition using functional magnetic resonance imaging. We found distinct peak activations for diverse object categories (sportsmen, animals, cars, and tools) in ventral occipito-temporal cortex. These regions showed a category preference not only during unimodal visual stimulation but also during unimodal auditory stimulation. As we could rule out visual imagery as a potential confound, this might reflect learned associations between frequently co-occurring visual and auditory object features.

Introduction

We perceive most common objects (for example, a hammer) via several senses simultaneously. Vision and audition are very often involved in this process. Although the respective visual and auditory object features are radically different, we experience audio-visual (AV) objects in a coherent fashion. What is the neural basis of this crossmodal integration? To address this issue in humans, we used functional magnetic resonance imaging (fMRI).

Several fMRI studies have recently reported object-related crossmodal convergence. Visuo-tactile convergence was observed in posterior dorsal parts of the lateral occipital complex (LOC), a large region of the visual 'what' pathway (Amedi et al. 2002; Pietrini et al. 2004). This pathway is dedicated to the perception of form and objects. According to the prevailing interpretation of the above data, natural sounds do not activate the visual 'what' pathway, because human form processing exclusively relies on vision and touch (Amedi et al. 2002). However, ventral occipito-temporal (VOT) regions of this pathway are largely invariant to the exact physical stimulus features and are sensitive to context information and object category (Grill-Spector and Malach 2004; Cox et al. 2004). Most studies that employed natural sounds did not allow testing for category-related AV convergence, because they either used an object category-spanning design (Amedi et al. 2002; Beauchamp et al. 2004) or stimulated only with objects of a single category.

For these reasons we examined the hypothesis that auditory object processing involves visual regions in VOT. We used a design that allowed us to compare fMRI activations in response to different categories of purely auditory stimuli. In each of three independent main experiments (experiments 1-3) we presented sounds and gray-scaled pictures of two object categories in separate stimulation blocks (exp. 1: sportsmen and animals; exp. 2: tools and cars; exp. 3: animals and tools), thus realizing a 2 by 2 factorial design (for each experiment). To increase statistical sensitivity for group analyses, we corrected for the inter-individual variation in cortical gyrification using the cortex-based intersubject alignment method. We performed conjunction analyses, which allowed the search for cortical regions that showed an object-category preference ($a > b$) during both unimodal visual (V) and unimodal auditory (A) stimulation ($V [a > b] \cap A [a > b]$).

Methods

Subjects

Thirteen subjects (ten male) participated in the fMRI experiments; their mean age was 28.2 years (range, 21 to 38 years). All subjects had normal or corrected-to-normal (four subjects) vision. Of these subjects, ten participated in experiments 1 and 2, nine in experiment 3, seven in experiment 4, and five in our imagery control experiment. All participants received information on MRI and a questionnaire to check for potential health risks and contraindications. Volunteers gave their informed consent after having been introduced to the procedure in accordance with the declaration of Helsinki.

Stimuli

The stimulus image signal was generated by a notebook (DELL 7500, 750 MHz) at a frame rate of 60 Hz. The image was projected onto a vertical screen positioned inside the scanner (above the head coil) with an LCD projector (Sony, VPL PX 20) equipped with a custom-made lens. Subjects viewed the screen through a mirror. Mirror and projection screen were fixed onto the head coil. The subjects' field of view was 40 degrees visual angle (maximum horizontal distance). Visual stimulation consisted of eight gray-scaled object photographs (mean stimulus size, 12.8 degrees visual angle) belonging to four different categories of natural or common objects (sportsmen, four-legged mammals, manipulable tools, and cars). Each visual stimulation block consisted of two photographs showing exemplars of a single object category that were alternately presented in the center of the screen at a rate of one every 2000 ms. In the center of the white-colored screen, a fixation cross was positioned which was present during the whole experiment.

Auditory stimuli were generated on the same notebook and passed through an amplifier (JVC AX R5BK). During experiments 1-3 and the imagery control experiment, subjects received these sounds through MR-compatible headphones, which were fitted into an earmuff (Bilsom 727, sound insulation of 33 dB at 1000 Hz). We built these headphones by removing the permanent magnets of standard stereo headphones (impedance of 32 ohm, frequency ranging from 20 till 20000 Hz) and using the external magnetic field of the MR scanner. Sound quality was in general superior to conventional MR-compatible stimulation systems working with sound transmission based on tubes. During experiment 4, we employed a commercially available MRI audio system (Commander XG, Resonance Technology, Northridge,

CA, U.S.A.). Auditory stimulation consisted of natural sounds related to the same eight objects (e.g., the sound of a tennis racket hitting a ball). Each auditory stimulation block consisted of two of these sounds that belonged to a single category of natural or common objects. Sounds were alternately presented at a rate of one every 2000 ms. The fixation cross was present even during purely auditory stimulation.

Procedure

During experiments 1-3, each of four conditions was presented for approximately 16 s (eight volumes) in a block design, separated from the next block by a fixation period of the same length. A complete presentation cycle consisted of the following sequence of conditions (with intervening fixations): natural sounds of one object category (A), sounds of another category (B), gray-scaled pictures of A, pictures of B (64 volumes). The imagery control experiment comprised two additional auditory conditions (sounds of animals and tools), during which subjects had to engage in visual imagery of the respective objects (cued by a green-colored fixation cross; 96 volumes). In the following cycle, the presentation order of the conditions was reversed with respect to the preceding cycle, providing a control for serial-order effects. A complete scan comprised two cycles of experimental conditions plus an additional eight volumes of fixation at the beginning of the scan (136 and 200 volumes for the main and imagery control experiments, respectively). During experiment 4, all eight experimental conditions (i.e. four object categories presented either visually or auditorily in separate stimulation blocks and randomized order) were employed within each of four experimental runs per subject. Subjects were asked to fixate and be attentive during the scans.

Imaging

fMRI scanning of experiments 1-3 and the imagery control experiment was performed on a 1.5 Tesla Siemens Magnetom Vision scanner (Siemens, Erlangen, Germany) at the University Clinic in Frankfurt am Main. A gradient-recalled echo-planar-imaging (EPI) sequence was used with the following parameters: 16 slices, oriented approximately in parallel to the AC-PC plane (AC, anterior commissure; PC, posterior commissure); TR, 2081 ms; TE, 60 ms; FA, 90°; FOV, 200 mm; in-plane resolution, 3.13 x 3.13 mm²; slice thickness, 5 mm; gap thickness, 1 mm. In addition, a detailed T1-weighted anatomical scan was acquired for all subjects using a

Siemens fast low-angle-shot (FLASH) sequence (isotropic voxel size 1 mm^3). For each subject, a magnetization-prepared rapid-acquisition gradient-echo (MP-RAGE) sequence was used (TR = 9.7 ms, TE = 4 ms, FA = 12° , matrix = 256×256 , voxel size $2.0 \times 1.0 \times 1.0 \text{ mm}^3$) for realignment with the detailed anatomical scan that had been measured in a previous session. Experiment 4 was performed on a 3 Tesla Siemens Magnetom Trio scanner (Siemens, Erlangen, Germany) at the Brain Imaging Center (BIC) in Frankfurt am Main. A gradient-recalled echo-planar-imaging (EPI) sequence was used with the following parameters: 34 slices; TR, 2080 ms; TE, 30 ms; FOV, 220 mm; in-plane resolution, $3.4 \times 3.4 \text{ mm}^2$; slice thickness, 3.2 mm; gap thickness, 0.3 mm.

Data Analysis

Data were analyzed using the BrainVoyager™ 2000 (version 4.9) and BrainVoyager™ QX (version 1.1) software packages (Brain Innovation, Maastricht, The Netherlands, <http://www.brainvoyager.com>). The first four volumes of each scan were discarded to preclude T1 saturation effects. Preprocessing of the functional data included the following steps: (i) three-dimensional motion correction using the Levenberg-Marquardt algorithm, (ii) linear-trend removal and temporal high-pass filtering at 0.01 Hz, (iii) slice-scan-time correction with sinc interpolation, and (iv) spatial smoothing using an 8 mm Gaussian kernel. Cortex-based statistical analysis was performed with multiple linear regression. For every cortical surface vertex, the time course was regressed on a set of dummy-coded predictors representing the experimental conditions. To account for the shape and delay of the hemodynamic response (Boynton et al. 1996), the predictor time courses (box-car functions) were convolved with a gamma function. We used group-based conjunction analyses to effectively restrict our search for cortical regions-of-interest (ROIs) with a specific functional profile. To find regions responding preferentially to one object category (A) as compared to another (B) during both unimodal visual and unimodal auditory stimulation, the activation produced by pictures of A had to be higher than the response to pictures of B, and a similar category preference profile had to be found during presentation of the respective natural sounds. The threshold was set to a P value of 0.05 (after correction for serial correlation). The activation patterns for the ROIs were visualized by plotting the beta weights from the GLM for the different experimental conditions. The data of experiments 1-3 were analyzed separately. In addition, we performed a two-factorial analysis of object category * sensory modality

to separate the respective main effects and to test for statistical interactions between these two factors in each of our ROIs (of experiments 1-3).

