Behavioral/Systems/Cognitive

Supramodal Representation of Emotions

Martin Klasen,[1,2] Charles A. Kenworthy,[3] Krystyna A. Mathiak,[1,2,4] Tilo T. J. Kircher,[5] and Klaus Mathiak[1,2,6]

[1]Department of Psychiatry, Psychotherapy, and Psychosomatics, Medical School, RWTH Aachen University, 52074 Aachen, Germany, [2]JARA-Translational Brain Medicine, Aachen/Jülich, Germany, [3]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, [4]Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, Medical School, RWTH Aachen University, 52074 Aachen, Germany, [5]Department of Psychiatry and Psychotherapy, Philipps University Marburg, 35043 Marburg, Germany, and [6]Institute of Neuroscience and Medicine (INM-1), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Supramodal representation of emotion and its neural substrates have recently attracted attention as a marker of social cognition. However, the question whether perceptual integration of facial and vocal emotions takes place in primary sensory areas, multimodal cortices, or in affective structures remains unanswered yet. Using novel computer-generated stimuli, we combined emotional faces and voices in congruent and incongruent ways and assessed functional brain data (fMRI) during an emotional classification task. Both congruent and incongruent audiovisual stimuli evoked larger responses in thalamus and superior temporal regions compared with unimodal conditions. Congruent emotions were characterized by activation in amygdala, insula, ventral posterior cingulate (vPCC), temporo-occipital, and auditory cortices; incongruent emotions activated a frontoparietal network and bilateral caudate nucleus, indicating a greater processing load in working memory and emotion-encoding areas. The vPCC alone exhibited differential reactions to congruency and incongruency for all emotion categories and can thus be considered a central structure for supramodal representation of complex emotional information. Moreover, the left amygdala reflected supramodal representation of happy stimuli. These findings document that emotional information does not merge at the perceptual audiovisual integration level in unimodal or multimodal areas, but in vPCC and amygdala.

Introduction

Stimulation in natural settings usually recruits several sensory channels simultaneously (Stein and Meredith, 1993). This is particularly important for the perception of emotional cues in social interactions, in which congruency between facial expression and emotional prosody facilitates emotion recognition (de Gelder and Vroomen, 2000). Emotional prosody can alter facial emotion perception (Massaro and Egan, 1996) independent from attention and even with the explicit instruction to ignore one modality (Ethofer et al., 2006a).

Extending well established findings on integration of low-level audiovisual cues (Calvert and Thesen, 2004; Stein and Stanford, 2008), various studies have addressed the neurobiology of supramodal emotion representation. Studies in nonhuman primates revealed an ability to integrate socially relevant multimodal cues from conspecifics (Ghazanfar and Logothetis, 2003), which is characterized by responsiveness of amygdala and auditory cortex (Ghazanfar et al., 2005; Remedios et al., 2009), superior temporal sulcus (STS) (Ghazanfar et al., 2008), and ventrolateral prefrontal cortex (Sugihara et al., 2006). However, it remained ambiguous whether this pattern was driven by emotional integration or low-level stimulus features. In humans, EEG studies reported interactions of facial and vocal emotions 110–220 ms after stimulus (de Gelder et al., 1999; Pourtois et al., 2000, 2002), suggesting a convergence already in primary sensory cortices. In contrast, fMRI investigations reported temporal structures as candidates for emotion integration (Pourtois et al., 2000, 2005; Ethofer et al., 2006b; Kreifelts et al., 2007, 2010; Robins et al., 2009). A review from emotional confluence in pain (Vogt, 2005) and a recent MEG study (Chen et al., 2010) suggested that emotional information does not merge at primary sensory cortices, but in higher association areas such as the frontal or the posterior cingulate cortex (PCC). Furthermore, there is evidence that affective structures such as the amygdala play a central role in emotional convergence (Dolan et al., 2001; Ethofer et al., 2006a; Chen et al., 2010).

Studies of supramodal emotion representation typically compared congruent audiovisual stimuli with purely auditory or visual ones (Kreifelts et al., 2007, 2010; Robins et al., 2009). Dolan et al. (2001) compared congruent and incongruent audiovisual stimuli, but combined emotional sentences with static faces. To achieve valid comparisons, two major challenges must be mastered in supramodal emotion representation paradigms. First, multisensory stimulus integration relies on spatial and temporal coincidence (King and Palmer, 1985; Stein and Wallace, 1996). Matched audiovisual speech enhances understanding (Campbell et al., 1998) and evokes distinctive neural patterns (Calvert et al., 2000; Miller and D'Esposito, 2005). Respective paradigms therefore call for precisely matched dynamic stimuli. Second, corre-

Received June 7, 2011; accepted July 12, 2011.

