Neuroscience 256 (2014) 230–241

A NETWORK ANALYSIS OF AUDIOVISUAL AFFECTIVE SPEECH PERCEPTION

H. JANSMA,[a,b] A. ROEBROECK[b] AND T. F. MÜNTE[a]*

[a] *Department of Neurology, University of Lübeck, Lübeck, Germany*

[b] *Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands*

Abstract—In this study we were interested in the neural system supporting the audiovisual (AV) integration of emotional expression and emotional prosody. To this end normal participants were exposed to short videos of a computer-animated face voicing emotionally positive or negative words with the appropriate prosody. Facial expression of the face was either neutral or emotionally appropriate. To reveal the neural network involved in affective AV integration, standard univariate analysis of functional magnetic resonance (fMRI) data was followed by a random-effects Granger causality mapping (RFX-GCM). The regions that distinguished emotional from neutral facial expressions in the univariate analysis were taken as seed regions. In trials showing emotional expressions compared to neutral trials univariate analysis showed activation primarily in bilateral amygdala, fusiform gyrus, middle temporal gyrus/superior temporal sulcus and inferior occipital gyrus. When employing either the left amygdala or the right amygdala as a seed region in RFX-GCM we found connectivity with the right hemispheric fusiform gyrus, with the indication that the fusiform gyrus sends information to the Amygdala. These results led to a working model for face perception in general and for AV-affective integration in particular which is an elaborated adaptation of existing models.

Key words: audiovisual speech, emotion, facial affect perception, amygdala, Granger causality.

INTRODUCTION

Audiovisual (AV) integration in the perception of speech is the rule rather than the exception. For example, the presence of congruent visual (V) information leads to a considerable improvement of intelligibility under noisy conditions (Sumby and Pollack, 1954; Schwartz et al., 2004; Ross et al., 2007) which can be in the range of 10 dB. On the other hand, incongruent V information may induce the striking McGurk illusion (McGurk and MacDonald, 1976) during which syllables are perceived that are neither heard nor seen (e.g., percept/da/, auditory information: /ba/, visual information/ga/). This illusion suggests that AV integration during speech perception is a rather automatic process.

A number of recent studies have addressed the neural underpinnings of AV integration in speech perception and consistently found two different brain areas, the inferior frontal gyrus (IFG) and the superior temporal sulcus (STS) (Calvert et al., 2000; Calvert and Campbell, 2003; Sekiyama et al., 2003; Wright et al., 2003; Barraclough et al., 2005; Szycik et al., 2008, 2009, 2012). Interactions of the STS and IFG are captured by the AV-motor integration model of speech perception (Skipper et al., 2007). This model posits the formation of a sensory hypothesis in STS which is further specified in terms of the motor goal of the articulatory movements established in the pars opercularis of the IFG. While the advantage of AV information over unimodal A or V information for the perception of speech is clearly established, the question arises whether bimodal AV information is also advantageous for other types of information. Indeed, it has been shown that the identity of a speaker might be recognized from both visual and auditory information but that there is a strong interaction between these kinds of information (von Kriegstein et al., 2005; Blank et al., 2011).

The topic of the present study is AV integration of affective information transmitted by the voice or face of a speaker. (Scherer, 2003) compared the recognition accuracy for vocal and facial emotions as they had been obtained in a number of studies either using unimodal vocal expression (previously reviewed by Scherer et al. (2001)) or unimodal facial expression (previously reviewed by Ekman (1994)). If one limits the analysis to studies of Western faces and voices, recognition accuracy for the emotions anger, fear, joy, sadness, disgust, and surprise ranged between 31% (disgust) and 77% (anger) for vocal emotions and between 77% (fear) and 95% (joy) for facial expressions. Thus, for both modalities recognition for most emotions is far from perfect. The question therefore arises whether the combination of both types of information will increase recognition accuracy. Indeed, a number of studies have revealed clear behavioral face–voice integration effects for affective stimuli: for example, de Gelder and

*Correspondence to: T. F. Münte, Department of Neurology, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Gremany. Tel: +49-45150052925; fax: +49-4515005457.
E-mail address: Thomas.muente@neuro.uni-luebeck.de (T. F. Münte).
Abbreviations: AV, audiovisual; BOLD, blood oxygen level dependent; EMO, emotional; fMRI, functional magnetic resonance; FOV, field of view; IFG, inferior frontal gyrus; MTG, middle temporal gyrus; NEU, neutral; RFX-GCM, random effects Granger causality mapping; STS, superior temporal sulcus; V, visual; VOI, volume of interest.