Cortical-surface reconstruction and visualization

The high-resolution T1-weighted 3D recordings were used for surface reconstruction of both cortical hemispheres of each subject (Kriegeskorte and Goebel 2001). The white/gray-matter border was segmented with a region-growing method preceded by inhomogeneity correction of signal intensity across space. The borders of the two resulting segmented subvolumes were tessellated to produce a surface reconstruction of each hemisphere.

Cortex-based inter-subject alignment

To improve the spatial correspondence between subjects' brains beyond Talairach space matching, the reconstructed hemispheres were aligned using curvature information reflecting the gyral/sulcal folding pattern (van Atteveldt et al. 2004). Folded cortical representations of each subject and hemisphere were morphed into a spherical representation that provided a parameterizable surface for nonrigid alignment across subjects. The curvature information of the folded representation was preserved as a curvature map on the spherical representation. This folding pattern was smoothed along the sphere surface to provide spatially extended gradient information driving intersubject alignment. Following a coarse-to-fine matching strategy, the alignment started with highly smoothed curvature maps and progressed to only slightly smoothed representations. While the alignment of major gyri and sulci was achieved reliably using this method, smaller structures, reflecting idiosyncratic differences between the subjects' brains, were not completely aligned.

Cortex-based intersubject alignment enabled us to align the time courses for multisubject GLM data analysis. Group-averaged functional data were then projected on inflated representations of the left and right cerebral hemispheres of the target subject. As a morphed surface always possesses a link to the folded reference mesh, functional data can be shown at the correct location of folded as well as inflated representations. This link was also used to keep geometric distortions to a minimum during inflation through inclusion of a morphing force that keeps the distances between vertices and the area of each triangle of the morphed surface as close as possible to the respective values of the folded reference mesh.

Figure 1. Experiments 1-3: BOLD responses to sounds and pictures of natural objects. **(A)** Inflated representation of one subject's cerebral hemispheres in a ventral view (middle panel) indicating right (dotted lines) and left VOT (solid lines) as enlarged in the left and right panels, respectively. Colored activation maps were based on conjunction analyses ($P < 0.05$, corrected), which allowed the search for cortical regions that showed an object-category preference ($a > b$) during both unimodal visual (V) and unimodal auditory (A) stimulation (resulting contrast: $V[a > b] \cap A[a > b]$). Group-averaged maps ($n=10$) are shown for sportsmen>animals (yellow; from exp. 1), cars>tools (green; from exp. 2), and tools>animals (blue; from exp. 3). **(B)** and **(C)** show mean BOLD signal intensity changes (beta weights) during presentation of auditory and visual objects. Colors of the dominant category match those of the respective cortical patches in **(A)**. Contrasts marked with * indicate both a significant conjunction analysis, and main effects of object category and sensory modality ($P < 0.01$, uncorrected) without any statistical interaction between these two factors. Contrasts marked with ▲ indicate a significant post hoc contrast, but did not reach significance level during main effect analysis.

Results

Conjunction analyses of group-averaged data from ten healthy adults revealed distinct peak activations in VOT for sportsmen, cars, and tools, thus indicating category-related processing (**Fig. 1A**). In general, the mean fMRI signal increases (beta weights) in these regions were found to be significantly larger during presentation of visual as compared to auditory objects (**Figs. 1B and 1C**). Interestingly, these 'visual' regions did not only show the respective category preference during unimodal visual stimulation, but also during unimodal auditory stimulation with natural sounds.

We could corroborate this finding using a two-factorial analysis of object category * sensory modality in each of these regions of interest (ROIs). All VOT ROIs in the left hemisphere (and one right-hemispheric ROI) showed main effects of sensory modality and object category ($P < 0.01$, uncorrected) without any statistical interaction between these two factors (**Figs. 1B and 1C**).

In addition, we performed a control experiment (experiment 4) in which we re-tested seven of the previous subjects in a 3 Tesla MRI scanner employing a revised design. This design was optimized to allow both the direct comparison of all four object categories within the same experimental runs and thorough analyses at the single-subject level. Thus, directly testing for all possible comparisons between our four object categories, we were able to replicate and extend the above reported findings (**Fig. 2A**). Moreover, we found examples of category-related auditory responses at the single-subject level that were even of the same amplitude than the category-related visual responses (**Fig. 2B**).

Figure 2. Experiment 4: left VOT response to sounds and pictures of natural objects. **(A)** (left panel) Auditory-visual activation maps were based on conjunction analyses that tested for all possible comparisons between our four object categories ($n=7$). Data are shown for sportsmen (yellow), animals (red), and cars (green). (Right panels) Mean BOLD signal intensity changes (beta weights) are shown in response to auditory (left half of each panel) and visual objects (right half of each panel). **(B)** Individual layout (for subjects 1, 2, and 3) of category-related audio-visual peak activations (upper panels) and BOLD response profiles (of exemplary regions; lower panels), indicating category-related audio-visual activation at the single-subject level ($P<0.05$, uncorr.).

Figure 3. Imagery control experiment: category-related responses of left VOT during passive viewing, passive listening, and visual imagery. **(A)** Group-averaged ($n=5$) category preference maps are shown separately for the passive viewing (animals>tools in red, tools>animals in dark blue), passive listening (animals>tools in yellow, tools>animals in light blue), and visual imagery conditions (in green). **(B)** Regions of overlap between visual and auditory category preference maps (animals>tools in orange; tools>animals in blue) do not overlap with the respective category preference maps obtained for visual imagery (green colored regions in **A**). **(C)** Mean BOLD signal intensity changes (beta weights; for the regions shown in **B**) indicate category-related audio-visual activation in VOT. BOLD response profiles are shown for animals>tools (upper panel) and tools>animals (lower panel; ** $P<0.001$, uncorr.; * $P<0.05$, uncorr.). Please note that these regions did not show any category-related effect during visual imagery.

Discussion

Based on our findings, predominantly left VOT regions are involved in AV integration of object-related information. This crossmodal convergence is organized according to object category. One possible objection against this interpretation is that the natural sounds might have triggered visual imagery of the respective objects. It is well-known

that VOT regions show category-related processing during both visual perception and imagery (Ishai et al. 2000). However, a category-related VOT activation does not necessarily involve an explicit utilization of prior visual experience, as has been demonstrated recently in a study on tactile object perception in congenitally blind subjects (Pietrini et al 2004). In contrast to the imagery literature, our subjects were neither instructed to engage in visual imagery (during experiments 1-4) nor did they have to perform a task for which visual imagery would have been beneficial. Most importantly, the explicit instruction to engage in visual imagery during presentation of the respective sounds resulted in a loss of category-related processing in our AV ROIs (**Fig. 3**). We therefore assume that auditorily triggered visual imagery did not contribute substantially to VOT activation.

We favor the alternative interpretation that AV integration of object information involves regions of the visual ‘what’ pathway. We think that vision, because of its superior spatial resolution, dominates audition during the perception of natural or common objects. While the exact mechanisms that constrain the organization of the visual ‘what’ pathway are still under vivid debate, it is widely accepted that VOT activation depends on category-related or context information rather than the exact physical stimulus properties (Grill-Spector und Malach 2004; Cox et al. 2004). The category-related auditory activation in VOT, as reported here, allows us to extend this view. This convergence of visual and auditory activations (potentially due to feedback from higher-order auditory regions in lateral temporal cortex; Lewis et al. 2004) reflects learned associations between radically different visual and auditory object features that co-occur very often in everyday life.

Acknowledgements

This research received support from the Bundesministerium für Bildung und Forschung (BMBF; Brain Imaging Center Frankfurt, DLR 01GO0203), and the Max Planck Society (MPS). We thank Ariane Meienbrock, Lutz Fiesser (for developing the auditory stimulation setup), Therese Lennert, Kathrin Utz (for assisting cortex-based intersubject alignment), Oliver Döhrmann (for help with the figures), and Axel Kohler (for helpful comments on earlier versions of this paper).

References

- Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex*, 12(11):1202-12.
- 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):809-23.
- Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207-21.
- Cox, D., Meyers, E., & Sinha, P. (2004). Contextually evoked object-specific responses in human visual cortex. *Science*, 304(5667):115-7.
- Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27:649-77.
- Ishai, A., Ungerleider, L. G., & Haxby, J. V. (2000). Distributed neural systems for the generation of visual images. *Neuron*, 28(3):979-90.
- Kriegeskorte, N., Goebel, R. (2001). An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes. *Neuroimage*, 14(2):329-46.
- Lewis, J. W., Wightman, F. L., Brefczynski, J. A., Phinney, R. E., Binder, J. R., & DeYoe, E. A. (2004). Human brain regions involved in recognizing environmental sounds. *Cerebral Cortex*, 14(9):1008-21.
- Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W. H., Cohen, L., Guazzelli, M., & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences U S A*, 101(15):5658-63.
- van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, 43(2):271-82.