Author contributions: M.K., C.A.K., K.A.M., T.T.J.K., and K.M. designed research; M.K. and C.A.K. performed research; M.K., C.A.K., and K.M. analyzed data; M.K., K.A.M., T.T.J.K., and K.M. wrote the paper.

This work was supported by German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) Grants IRTG 1328 and MA 2631/4-1, and the Interdisciplinary Center for Clinical Research Aachen (N2-3 and N4-2).

Correspondence should be addressed to Martin Klasen, Department of Psychiatry, Psychotherapy, and Psychosomatics, RWTH Aachen University, Pauwelsstrasse 30, 52074 Aachen, Germany. E-mail: mklasen@ukaachen.de.

DOI:10.1523/JNEUROSCI.2833-11.2011

Copyright © 2011 the authors 0270-6474/11/3113635-09$15.00/0

lates of successful emotion integration have to be separated from those of audiovisual speech integration in general. The latter can be solved by including emotionally incongruent trials. This may interfere, however, with the first issue since both emotions have to be unambiguous when considered in isolation but have to be combined in one perfectly synchronized stimulus, which is impossible to do even for a professional actor. Consequently, there have so far been no studies investigating emotion integration with congruent and incongruent dynamic stimuli.

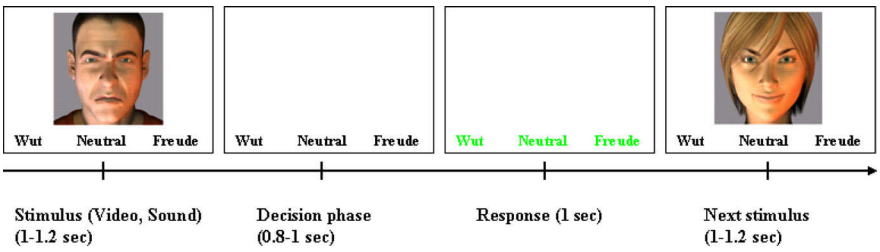

Figure 1. Exemplary trial scheme. Trials consisted of three phases (total, 3 s): presentation of the stimulus (duration, 1–1.2 s), decision phase (duration, 0.8–1 s), and response phase (duration, 1 s). The example shows a video trial. In purely auditory trials, only the three emotion choices were visible during the stimulation. Trial order was randomized across subjects.

Materials and Methods

The effectiveness of virtual characters for emotion recognition tasks has been successfully validated in patient and control populations (Dyck et al., 2008, 2010; Wallraven et al., 2008). We used dynamic virtual characters (avatars) exhibiting angry, neutral, and happy facial emotions and combined them with pseudowords with angry, neutral, and happy prosody in congruent and incongruent trials, assuring standardized facial expressions and perfect lip–speech synchronization. In addition to the congruent and incongruent trials, visual and auditory stimuli were also presented in isolation. The subjects were always to rate the emotion of the stimulus as a whole [i.e., to consider the most salient emotion (happy, neutral, or angry)]. Trials started with a presentation phase [audio (A), visual (V), or audiovisual (AV) stimulus; 1–1.2 s] followed by a decision phase to allow for integrative processing (summing up to 2 s), and a reaction phase (1 s) indicated by a color change (Fig. 1).

Subjects. Twenty-four right-handed subjects [12 females, 12 males; age, 18–34 years (25.0 ± 3.3); German native language] participated in the experiment. All participants had normal or corrected to normal vision, normal hearing, no contraindications against MR investigations, and no history of neurological or psychiatric illness. The experiment was designed according to the Code of Ethics of the World Medical Association (Declaration of Helsinki, 1964), and the study protocol was approved by the local ethics committee. Informed consent was obtained from all subjects.

Stimuli. Audiovisual stimuli were dynamic neutral, angry, and happy virtual character (avatar) faces combined with disyllabic pseudowords with neutral, angry, or happy prosody (Thönnessen et al., 2010). Stimuli had 1 s duration and were created with the virtual reality software Poser Pro (Smith Micro Software) using the lip synchronization tool that allows for a precise matching of speech and lip movements. Congruent trials had matching facial and prosodic emotions, and incongruent runs had different expressions in voice and face. Four different avatars (two males, two females) were combined with the voices of four speakers (two males, two females). To create an unambiguous identity of each character the voice–avatar assignment was unique (i.e., each avatar was only combined with the voice of one speaker). Ninety-six different audiovisual stimuli were used: 48 congruent (3 emotions × 4 speakers × 4 pseudowords) and 48 incongruent ones, with the six possible emotion combinations being equally represented.

The avatar faces were rated for their emotion by 28 subjects not included in the neuroimaging study. A high percentage (90.4%) of the face stimuli were classified correctly according to their emotion, with the recognition rates for angry faces (97.3%) being higher than those for neutral (86.4%) and happy faces (87.4%; $\chi^2 = 5.4$; $p = 0.04$). The videos without sound were also presented as purely visual stimuli in the fMRI experiment.

