The Face in Your Voice–How Audiovisual Learning Benefits Vocal Communication

Dissertation

Zum Erlangen des akademischen Grades

Doctor rerum naturalium (Dr. rer. Nat.)

im Fach Psychologie

Eingereicht am 27.09.2013 an der

Mathematisch-Naturwissenschaftlichen Fakultät II

der Humboldt-Universität zu Berlin

von **M.Sc., Sonja Schall**

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II

Prof. Dr. Elmar Kulke

Gutachter/Gutachterinnen:

1. Prof. Dr. Rasha Abdel Rahman

2. Prof. Dr. Katharina von Kriegstein

3. Prof. Dr. Elia Formisano

Tag der Verteidigung: 26.08.2014

Abstract

Face and voice of a person are strongly associated with each other and usually perceived as a single entity. Despite the natural co-occurrence of faces and voices, brain research has traditionally approached their perception from a unisensory perspective. This means that research into face perception has exclusively focused on the visual system, while research into voice perception has exclusively probed the auditory system. In this thesis, I suggest that the brain has adapted to the multisensory nature of faces and voices and that this adaptation is evident even when one input stream is missing, that is, when input is actually unisensory. Specifically, the current work investigates how the brain exploits previously learned voice-face associations to optimize the auditory processing of voices and vocal speech. Three empirical studies providing spatiotemporal brain data—via functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG)—constitute this thesis. All data were acquired while participants listened to auditory-only speech samples of previously familiarized speakers (with or without seeing the speakers' faces). Three key findings demonstrate that previously learned visual speaker information support the auditory analysis of vocal sounds: (i) face-sensitive areas were part of the sensory network activated by voices, (ii) the auditory analysis of voices was temporally facilitated by learned facial associations and (iii) multisensory interactions between face- and voice/speech-sensitive regions were increased. The current work challenges traditional unisensory views on vocal perception and rather suggests that voice and vocal speech perception profit from a multisensory neural processing scheme.

Abstrakt

Gesicht und Stimme einer Person sind stark miteinander assoziiert und werden normalerweise als eine Einheit wahrgenommen. Trotz des natürlichen gemeinsamen Auftretens von Gesichtern und Stimmen, wurden deren Wahrnehmung in den Neurowissenschaften traditionell aus einer unisensorischen Perspektive untersucht. Das heißt, dass sich Forschung zu Gesichtswahrnehmung ausschließlich auf das visuelle System fokusierte, während Forschung zu Stimmwahrnehmung nur das auditorische System untersuchte. In dieser Arbeit schlage ich vor, dass das Gehirn an die multisensorische Beschaffenheit von Gesichtern und Stimmen adaptiert ist, und dass diese Adaption sogar dann sichtbar ist, wenn nur die Stimme einer Person gehört wird, ohne dass das Gesicht zu sehen ist. Im Besonderen, untersucht diese Arbeit wie das Gehirn zuvor gelernte Gesichts-Stimmassoziationen ausnutzt um die auditorische Analyse von Stimmen und Sprache zu optimieren. Diese Dissertation besteht aus drei empirischen Studien, welche raumzeitliche Hirnaktivität mittels funktionaler Magnetresonanztomographie (fMRT) und Magnetoenzephalographie (MEG) liefern. Alle Daten wurden gemessen, während Versuchspersonen auditive Sprachbeispiele von zuvor familiarisierten Sprechern (mit oder ohne Gesicht des Sprechers) hörten. Drei Ergebnisse zeigen, dass zuvor gelernte visuelle Sprecherinformationen zur auditorischen Analyse von Stimmen beitragen: (i) gesichtssensible Areale waren Teil des sensorischen Netzwerks, dass durch Stimmen aktiviert wurde, (ii) die auditorische Verarbeitung von Stimmen war durch die gelernte Gesichtsinformation zeitlich faszilitiert und (iii) multisensorische Interaktionen zwischen gesichtsensiblen und stimm-/sprachsensiblen Arealen waren verstärkt. Die vorliegende Arbeit stellt den traditionellen, unisensorischen Blickwinkel auf die Wahrnehmung von Stimmen und Sprache in Frage und legt nahe, dass die Wahrnehmung von Stimme und Sprache von von einem multisensorischen Verarbeitungsschema profitiert.

Acknowledgements

First, I would like to thank my supervisors and colleagues at the MPI for their scientific advice, support and friendship over the last four years. In particular, I would like to thank Katharina von Kriegstein and Stefan Kiebel for their excellent supervision, guidance and confidence in my work. I would also like to thank everyone who journeyed with me these past years as members of the „Neural Mechanisms of Communication" and the „Dynamic Action and Perception" groups, especially Helen Blank for many valuable discussions, Stefanie Schelinski for always being ready to help, as well as Begona Diaz and Claudia Roswandowitz for being the most fabulous office mates. I am happy to share so many memorable moments with all of you! Further, I would like to thank members and friends of the Bennewitz MEG-group, in particular Burkhard Maess for his patience and goodwill in discussing the, sometimes harsh, reality of MEG, Björn Herrmann for helpful advice and Yvonne Wolff for perfect participant-handling and data acquisition.

I would also like to thank everyone I have worked and studied with at the Neurobiopsychology department in Osnabrück. I am deeply grateful for having had the chance to study in such a warm and inspiring place. In particular, I would like to thank Peter Koenig for teaching me that science is fun, Saskia Nagel for her ability to sense the difference, Sonja Engmann for living the difference, Alper Acik for being so smart, Daniel Weiller for being himself and Cliodhna Quigley for a perfect symbiosis in coffee consumption and work as well as for genuine friendship. I am thankful to Selim Onat whose mind and true scientific heart has shaped me in countless ways.

I would like to thank my dear friends Felix Faber and Sebastian Zösch–without you my life would be only half as much fun–as well as my oldest friends Maria Vo and Tatjana Petojevic for being there through thick and thin. I could not wish for better friends. Most of all, I would like to express my deepest gratitude to my parents, my brother, my grandmother and my aunt for their unconditional love and support. Thank you for always having encouraged me to study and to explore the world and for having provided me with the means to do so. Danke.

Table of Contents

1 Introduction

In our everyday life we communicate constantly. We greet our neighbor in the street, we talk to our friends over a coffee and we raise our eyebrows to signal a colleague how we feel about work. As humans, we are communication experts and the skill to recognize and understand others is essential for a healthy social life.

There are many different implicit and explicit communication forms, but a lot of them, if not most, heavily rely on the perception of others' faces and voices. Faces and voices both allow us to recognize a person, to understand speech or read emotional states. We master the challenge of understanding vocal and facial patterns, speech and gestures with ease and we start doing so from a very young age. Even before birth, we can recognize our mother's voice and we start becoming sensitive to our native language s. Right after birth, newborns find faces more interesting than other visual objects (for review see Simion et al., 2007) and may imitate facial gestures only minutes after birth (Meltzoff and Moore, 1977). Only hours after birth, newborns can discriminate their mother's face from other faces (Sai, 2005). The ability to read and recognize others from facial and vocal signals is not only found in humans, but in a wide range of species including mammals, birds and amphibians; for reviews see (Leopold and Rhodes, 2010; Sidtis and Kreiman, 2012). It is rather astonishing that even frogs can recognize their neighbor's voice (Bee and Gerhardt, 2002; Feng et al., 2009; Gasser et al., 2009). These findings underline the importance of face and voice perception, not only for our adult social life but also for our phylogenetic and ontological history. It is therefore not surprising that our brain has developed specialized sensory structures to accommodate the perception of others' faces and voices; for a review see (Haxby et al., 2000; Belin et al., 2004).