CHAPTER 4

Audio-visual integration of artificial and natural object features in the human brain

Based on manuscript:

Hein, G. *, Doehrmann, O., Müller, N.G., Kaiser, J., Muckli, L., Naumer, M.J.* (2006).
Audio-visual integration of artificial and natural object features in the human brain.
Under revision.

*equal contribution

Abstract

Object recognition often requires the neuronal integration of auditory and visual features. In our study we addressed the question whether neuronal correlates of audio-visual (AV) object recognition are modulated by semantic congruency and familiarity of the object features being integrated. Using functional magnetic resonance imaging (fMRI), we compared AV integration of congruent and incongruent animal sounds and pictures with arbitrary pairings of complex artificial sounds and object images. Our results revealed AV integration effects in different regions, including inferior frontal cortex (IFC), superior temporal sulcus/middle temporal gyrus (pSTS/MTG), and superior temporal gyrus (STG). For natural objects, activations in superior and lateral temporal areas were modulated by semantic congruency. While pSTG was stronger activated by congruent than incongruent AV pairings, pMTG preferably integrated incongruent animal sounds and images. Most importantly, we revealed overlapping AV integration effects in pSTS and IFC for natural and artificial material. We conclude that AV integration in these areas is independent of stimulus familiarity, which indicates a core role of pSTS and IFC in object-related AV integration.

Introduction

Many objects are characterized by auditory and visual features, which have to be integrated during object recognition. The neural correlates of AV object recognition have mainly been investigated with sounds and images of familiar objects. Previous fMRI studies on AV perception of natural or common objects have reported activations in inferior frontal cortex (IFC) (Belardinelli et al., 2004; Taylor et al., 2006), superior temporal sulcus/middle temporal gyrus (pSTS/MTG) (Beauchamp et al., 2004; 2004b; Taylor et al., 2006), and higher-level visual regions (Belardinelli et al., 2004; Taylor et al., 2006). In addition, there is recent evidence for an involvement of higher-level auditory regions in AV object recognition (Lewis et al., 2004, 2005).

Integration of familiar object features in higher-level auditory and visual regions as well as in frontal cortex seems to be influenced by the semantic properties of the stimulus material. With respect to object category, Lewis et al. (2005) found stronger activation in middle superior temporal gyrus (mSTG) for animal sounds as compared to sounds of tools. Similarly, Taylor et al. (2006) reported stronger effects in higher-level visual regions during AV perception of living as compared to non-living objects. Moreover, occipital visual regions were stronger involved in the integration of semantically congruent (e.g. dog image and barking sound) than incongruent AV object features (Belardinelli et al., 2004). These results indicate that AV integration in higher-level auditory and visual regions is sensitive to both object category and semantic congruency. In addition, effects of semantic congruency have also been found in frontal regions. AV integration sites in frontal cortex were mainly involved in the integration of semantically incongruent object features (e.g. a dog image and a meowing sound) (Belardinelli et al., 2004; Taylor et al., 2006).

Compared to the auditory, visual, and frontal regions described above, AV processing of object features in pSTS/MTG seemed to be less sensitive to semantic properties. For example, pSTS/MTG showed a slightly higher activation under semantically congruent conditions (Beauchamp et al., 2004; see also van Atteveldt et al., 2004), but was also strongly involved in the integration of incongruent AV pairings (see also Taylor et al., 2006). Taken together, these studies have demonstrated that the different cortical components of the AV object recognition network are modulated individually by the semantic properties of the object features being integrated.

The present fMRI study addressed the question in how far responses of this AV object recognition network are modulated by semantic congruency and stimulus familiarity. To test this, we compared AV integration of congruent and incongruent combinations of animal vocalizations and pictures with pairings of complex artificial sounds and object images (“fribbles”; Fig. 1a, see Experimental Procedures for details). To balance the experimental design and control for potential learning effects, artificial sounds and object images were presented either in fixed or random combinations. We hypothesized that higher-level auditory and visual regions preferentially integrate highly familiar (i.e. congruent natural) rather than unfamiliar (i.e. incongruent natural or even artificial) object features, as these regions have been shown to prefer semantically congruent natural stimuli (Belardinelli et al., 2004; Lewis et al., 2004; 2005; Taylor et al., 2006). In contrast, we expected lateral temporal and frontal regions to integrate features of complex AV objects irrespective of either semantic congruency or stimulus familiarity. This expectation was based on reports of an involvement of human frontal regions in the integration of incongruent natural object features (Belardinelli et al., 2004; Taylor et al., 2006), and on single-cell results which showed a multisensory response to tones and colors in monkey prefrontal cortex (Fuster et al., 2000). Similarly, human pSTS/MTG is known to be involved in AV integration of a broad variety of different stimuli (e.g., Amedi et al., 2005; Beauchamp, 2005b; Calvert et al., 2001; Taylor et al., 2006).

Methods

Subjects

We measured eighteen subjects (seven female, mean age = 29.8, range 23–41 years, one left-handed). All subjects had normal or corrected-to-normal vision and hearing. All participants received information on MRI and a questionnaire to check for potential health risks and contraindications. Volunteers gave their informed consent after having been introduced to the procedure in accordance with the declaration of Helsinki.

Stimuli

The stimuli were presented using Neurobehavioral Systems Presentation software (<http://nbs.neuro-bs.com>) running on a PC (Miditower Celeron) at a frame rate of 60 Hz. Sounds and images were presented in stimulation blocks at a rate of one every

2000 ms. Stimulation blocks consisted of eight stimulus events with a fixation cross being constantly present. Images were projected onto a vertical screen positioned inside the scanner with an LCD projector (Sony, VPL PX 20) equipped with a custom-made lens. Subjects viewed the screen through a mirror. Mirror and projection screen were fixed on the head coil. The subjects' field of view was 52.5 degrees visual angle (maximum horizontal distance). Visual stimulation consisted of gray-scaled photographs (mean stimulus size, 14.6 degrees visual angle) of animals and so-called "fribbles" (<http://www.cog.brown.edu/~tarr/stimuli.html>; eight exemplars per category), presented in the center of a black screen.

Auditory stimuli were presented through an MRI audio system (Commander XG, Resonance Technology, Northridge, CA, U.S.A.). Subjects received them via headphones simultaneously to both ears. Auditory stimulation consisted either of animal sounds or abstract sounds, created through distortion (played backwards and with underwater effect) of the same eight animal sounds. In a behavioral pre-test, distorted sounds were played to eight subjects who did not take part in the fMRI sessions. According to this pre-test, none of the distorted sounds was associated to any natural object.

Procedure

Animal sounds, animal images, distorted sounds, and fribble images were presented in eight experimental conditions (unimodal auditory animal / distorted; unimodal visual animal / fribble; AV congruent natural; AV incongruent natural; AV artificial - fixed order; AV artificial - random order). During bimodal conditions, subjects simultaneously perceived sounds and pictures. In the congruent natural AV condition, animal pictures matched with animal sounds (e.g., dog-barking; cat-meowing etc.). In the incongruent natural AV condition, there was no semantic match between animal picture and sound (e.g., dog-meowing etc.) This allowed us to test the effects of both semantic congruency (congruent versus incongruent) and stimulus familiarity (artificial versus natural). To balance the experimental design and control for potential learning effects, we presented fribble pictures and distorted sounds either in fixed (fixed order) or random pairings (random order).

Each of these eight experimental conditions was presented for approximately 16 s (eight measurement volumes) in a block design, separated from the next block by a fixation period of equal length. A complete experimental run comprised two cycles of experimental conditions plus an additional eight volumes of fixation at the beginning

of the run (280 volumes). The session had five experimental runs, including all experimental conditions in randomized order. Subjects were asked to fixate and be attentive during the measurements. We have chosen a passive paradigm to minimize task-related frontal activity (see also Belardinelli et al., 2004; Calvert et al., 2000; van Atteveldt et al., 2004; 2006). Given the sluggishness of the BOLD signal, otherwise it is hard to compellingly disentangle task-related frontal activations from frontal involvement in AV integration.

Visuo-tactile (VT) control experiment

We measured thirteen subjects, who had also participated in our AV main experiment. Subjects saw and/or touched artificial objects (wooden fribbles) and natural objects (toy animals). Artificial and natural objects were presented in both unimodal (visual and tactile) as well as bimodal (VT) conditions. Natural VT combinations consisted of touchable animals and their respective photographs that were presented either canonically (VT_con) or horizontally mirrored (VT_incon). Employing a block-design, each experimental condition was presented for approximately 20 s, (10 measurement volumes; TR = 2 s) separated from the next block by a fixation period of equal length. Each single stimulus was presented for 2s and re-opening of the right hand was cued by a color change of the fixation cross. A complete experimental run comprised two cycles of experimental conditions plus an additional ten volumes of fixation at the beginning of each run (350 volumes). The session consisted of four experimental runs each including all experimental conditions in randomized order.