The pseudowords followed German phonotactical rules but had no semantic content. They were validated in a prestudy on 25 subjects who did not participate in the fMRI study; 91.7% of the neutral, 87.0% of the happy, and 96.7% of the angry stimuli were classified correctly, with no significant differences in the recognition rates of the three emotions ($\chi^2 = 1.5$; $p = 0.47$). The pseudowords without accompanying faces were used as purely auditory stimuli as well.

Experimental design. A hybrid fMRI design encompassed blocks for modality and events for the distinct emotion. Functional data were recorded in one session comprising 36 blocks (8 visual, 8 auditory, and 16 audiovisual), each of which included 12 trials in pseudorandomized order. The experiment included 384 trials with a duration of 3 s each (96 auditory, 96 visual, 96 congruent, 96 incongruent, with all emotions equally represented). Blocks were separated by resting phases without a specific task showing a fixation cross (20 s). Audiovisual blocks comprised both congruent and incongruent trials. Subjects were instructed to rate the emotion of the stimulus in a forced-choice task (i.e., to consider the stimulus as a whole including both auditory and visual modalities and to judge which emotion they perceived as being expressed: neutral, angry, or happy). This instruction was identical for all stimulus types; responses were given by pressing one of three buttons with one of three fingers of the right hand. During stimulation, the three emotion choices were always visible in the lower part of the screen (Fig. 1). Each trial started with the presentation phase (audio, visual, or audiovisual; 1–1.2 s) followed by a decision phase (0.8–1 s), and a reaction phase (1 s) indicated by the letters turning green (Fig. 1). Subjects were instructed to respond only during the reaction phases (delayed response task).

Image acquisition. Whole-brain functional magnetic resonance imaging was conducted with EPI sequences (TE, 28 ms; TR, 2000 ms; flip angle, 77°; voxel size, 3 × 3 mm; matrix size, 64 × 64; 34 transverse slices; 3 mm slice thickness; 0.75 mm gap) on a 3 T Siemens Trio whole-body scanner (Siemens Medical) with a standard 12-channel head coil. A total of 920 functional images were acquired. After the functional measurements, high-resolution T1-weighted anatomical images were performed using a MPRAGE sequence (TE, 2.52 ms; TR, 1900 ms; flip angle, 9°; FOV, 256 × 256 mm^2; 1 mm isotropic voxels; 176 sagittal slices). Total time for functional and anatomical scans was 35 min.

Image analysis. Image analyses were performed with BrainVoyager QX 2.0.8 (Brain Innovation). Preprocessing of the functional MR images included slice scan time correction, 3D motion correction, Gaussian spatial smoothing (4 mm FWHM), and high-pass filtering including linear trend removal. The first three images of each functional run were discarded to avoid T1 saturation effects. Functional images were coregistered to 3D anatomical data and transformed into Talairach space (Talairach and Tournoux, 1988). Statistical parametric maps were created by using a random-effects general linear model (RFX-GLM) with multiple predictors according to the stimulus types. Events were defined stimulus-locked (i.e., beginning with the onset of the presentation phase and for the duration of the entire trial). Trials without response in the time of the response phase were omitted from the analysis. For auditory, visual, and congruent audiovisual trials, only trials with correct responses were included. For incongruent audiovisual stimuli, all trials were included; separate predictors coded whether the subject decided for face, voice, or neither of them. Task contrasts were investigated using t statistics. Activations were corrected for multiple comparisons and thresholded at $p < 0.05$, false discovery rate (FDR). Moreover, an uncorrected threshold of $p < 0.001$ was applied to two of the conjunction analyses for further descriptive data evaluation. All reported conjunction analyses tested the conservative conjunction null hypotheses (Nichols et al., 2005).

Table 1. Response patterns for incongruent trials

	Neutral		Angry		Happy	
	Face (%)	Voice (%)	Face (%)	Voice (%)	Face (%)	Voice (%)
Decision for face emotion	45.3	65.1	72.9	56.1	65.7	72.2
Decision for voice emotion	44.9	33.5	21.5	45.1	29.0	15.7
Decision for neither	9.8	0.6	6.6	8.8	5.3	12.1

Incongruent trials exhibited different emotions in voice and face. The table shows the influence of the three emotions (neutral, angry, happy) and the modality in which they were presented (face, voice) on the perceived emotion of the incongruent stimuli. As an example, incongruent trials with a neutral face led to a decision for the facial emotion in 45.3% of all trials, whereas in 44.9% of all trials with neutral faces subjects decided for the emotion of the voice.