One of the most well-known brain regions is the fusiform face area (Kanwisher et al., 1997). This area is part of the visual pathway and has been named due its vigorous neural response during visual stimulation with faces. More recently, a region similarly sensitive to the auditory presentation of voices has been detected in the auditory pathway (Belin et al., 2000). Probably owing to the tradition of studying the visual and auditory sensory systems in unisensory settings (i.e. with input from only one sensory modality), face and voice perception have mostly been studied in isolation from each other. Communication,

however, is in essence a multisensory experience. Facial and vocal characteristics are highly dependent on each other (Chandrasekaran et al., 2009). Vocal pitch is, for instance, informative about facial gender and speech sounds are highly dependent on lip-postures. We have a lifelong training in perceiving these audiovosial regularities and our brain undoubtedly exploits these learned regularities to make voice and speech perception more robust. For example, viewing a speaker's facial movements improves the comprehension of speech (Sumby and Pollack, 1954; Grant and Seitz, 2000; Ross et al., 2007). Also, being familiar with a person's face helps us to identify that person from voice alone, for example, when we communicate via the phone (Sheffert and Olson, 2004; von Kriegstein et al., 2008). These findings suggest that the neural mechanisms of communication can, to their fullest extent, only be understood when approached from a multisensory perspective.

The current thesis is based on the assumption that the audiovisual nature of communication signals, in other words, the usual co-occurrence of faces and voices, is reflected in the brain's perceptual mechanisms during communication and that this is in particular the case when we are familiar with a person. This assumption will be investigated using the example of auditory-only voice and speech recognition. Specifically, this thesis is concerned with the engagement of face-sensitive brain areas during vocal (i.e. auditory-only) communication and how previously learned visual speaker information is incorporated into the auditory analysis of voice and speech.

In the following chapter, the basic neural mechanisms of face and voice perception will be briefly described. Chapter 2 will introduce traditional and current models of communication with a particular emphasis on the recently developed audiovisual model of communication which was the underlying theoretical framework of this thesis. The empirical studies (studies 1-3) are summarized in chapter 3. The complete manuscripts of these studies are appended as individual chapters.

2 The Neural Basis of Voice, Speech and Face Perception

In order to unveil the neuronal mechanisms of face and voice recognition, numerous studies have aimed to identify brain responses that are particularly sensitive to faces or voices. Mostly, studies have approached this question from a purely unisensory viewpoint and asked the question of where or when in the cortical hierarchy face- or voice-sensitive responses occur. Whereas neuropsychological lesion studies and functional magnetic resonance imaging (fMRI) studies are well suited to address the question of where in the brain face- and voice-sensitive regions exist, electroencephalography (EEG) and magnetoencephalography (MEG) studies are informative about when face- and voice-sensitive responses emerge. In the following sections, a short review on empirical findings will be given and what we know from functional imaging as well as from EEG/MEG studies will be summarized. The focus will be on the brain areas and brain responses most critical to the current thesis. In addition to general voice- and face-sensitive regions and their involvement in identity-processing, the following sections will also touch on brain areas particularly sensitive to vocal and facial speech.

2.1 Voice and Auditory Speech Perception

The human voice is a rich acoustic signal and a lot of research has been conducted in an attempt to understand how the brain extracts meaning from a vocal speech signal. Far fewer experiments have looked into the nature of non-linguistic voice perception including the perception of voice-identity. One major finding was the discovery of voice-sensitive regions in the right superior temporal sulcus/superior temporal gyrus (STS/STG; (von Kriegstein et al., 2003; Formisano et al., 2008). This is in opposition to the classic finding of lateralization of auditory speech-sensitive regions which are predominantly found in the left STS/STG.

2.1.1 Voice Perception–Findings from Lesions and Functional Imaging

The first empirical evidence that voice recognition is supported by specialized brain regions came from early clinical studies investigating patients with right-hemispheric lesions in temporal and parietal areas (Van Lancker and Canter, 1982; Van Lancker and Kreiman, 1987; Van Lancker et al., 1989). These patients had specific difficulty in recognizing identity from voices despite showing normal speech comprehension and production abilities (Van Lancker and Canter, 1982; Lang et al., 2009). A range of fMRI studies confirmed the existence of voice-sensitive brain areas in human adults (Belin et al., 2000; Belin et al., 2002; Belin and Zatorre, 2003; von Kriegstein et al., 2003; Kriegstein and Giraud, 2004; Andics et al., 2010) as well as in infants (Grossmann et al., 2010; Blasi et al., 2011) and in non-human primates (Petkov et al., 2008; Perrodin et al., 2011).

Using fMRI in the adult human brain, several voice-sensitive regions have been identified. These are typically found with a right-hemispheric dominance and are located, depending on the specific design of the study, in differing portions of the STS (Belin et al., 2000; Belin et al., 2002; Belin and Zatorre, 2003; von Kriegstein et al., 2003; Kriegstein and Giraud, 2004; Formisano et al., 2008; Andics et al., 2010). For instance, Belin et al. (2000) identified voice-sensitive areas in bilateral STS regions (with a right hemispheric dominance) by contrasting human vocal sounds and non-vocal, environmental sounds. Kriegstein and Giraud (2004) identified several voice-sensitive areas along the right STS by selectively engaging participants in a voice-identification or speech-recognition task. Taken together, the current empirical evidence indicates the existence of at least two functionally different voice-sensitive regions: the right posterior and the right anterior STS. The posterior STS has been suggested to be more closely related to the lower-level acoustic processing of voices (Kriegstein and Giraud, 2004; Andics et al., 2010), while the anterior STS is specifically involved in voice-identity recognition (Belin and Zatorre, 2003; Kriegstein and Giraud, 2004; Andics et al., 2010).

2.1.2 Vocal Speech Perception–Findings from Functional Imaging

Understanding speech entails that the brain extracts phonetic, lexical, syntactic, and semantic structures from the ongoing modulations of spoken speech. A substantial body of research has shown that the mapping of sound to meaning is supported by a wide

network of brain regions including auditory, motor and frontal regions see for review (Hickok and Poeppel, 2007).

The speech perception network is typically strongly left-lateralized. It is, for example, a common finding that aphasic patients with problems in speech comprehension and production show lesions in the left rather than the right hemisphere. This is also confirmed by functional imaging studies that revealed a particular role of the left STG/STS during speech comprehension. The left STG/STS is, for example, more strongly engaged when listening to intelligible speech compared to non-intelligible acoustic control stimuli (Scott et al., 2000; Obleser et al., 2007; Rosen et al., 2011) or when speech recognition is behaviorally relevant (von Kriegstein et al., 2003). Specifically, there is evidence for an anterior stream along the left STG/STS (Davis and Johnsrude, 2003; Leff et al., 2008; DeWitt and Rauschecker, 2012) that becomes–with increasing distance to the primary auditory cortex–increasingly invariant to low-level acoustic properties (Davis and Johnsrude, 2003; Rosen et al., 2011; DeWitt and Rauschecker, 2012). At the anterior tip of the left anterior STS, activity levels have been associated with successful speech comprehension (Davis and Johnsrude, 2003). Also, auditory speech comprehension has been shown to be compromised after transient electric stimulation of the left anterior STS (Matsumoto et al., 2011).

2.1.3 Voice Perception–Findings from EEG and MEG

Concerning the timing of voice-sensitive brain responses, there is accumulating evidence from EEG and MEG studies that the time frame around the auditory P2/M200 component is particularly sensitive to vocal features (Schweinberger, 2001; Charest et al., 2009; Zaske et al., 2009; Altmann et al., 2010; De Lucia et al., 2010; Renvall et al., 2012; Capilla et al., 2013). The P2 (as referred to in EEG literature) or M200 (the magnetic analogue of the P2) is considered a late auditory evoked response which peaks around 200 ms peri-stimulus following the well known auditory N1/M100 component; see for review (Crowley and Colrain, 2004). Similar to 100 ms latency response (N1/M100), the 200 ms latency response (P2/M200) is considered to reflect the sensory encoding of the auditory stimulus (see for review (Martin et al., 2008). Later, but not earlier voice-sensitive responses have also been reported (Levy et al., 2001; Gunji et al., 2003; Levy et al., 2003).