Imaging

fMRI was performed on a 3 Tesla Siemens Magnetom Allegra scanner (Siemens, Erlangen, Germany) at the Brain Imaging Center (BIC) in Frankfurt/Main. A gradient-recalled echo-planar-imaging (EPI) sequence was used with the following parameters: 34 slices; TR, 2000 ms; TE, 30 ms; FOV, 192 mm; in-plane resolution, 3 x 3 mm²; slice thickness, 3 mm; gap thickness, 0.3 mm. For each subject, a magnetization-prepared rapid-acquisition gradient-echo (MP-RAGE) sequence was used (TR = 2300 ms, TE = 3.49 ms, FA = 12°, matrix = 256 x 256, voxel size 1.0 x 1.0 x 1.0 mm³) for detailed anatomical imaging.

Data Analysis

Data were analyzed using the BrainVoyager™ QX (version 1.7) software package (<http://www.brainvoyager.com>). The first four volumes of each experimental run were discarded to preclude T1 saturation effects. Preprocessing of the functional data included the following steps: (i) three-dimensional motion correction, (ii) linear-trend removal and temporal high-pass filtering at 0.0054 Hz, (iii) slice-scan-time correction with sinc interpolation, and (iv) modest spatial smoothing using a Gaussian filter of 8 mm (FWHM) that was applied only during volume-based but not during cortex-based analysis.

Volume-based as well as cortex-based statistical analyses were performed using multisubject fixed-effects general linear models (GLMs). These GLMs took into account the numbers of both subjects (18) and experimental conditions (8) by means of employing $18 \times 8 = 144$ regressors (so-called “separate subject predictors”). For every voxel or cortical surface vertex, the time course was regressed on a set of dummy-coded predictors representing the experimental conditions. To account for the shape and delay of the hemodynamic response (Boynton et al., 1996), the predictor time courses (box-car functions) were convolved with a gamma function. Neural correlates of object-related AV integration were assessed separately for artificial, incongruent natural, and congruent natural material. In each case, we searched for regions that were a) significantly activated during each of the unimodal conditions (A; V), and b) showed a stronger activation during bimodal as compared to each of the unimodal conditions. Accordingly, the identification of brain regions as regions of object-related AV integration was based on a conjunction analysis ($0 < A < AV < V < 0$). In a recent debate, these criteria have been suggested as strict and appropriate for the definition of multisensory integration effects in human brain imaging studies (Beauchamp, 2005a; Beauchamp et al., 2004a; Laurienti et al., 2005; but see Calvert et al., 2001). During analyses of group-averaged data, statistical thresholds were set to $q(\text{FDR}) < 0.05$ (unless differently indicated in the respective figure captions). During single-subject data analysis, statistical thresholds were set individually and are reported in Table 1. Event-related averaged time courses indicate the average percentage of BOLD signal change during the respective experimental conditions. Error bars indicate standard errors.

The high-resolution T1-weighted 3D recordings were used for surface reconstruction of both cortical hemispheres of each subject (Kriegeskorte & Goebel, 2001). The white/gray-matter border was segmented with a region-growing method preceded by inhomogeneity correction of signal intensity across space. The borders of the two resulting segmented subvolumes were tessellated to produce a surface reconstruction of each hemisphere.

To improve the spatial correspondence between subjects' brains beyond Talairach space matching, the reconstructed hemispheres were aligned using curvature information reflecting the gyral/sulcal folding pattern (see van Atteveldt et al. 2004 for details). As demonstrated recently for a similar analysis tool (Argall et al., 2006), such a surface-based approach improves t-statistics in the average activation map as compared to the volume average. Folded cortical representations of each subject and hemisphere were morphed into a spherical representation that provided a parameterizable surface for nonrigid alignment across subjects. The curvature information of the folded representation was preserved as a curvature map on the spherical representation. This folding pattern was smoothed along the sphere surface to provide spatially extended gradient information driving intersubject alignment. Following a coarse-to-fine matching strategy, the alignment started with highly smoothed curvature maps and progressed to only slightly smoothed representations. While the alignment of major gyri and sulci was achieved reliably using this method, smaller structures, reflecting idiosyncratic differences between the subjects' brains, were not completely aligned.

Cortex-based intersubject alignment enabled us to align the time courses for multisubject GLM data analysis. Group-averaged functional data were then projected on inflated representations of the left and right cerebral hemispheres of one subject. As a morphed surface always possesses a link to the folded reference mesh, functional data can be shown at the correct location of folded as well as inflated representations. This link was also used to keep geometric distortions to a minimum during inflation through inclusion of a morphing force that keeps the distances between vertices and the area of each triangle of the morphed surface as close as possible to the respective values of the folded reference mesh.

Figure 1. Visual (VIS) and auditory (AUD) example stimuli, cortical regions of object-related audio-visual (AV) integration (defining contrast: $0 < A < AV > V > 0$; $q(\text{FDR}) < .05$). Unimodal auditory (yellow), and unimodal visual (light blue) activations ($p(\text{Bonf. corr.}) < .001$) for (a) artificial (b) incongruent natural, and (c) congruent natural conditions. LH = left hemisphere, RH = right hemisphere.

Results

Unimodal activations

Fig. 1 shows the pattern of activation during unimodal visual (light blue) and auditory stimulations (yellow). Both unimodal contrasts ($A > V$; $V > A$) revealed activations of the respective modality-specific cortices. Strength and distribution of activation was comparable for animal sounds (**Fig. 1b** and **c**) and distorted sounds (**Fig. 1a**). The same was true for animal pictures (**Fig. 1b** and **c**) and pictures of fribbles (**Fig. 1a**).

Figure 2. Overlap of significantly activated AV integration regions ($q(\text{FDR}) < .05$) for artificial (blue), incongruent natural (red), and congruent natural object features (green), and average percentage of BOLD signal change during bimodal and unimodal conditions. RH = right hemisphere; p = posterior; STG = superior temporal gyrus, STS = superior temporal sulcus; MTG = middle temporal gyrus; IFC = inferior frontal cortex. A = unimodal auditory stimulation; V = unimodal visual stimulation; AV = bimodal audio-visual stimulation.

Bimodal activations

Neural correlates of object-related AV integration for artificial, incongruent natural, and congruent natural AV combinations are depicted in **Fig. 1a-c** (indicated by blue, red, and green, respectively). As no significant differences ($q(\text{FDR}) < .05$) were found between fixed and random combinations of artificial object pictures and sounds, we pooled the data from these two experimental conditions.

Fig. 1a shows regions of AV integration of artificial object features. A conjunction analysis comparing the respective bimodal and unimodal conditions ($0 < A < AV > V > 0$;

$q(\text{FDR}) < 0.05$) revealed an IFC activation that extended ventrally along the inferior frontal gyrus (IFG) and dorsally along the precentral sulcus (PreCS). In addition, we found an activation of pSTS, overlapping with unimodal auditory activation (**Fig. 1a**, dark blue).

Object-related AV integration effects for incongruent natural features are depicted in **Fig. 1b** (red; $0 < A < AV > V > 0$; $q(\text{FDR}) < 0.05$). Again, we found strong IFC activations, extending dorsally along PreCS and showing a strong involvement of ventral regions along the IFG. Apart from these frontal effects, AV integration of incongruent natural object features activated pSTS/MTG.

Testing for AV integration of congruent natural object features, we identified a strong bilateral activation of pSTS (**Fig. 1c**; green; $0 < A < AV > V > 0$; $q(\text{FDR}) < 0.05$). In the right hemisphere, activation extended further anterior into the auditory STG region. Accordingly, **Fig. 1c** shows a substantial overlap between this AV integration effect (green) and auditory activations (yellow). In addition, AV integration of congruent natural material also involved an activation of right IFC. Compared to the results shown in **Fig. 1a** and **b**, this IFC activation was focused along PreCS and showed no ventral extension along IFG.

Fig. 2 shows a striking overlap of AV integration effects for artificial, incongruent natural, and congruent natural object features in IFC (along the PreCS) and in pSTS (gray). Integration of artificial and incongruent natural, but not congruent natural AV pairings further elicited overlapping activations in the ventral portion of IFC along IFG (purple). The pSTS/MTG region showed an interesting pattern of AV integration effects. Posterior dorsal to pSTS, which was activated during AV integration under all three conditions, we found a region activated only during integration of natural material (congruent as well as incongruent; dark green). Extending further posterior into MTG, its function becomes more specialized and is solely related to integration of incongruent natural material (red). The most anterior portion of STS activation spreads into auditory STG and is correlated to integration of congruent natural AV combinations (light green). As depicted in **Fig. 2**, there are joint cortical integration sites for artificial, incongruent natural, and congruent natural object features in IFC and pSTS. In addition, our data point to a more specialized role of pMTG for integrating incongruent natural, and of pSTG for integrating congruent natural AV pairings.