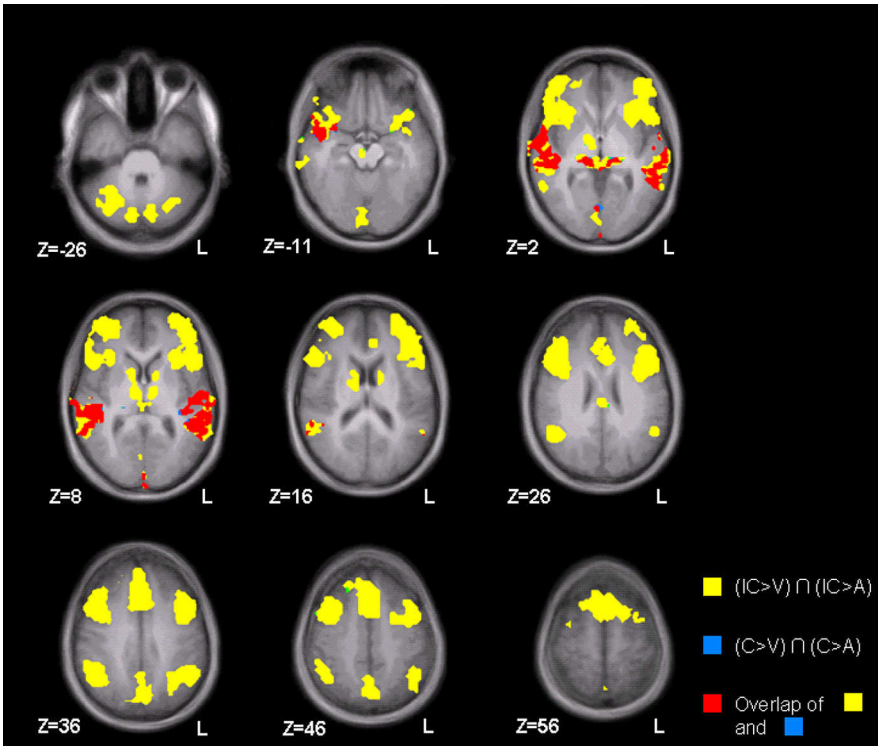

Figure 2. Putative AV integration network. AV stimuli elicited stronger activation in a widespread network of frontal, parietal, occipital, and temporal cortex areas, along with cerebellum, thalamus, and basal ganglia compared with A and V conditions. Thalamic, temporal, and occipital regions were activated both by congruent (C_{AV}) and incongruent (IC_{AV}) runs (overlap shown in red), whereas activity in the basal ganglia, frontal, and parietal cortex was exclusively elicited by incongruent emotions (yellow; all $p < 0.05$, FDR-corr., $k > 10$ voxels). Virtually no area was attributable to congruent emotions only (blue).

Results

Behavior

Subjects responded in the reaction phase in 98.4% (±0.6%) of all trials, and only these trials were considered for behavioral and fMRI analyses. Emotions were correctly classified in 82.7% (±7.1%) of the auditory trials (neutral, 98.6 ± 2.0%; angry, 73.1 ± 16.3%, happy, 75.9 ± 16.6%) as well as in 93.7% (±5.7%) of the visual trials (neutral, 91.0 ± 7.6%, angry, 95.4 ± 6.8%, happy, 94.7 ± 9.3%) and 94.7% (±5.2%) of the congruent audiovisual trials (neutral, 94.9 ± 6.9%; angry, 97.7 ± 3.0%; happy, 91.5 ± 10.7%). The integration of prosody and facial expression led to a better emotion recognition for the matched emotional expressions: Congruent AV stimuli (C_{AV}) achieved better recognition rates than each of both unimodal presentations (C_{AV} vs A: parameter $\beta = 1.70 \pm 0.05$ in binominal generalized linear model, $p < 0.001$; C_{AV} vs V: $\beta = 0.12 \pm 0.06$, $p < 0.05$). In incongruent audiovisual trials, the subjects decided for the prosodic emotion in 31.6% (±12.5%) and for the facial emotion in 61.6% (±14.1%) of all trials; in 6.8% (±4.0%), they decided for the emotion neither of the face nor of the voice. A detailed description of the different emotions, the modality in which they were presented in incongruent trials, and their influence on the response patterns can be found in Table 1.