Voice-sensitive responses around the time frame of the 200 ms latency component have, for instance, been reported when comparing human vocal with non-vocal stimuli (Charest et al., 2009; Capilla et al., 2013) or during voice-priming (Schweinberger, 2001). Using a mismatch negativity design, this time frame has also been shown to be sensitive to voice-familiarity (Beauchemin et al., 2006). Concerning the topography of voice-sensitive responses, there is, so far, little consistency across studies and widely differing designs have resulted in differing topographical patterns; however, topographical distributions were typically bilateral with no apparent hemispheric bias. Knowledge about the underlying neural sources of these 200ms latency responses is scarce. Currently, only a few studies have performed source analyses, two of which have employed bottom-up designs that used different stimulation conditions, in particular human vocalizations and non-vocal sounds or animal vocalizations (De Lucia et al., 2010; Capilla et al., 2013) and one using an adaptation design (Renvall et al., 2012). Similar to the above mentioned fMRI studies, these studies identified the right STS as the underlying neural source of the voice-sensitive 200ms latency response. The extent to which these results translate to situations that emphasize voice-identity perception is currently unclear (see Chapter 7 for recent progress on this question).

2.2 Face Perception

Face perception is thought to be accomplished by a distributed, bilateral neural network of brain regions including, but not limited to, the inferior occipital gyrus, the fusiform gyrus, the STS and the amygdala (see for review (Haxby et al., 2000; Ishai, 2008). The following sections will summarize findings with an emphasis on the fusiform face area (FFA) with respect to face-sensitive regions and the M170 with respect to timing. Brain areas sensitive to facial speech will be described shortly.

2.2.1 Face Perception–Findings from Functional Imaging

The best-known face-sensitive region is probably FFA, which is part of the visual ventral stream. The FFA responds vigorously to the visual presentation of human faces (see for review (Kanwisher and Yovel, 2006) and can, using fMRI, be located by comparing the

blood oxygenation level dependent (BOLD) response in the brain to faces and to other visual stimuli, such as objects or houses. Typically, the differential activation patterns reveal bilateral activity in the fusiform gyri with a stronger and more consistent pattern in the right hemisphere. Functionally, the FFA has been suggested to play a role in the perception of time-invariant facial features (Haxby et al., 2000; Ishai, 2008) such as identity (Sergent et al., 1992; Eger et al., 2004; Rotshtein et al., 2005). This has, for example, been shown by (Rotshtein et al., 2005) who used face stimuli from a morph-continuum to show that the FFA is more sensitive to identity changes than to physical changes.

2.2.2 Facial Speech Perception–Findings from Functional Imaging

The posterior STS, in contrast, is thought be involved in the perception of time-varying facial information such as mouth-movements (Puce et al., 1998; Pelphrey et al., 2005). The left posterior STS, more than the right, has in particular been implicated in the processing of speech-related mouth-movements (Calvert and Campbell, 2003) and the extent of activation in left STS/STG when viewing visual speech has been shown to predict individual speech-reading ability (Hall et al., 2005).

2.2.3 Findings from EEG and MEG

With regard to time, face-sensitive event-related responses have been found around at various latencies including 100ms (Liu et al., 2002), 170ms (Eimer, 2011), 250ms (Schweinberger et al., 2002; Schweinberger et al., 2004) and 400ms (e.g. Bentin and Deouell, 2000). Although the earliest face-sensitive response within the visual evoked responses has been revealed around the 100ms latency (Liu et al., 2002), the most prominent, and also best-studied face-sensitive response occurs around the 170ms component and has accordingly been termed the N170 (in the EEG literature, 'N' due to its negative deflection) or M170 (in MEG reports) (Bentin et al., 1996; Liu et al., 2000; for review see Eimer, 2011). Although the N170/M170 is also evoked by other visual objects its amplitude is considerably stronger for faces. The topographical distribution of the M170 is typically bilateral with a right-hemispheric bias and source localization results show that the neural origin of the M170 is located in the fusiform gyrus (Sams et al., 1997; Halgren et al., 2000; Deffke et al., 2007; Henson et al., 2009; Taylor et al., 2011). This result is also confirmed by an intracranial study showing that the earliest face-sensitive response in the

posterior and middle fusiform gyrus occurs around 110ms, followed by 170ms and 240ms latency responses (Barbeau et al., 2008). In the past, the N170/M170 has mostly been described as a category-selective response that marks the structural encoding of faces, preceding stages of identity recognition. However, there is now some evidence suggesting that the N170/M170 is sensitive to face familiarity and identity (Kloth et al., 2006; Harris and Aguirre, 2008; Caharel et al., 2009b; Caharel et al., 2009a). A more prominent response to face-identity is, however, seen in the later N250r component- an EEG response evoked by repetition priming (Schweinberger et al., 2002; Schweinberger et al., 2004) and during the N400 (Bentin and Deouell, 2000).

2.3 Conclusion

In summary, there is good evidence that our brain is tuned to facial and vocal communication signals. An overarching principle in the neural architecture supporting facial and vocal communication signals consists in the hemispheric dissociation of identity- and speech-recognition related processes; while facial and vocal identity are preferentially processed in the right hemisphere, speech-related aspects of face and voice are preferentially processed in the left. As for timing, the most dominant face- and voice-sensitive responses have both been observed within similar time frames, around 170 and 200ms.

3 Communication is Multisensory

We mostly talk to others face-to-face. Under these natural conditions, communication is multisensory. However, as described in the previous sections, the perceptions of voices and faces have mostly been studied in isolation from each other. Models that incorporate the integration of auditory and visual communication signals exist (Ellis et al., 1997; Calvert, 2001). Yet, they adhere to the traditional multisensory view (Mesulam, 1998) which suggests that it is only after the initial sensory processing stages of auditory and visual information that the two streams are integrated in heteromodal brain areas. In contrast to these approaches, the recently developed audiovisual model of communication presumes that communication is 'truly' multisensory (von Kriegstein et al., 2008). This model is based on the idea that information from face and voice is integrated via early interactions of presumably unisensory face- and voice-sensitive regions and that these interactions take place even when concurrent information about one modality is missing (von Kriegstein et al., 2008). The following sections will elaborate in more detail on traditional and recent perspectives of audiovisual communication. In section 2.1 I traditional perspectives on person and speech recognition will be described in brief and in section 2.2 a detailed account of the audiovisual model of communication will be given. Section 2.3 lists currently unresolved empirical questions concerning the audiovisual model that are the subject of this thesis.

3.1 Traditional Models of Communication

Traditional models of human communication assume that vocal and facial information is first processed independently in dedicated unisensory areas before information from both streams is integrated in higher-order, heteromodal brain regions. This traditional view is most explicitly described in the person-identity recognition model by (Ellis et al., 1997)(Figure 3-1).

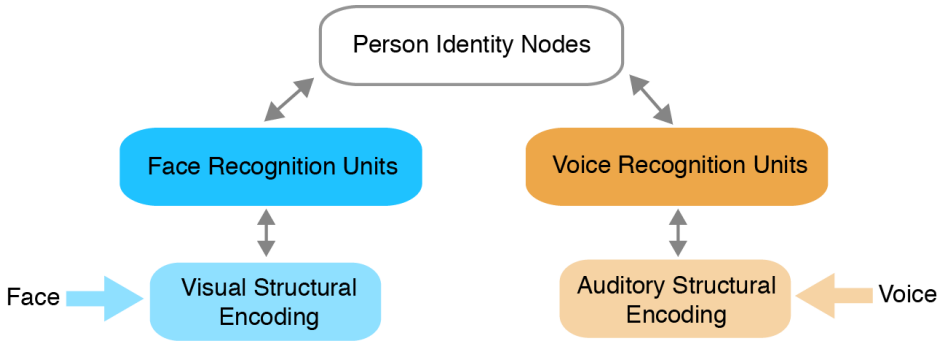

Figure 3-1 The Person-Identity Recognition Model. The traditional view on multisensory person recognition holds that information about voice and face is, after an initial structural encoding phase, passed on to face- and voice-recognition units, where identity is processed separately for face and voice. These units feed to heteromodal person-identity nodes (PINs) that integrate person-related information from face and voice. There is no interaction of facial and vocal identity processing before the heteromodal PIN (adapted from (Ellis et al., 1997)

Inspired by an influential face-recognition model (Bruce and Young, 1986) and empirical results from priming experiments, the authors postulate that the face and voice of a person are processed by parallel, separated information streams that culminate in so-called person-identity nodes (PIN). According to their model, face and voice are first structurally encoded at modality-specific processing stages followed by a familiarity assessment of face and voice in face-recognition units and voice-recognition units, respectively. Their output is then forwarded to PINs, which constitutes the first processing stage where person-identity information from different modalities is integrated. These heteromodal PINs integrate identity information from face and voice as well as other person-related details like name or gait. Importantly, this model explicitly negates a linkage between modality-specific face- and voice-recognition units but assumes that vocal and facial identity information is only integrated via the heteromodal PINs.