Figure 3. Direct contrasts between bimodal conditions ($q(\text{FDR}) < .05$). (a) Regions activated more strongly by congruent natural as compared to artificial AV object features (green), and vice versa (blue; additional criteria: $A_{\text{nat}} > 0 < V_{\text{nat}}$, and $A_{\text{art}} > 0 < V_{\text{art}}$, respectively). (b) Regions activated more strongly by incongruent natural as compared to artificial AV object features (red), and vice versa (blue; additional criteria: $A_{\text{nat}} > 0 < V_{\text{nat}}$, and $A_{\text{art}} > 0 < V_{\text{art}}$, respectively). (c) Regions activated more strongly by congruent as compared to incongruent natural AV object features (green), and vice versa (red; additional criterion: $A_{\text{nat}} > 0 < V_{\text{nat}}$). LH = left hemisphere, RH = right hemisphere.

The results of direct contrasts between our three different AV conditions (**Fig. 3**) support these findings. The contrast between congruent natural and artificial AV combinations (**Fig. 3a**) showed a strong activation of pSTS/MTG, extending to regions of higher-level auditory processing in mSTG (green; Lewis et al., 2005). The contrast between incongruent natural and artificial AV pairings revealed strong bilateral activations of IFC, pSTS/MTG, and mSTG regions (**Fig. 3b**). The results of the contrasts between congruent and incongruent natural AV combinations indicate that mSTG activations are stronger for congruent (green) and pMTG activations are stronger for incongruent natural pairings (red). In addition, incongruent natural combinations led to stronger activations of parietal regions (**Fig. 3c**). The contrast between artificial (blue) and both natural AV conditions revealed dorsal occipital activations extending into posterior parietal cortex (PPC), two regions that are well-known as parts of the ‘dorsal stream’, associated with action-related visual processing (Milner and Goodale, 1995). **Fig. 4** contains a fusion of the maps in **Fig. 3** and shows the respective event-related averaged BOLD signal time courses for our three AV conditions.

There is growing evidence that deactivation of brain regions (often referred to as “negative BOLD”) can be as functionally important as positive BOLD activations (Amedi et al., 2005; Shmuel et al., 2006). We examined effects of BOLD deactivation in a further analysis. Its results did not show any systematic effects.

Single-subject results

The group-averaged results indicate that AV integration of artificial object features correlates to activations of IFC and pSTS, which are also involved in AV integration of both incongruent and congruent natural objects. To test the robustness of these effects, we examined our data also at the single-subject level. **Fig. 5** provides parts of our single-subject results by showing overlapping AV integration sites for artificial, incongruent natural, and congruent natural object features in lateral frontal cortex (for 9/18 subjects; see also **Table 1**). In addition, **Table 1** provides the individual Talairach coordinates, p-, and t-values for lateral temporal regions of overlap that were found in 7/18 subjects.

Figure 4. Average percentage of BOLD signal change in left hemisphere regions with different bimodal response preferences. LH = left hemisphere, p = posterior; STG = superior temporal gyrus, STS = superior temporal sulcus; MTG = middle temporal gyrus; PreCS = precentral sulcus; IFC = inferior frontal cortex; PPC = posterior parietal cortex; dOC = dorsal occipital cortex.

Discussion

Summary of main results

This study addressed the impact of semantic congruency and stimulus familiarity on neuronal correlates of AV object recognition. AV integration of congruent and incongruent combinations of animal vocalizations and pictures was compared with arbitrary pairings of complex artificial sounds and 'fribble' images. Our results revealed AV integration effects in IFC and in both lateral and superior temporal

cortex. For natural objects, activations in superior and lateral temporal areas were modulated by semantic congruency. While pSTG was activated more strongly by congruent than incongruent AV pairings, pMTG showed the opposite response pattern. Most importantly, we found overlapping AV integration effects in pSTS and IFC for natural and artificial material. Apparently AV integration in these areas is independent of stimulus familiarity. The functional parallels between IFC and pSTS are in line with primate data, which show close anatomical connections between these prefrontal and lateral temporal regions (e.g., Barbas et al., 2005; Keysers and Perrett, 2004; Petrides and Pandya, 1988; Romanski et al., 1999).

Figure 5. AV integration effects for artificial (blue), incongruent natural (red), and congruent natural object features (green) overlap in frontal cortices of individual subjects. (9/18 subjects; **Table 1** provides the individual Talairach coordinates, p and t values also for lateral temporal regions of overlap that were found in 7/18 subjects).

AV integration in frontal cortex

Compared to pSTS/MTG, current knowledge about the role of IFC in AV object recognition is sparse. There is convincing evidence for multisensory response properties of prefrontal neurons in primates (Fuster et al., 2000), but very few studies have assessed frontal integration sites in humans. One main reason for this might be the multi-functionality of frontal regions, which makes it difficult to disentangle AV integration from other frontal functions. Particularly critical confounds are frontal activations caused by explicit and even implicit task performance (e.g., Miller and Cohen, 2001), and the increase of frontal activations with increasing complexity of stimulation (e.g., Duncan and Owen, 2000). To minimize effects of task performance, we used a passive paradigm. Nevertheless, the observed frontal effects could still reflect differences in stimulation complexity, because we contrasted single input during unimodal conditions with dual inputs during bimodal conditions.

Visuo-tactile control experiment

To test this alternative explanation, we conducted a control experiment with 13 of our subjects and employed the same overall experimental design. Again we used animal and fribble stimuli in unimodal and bimodal conditions, but this time presented them in the visual and tactile modalities (see Experimental Procedures). Activations during unimodal and bimodal conditions of the visuo-tactile (VT) control experiment were analyzed region-of-interest-based in the same IFC regions which were identified as integration sites in our main AV experiment (cf. **Fig. 2**). If the AV effects in IFC were mainly caused by the higher complexity of bimodal as compared to unimodal stimulation, bimodal VT stimulation should also elicit stronger activations than unimodal visual and tactile conditions. Our results argue against this interpretation. In contrast to our AV main experiment, IFC activations in the VT control experiment were comparable for bimodal and unimodal conditions (**Fig. 6**). Based on this finding, it is unlikely that differences in stimulation complexity between bimodal and unimodal conditions are the main source of the increased AV activations in IFC. Thus, we assume that the IFC effects observed in our main experiment are specific to the integration of AV object features.

Table 1. Individual Talairach coordinates, t-, and p-values of overlapping AV integration regions for artificial, incongruent natural, and congruent natural objects features. Lat. front. = lateral frontal cortex; lat. temp. = lateral temporal cortex; LH = left hemisphere.

subject	Region	X	Y	Z	t	p
S1	Lat. front	37	13	25	2.6	<0.01
S3		47	16	28	3.3	<0.001
S6		43	26	23	2.0	<0.05
S8		38	1	27	2.6	<0.01
S11		49	21	33	3.3	<0.001
S12		49	15	19	3.3	<0.001
S15		55	9	40	2.0	<0.05
S16		49	1	43	2.0	<0.05
S17	LH	-50	-9	49	2.6	<0.01
S1	Lat. temp.	62	-32	13	2.0	<0.05
S3		40	-58	7	2.0	<0.05
S4		55	-26	13	2.6	<0.01
S6		61	-23	16	3.3	<0.001
S12		50	-39	14	3.3	<0.001
S12	LH	-44	-41	11	3.3	<0.001
S13	LH	-58	-59	0	2.0	<0.05
S15	LH	-60	-46	-2	2.0	<0.05

Frontal incongruency effects

IFC activation related to AV integration was strongest and most widespread for incongruent natural material. This is in line with studies showing stronger frontal activation during processing of incongruent combinations of visual objects as compared to congruent object stimuli (e.g., Michelon et al., 2003). The AV integration effect for incongruent natural material was more pronounced in the ventral portion of IFC, along IFG, whereas dorsal IFC along PreCS was also activated during AV integration of both congruent natural and artificial object features. As PreCS is also discussed in the context of action-related integration of visual, auditory, proprioceptive, and motor cues (Ehrsson et al., 2004; Fogassi and Gallese, 2004; Lloyd et al., 2003), it is possible that this region integrates complex sounds and images of three-dimensional objects in a more grasping-related way, even if they are passively heard and seen.

Figure 6. Results of the visuo-tactile (VT) control experiment. **(a)** Overlap of significantly activated AV integration regions for artificial (blue), incongruent natural (red), and congruent natural object features (green; from Fig. 2). **(b)** Average percentage of BOLD signal change in IFC regions during unimodal and bimodal AV and **(c)** VT conditions. In contrast to the results of the AV main experiment there was no activation increase during bimodal VT stimulation as compared to the respective unimodal conditions. Thus, it is unlikely that the observed AV activation increase in IFC (main experiment) reflects the higher number of inputs in bimodal as compared to unimodal stimulation. RH = right hemisphere; IFC = inferior frontal cortex. A = unimodal auditory stimulation; V = unimodal visual stimulation; T = unimodal tactile stimulation; AV = bimodal audio-visual stimulation; VT = bimodal visuo-tactile stimulation.