A repeated-measures ANOVA ($F_{(3,21)} = 14.7$; $p < 0.001$) revealed that reaction times (RTs) after the onset of the response phase were longer for incongruent trials (430.1 ± 11.6 ms) compared with congruent (397.2 ± 9.8 ms; $p < 0.001$), auditory (384.8 ± 10.0 ms; $p < 0.001$), and visual stimuli (389.5 ± 10.4 ms; $p < 0.001$). No significant differences in RTs emerged between congruent audiovisual, auditory, and visual conditions. Purely visual happy stimuli (376.7 ± 10.8 ms) were recognized faster than angry (400.2 ± 11.2 ms; $p < 0.01$) and neutral videos (391.8 ± 10.5 ms; $p < 0.05$). For congruent audiovisual stimuli, neutral trials (411.5 ± 11.9 ms) had longer RTs than happy (386.5 ± 8.8 ms; $p < 0.01$) and angry trials (393.5 ± 10.9 ms; $p < 0.05$). For purely auditory trials, no effect of emotion on RTs was observed.

fMRI

For the neurophysiological data, a conjunction analysis (AV > A) ∩ (AV > V) revealed structures that were more involved in the bimodal than in both unimodal conditions (Nichols et al., 2005): Compared with either unimodal condition, all AV stimuli together yielded higher thalamic, visual, and auditory activity including the superior temporal gyrus (STG) as well as bilateral frontal and prefrontal areas, temporal poles, inferior parietal cortex, cerebellum, and the basal ganglia compared with either unimodal condition [Fig. 2; $p < 0.05$, FDR-corrected (corr.)]. However, the AV stimuli comprised congruent (C_{AV}) and incongruent (IC_{AV}) emotion displays. We therefore categorized the AV integration network into regions driven by congruent AV stimuli [conjunction (C_{AV} > V) ∩ (C_{AV} > A)] and regions driven by incongruent stimuli [conjunction (IC_{AV} > V) ∩ (IC_{AV} > A), both masked with (AV > A) ∩ (AV > V)]. For congruent runs, only sensory structures (thalamus, auditory, and visual cortices) emerged (Fig. 2, blue and red areas), whereas the entire network was significantly affected by incongruent trials (Fig. 2, yellow and red areas; red areas indicate the activation overlap of congruent and incongruent runs). This activity therefore reflected analysis of AV stimulation independent from emotional congruency and thus rather conflict processing of incongruent emotions than a processing supporting the integration of emotion information from the two modalities.

To rule out the possibility that parts of the audiovisual integration mask used in Figure 2 were driven by incongruency-specific patterns, we additionally conducted explorative unmasked conjunction analyses. The fourfold conjunction (C_{AV} > A) ∩ (C_{AV} > V) ∩ (IC_{AV} > A) ∩ (IC_{AV} > V) [thresholded at a descriptive $p < 0.005$, uncorrected (uncorr.)] confirmed the thalamus and posterior superior temporal cortices as audiovisual integration areas

that were independent from emotional congruency (Fig. 3*A*). Separate analyses for both congruent [*B*; ($C_{AV} > A$) ∩ ($C_{AV} > V$); $p < 0.005$, uncorr.] and incongruent emotions [*C*; ($IC_{AV} > A$) ∩ ($IC_{AV} > V$); $p < 0.005$, uncorr.] additionally supported this notion. Specifically for congruent audiovisual emotions, supraadditive activation could be observed in bilateral amygdala. The frontoparietal network along with basal ganglia and cerebellum was exclusively attributed to incongruent audiovisual emotions.

Additive integration of emotion information can emerge to the congruent stimuli but not to the incongruent trials. The direct comparison ($C_{AV} > IC_{AV}$) revealed bilateral ventral posterior cingulate cortex (vPCC), right inferior temporal gyrus, left insula, left superior temporal gyrus, and bilateral amygdala as candidate areas for AV emotion integration (Fig. 4, hot colors; $p < 0.05$, FDR-corr.; Table 2). Similar to the results depicted in Figures 2 and 3, emotional incongruency was associated with activity in bilateral frontal and prefrontal areas including operculum and dorsolateral prefrontal cortex, anterior insula, medial prefrontal and anterior cingulate cortex (ACC), intraparietal sulcus (IPS), inferior parietal lobe, precuneus, cerebellum, and bilateral caudate nucleus (Fig. 4, cold colors; Table 2).