In contrast to person-identity recognition, models or perspectives on speech recognition are more implicitly grounded in empirical work. Probably owing to the complexity of speech, speech perception has mostly been approached from an auditory-only perspective and such work has been focused on disentangling phonetic, lexical and syntactic phenomena (see for review (Hickok and Poeppel, 2007) without taking facial information into consideration. However, a lot of research investigating the brain mechanisms of multisensory integration has been conducted using speech material.

Common underlying assumptions about the integration of vocal and facial information during speech recognition can be deduced from such experiments. Indeed, some of the most influential multisensory studies used speech-related stimulus material and could thereby show that viewing of silent lip-forms leads to an activation of auditory cortices (Calvert et al., 1997) and visual and auditory responses showed increased response levels during audiovisual speech perception when compared to auditory- or visual-only speech (Calvert et al., 1999). The authors speculated, however, that these activations are caused by feedback from heteromodal brain regions, rather than by direct interactions between auditory and visual areas (Calvert et al., 1999). Consequently, subsequent research on audiovisual speech was focused on the identification of heteromodal brain regions. Heteromodal regions have since been identified in various brain areas including the posterior STS, the inferior frontal gyrus (IFG) and the inferior parietal lobule (for review see (Calvert, 2001; Romanski, 2012)

In conclusion, research on person identification and speech perception has been dominated by the assumptions that vocal and facial information is solely processed in modality-specific areas before being integrated in heteromodal brain regions.

3.2 The Audiovisual Model of Communication

In contrast to traditional models, the recently developed audiovisual model of communication assumes that the neural architecture and mechanisms supporting communication in humans is intricately shaped by and adapted to the audiovisual nature of communication signals (von Kriegstein et al., 2008)(Figure 3-2). The audiovisual model has been motivated by behavioral and functional imaging findings investigating speaker and speech recognition under auditory-only listening conditions. Behaviorally, these experiments have shown that prior experience with a speaker's face—facial familiarity with the speaker—improves subsequent auditory-only voice and speech recognition, even when concurrent facial information is missing (Sheffert and Olson, 2004; von Kriegstein et al., 2008). On the neural level, this behavioral benefit is paralleled by and correlated to an increase of the BOLD signal in the task-relevant face-sensitive brain regions, in particular the face-identity sensitive FFA during speaker recognition and the face-movement-sensitive left posterior STS during speech-recognition (von Kriegstein et al., 2008). Also, the

FFA has been shown to be functionally connected with voice-identity sensitive anterior STS during voice recognition (von Kriegstein et al., 2005; von Kriegstein and Giraud, 2006). These results are astonishing given that face-sensitive regions are part of the visual pathway and considered to respond only to modality-specific visual input (Kanwisher and Yovel, 2006). Following these results, von Kriegstein et al. developed the audiovisual model of communication (von Kriegstein et al., 2008) within a predictive framework (Rao and Ballard, 1999; Friston, 2005).

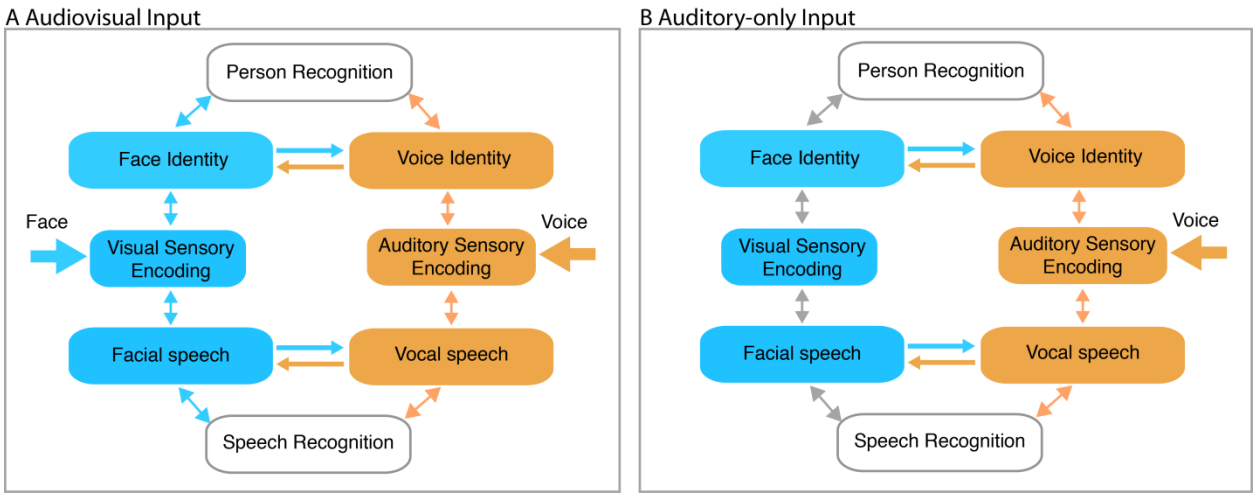

Figure 3-2 The audiovisual model of communication is a unified framework for person and speech recognition. Visual (blue) and auditory (orange) information from face and voice is first encoded in low-level sensory areas, before it diverges into specialized sensory areas that preferentially process identity or speech. Identity-specific information from face and voice is already integrated at a sensory level, in particular visual face-identity and auditory voice-identity sensitive areas interact with each other. In addition, the face-identity and voice-identity sensitive sensory areas interact with heteromodal areas supporting person recognition. Analogously, facial and vocal speech is integrated via interactions between the specialized sensory areas as well as via interactions with higher-order speech areas. The model assumes that the same lateral connectivity exists between the according face- and voice/speech-sensitive areas during auditory-only communication with familiar persons. In this case, the model predicts that learned facial features and dynamics of the person are 'simulated' to make auditory recognition more robust (adapted from (von Kriegstein et al., 2008).

The core idea of the audiovisual model of communication is that facial and vocal signals are integrated via functional connections between the relevant unisensory regions, in particular between voice-identity and face-identity sensitive regions and between facial- and vocal-speech sensitive regions (Figure 3-2A). The model furthermore assumes that these functional interactions are modulated by their behavioral relevance, that is to say functional connectivity between voice-identity and face-identity sensitive regions is

increased during person recognition, while functional connections between facial-speech and vocal-speech sensitive regions are increased during speech recognition. Importantly, these functional interactions are also at work under auditory-only listening conditions, in particular when we are familiar with a person (Figure 3-2B). In this case, the model assumes that learned facial and vocal speaker dynamics can be exploited to make communication under auditory-only listening conditions more robust. According to the model, this is instantiated by a 'simulation' of the speaker's face within the task-relevant face-sensitive regions. This way, face-sensitive regions provide the auditory sensory system with predictions about the incoming vocal signal leading to a facilitated auditory analysis of the vocal signal and to the optimized recognition of voice identity and speech.