0306-4522/13 $36.00
http://dx.doi.org/10.1016/j.neuroscience.2013.10.047

Vroomen (2000) obtained affective ratings of facial stimuli that were morphed to represent a continuum between two facial expressions. These ratings were clearly influenced by the concurrent presentation of an affective vocalization, even under instructions to ignore the voice. A comparable effect was also obtained for ratings of affect for the vocalizations. Further research revealed that crossmodal interaction occurs even, if facial expressions are presented subliminally (de Gelder et al., 2002). In a similar vein, Collignon et al. (2008) found that the irrelevant information affected processing, even if participants were asked to ignore one sensory modality, thus further suggesting mandatory integration of visual and auditory emotional information. Further evidence for early interaction comes from electrophysiological studies. Incongruent pairings of an affective vocalization and a facial emotion have been found to evoke a negativity akin to the mismatch negativity around 180 ms (de Gelder et al., 1999), whereas in another study affectively congruent voice/face pairings gave rise to an enhanced amplitude of the auditory N1 response (Pourtois et al., 2000).

Neuroimaging has been used to shed light on the functional neuroanatomy of the processing of affective facial and vocal information. Facial expressions, even when presented subliminally, have been shown to activate the amygdala with the greatest responses observed for expressions of fear (Breiter et al., 1996;Morris et al., 1996; Gelder et al., 1997; Whalen et al., 2001; Williams, 2002; Noesselt et al., 2005; Vuilleumier and Pourtois, 2007). Interestingly, robust activations for the amygdala have been observed when emotional processing is implicit, whereas explicit emotion recognition often leads to a deactivation of the amygdala (Critchley et al., 2000). Other regions that have been found with regard to the processing of facial expressions include the orbitofrontal cortex which is activated by fearful (Vuilleumier et al., 2001), angry but not sad expressions (Blair et al., 1999). The latter dissociation of the processing of angry and sad expressions has also been found for the anterior cingulate cortex (Blair et al., 1999). Adolphs (2002a,b) has summarized the imaging and lesion findings and has suggested a neuroanatomical model of affect recognition from facial expressions.

With regard to the processing of voice, the seminal study of Belin et al. (2002) suggested a prominent role of the STS. Even earlier, the amygdala has been implicated by a number of studies for the processing of affective vocalizations (Phillips et al., 1998; Morris et al., 1999). Buchanan et al. (2000) compared the detection of emotional and semantic properties of stimuli with the former giving rise to activity in the right inferior frontal lobe. Adolphs et al. (2002) found that the right frontoparietal cortex, left frontal operculum and bilateral frontal polar cortex (area 10) are critical to recognizing emotion from prosody. Investigating vocal attractiveness as a paralinguistic cue during social interactions, Bestelmeyer et al. (2012) similarly found that inferior frontal regions in addition to voice-sensitive auditory areas were strongly correlated with implicitly perceived vocal attractiveness. In an effort to distinguish the neural representation of different kinds of emotion, Ethofer et al. (2009) presented pseudowords spoken with different affective connotation (anger, sadness, neutral, relief, and joy) and subjected their functional magnetic resonance (fMRI) activations to multivariate pattern analysis. These authors successfully decoded the different vocal emotions from fMRI in bilateral voice-sensitive areas.

With regard to crossmodal integration of emotional information, a first fMRI study required participants to categorize static facial expression as fearful or happy while simultaneously presented emotional vocalizations were to be ignored (Dolan et al., 2001). Activation of the left amygdala was stronger when both, facial expression and voice signaled fear, thus suggesting a role of the amygdala in the crossmodal integration of fear. Building on this early study, Ethofer et al. (2006a,b) found that the crossmodal bias observed in affective ratings of fear correlated with activity in the amygdalae. Applying the criterion of supra-additivity (i.e., the response to face–voice pairings in combination is greater than the sum of the activations to each of the modalities presented separately), Pourtois et al. (2005) delineated the middle temporal gyrus (MTG) as a core region for the crossmodal integration of a variety of emotions. While these earlier studies used static facial expressions, Kreifelts et al. (2007) employed dynamic video-clips and tested a number of different emotional expressions. The bilateral posterior STS region, which has also been highlighted for AV integration in general, was found to be important for affective integration in this study. In a further study (Kreifelts et al., 2009) these authors found evidence for a segregation of the STS region into a voice-sensitive region in the trunk section, a region with maximum face sensitivity in the posterior terminal ascending branch, and an AV integration area for emotional signals at the bifurcation of the STS. Similar to the present study, Klasen et al. (2011) used computer-generated emotional faces and voices to assess the neural effects of emotional congruency during an explicit emotional classification task. Congruent AV stimuli led to activation in amygdala, insula, ventral posterior cingulate, temporo-occipital, and auditory cortex, whereas incongruent stimuli gave rise to activations in frontoparietal regions as well as the caudate nucleus bilaterally.

In the present investigation we were interested in the neural system supporting the AV integration of emotional face information and emotional prosody. Normal participants were exposed to short videos of a computer-animated face voicing emotionally positive or negative words with the appropriate prosody. Facial expression was either neutral or emotionally appropriate. To reveal the neural network involved in AV integration, standard univariate analysis of fMRI data was followed by a connectivity analysis (Granger causality mapping (GCM); Roebroeck et al., 2005; Valdes-Sosa et al., 2011; Stephan and Roebroeck, 2012). GCM was introduced as it allows, similar to dynamic causal modeling (Stephan and Roebroeck,