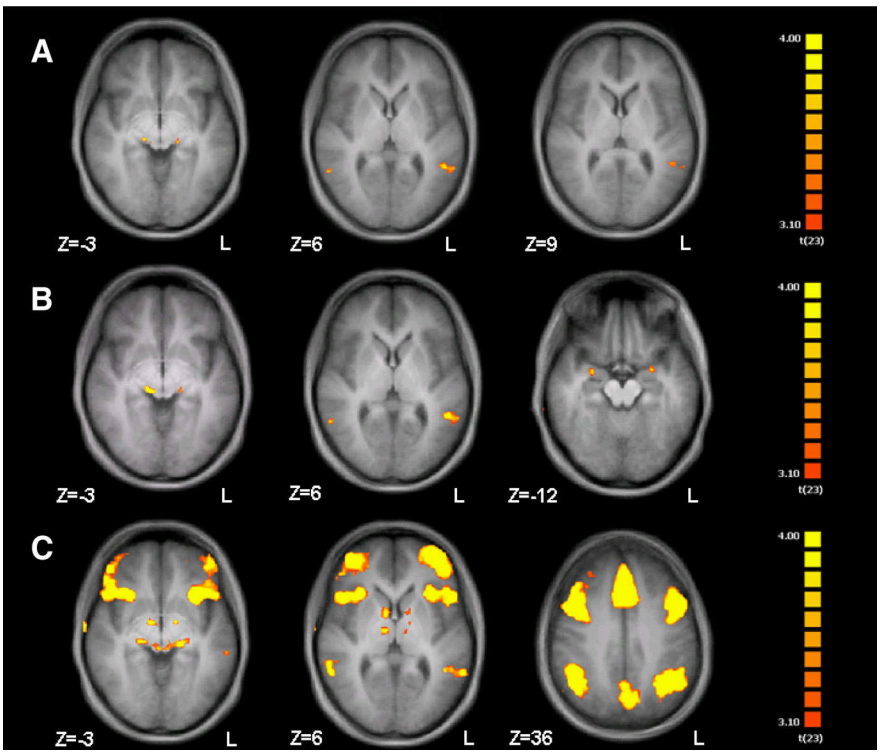

Figure 3. AV integration independent from congruency. The role of thalamus and posterior superior temporal areas in congruency-independent audiovisual integration was confirmed by the explorative conjunction analysis ($C_{AV} > A$) ∩ ($C_{AV} > V$) ∩ ($IC_{AV} > A$) ∩ ($IC_{AV} > V$) (***A***) ($p < 0.05$, uncorr.). The pattern was observed independently from emotional congruency for both congruent (***B***) [($C_{AV} > A$) ∩ ($C_{AV} > V$); $p < 0.005$, uncorr.] and incongruent emotions (***C***) [($IC_{AV} > A$) ∩ ($IC_{AV} > V$); $p < 0.005$, uncorr.]. In addition, bilateral amygdala was observed only for the integration of congruent emotions. Basal ganglia, frontal, and parietal cortex were again only activated by incongruent emotions.

In general, brain structures with higher response to congruent compared with incongruent emotion display showed higher sensitivity to happy compared with angry or neutral emotions (Fig. 4, insets). This is analogous to the behavioral finding that emotion recognition in incongruent stimuli was dominated by anger (72.9% in visual and 45.1% in auditory display of anger; Table 1) but congruent over unimodal display was most effective for happy expressions (e.g., voice only vs congruent: 15.7 vs 91.5%). The left amygdala—being an area well known to reflect stimulus arousal and social relevance (Sander et al., 2003)—was the only region from the AV integration network with significantly stronger responses to congruent happy AV stimuli compared with the incongruent ones even at a liberal uncorrected threshold of $p < 0.001$ (Fig. 5*A*); neither angry nor neutral emotion yielded such an effect. Structures for multimodal emotion encoding, however, can be expected to integrate information from voice and face for each emotional category and should thus show a stronger reactivity for all types of congruent emotions when compared with incongruent runs. Such a pattern was observed in the vPCC (Fig. 3, left bottom insets), and a refined mapping of congruency effects common to all three emotions confirmed this specificity in the conjunction analysis [C_{AV} (happy) > IC_{AV}] ∩ [C_{AV} (angry) > IC_{AV}] ∩ [C_{AV} (neutral) > IC_{AV}], which identified the vPCC only (right vPCC at $p < 0.05$, FDR-corrected; additionally left PCC at an uncorrected $p < 0.001$; Fig. 5*B*, *C*; Table 2).

Most of the putative AV integration network emerged in response to incongruent stimuli alone (Figs. 2, 3). These responses thus may reflect conflict resolution or prioritizing of one sensory input channel over another. The latter mechanism may be segregated by analyzing the brain activity of the AV integration network with respect to the reported percept (decision for either the emotion of the face or the voice). The network for emotional incongruency (Fig. 4) appeared to be mainly independent from the decision for either facial or vocal emotion [Fig. 6*A*; (Decision for face > Congruent) ∩ (Decision for voice > Congruent); masked with $IC_{AV} > C_{AV}$; $p < 0.05$, FDR-corr.; for details, see Table 2]. However, a direct comparison of incongruent trials that were categorized according to the prosodic information with those categorized according to the facial expression revealed differences in bilateral anterior insula and right frontal operculum (Fig. 6*B*; Decision for voice > Decision for face; masked with $IC_{AV} > C_{AV}$; $p < 0.05$, FDR-corr.). The reversed contrast (trials categorized by face) yielded no significant effect.