3.3 Open Questions Addressed in this Thesis

The audiovisual model of communication is supported by behavioral as well as several functional imaging findings. However, there are several aspects which so far lack empirical evidence. The current thesis aims to fill in some of this missing information. The following sections will introduce the scientific questions addressed in this thesis and how they relate to the audiovisual model. Note that the empirical studies (Studies 1-3) each address one or more of these questions. For each question, the relevant study is referred to in bold letters.

3.3.1 Do Face-Sensitive Regions Respond During Early Sensory Processing Stages of Vocal Stimuli?

One of the most critical questions is whether the activation of face-identity sensitive regions during voice-recognition reflects an essential involvement in the auditory recognition process or whether it is merely a side effect after successful speaker recognition. It is, for example, conceivable that the activation of the face-sensitive regions reflects the listener's mental imagery of the speaker's face, once the voice has been recognized rather than an involvement of face-sensitive areas in the early sensory analysis of the vocal signal. One way to shed light into this question is to investigate the response latency of face-identity sensitive regions. Whereas an early effect (<200ms) would speak for an involvement during early sensory processes, a late effect (>200ms) would rather speak for visual imagery following speaker identification. To adjudicate between these two possibilities MEG data were analyzed and the response timing of face-identity sensitive

areas during auditory-only voice-recognition was determined. The results of this analysis are reported in **Study 1**.

3.3.2 Is the Sensory Encoding of Voices from Facially-Familiar Speakers Facilitated?

The audiovisual model postulates that face-sensitive regions provide the auditory system with predictions that optimize the auditory analysis of the vocal stimulus. So far, evidence for an optimized auditory encoding of voices existed only in the form of improved behavioral voice recognition accuracies (Sheffert and Olson, 2004; von Kriegstein et al., 2008). This behavioral benefit, however, may also arise due to improved memory recall of face-learned voices and not due to an optimized auditory analysis of the incoming voice. This question is addressed in **Study 1** by investigating whether there is evidence for a sensory facilitation mechanism in the auditory system and how it is linked to the behaviorally measured face-benefit.

3.3.3 When Do the First Voice-Identity Sensitive Brain Responses Occur?

Currently, there is little general knowledge about the timing of voice-identity sensitive neural responses and their dissociation from speech-related processes. These questions are of relevance to the audiovisual model, as they may help to establish a temporal framework for the neural processing of vocal sounds and their mapping to identities. To address these questions an MEG study optimized to disentangle voice-identity and speech-sensitive neural processes was done. Using source-localization and behavioral voice recognition performance data, it was possible to isolate brain regions sensitive to voices and assess their contribution to voice-identity recognition. These results are reported in **Study 2**.

3.3.4 Do the Functional Connectivity Findings Found During Voice Recognition Generalize to Speech Recognition?

The audiovisual model assumes that face- and voice/speech-sensitive regions engage in the exchange of predictions and prediction-errors during the sensory analysis of the vocal stimulus. An important pre-requisite for this exchange is that face- and voice/speech-sensitive areas communicate during vocal perception, or in other words that they are

functionally connected. Previous functional imaging studies could show that, during voice-identity recognition, such a functional connectivity exists between FFA and voice-identity sensitive STS. According to the audiovisual model, this connectivity finding should also translate to speech recognition where face-movement-sensitive pSTS and auditory areas supporting speech recognition are supposed to interact. This question is addressed in **Study 3** by performing a functional connectivity analysis on fMRI data.

3.3.5 Is Functional Connectivity Between Face- and Voice-Sensitive Regions Direct?

The audiovisual model assumes that the functional connection between face-identity and voice-identity sensitive regions is direct, rather than mediated via a heteromodal brain region like the PIN. A general shortcoming of functional connectivity findings is that they cannot establish directionality. Therefore, if two regions are functionally connected this might either mean that the two regions share direct functional connectivity or that there is a third, mediating region, through which the two regions are functionally connected. One way of finding supporting evidence that a functional connection is direct, is by investigating participants with a perceptual deficit that compromises the potential mediating region. In order to exclude the PIN as a potential mediating region, functional connectivity of face-identity and voice-identity sensitive regions were compared between a group of participants with a face-identity recognition deficit (prosopagnosia) and a group of participants with normal face recognition abilities. Results are reported in **Study 3.**

4 Summary of Empirical Studies

4.1 Study 1

The audiovisual model postulates that face-identity sensitive regions optimize the auditory sensory analysis of vocal signals. Two testable hypotheses follow from this assumption: (i) face-sensitive regions respond already during early sensory processing stages of facially-familiar voices; (ii) the auditory analysis is facilitated during recognition of facially-familiar voices. In order to investigate these questions, an MEG experiment designed to reveal sensory processes during auditory-only voice-identity recognition of facially-familiar speakers was performed. Before MEG recording, participants were trained to recognize the voices of six male speakers. The crucial experimental manipulation consisted of the type of training: half of the speakers learned the speaker's face in association with voice in the form of video-clips of the talking speakers (face-learned voices) while the other half of the speakers learned with a visual control stimulus in the form of an accompanying symbol visualizing the speaker's occupation (occupation-learned voices). In addition, all speakers were learned with a name. During MEG recording, participants listened to speech samples from these, now familiar, speakers and performed a voice-identity recognition task. Data analysis was based on the comparison of event-related activity during recognition of face-learned and occupation-learned voices. To address the first question, regarding when FFA responds to face-learned voices, source activity was estimated using a minimum-norm approach and activity in the fusiform gyrus was statistically examined. It was found that the posterior fusiform gyrus showed significantly increased activity to face-learned compared to occupation-learned voices as soon as 110ms peri-stimulus ($t=2.71$, $df=18$, $p=0.004$). To address the second question, regarding whether the auditory sensory analysis is facilitated for face-learned speakers, a peak latency analysis was performed on the two major auditory components (M100, M200). It was found that the M200 peaked systematically earlier for face-learned than occupation-learned voices ($n=17$, $t=2.24$, $df=16$, $p=0.02$), with an average latency difference of 11ms. For the M100, peak latencies did not differ significantly between conditions. In addition to these neural effects, a corresponding behavioral benefit was found: consistent with previous studies (Sheffert and Olson, 2004; von Kriegstein et al., 2008), participants had an average face benefit (% correct face-

learned voices - % correct occupation-learned voices) of 3.85%. To examine whether this behavioral benefit is linked to the temporal facilitation of the M200, the correlation between individual face benefits and M200 peak latency differences was additionally investigated. This analysis revealed that participants who profited most from face-voice training, showed stronger M200 facilitation (i.e. faster peak latencies to face-learned compared to occupation-learned voices; multiple regression: n=17, t=-2.58, df=14, p=0.011). Together, these results show that familiarity with a speaker's face alters the sensory processing of vocal sounds. Within the framework of the audiovisual model, they clarify two central questions: (i) face-sensitive areas respond during sensory encoding stages of voices, which strongly suggests that they are actively involved in the auditory analysis rather than being a side effect of successful person recognition; and (ii) the activation of face-identity sensitive regions is followed by a temporal facilitation of the auditory analysis indicating that the available visual information about the speaker optimizes the auditory analysis of the vocal sound.

