Neuropsychologia 66 (2015) 134–143

Contents lists available at ScienceDirect

Neuropsychologia

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

Affect differentially modulates brain activation in uni- and multisensory body-voice perception

Sarah Jessen [a,*], Sonja A. Kotz [a,b]

[a] Max-Planck-Institute for Human Cognitive and Brain Sciences, Stephanstr. 1A, 04103 Leipzig, Germany
[b] School of Psychological Sciences, University of Manchester, Brunswick Street, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 21 July 2014
Received in revised form
22 September 2014
Accepted 30 October 2014
Available online 4 November 2014

Keywords:
Audiovisual
Crossmodal prediction
Emotion
fMRI
Voice
Body

ABSTRACT

Emotion perception naturally entails multisensory integration. It is also assumed that multisensory emotion perception is characterized by enhanced activation of brain areas implied in multisensory integration, such as the superior temporal gyrus and sulcus (STG/STS). However, most previous studies have employed designs and stimuli that preclude other forms of multisensory interaction, such as crossmodal prediction, leaving open the question whether classical integration is the only relevant process in multisensory emotion perception. Here, we used video clips containing emotional and neutral body and vocal expressions to investigate the role of crossmodal prediction in multisensory emotion perception.

While emotional multisensory expressions increased activation in the bilateral fusiform gyrus (FFG), neutral expressions compared to emotional ones enhanced activation in the bilateral middle temporal gyrus (MTG) and posterior STS. Hence, while neutral stimuli activate classical multisensory areas, emotional stimuli invoke areas linked to unisensory visual processing. Emotional stimuli may therefore trigger a prediction of upcoming auditory information based on prior visual information. Such prediction may be stronger for highly salient emotional compared to less salient neutral information. Therefore, we suggest that multisensory emotion perception involves at least two distinct mechanisms; classical multisensory integration, as shown for neutral expressions, and crossmodal prediction, as evident for emotional expressions.

1. Introduction

Emotion perception is a fundamental aspect of any social interaction. Hence it is not surprising that its neural underpinnings have received increasing interest in recent decades. One key feature of emotion perception is its multimodality; more often than not, congruent emotion expressions can be perceived from facial, vocal, and bodily expressions simultaneously. Previous studies have suggested that the mechanisms underlying multisensory emotion perception are akin to those observed in multisensory integration in general, but that the emotional content leads to enhanced multisensory integration (Kreifelts et al., 2007). The claim arises from the observation that emotional compared to neutral multisensory stimuli lead to enhanced activation of the posterior superior temporal gyrus (pSTG, Ethofer et al., 2006; Kreifelts et al., 2007; Robins et al., 2009; Park et al., 2011) and the middle temporal gyrus (MTG, Park et al., 2011).

While these studies address one important mechanism underlying multisensory emotion perception, namely classical multisensory integration, this is not necessarily the only mechanism involved in perceiving emotions from multiple modalities. A different mechanism that has received increasing interest in other fields of multisensory perception (see e.g., Schroeder et al., 2008), namely crossmodal prediction, may also play an important role in multisensory emotion processing. In fact, multisensory emotion perception, where visual information naturally precedes the auditory one, seems to be a prime candidate for crossmodal prediction. There are several reasons why this mechanism has been largely neglected in previous studies on multisensory emotion perception.

Most importantly, many previous studies on multisensory emotion perception relied on stimulus material that differs drastically from real life expressions. Many studies use static visual information (Dolan et al., 2001; Pourtois et al., 2005; Ethofer et al., 2006; Müller et al., 2011; Park et al., 2011), which is not naturally occurring and has shown to elicit strong processing differences when compared to dynamic visual information (LaBar et al., 2003; Sato et al., 2004; Yoshikawa and Sato, 2006; Trautmann et al., 2009). Furthermore, predictive visual information (i.e. visual information containing accurate information about the onset and the

* Corresponding author. Tel.: +49 341 9940 2475; fax: +49 341 9940 2204.
E-mail addresses: jessen@cbs.mpg.de (S. Jessen), sonja.kotz@manchester.ac.uk (S.A. Kotz).

http://dx.doi.org/10.1016/j.neuropsychologia.2014.10.038
0028-3932/

content of the following auditory information) preceding the auditory onset is essential for the optimal integration between two modalities (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010).

In addition, previous studies mainly focused on face-voice interaction. However, faces are not the only visual source of emotional information; bodies offer equally reliable information regarding someone's emotional state (Atkinson et al., 2004). Furthermore, body expressions are particularly important for the perception of emotions at large distances and the link between emotional experience and an intended action (de Gelder, 2009). Hence, they provide an important source of emotional information complementary to facial expressions.

Finally, many previous studies have used semantically neutral words spoken with affective intonation (Ethofer et al., 2006; Kreifelts et al., 2007; Robins et al., 2009), creating a potential conflict between semantic and prosodic information, as suggested in work investigating the interaction between prosodic and semantic content (Kotz and Paulmann, 2007; Paulmann and Kotz, 2008). Therefore, it is difficult to disentangle the processing of conflict information (Wittfoth et al., 2010; Kotz et al., 2014) from multisensory integration per se.