Lateral temporal activations

The important role of lateral temporal cortex for multisensory integration in general (Amedi et al., 2005, Calvert et al., 2001) and AV object recognition in particular (Beauchamp, 2005b) is well known. Apart from the joint integration sites in IFC and pSTS, we revealed more specialized sites for the integration of natural object

features. For the integration of incongruent natural stimuli we found an additional region in pMTG. Together with the pSTS region, which was involved in AV integration under all three bimodal conditions, this finding further supports the notion of a “patchy organization” of pSTS/MTG (Beauchamp et al., 2004b). Beauchamp et al. (2004b) reported evidence that different pSTS patches are involved in the processing of incoming visual and auditory inputs. Extending Beauchamp et al.’s (2004b) findings, our results indicate a functional subdivision in superior to lateral temporal cortex based on semantic AV congruency and stimulus familiarity; we identified a region in pSTS, which integrates AV object features independently of these two factors, whereby pMTG was activated more strongly during AV integration of incongruent natural stimuli.

Superior temporal activations

To our knowledge, here we present the first data indicating a prominent role of higher-level auditory regions in AV integration of natural objects. We found a pSTG region preferably integrating semantically congruent natural material. It is located in the close vicinity of an mSTG region that is well known to be involved in the processing of animal sounds (Lewis et al., 2005). Previous studies on object-related AV integration have reported effects in visual, but not auditory regions (Belardinelli et al., 2004; Taylor et al., 2006). Both Belardinelli et al. (2004) and Taylor et al. (2006) have employed a mixture of living and non-living objects, while in our study all meaningful objects belonged to the animal category. As the findings of Lewis et al. (2005) indicate differences in neural representation between animal and tool sounds, AV objects from different object categories might be integrated in different sensory-specific regions. A similar involvement of auditory cortex has also been shown during the integration of written letters and speech sounds (van Atteveldt et al., 2004; 2006). These findings thus support the assumption that higher-level auditory regions preferably integrate semantically congruent AV features of natural or common objects.

Conclusion

The present study of semantic congruency and stimulus familiarity effects on object-related AV integration revealed effects in IFC and in both lateral and superior temporal cortex. Selective AV integration effects for congruent natural objects were observed in superior temporal areas (pSTG). In addition, pSTG responded preferentially to congruent as compared to incongruent AV pairings, whereas pMTG exhibited the inverse activation pattern and showed selective AV integration effects for incongruent natural material. Most importantly, we showed overlapping AV integration effects in pSTS and IFC for both natural and unfamiliar artificial material. As we were able to rule out a more general preference for dual as compared to single input, IFC activation was clearly related to object-related AV integration. Apparently AV integration in these regions is independent of stimulus familiarity, indicating a core role of pSTS and IFC in object-related AV integration. It would be interesting to investigate whether sensory-specific regions get also involved in AV integration of artificial object features if artificial AV associations are trained, i.e., a semantic AV relationship is established and familiarity is substantially increased.

Acknowledgements

We are grateful to three anonymous reviewers for constructive comments on a previous version of the manuscript, Petra Janson and Isabella Nowak for help with preparing the visual stimuli, and Tim Wallenhorst for data acquisition support. Funding was obtained from the German Ministry of Education and Research (BMBF), the Volkswagen Foundation (G.H.), the German Research Foundation (HE 4566/1-1; G.H.), and the Max Planck Society (MPS; M.J.N.).

References

- Amedi, A., Malach, R., and Pascual-Leone, A. (2005). Negative BOLD differentiates visual imagery and perception. *Neuron* 48, 859-872.
- Amedi, A., von Kriegstein, K., van Atteveldt, N.M., Beauchamp, M.S., and Naumer, M.J. (2005). Functional imaging of human crossmodal identification and object recognition. *Exp. Brain Res* 166, 559-571.
- Argall, B.D., Ziad, S.S., and Beauchamp, M.S. (2006). Simplified intersubject averaging on the cortical surface using SUMA. *Human Brain Mapp* 27, 14-27.
- Barbas, H., Medalla, M., Alade, O., Suski, J., Zikopoulos, B., and Lera, P. (2005). Relationship of prefrontal connections to inhibitory systems in superior temporal areas in the rhesus monkey. *Cereb. Cortex* 15, 1356-1370.
- Beauchamp, M.S. (2005a). Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics* 3, 93-113.
- Beauchamp, M.S. (2005b). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr Opin Neurobiol* 15, 145-53.
- Beauchamp, M.S., Lee, K.E., Argall, B.D., and Martin, A. (2004a). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 4, 809-823.
- Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., and Martin, A. (2004b). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat Neurosci* 11, 1190-1192.
- Belardinelli, M.O., Sestieri, C., Di Matteo, R., Delogu, F., Del Gratta, C., Ferretti, A., Caulo, M., Tartaro, A., and Romani, G.L. (2004). Audio-visual crossmodal interactions in environmental perception: an fMRI investigation. *Cogn Process* 5, 167-174.
- Boynton, G. M., Engel, S. A., Glover, G. H., Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* 16, 4207-4221.

- Calvert, G.A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110-1123.
- Calvert, G.A., Campell, R., and Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in human heteromodal cortex. *Curr Biol* 10, 649-657.
- Calvert, G.A., Hansen, P.C., Iversen, S.D, and Brammer, M.J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage* 14, 427-438.
- Cappe, C., and Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *Eur. J. Neurosci* 22, 2886-2902.
- Duncan, J., and Owen, A.M. (2000). Common regions in the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci* 23, 475-483.
- Ehrsson, H.H., Spence, C., and Passingham, R.E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science* 305, 875-877.
- Fogassi, L., and Gallese, V. (2004). Action as a binding key to multisensory integration. In *The handbook of multisensory processes*, G. Calvert, C. Spence, and B. E. Stein, eds. (Cambridge, MA: The MIT Press), pp. 425-441.
- Fuster, J.M., Bodner, M., and Kroger, J.K. (2000). Cross-modal and cross-temporal associations in neurons of frontal cortex. *Nature* 405, 347-351.
- Keysers, C., and Perrett, D. I. (2004). Demystifying social cognition: a hebbian perspective. *Trends Cogn Sci* 8, 501-507.
- Kriegeskorte, N., Goebel, R. (2001). An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes. *Neuroimage* 14, 329-346.
- Laurienti, P.J., Perrault, T.J., Stanford, T.R., Wallace, M.T., and Stein, B.E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp Brain Res* 166, 289-297.

- Laurienti, P.J., Wallace, M.T., Maldjian, J.A., Susi, C.M., Stein, B.E., and Burdette, J.H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Hum Brain Mapp* 19, 213-223.
- Lewis, J.W., Brefczynski, J.A., Phinney, R.E., Janik, J.J., and DeYoe, E.A. (2005). Distinct cortical pathways for processing tool versus animal sounds. *J Neurosci.* 25, 5148-5158.
- Lewis, J.W., Wightman, F.L., Brefczynski, J.A., Phinney, R.E., Binder, J.R., and DeYoe, E.A. (2004). Human brain regions involved in recognizing environmental sounds. *Cereb. Cortex* 14, 1008-1021.
- Lloyd, D.M., Shore, D.I., Spence, C., and Calvert, G.A. (2003). Multisensory representation of limb position in human premotor cortex. *Nat. Neurosci* 6, 17-18.
- Michelon, P., Snyder, A.Z., Buckner, R.L., McAvoy, M., and Zacks, J.M. (2003). Neural correlates of incongruous visual information. An event-related fMRI study. *Neuroimage* 19, 1612-1626.
- Miller E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24, 167-202.
- Milner, A.D., and Goodale, M.A. (1995). *The visual brain in action*. Oxford Univ. Press, Oxford.
- Petrides, M., and Pandya, D.N. (1988). Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *J Comp Neurol* 273, 52-66.
- Taylor, K.I., Moss, H.E., Stamatakis, E.A., and Tyler, L.K. (2006). Binding crossmodal object features in perirhinal cortex. *Proc Natl Acad Sci U S A* 103, 8239-8244.
- Romanski, L.M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P.S., Rauschecker, J.P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat Neurosci* 2, 1131-1136.
- Shmuel, A., Augath, A., Oeltermann, A., and Logothetis, N.K. (2006). Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. *Nat. Neurosci.* 9, 569-577.

van Atteveldt, N.M., Formisano, E., Blomert, L., and Goebel, R. (2006). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cereb. Cortex*, June 2, online.

van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron* 43, 271-282.