4.2 Study 2

In line with other voice perception models (Belin et al., 2004), the audiovisual model assumes that voice-identity and speech are processed along partially different neural pathways. In contrast to speech-sensitive regions that have been shown to be predominantly located in the left hemisphere, voice-sensitive regions have been primarily found in the right hemisphere. In particular, the right anterior STS has proven to be sensitive to voice identities. However, its behavioral relevance for voice-identity recognition has not yet been established. Here, two, so far unresolved, questions were addressed relating to voice-identity perception; (i) when and where does the neural processing of voice- and speech-specific information dissociate? And, (ii) what is the role of right anterior STS during voice-identity processing? To investigate these questions, MEG was utilized together with behaviorally assessed voice-recognition performance data. Participants were first familiarized with the voices and names of six male speakers. During MEG recording, participants listened to speech samples from the familiarized speakers and performed either a voice-identity or speech recognition task. In order to disentangle voice-identity related processes from speech perception, data analysis was based on the comparison of neural activity between the two tasks (i.e. voice task – speech task). The

timing of task-related neural activity modulations was first determined on the sensory-level using cluster-permutation statistics. Source analysis was performed to determine the involvement of the right and left STS during both tasks. At the sensory-level, It was found that voice- and speech-sensitive neural processes diverged around the 200ms latency ($p<0.05$ for two clusters). The source analysis revealed that this dissociation was due to the differential involvement of the right STS and left STS. Two different voice-sensitive responses were found along the right STS: (i) in the posterior STS, a general increase of source activity during the voice-identity compared to the speech-recognition was found (n=19, t=2.7263, df=18, p=0.0069) and (ii) in the anterior STS a correlation of task-related source activity and voice-recognition performance was found (n=19, r=0.6835, p=0.0013); participants with higher activity during the voice compared to the speech task performed better in the voice-recognition task. Both voice-sensitive effects occurred within the same 200ms time frame. In contrast to the right STS, the left STS showed the reverse pattern: the left middle STS was significantly more activated during speech compared to voice recognition (t=2.3416, df=18, p=0.0155). These results suggest that voice-identity and speech perception diverge around the 200ms latency and that it is during this particular time frame that activity in the right anterior STS is crucial for successful voice recognition.

4.3 Study 3

The existence of functional connections between task-relevant face- and voice-sensitive areas is a pre-requisite of the audiovisual model. Previous research has shown that functional connectivity during voice-identity recognition exists between face-identity sensitive FFA and the voice-sensitive STS. Also, a case-study on a participants with prosopagnosia (i.e. a deficit recognizing person identity from faces) showed that this functional connectivity is not necessarily dependent on an available face-identity representation. This suggests that the functional connectivity between FFA and STS is direct, rather than being mediated by a PIN (as it is unlikely that prosopagnosics have visual imagery of person's face after having identified the person from voice). Here, two questions related to the connectivity of the task-relevant face-sensitive and voice/speech-sensitive regions were addressed: (i) do the connectivity findings during voice-recognition generalize to the speech domain, specifically, is there a functional connection between facial- and vocal-speech sensitive areas during speech recognition? And, (ii) Can the

functional connectivity findings observed in a single prosopagnosic participant be replicated in a group study? To answer these questions, a functional connectivity analysis on a previously acquired fMRI data set was performed (von Kriegstein et al., 2008). This data set was acquired on a group of 17 prosopagnosics and 17 normal participants. Before scanning, participants underwent an audiovisual training in which they learned to recognize six male voices. Half of the voices were learned together with the dynamic face of the speaker, while the other half of the voices were learned in association with a symbol visualizing the speaker's occupation. During scanning, participants listened to voice samples and performed either a voice-recognition or speech-recognition task. In short, the data set was based on a 2x2x2 factorial design, with the factors learning type (voice-face training or occupation-face training), task (voice recognition or speech recognition) and group (prosopagnosics and matched typical participants). To investigate whether facial- and vocal-speech sensitive areas are functionally connected during speech recognition, a functional connectivity analysis (PPI) was performed using the face-movement-sensitive left posterior STS (i.e. an area involved in the recognition of facial speech) as a seed region and the speech-intelligible left anterior STS (i.e. an area involved in the recognition of vocal speech). To test whether functional connectivity was increased during speech compared to voice-identity recognition as well as for face-learned voices compared to occupation-learned voices the following interaction contrast was used: ((speech task/ voice-face > speech task/ voice-occupation) > (voice task/ voice-face > voice task/ voice-occupation)). A significant functional interaction between face-movement sensitive pSTS and the speech-intelligibility sensitive aSTS was found ($p < 0.05$ FWE corrected for aSTG/S), indicating that these two areas interacted when participants recognized speech from face-learned speakers. To investigate the second research question, regarding whether functional connectivity between FFA and voice-sensitive STS is preserved in prosopagnosia, A functional connectivity analysis was performed using the FFA as a seed region and voice-sensitive right STS as a target region. To test whether functional connectivity was increased during voice compared to speech recognition and for face-learned voices compared to occupation-learned voices the following interaction contrast was used: ((voice task/ voice-face > voice task/ voice occupation) > (speech task/ voice-face > speech task/ voice-occupation)). A significant functional connection between FFA and voice-sensitive right STS was found in prosopagnosics during recognition of face-learned

speakers ($p < 0.042$ FWE corrected for the voice-sensitive right STS). This functional connectivity was not significantly different in prosopagnosics and controls, even at a lenient statistical threshold ($p < 0.01$, uncorrected). These two connectivity findings, regarding the functional connectivity between face-movement- and speech-intelligibility-sensitive areas and between face-identity and voice-identity sensitive areas in prosopagnosics support and validate the audiovisual model of communication. They suggest that a modulation of functional connectivity between the task-relevant face- and voice/speech-sensitive regions is a general mechanism in operation during auditory-only communication. The results furthermore suggest that this connection is direct and occurring on a sensory level, rather than being mediated via heteromodal brain regions.

5 Conclusions

The current work demonstrates that the brain adapts to speaker-specific regularities of faces and voices and exploits these to make vocal communication more robust. This is not only visible in an extended sensory network (i.e. the inclusion of face-sensitive areas during auditory-only vocal communication), but also in a change of network dynamics and an optimized sensory processing scheme of vocal stimuli. These findings strongly suggest that the way we perceive voices and speech is critically dependent on whether or not we are familiar with a speaker. These findings also imply that voice- and speech perception can be best understood when taking their natural co-occurrence with faces into account.

5.1 Implications for the Audiovisual Model

With respect to the audiovisual model of communication, the current work lends important evidence to some of its core claims. Specifically, this work shows that knowledge about a speaker's face-voice dynamics lead—during subsequent auditory-only vocal communication—to (i) the early involvement of face-sensitive areas even before the speaker's identity has been recognized, to (ii) a facilitation of the auditory analysis of the vocal sound and to (iii) the integration of facial information into the auditory analysis via functional connections between the task-relevant face- and voice/speech-sensitive regions. In addition to these affirmative findings, the current work challenges the audiovisual model of communication in one critical aspect. The audiovisual model suggests that face-sensitive regions are, under auditory-only conditions, activated by functional connections via voice-identity sensitive regions. This scenario predicts that face-sensitive regions become activated, together with, or shortly after the activation of voice-identity sensitive regions. However, this prediction is not confirmed by the current empirical findings. While Study 1 showed that face-identity sensitive regions are activated roughly 100ms after voice-onset, Study 2 suggests that voice-identity sensitive areas are not activated before 200ms after voice-onset. There are several possibilities explaining this discrepancy. A likely alternative pathway might consist of the activation of face-identity sensitive areas by lower-level auditory cortices which are involved in the general sensory encoding of sounds, rather than by voice-identity sensitive regions.

5.2 Implications for Multisensory Brain Research

In more general terms, the presented data is highly relevant for the understanding of the multisensory brain. Research into this topic has mostly been performed in settings where information from two senses, for example visual and auditory information, is presented simultaneously (Ghazanfar and Schroeder, 2006; Kayser, 2010). Such research has, for example, shown that activity in early unisensory regions (i.e. early in the cortical hierarchy) can be modulated by multisensory input, that is by input from a non-preferred modality (Ghazanfar and Schroeder, 2006; Kayser, 2010). Recent research also suggests that within unisensory brain areas, the functionality of non-preferred input may consist in a facilitation or enhancement of the sensory processing of the (modality-)preferred stimulus (van Wassenhove et al., 2005; Lakatos et al., 2007; Kayser et al., 2010). Additionally, there is evidence that the integration of information from different modalities is, at least partially, instantiated by direct interactions between the involved unisensory cortices (Ghazanfar et al., 2005; Ghazanfar et al., 2008). The results presented in the current thesis indicate that these brain mechanisms, also hold under unisensory conditions given that prior knowledge about the missing modality is available and can be 'filled in'. These findings challenge unisensory approaches to brain function and indicate that perception of naturally multisensory stimuli, like faces and voices, may always be multisensory.