Therefore, the main aim of the present study was to investigate the role of crossmodal prediction in multisensory emotion perception, controlling for the above-mentioned possible confounds such as an audiovisual mismatch. Hence, we used a stimulus set consisting of dynamic videos containing emotional body expressions along with matching emotional interjections (Jessen and Kotz, 2011; Jessen et al., 2012). By using comparably long stimuli (on average above 4 s) and a naturally occurring delay between the auditory and the visual onset (on average larger than 600 ms), we were able to investigate the influence of preceding and ongoing visual information on auditory processing in its natural evolution. Interjections (such as “ah” and “oh”) are especially well suited to investigate emotional voice processing, as they contain close to no semantic information, naturally express different emotional states, and yet allow for acoustic control of stimulus properties (Dietrich et al., 2008). The combination of these features – dynamic visual stimuli, interjections, congruent voice and body information – allows investigating multisensory interaction and, in particular, crossmodal prediction under settings closely approaching an ecologically valid, natural situation.

We suggest that such a setting may lead to a different pattern of activation than previously reported. We expect stronger crossmodal predictions for emotional compared to neutral stimuli, as emotional content commonly leads to preferred information processing (e.g., Hansen and Hansen, 1988; Burton et al., 2005). As borne out clearly in priming studies, successful prediction leads to reduced activation in processing-relevant areas (Rissman et al., 2003; Noppeney et al., 2008; Liu et al., 2010). Hence, if visual information provided in neutral stimuli is less predictive than in emotional stimuli, this should – paradoxically – lead to stronger activation for neutral than for emotional auditory stimuli.

In sum, we hypothesized that if emotional bodies and voices under ecologically valid conditions interact in ways similar to what has been observed previously for faces and voices, emotional multisensory stimuli should lead to an enhanced activation in the STG/MTG compared to neutral stimuli. If, however, dynamics, the use of bodies, and the differences in auditory stimuli shift the interaction between the modalities from a classical integration to crossmodal prediction, we would predict a stronger activation in the STG/MTG for neutral stimuli.

In order to ensure that vocal and body information in itself was processed as intended, we also analyzed the unisensory conditions, before contrasting audiovisual emotional and neutral stimuli. Here, we expected an enhanced activation for emotional compared to neutral interjections in voice specific areas (e.g. superior temporal sulcus and gyrus, STS/STG) as well as cortical (e.g. ventral anterior cingulate cortex, ACC) and subcortical areas (e.g. amygdala, insula) associated with more elaborate processing of emotional information (Grandjean et al., 2005; Beaucousin et al., 2007; Dietrich et al., 2008). Regarding emotional body expressions, we expected an increased blood-oxygen level dependent (BOLD) response in areas linked to the processing of body expressions (e.g. fusiform gyrus, extrastriate body area) and areas implied in more general, modality-independent emotion processing (e.g. amygdala, insula) (Hadjikhani and de Gelder, 2003; de Gelder et al., 2004; Grèzes et al., 2007; Pichon et al., 2008, 2009).

Finally, we computed a conjunction analysis ($AV > V \cap AV > A$) as a measure of multisensory integration (Kreifelts et al., 2007; Park et al., 2011). If the multisensory interaction in the present study is predominantly characterized by classical multisensory integration, we expect an increased activation in the MTG/STG for emotional compared to neutral settings. In contrast, if the interaction is characterized by crossmodal prediction, we expect an increased activation in the MTG/STG for neutral compared to emotional stimuli.

2. Materials and methods

2.1. Participants

Seventeen native speakers of German (8 female) participated in the experiment. Their mean age was 25.9 (standard deviation (SD)=3.9). All were right-handed, had normal or corrected-to-normal vision, and reported no hearing deficit. They gave written informed consent and were compensated financially for their participation. The study was approved by the local ethics committee of the University of Leipzig.

2.2. Stimuli and design

To investigate multisensory emotion perception of complex dynamic stimuli, we created a stimulus set consisting of short video clips displaying one of four semi-professional actors (two female) portraying anger, fear, and a non-emotional (neutral) state in vocal and body expressions. Each actor was standing on an indicated spot in front of a gray screen and was instructed to express the intended emotion in a way he/she thought best fitting. For neutral stimuli, actors were asked to perform a variety of different movements, such as grooming gestures or speech accompanying gestures. In addition, they were instructed to take a step in any direction to increase the similarity to the emotional stimuli, which commonly included leg movements. Synchronized with the body motion, he/she was instructed to express a vocal emotion in form of interjections “ah”, “oh”, and “mh” (see Fig. 1 for an example of the stimulus material, Supplementary material for additional examples, and Jessen and Kotz (2011) for further details). Voice onset naturally occurred at an average delay of 627 ms (SD=431) with respect to body motion onset (see Supplementary material for a more detailed overview of the physical stimulus parameters). As we used the same material as previously tested in two electroencephalographic (EEG) studies (Jessen and Kotz, 2011; Jessen et al., 2012), each video started with a 500 ms still frame of the respective actor standing in a neutral position. We did not manipulate the delay between video and sound onset, and hence differences between emotional and neutral stimuli were present in the stimulus material (mean audiovisual delay: anger: 969 ms (SD=297 ms), fear: 845 ms (SD=140 ms), neutral: 1568 ms (SD=389 ms)). As each video started with a 500 ms still frame, this resulted in a delay of 1127 ms (SD=431) with respect to the video onset.

To control for the impact of facial expressions on the perception of body language, the faces of the actors were blurred using the