A potential confound in the comparison of congruent and incongruent trials might arise if the subjects just attended one sensory channel and ignored the other one. Although the behavioral response pattern did not support this notion, we calculated a subject-wise modality preference (MP) coefficient for incongruent trials, which was defined as $MP = (\%_{\text{Decision for face}} - \%_{\text{Decision for voice}})$. To rule out a possible influence of modality preference, we calculated a voxelwise correlation of MP with the contrast ($C_{AV} > IC_{AV}$); no suprathreshold clusters could be observed for this correlation ($p < 0.05$, FDR-corr., no mask), indicating a negligible influence of modality preference on the networks and a successful attendance of both modalities. In a similar vein, we calculated correlations for the contrast ($C_{AV} > IC_{AV}$) with subject-wise accuracy rates in the recognition of congruent audiovisual emotions. Only negative correlations emerged in parts of the incongruency network (Fig. 6*C*; $p < 0.05$, FDR-corr.; incongruency network shown in yellow). Neither correlations with the congruency network nor any

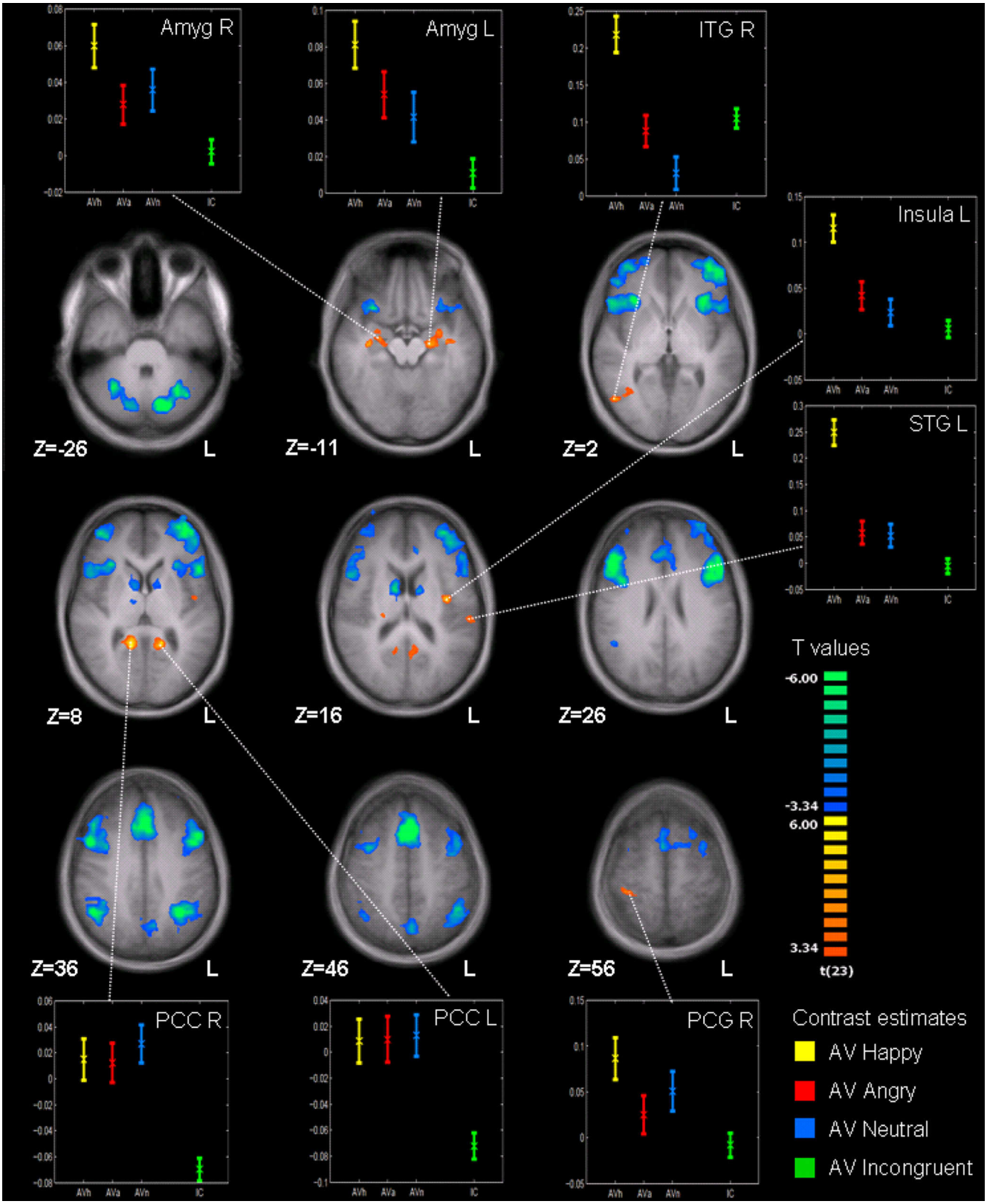

Figure 4. Congruent versus incongruent emotion. Incongruent trials (cold colors) activated a widespread frontoparietal network, cerebellum, and bilateral caudate nucleus. Congruent emotions (warm colors) in turn activated insula and a temporoparietal network along with bilateral amygdala ($p < 0.05$, FDR-corr., $k > 10$ voxels; for details, see Table 2). The insets show %-BOLD signal changes (±SE) to happy (yellow), angry (red), neutral (blue), and incongruent (green) audiovisual trials.