In conclusion, the current thesis does not only provide relevant insight into the neural mechanisms of vocal communication, but also crucially adds to our understanding of how perceptual mechanisms are shaped by prior experience and how the brain is tuned to natural stimuli and their co-occurrence in the world..

6 Published Manuscript 1

Early auditory sensory processing of voices is facilitated by visual mechanisms

Sonja Schall[1], Stefan Kiebel[1,2], Burkhard Maess[1] and Katharina von Kriegstein[1,3]

[1]Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, Leipzig, Germany

[2]University Clinic Jena, [2]Biomagnetic Centre, Dept of Neurology, Erlanger Allee 101, Jena, Germany

[3]Humboldt University Berlin, ,Institute of Psychology, Rudower Chaussee 18, Berlin, Germany

Published in:

Schall, S., Kiebel, S.J., Maess, B., von Kriegstein, K., 2013. Early auditory sensory processing of voices is facilitated by visual mechanisms. Neuroimage 77, 237-245.

http://dx.doi.org/10.1016/j.neuroimage.2013.03.043

7 Published Manuscript 2

Voice identity recognition: Functional division of the right superior temporal sulcus and its behavioral relevance

Sonja Schall[1], Stefan J. Kiebel[2,,3], Burkhard Maess[1] and Katharina von Kriegstein[1,4]

[1]*Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, Leipzig, Germany*

[2]*University Clinic Jena, Biomagnetic Centre, Dept of Neurology, Erlanger Allee 101, Jena, Germany*

[3]*Technical University Dresden, Institute of Psychology, , Dresden, Germany*

[4]*Humboldt University Berlin, ,Institute of Psychology, Rudower Chaussee 18, Berlin, Germany*

Published in:

Schall, S., Kiebel, S.J., Maess, B., von Kriegstein, K., 2014. Voice Identity Recognition: Functional Division of the Right STS and Its Behavioral Relevance. J Cogn Neurosci, 1-12.

http://dx.doi.org/10.1162/jocn_a_00707

8 Published Manuscript 3

Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception.

Sonja Schall[1] and Katharina von Kriegstein[1,2]

[1] *Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, 04103 Leipzig, Germany*

[2] *Humboldt University of Berlin, Institute for Psychology, Rudower Chaussee 18, 12489 Berlin, Germany*

Published in:

Schall, S., von Kriegstein, K., 2014. Functional Connectivity between Face-Movement and Speech-Intelligibility Areas during Auditory-Only Speech Perception. PLoS One 9, e86325.

http://dx.doi.org/10.1371/journal.pone.0086325

9 References

Altmann CF, Gomes de Oliveira Junior C, Heinemann L, Kaiser J (2010) Processing of spectral and amplitude envelope of animal vocalizations in the human auditory cortex. Neuropsychologia 48:2824-2832.

Andics A, McQueen JM, Petersson KM, Gal V, Rudas G, Vidnyanszky Z (2010) Neural mechanisms for voice recognition. Neuroimage 52:1528-1540.

Barbeau EJ, Taylor MJ, Regis J, Marquis P, Chauvel P, Liegeois-Chauvel C (2008) Spatio temporal dynamics of face recognition. Cereb Cortex 18:997-1009.

Beauchemin M, De Beaumont L, Vannasing P, Turcotte A, Arcand C, Belin P, Lassonde M (2006) Electrophysiological markers of voice familiarity. Eur J Neurosci 23:3081-3086.

Bee MA, Gerhardt HC (2002) Individual voice recognition in a territorial frog (Rana catesbeiana). P Roy Soc B-Biol Sci 269:1443-1448.

Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14:2105-2109.

Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13:17-26.

Belin P, Fecteau S, Bedard C (2004) Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8:129-135.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309-312.

Bentin S, Deouell LY (2000) Structural encoding and identification in face processing: erp evidence for separate mechanisms. Cogn Neuropsychol 17:35-55.

Bentin S, Allison T, Puce A, Perez E, McCarthy G (1996) Electrophysiological Studies of Face Perception in Humans. Journal of cognitive neuroscience 8:551-565.

Blasi A, Mercure E, Lloyd-Fox S, Thomson A, Brammer M, Sauter D, Deeley Q, Barker GJ, Renvall V, Deoni S, Gasston D, Williams SC, Johnson MH, Simmons A, Murphy DG (2011) Early specialization for voice and emotion processing in the infant brain. Curr Biol 21:1220-1224.

Bruce V, Young A (1986) Understanding face recognition. Br J Psychol 77 (Pt 3):305-327.

Caharel S, Jiang F, Blanz V, Rossion B (2009a) Recognizing an individual face: 3D shape contributes earlier than 2D surface reflectance information. Neuroimage 47:1809-1818.

Caharel S, d'Arripe O, Ramon M, Jacques C, Rossion B (2009b) Early adaptation to repeated unfamiliar faces across viewpoint changes in the right hemisphere: evidence from the N170 ERP component. Neuropsychologia 47:639-643.

Calvert GA (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex 11:1110-1123.

Calvert GA, Campbell R (2003) Reading speech from still and moving faces: the neural substrates of visible speech. Journal of cognitive neuroscience 15:57-70.

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.

Capilla A, Belin P, Gross J (2013) The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. Cereb Cortex 23:1388-1395.

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5:e1000436.

Charest I, Pernet CR, Rousselet GA, Quinones I, Latinus M, Fillion-Bilodeau S, Chartrand JP, Belin P (2009) Electrophysiological evidence for an early processing of human voices. BMC Neurosci 10:127.

Crowley KE, Colrain IM (2004) A review of the evidence for P2 being an independent component process: age, sleep and modality. Clin Neurophysiol 115:732-744.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423-3431.

De Lucia M, Clarke S, Murray MM (2010) A temporal hierarchy for conspecific vocalization discrimination in humans. J Neurosci 30:11210-11221.

Deffke I, Sander T, Heidenreich J, Sommer W, Curio G, Trahms L, Lueschow A (2007) MEG/EEG sources of the 170-ms response to faces are co-localized in the fusiform gyrus. Neuroimage 35:1495-1501.

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109:E505-514.

Eger E, Schyns PG, Kleinschmidt A (2004) Scale invariant adaptation in fusiform face-responsive regions. Neuroimage 22:232-242.

Eimer M (2011) The face-sensitive N170 component of the event-related brain potential. In: The Oxford Handbook of Face Perception. (Calder A, Rhodes G, Johnson MH, Haxby J, eds), pp 329-344: Oxford University Press.

Ellis HD, Jones DM, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. British Journal of Psychology 88:14.

Feng AS, Arch VS, Yu ZL, Yu XJ, Xu ZM, Shen JX (2009) Neighbor-Stranger Discrimination in Concave-Eared Torrent Frogs, Odorrana tormota. Ethology 115:851-856.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970-973.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815-836.

Gasser H, Amezquita A, Hodl W (2009) Who is Calling? Intraspecific Call Variation in the Aromobatid Frog Allobates femoralis. Ethology 115:596-607.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278-285.

Ghazanfar AA, Chandrasekaran C, Logothetis NK (2008) Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J Neurosci 28:4457-4469.

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J Neurosci 25:5004-5012.

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108:1197-1208.