Summary

Invasive electrophysiological studies in animals and non-invasive investigations in human subjects provide impressive evidence that even single cognitive functions involve widely distributed regions of the brain. This is true not only for crossmodal and sensory-motor integration that, per definition, require coordination between diverse sensory and executive structures. It also holds for unimodal perception because sensory systems also consist of multiple subsystems that operate in parallel and are specialized for the analysis of particular features of objects. A core question of cognitive neuroscience is, how these multiple, widely distributed processes are coordinated and how the respective computational results are bound together in order to generate coherent unisensory and multisensory percepts. The functional magnetic resonance imaging (fMRI) studies reported in this dissertation had the common goal to contribute to a better understanding of the mechanisms that underlie crossmodal audio-visual integration. Particular emphasis has been laid on the investigation of two dimensions that are of prime importance for audio-visual integration: spatial (**chapter 2**) and semantic congruency (**chapters 3 and 4**).

Chapter 1 provides an introduction to both unisensory (auditory, visual, and tactile) as well as multisensory (audio-visual and visuo-tactile) object processing. We reviewed human fMRI studies that showed that visual, tactile, and auditory information about objects can activate cortical association areas that were once believed to be modality-specific. Processing converges either in multisensory zones or via direct crossmodal interaction of modality-specific cortices without relay through multisensory regions. We integrate these findings with existing theories about semantic processing and propose a general mechanism for crossmodal object recognition: The recruitment and location of multisensory convergence zones varies depending on the information content and the modality that provides the most reliable sensory input.

For the successful integration of visual and auditory information it is important to identify whether visual and auditory signals originate from corresponding places in the external world. In **chapter 2**, we report crossmodal effects of spatially congruent and incongruent audio-visual stimulation. Visual and auditory stimuli were matched to one of four horizontal locations in external space. Subjects had to assess the spatial fit of a visual stimulus (i.e. a gray-scaled picture of a comic dog) and a

simultaneously presented auditory stimulus (i.e. a barking sound). We were able to reveal two distinct networks of cortical regions preferentially processing either spatially congruent or incongruent audio-visual stimulation. Whereas low-level visual areas responded preferentially to incongruent stimulation, higher-level visual areas of the temporal and parietal cortex and prefrontal regions responded preferentially to congruent audio-visual stimulation. A position-resolved analysis revealed three robust cortical representations for each of our four visual stimulus locations in retinotopic visual regions corresponding to the representation of the horizontal meridian in area V1, and at the dorsal and ventral borders between areas V2 and V3. While these regions of interest (ROIs) did not show any significant effect of spatial congruency, we found subregions within ROIs in the right hemisphere that showed an incongruency effect (i.e. an increased fMRI signal during spatially incongruent as compared to congruent audio-visual stimulation). We interpret this finding as a correlate of spatially dispersed recurrent feedback during mismatch processing: whenever a spatial mismatch is detected in multisensory regions (such as the IPS), processing resources are re-directed to low-level visual areas.

In **chapter 3**, we investigated audio-visual object recognition using fMRI. We found distinct peak activations for diverse object categories (sportsmen, animals, cars, and tools) in ventral occipito-temporal cortex. These regions showed a category preference not only during unimodal visual stimulation but also during unimodal auditory stimulation. As we could rule out visual imagery as a potential confound, this might reflect learned associations between frequently co-occurring visual and auditory object features.

In **chapter 4**, we addressed the question whether neuronal correlates of audio-visual object recognition are modulated by semantic congruency and familiarity of the object features being integrated. We compared audio-visual integration of congruent and incongruent animal sounds and pictures with arbitrary pairings of complex artificial sounds and object images. We revealed audio-visual integration effects in different regions, including inferior frontal cortex (IFC), superior temporal sulcus/middle temporal gyrus (STS/MTG), and superior temporal gyrus (STG). For natural objects, activations in superior and lateral temporal areas were modulated by semantic congruency. While posterior STG was stronger activated by congruent than incongruent audio-visual pairings, posterior MTG preferably integrated incongruent animal sounds and images. Most importantly, we revealed overlapping audio-visual

integration effects in posterior STS and IFC for natural and artificial material. We conclude that AV integration in these areas is independent of stimulus familiarity, which indicates a core role of posterior STS and IFC in object-related audio-visual integration.

Independent of their computational function, multisensory regions presumably interact with unisensory cortices via feedback connections and thus subserve a cortical network in which semantic knowledge about objects is represented in a distributed fashion. According to this model, brain regions that process a specific type of incoming sensory information also serve as a knowledge store about the attributes of particular objects. For instance, a hammer has a number of salient attributes (such as a characteristic color, shape, sound). Ventral temporal cortex processes incoming information about visual form and color. If the incoming form and color match the stored hammer template, ventral temporal cortex identifies the object as a hammer. Lateral temporal cortex processes information about object motion. If the incoming motion information matches the stored template (e.g. up, down, impact) then the object is identified as a hammer. Frontal and parietal regions store information about the hand orientation and arm posture that would be required to grasp the hammer. The representations in these different regions are linked, so that simply seeing a photograph of a static hammer also activates the respective visual motion and motor representations. Thus, networks of neurons in auditory association cortex that respond to an auditory object are also linked to visual, tactile, and motor information about the same object represented in temporal, parietal, and frontal regions.

To better understand the rapid computations performed during object perception, future studies should combine high-resolution fMRI with methods that enable examination of the temporal sequence of information processing (for example magnetoencephalography). Such an approach should deepen our understanding of crossmodal object recognition, while the use of more natural stimulation conditions should substantially contribute to human object recognition theories with increased ecological validity.

Samenvatting

Invasieve electrofysiologische studies in dieren en non-invasieve studies in mensen leveren indrukwekkend bewijs voor de betrokkenheid van zeer verspreide hersengebieden bij zelfs eenduidige cognitieve functies. Dit is niet alleen waar voor cross-modale en sensor-motor integratie, waarbij per definitie coördinatie tussen verschillende sensorische en executieve structuren vereist is. Het is ook het geval voor unimodale perceptie, omdat sensorische systemen uit meerdere subsystemen bestaan die parallel opereren en gespecialiseerd zijn in de analyse van specifieke kenmerken van objecten. Een kernvraag binnen de cognitieve neurowetenschap is hoe deze verschillende, wijd verspreide processen gecoördineerd worden en hoe de respectievelijke computationele resultaten gebonden worden tot coherente unisensorische en multisensorische waarnemingen. De functionele magnetic resonance imaging (fMRI) studies beschreven in dit proefschrift hadden als gemeenschappelijk doel om bij te dragen aan een beter begrip van de mechanismen die ten grondslag liggen aan cross-modale audio-visuele integratie. De nadruk is gelegd op het onderzoeken van twee dimensies die van primair belang zijn voor audio-visuele integratie: spatiële (**hoofdstuk 2**) en semantische (**hoofdstukken 3 en 4**) congruentie.

Hoofdstuk 1 geeft een inleiding in zowel unisensorische (auditive, visuele en tactiele) als multisensorische (audio-visuele en visuo-tactiele) objectverwerking. Er wordt een overzicht gegeven van humane fMRI studies waarin wordt aangetoond dat visuele, tactiele en auditive informatie over objecten corticale associatiegebieden kunnen activeren waarvan oorspronkelijk werd aangenomen dat ze modaliteitspecifiek waren. Verwerking convergeert óf in multisensorische zones, óf via directe cross-modale interactie tussen modaliteitspecifieke cortices zonder tussenkomst van multisensorische gebieden. Deze bevindingen worden geïntegreerd met bestaande theorieën over semantische verwerking tot een algemeen mechanisme voor cross-modale objectherkenning: de rekrutering en locatie van multisensorische zones varieert afhankelijk van de informatie-inhoud, en van welke modaliteit de meest betrouwbare sensorische input levert.

Voor succesvolle integratie van visuele en auditive informatie is het belangrijk om vast te stellen of visuele en auditive signalen uit corresponderende locaties in de omgeving komen. In **hoofdstuk 2** rapporteren we cross-modale effecten van spatieel

congruente en incongruente audio-visuele stimulatie. Visuele en auditieve stimuli kwamen uit één van vier horizontale locaties, en proefpersonen moesten beslissen of de visuele (b.v. een zwart-wit cartoon van een hond) en auditieve (b.v. een blaf-geluid) stimuli uit dezelfde locatie kwamen of niet. We hebben hiermee aan het licht gebracht dat twee verschillende netwerken van corticale gebieden elk hun eigen voorkeur hebben voor spatieel congruente of spatieel incongruente audio-visuele stimulatie. Terwijl lagere visuele gebieden voorkeur hadden voor incongruente stimulatie, lieten hogere visuele gebieden in de temporale en pariëtale cortex, en prefrontale gebieden een voorkeur voor spatieel congruente audio-visuele stimulatie zien. Een analyse gericht op stimulus-positie onthulde drie robuuste corticale representaties voor elk van de vier visuele stimulus locaties in retinotopisch georganiseerde visuele gebieden: corresponderend met de horizontale meridiaan in gebied V1, en met de dorsale en ventrale grenzen in V2 en V3. Terwijl deze “regions-of-interest” (ROIs) geen significante effecten van spatiele congruentie lieten zien, hebben we subgebieden binnen de rechter hemisfeer ROIs gevonden die een incongruentie-effect vertoonden (d.w.z. een verhoogd fMRI signaal tijdens spatieel incongruente stimulatie t.o.v. spatieel congruente audio-visuele stimulatie). We interpreteren deze bevinding als een correlaat van spatieel verspreide terugprojecties (feedback) tijdens mismatch verwerking: steeds als een spatiele mismatch wordt gedetecteerd in multisensorische gebieden (zoals de IPS), worden verwerkingsbronnen opnieuw toegewezen aan lagere visuele gebieden.