Grossmann T, Oberecker R, Koch SP, Friederici AD (2010) The developmental origins of voice processing in the human brain. Neuron 65:852-858.
Gunji A, Koyama S, Ishii R, Levy D, Okamoto H, Kakigi R, Pantev C (2003) Magnetoencephalographic study of the cortical activity elicited by human voice. Neurosci Lett 348:13-16.
Halgren E, Raij T, Marinkovic K, Jousmaki V, Hari R (2000) Cognitive response profile of the human fusiform face area as determined by MEG. Cereb Cortex 10:69-81.
Hall DA, Fussell C, Summerfield AQ (2005) Reading fluent speech from talking faces: typical brain networks and individual differences. J Cogn Neurosci 17:939-953.
Harris AM, Aguirre GK (2008) The effects of parts, wholes, and familiarity on face-selective responses in MEG. J Vis 8:4 1-12.
Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4:223-233.
Henson RN, Mouchlianitis E, Friston KJ (2009) MEG and EEG data fusion: simultaneous localisation of face-evoked responses. Neuroimage 47:581-589.
Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nature reviews Neuroscience 8:393-402.
Ishai A (2008) Let's face it: it's a cortical network. Neuroimage 40:415-419.
Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond B Biol Sci 361:2109-2128.
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.
Kayser C (2010) The multisensory nature of unisensory cortices: a puzzle continued. Neuron 67:178-180.
Kayser C, Logothetis NK, Panzeri S (2010) Visual enhancement of the information representation in auditory cortex. Curr Biol 20:19-24.
Kloth N, Dobel C, Schweinberger SR, Zwitserlood P, Bolte J, Junghofer M (2006) Effects of personal familiarity on early neuromagnetic correlates of face perception. The European journal of neuroscience 24:3317-3321.
Kriegstein KV, Giraud AL (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22:948-955.
Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279-292.
Lang CJ, Kneidl O, Hielscher-Fastabend M, Heckmann JG (2009) Voice recognition in aphasic and non-aphasic stroke patients. Journal of neurology 256:1303-1306.
Leff AP, Schofield TM, Stephan KE, Crinion JT, Friston KJ, Price CJ (2008) The cortical dynamics of intelligible speech. J Neurosci 28:13209-13215.
Leopold DA, Rhodes G (2010) A Comparative View of Face Perception. J Comp Psychol 124:233-251.
Levy DA, Granot R, Bentin S (2001) Processing specificity for human voice stimuli: electrophysiological evidence. Neuroreport 12:2653-2657.
Levy DA, Granot R, Bentin S (2003) Neural sensitivity to human voices: ERP evidence of task and attentional influences. Psychophysiology 40:291-305.
Liu J, Harris A, Kanwisher N (2002) Stages of processing in face perception: an MEG study. Nat Neurosci 5:910-916.
Liu J, Higuchi M, Marantz A, Kanwisher N (2000) The selectivity of the occipitotemporal M170 for faces. Neuroreport 11:337-341.

Martin BA, Tremblay KL, Korczak P (2008) Speech evoked potentials: from the laboratory to the clinic. Ear and hearing 29:285-313.
Matsumoto R, Imamura H, Inouchi M, Nakagawa T, Yokoyama Y, Matsuhashi M, Mikuni N, Miyamoto S, Fukuyama H, Takahashi R, Ikeda A (2011) Left anterior temporal cortex actively engages in speech perception: A direct cortical stimulation study. Neuropsychologia 49:1350-1354.
Meltzoff AN, Moore MK (1977) Imitation of facial and manual gestures by human neonates. Science 198:75-78.
Mesulam MM (1998) From sensation to cognition. Brain 121 (Pt 6):1013-1052.
Obleser J, Wise RJ, Alex Dresner M, Scott SK (2007) Functional integration across brain regions improves speech perception under adverse listening conditions. The Journal of neuroscience : the official journal of the Society for Neuroscience 27:2283-2289.
Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G (2005) Functional anatomy of biological motion perception in posterior temporal cortex: an FMRI study of eye, mouth and hand movements. Cereb Cortex 15:1866-1876.
Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice cells in the primate temporal lobe. Curr Biol 21:1408-1415.
Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK (2008) A voice region in the monkey brain. Nat Neurosci 11:367-374.
Puce A, Allison T, Bentin S, Gore JC, McCarthy G (1998) Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 18:2188-2199.
Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79-87.
Renvall H, Staeren N, Siep N, Esposito F, Jensen O, Formisano E (2012) Of cats and women: temporal dynamics in the right temporoparietal cortex reflect auditory categorical processing of vocalizations. Neuroimage 62:1877-1883.
Romanski LM (2012) Integration of faces and vocalizations in ventral prefrontal cortex: implications for the evolution of audiovisual speech. Proc Natl Acad Sci U S A 109 Suppl 1:10717-10724.
Rosen S, Wise RJ, Chadha S, Conway EJ, Scott SK (2011) Hemispheric asymmetries in speech perception: sense, nonsense and modulations. PLoS One 6:e24672.
Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17:1147-1153.
Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat Neurosci 8:107-113.
Sai FZ (2005) The Role of the Mother's Voice in Developing Mother's Face Preference: Evidence for Intermodal Perception at Birth. Infant and Child Development 14:22.
Sams M, Hietanen JK, Hari R, Ilmoniemi RJ, Lounasmaa OV (1997) Face-specific responses from the human inferior occipito-temporal cortex. Neuroscience 77:49-55.
Schweinberger SR (2001) Human brain potential correlates of voice priming and voice recognition. Neuropsychologia 39:921-936.
Schweinberger SR, Huddy V, Burton AM (2004) N250r: a face-selective brain response to stimulus repetitions. Neuroreport 15:1501-1505.

Schweinberger SR, Pickering EC, Jentzsch I, Burton AM, Kaufmann JM (2002) Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. Brain research Cognitive brain research 14:398-409.
Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain : a journal of neurology 123 Pt 12:2400-2406.
Sergent J, Ohta S, MacDonald B (1992) Functional neuroanatomy of face and object processing. A positron emission tomography study. Brain 115 Pt 1:15-36.
Sheffert SM, Olson E (2004) Audiovisual speech facilitates voice learning. Percept Psychophys 66:352-362.
Sidtis D, Kreiman J (2012) In the beginning was the familiar voice: personally familiar voices in the evolutionary and contemporary biology of communication. Integrative psychological & behavioral science 46:146-159.
Simion F, Leo I, Turati C, Valenza E, Barba BD (2007) How face specialization emerges in the first months of life. From Action to Cognition 164:169-185.
Sumby WH, Pollack I (1954) Visual Contribution to Speech Intelligibility in Noise. J Acoust Soc Am 26:212-215.
Taylor MJ, Bayless SJ, Mills T, Pang EW (2011) Recognising upright and inverted faces: MEG source localisation. Brain research 1381:167-174.
Van Lancker D, Kreiman J (1987) Voice discrimination and recognition are separate abilities. Neuropsychologia 25:829-834.
Van Lancker DR, Canter GJ (1982) Impairment of voice and face recognition in patients with hemispheric damage. Brain Cogn 1:185-195.
Van Lancker DR, Kreiman J, Cummings J (1989) Voice perception deficits: neuroanatomical correlates of phonagnosia. Journal of clinical and experimental neuropsychology 11:665-674.
van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102:1181-1186.
von Kriegstein K, Giraud AL (2006) Implicit multisensory associations influence voice recognition. PLoS Biol 4:e326.
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, Kleinschmidt A, Sterzer P, Giraud AL (2005) Interaction of face and voice areas during speaker recognition. J Cogn Neurosci 17:367-376.
von Kriegstein K, Dogan O, Gruter M, Giraud AL, Kell CA, Gruter T, Kleinschmidt A, Kiebel SJ (2008) Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A 105:6747-6752.
Zaske R, Schweinberger SR, Kaufmann JM, Kawahara H (2009) In the ear of the beholder: neural correlates of adaptation to voice gender. Eur J Neurosci 30:527-534.

Appendix. List of Figures

Erklärung über die selbstständige Verfassung der Arbeit

Hiermit erkläre ich, dass die vorliegende Arbeit ohne unzulässige Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde und dass die aus fremden Quellen direkt oder indirekt übernommenen Gedanken in der Arbeit als solche erkenntlich gemacht worden sind.

Sonja Schall

Leipzig, den 24. September 2013