In **hoofdstuk 3** hebben we audio-visuele objectherkenning onderzocht m.b.v. fMRI. We hebben verschillende activatiepieken gevonden voor verschillende objectcategorieën (sporters, dieren, auto's en gereedschappen) in de ventrale occipitaal-temporale cortex. Deze gebieden lieten een categorievoorkeur niet alleen zien tijdens unimodale visuele stimulatie maar ook tijdens unimodal auditieve stimulatie. Omdat we visuele inbeelding hebben kunnen uitsluiten als mogelijke verklaring, zou het de geleerde associaties tussen veel gebruikte, en samen optredende visuele en auditieve objectkenmerken kunnen reflecteren.

In **hoofdstuk 4** hebben we ons op de vraag gericht of neuronale correlaten van audio-visuele objectherkenning beïnvloed worden door semantische congruentie en bekendheid met de objectkenmerken die geïntegreerd worden. We hebben audio-visuele integratie van congruente en incongruente dierengeluiden en -plaatjes vergeleken met arbitrair gepaarde complexe artificiële objectgeluiden en -plaatjes.

We hebben audio-visuele integratie-effecten onthuld in verschillende gebieden, waaronder de inferieure frontale cortex (IFC), superieure temporale sulcus/middelste temporale gyrus (STS/MTG), en de superieure temporale gyrus (STG). Voor natuurlijke objecten werden de activaties in superieure en laterale temporale gebieden beïnvloed door congruentie. Terwijl de posterieure STG sterker geactiveerd werd door congruente dan door incongruente audio-visuele paren, integreerde de posterieure MTG voornamelijk incongruente dierengeluiden en -plaatjes. De belangrijkste bevinding was dat audio-visuele integratie-effecten voor natuurlijke en artificiële informatie overlap vertoonden in posterieure STS en IFC. We concluderen dat AV-integratie in deze gebieden onafhankelijk is van stimulusbekendheid. Dit indiceert dat deze gebieden een kernrol spelen in object-gerelateerde audio-visuele integratie.

Het is aannemelijk dat onafhankelijk van hun computationele functie, multisensorische gebieden met unisensorische gebieden interacteren via terugprojecties, en daarmee onderdeel zijn van een corticaal netwerk waarbij semantische kennis over objecten op een verspreide manier wordt gerepresenteerd. Volgens dit model dienen hersengebieden die een bepaald type sensorische informatie verwerken ook als kennisopslag voor de kenmerken van specifieke objecten. Een hamer bijvoorbeeld heeft verschillende opvallende kenmerken (een karakteristieke kleur, vorm en geluid). De ventrale temporale cortex verwerkt inkomende informatie over visuele vorm en kleur. Wanneer de inkomende vorm en kleur overeenkomen met de opgeslagen hamersjabloon wordt het object door de ventrale temporale cortex geïdentificeerd als hamer. De laterale temporale cortex verwerkt informatie over beweging. Als de inkomende bewegingsinformatie overeenkomt met het opgeslagen sjabloon (b.v. naar boven, naar beneden, de kracht) wordt het object als hamer geïdentificeerd. In frontale en pariëtale gebieden ligt de informatie over handoriëntatie en houding van de arm opgeslagen die nodig is om een hamer vast te pakken. De representaties in deze gebieden zijn verbonden zodat alleen het kijken naar een stilstaand plaatje van een hamer de respectievelijke visuele bewegings- en motorische representaties activeert. Kort samengevat zijn dus netwerken van neuronen in auditieve associatiecortex die reageren op een auditief object ook verbonden met visuele, tactiele en motorische informatie over hetzelfde object, gerepresenteerd in temporale, pariëtale en frontale hersengebieden.

Om de snelle computaties tijdens objectherkenning beter te kunnen begrijpen, zijn toekomstige studies nodig die fMRI op hoge resolutie combineren met methoden die de temporele sequentie van informatieverwerking kunnen bestuderen, bijvoorbeeld magneto-encefalografie. Een dergelijke benadering zou ons begrip van cross-modale objectherkenning kunnen verdiepen, terwijl het gebruik van natuurlijkere stimulatieomstandigheden substantieel zou bijdragen aan meer ecologische valide theorieën over objectherkenning.

Acknowledgements

For their support and advice during the realization of this thesis I thank my mentors and promotor Dr. Lars Muckli, Prof. Dr. Wolf Singer, and Prof. Dr. Rainer Goebel. It has been a great pleasure to work with these inspiring personalities. I greatly appreciated the academic freedom they encouraged, and I am indebted to them for their invaluable criticism and support. More personally, I wish to thank Lars Muckli for his great mentorship and constant enthusiasm. The diversity, productive atmosphere, and understanding in Wolf Singer's department I have very much appreciated.

For their contributions to the studies of this thesis I would like to thank again my coauthors and all colleagues, who have been acknowledged individually at the end of each chapter. My thanks for exciting cooperation go out to Dr. Grit Hein, Dr. Nienke van Atteveldt (special thanks for the Dutch translation of the summary!), Dr. Katharina von Kriegstein, Dr. Amir Amedi, and Dr. Michael S. Beauchamp as well as to my *paranimfen* Bettina Sorger and Dr. Vincent van de Ven.

I am indebted to my former roommates and colleagues at the MPIH, especially Axel Kohler (thank you very much for all your support!), Ariane Meienbrock, Michael Wibral, Petra Janson (special thanks for designing the cover!), Nikolaus Kriegeskorte, Sarah Weigelt, Stephan Schleim, Arjen Alink, Amanda Kaas, Katharina Melnikovic, Boris Ebisch, Sascha Gotthardt, and Matthias Munk for their invaluable assistance and the great time we had together.

Many thanks also to my new colleagues and students at the Institute of Medical Psychology (Frankfurt Medical School), namely Prof. Dr. Jochen Kaiser, Oliver Döhrmann, Christian Altmann, Christoph Bledowski, Monika Geis-Dogruel, Karen Spannhake, Michaela Henning, Leonie Ratz, Andrea Polony, Oliver Horn, and Yavor Yalachkov for their support and the joyful atmosphere at work.

For his friendship I wish to thank Stefan Kieß. Meinen Eltern Hans und Mary Naumer sowie meinen Geschwistern Esther, Lucas und Hannah möchte ich herzlich danken für ihre beständige und liebevolle Unterstützung. I am for ever grateful to my wife Anja and our wonderful sons Ben and Leo for all their love.

Curriculum Vitae

Marcus J. Naumer was born on September 1st 1973 in Neustadt an der Weinstrasse, Germany. There, he attended the Kurfürst-Ruprecht-Gymnasium and obtained his Abitur in 1992. From 1993 until 2001 he studied Biology, Philosophy, and Psychology at the Universities of Freiburg im Breisgau and Landau in der Pfalz, where he graduated in May 2001. In this period, he specialized in cognitive neuroscience and joined the Department of Neurophysiology (headed by Prof. Dr. Wolf Singer) at the Max Planck Institute for Brain Research in Frankfurt am Main. In his M.Sc.-thesis, he used fMRI to investigate interactions between visual motion and form processing under supervision of Prof. Dr. Rainer Goebel. From 2001 until 2005 he was Ph.D.-student at the Department of Neurophysiology (Max Planck Institute for Brain Research, Frankfurt). During this period, he conducted the experiments described in this thesis at the Frankfurt Brain Imaging Center (BIC, Frankfurt Medical School) under supervision of Dr. Lars Muckli. Since May 2005, he continues his work as a post-doc researcher at the Institute of Medical Psychology (headed by Prof. Dr. Jochen Kaiser), where he heads the Multisensory Research Group.

Contact information

Marcus J. Naumer
Institut für Medizinische Psychologie
Klinikum der J. W. Goethe-Universität
Heinrich-Hoffmann-Str. 10
Haus 93, Raum C230
D-60528 Frankfurt am Main
Tel.: +49-(0)69-6301-6581
Fax: +49-(0)69-6301-7606
Email: M.J.Naumer@med.uni-frankfurt.de