Table 2. Activation clusters for Figures 3–6

Contrast (figure)	Anatomical region	Hemisphere	Talairach coordinates			Peak *t* value	Cluster size (mm^3)
			X	*Y*	*Z*		
Figure 3*A*	Posterior STS*	R	56	−50	6	3.69	25
	Thalamus*	R	14	−23	−3	3.91	48
	Thalamus*	L	−16	−26	0	3.71	28
	Posterior STG/STS*	L	−49	−47	6	3.77	226
Figure 3*B*	Posterior STS*	R	56	−50	6	4.02	42
	Posterior STG/STS*	L	−49	−44	6	4.61	515
	Amygdala*	R	26	−2	−12	3.87	111
	Amygdala*	L	−25	−2	−12	4.04	39
	Thalamus*	R	8	−26	−3	5.22	376
	Thalamus*	L	−16	−26	0	3.86	52
Figure 3*C*	Posterior STG/STS*	R	56	−47	6	5.12	242
	Posterior STG/STS*	L	−58	−47	6	4.49	741
	Posterior STG/STS*	L	−58	−29	−6	3.70	138
	Thalamus*	L	−16	−26	−3	4.91	334
	Thalamus*	R	20	−23	−3	4.14	177
Figure 4	Posterior cingulate	R	12	−49	10	9.20	1123
	Posterior cingulate	L	−15	−46	7	6.54	1061
	Insula	L	−45	−7	10	5.81	553
	Inferior temporal/occipital	R	48	−64	4	5.23	533
	Superior temporal gyrus	L	−57	−25	13	4.92	606
	Amygdala	L	−21	−16	−11	4.90	774
	Amygdala	R	30	−7	−14	5.20	426
	Postcentral gyrus	R	39	−31	58	4.75	456
	Caudate nucleus (head)	R	9	2	13	−5.83	1843
	Precuneus	L/R	−9	−64	37	−5.83	2429
	Cerebellum	R	6	−70	−20	−6.15	3688
	Inferior parietal lobule/IPS	R	45	−52	37	−7.73	4970
	Frontal/cingulate/insula	L/R	−3	20	49	−7.90	53,223
	Frontal/insula	R	48	11	43	−8.08	29,381
	Inferior parietal lobule/IPS	L	−33	−52	34	−8.62	8227
	Caudate nucleus (head)	L	−15	−70	−26	−10.29	838
	Cerebellum	L	−15	−70	−26	−10.30	5176
Figure 5*A*	Amygdala**	L	−30	−7	−14	4.98	144
Figure 5*B*	Ventral posterior cingulate	R	12	−46	10	5.32	81
Figure 5*C*	Ventral posterior cingulate**	L	−12	−49	7	4.65	95
Figure 6*B*	Frontal/insula	R	27	20	4	5.47	204
	Frontal operculum	R	45	14	7	5.25	347
	Frontal/insula	L	−33	20	4	5.16	264
Figure 6*C*	Inferior parietal lobule/IPS	R	47	−56	39	−0.69†	77
	Middle frontal gyrus	R	44	4	39	−0.74†	30
	Middle frontal gyrus	R	44	19	30	−0.66†	31
	Middle frontal gyrus	R	41	31	27	−0.72†	51
	Cerebellum	R	38	−53	−39	−0.74†	285
	Precuneus	R	23	−65	24	−0.66†	48
	Inferior frontal gyrus	L	−37	10	27	−0.68†	32
	Inferior parietal lobule	L	−43	−53	42	−0.70†	65

For Figure 3*C*, only the activation clusters in superior temporal areas and thalamus are listed (see section for Fig. 4 for the incongruency network). Negative *t* values for Figure 4 indicate activation in the incongruent condition.
†Correlation coefficients (*r* values, Fig. 6*C*). All clusters achieved $p < 0.05$, FDR-corr.; except **uncorr. $p < 0.001$, *uncorr. $p < 0.005$.

correlations with accuracy rates for unimodal visual and auditory stimuli were observed.

RTs reflected task difficulty and response preparation. To identify the underlying neural processes and their influence on supramodal representation of emotions, we introduced additional regressors based on trial-wise RTs separately for auditory, visual, and audiovisual stimuli. In accordance with the behavioral data and the observed networks for congruent and incongruent emotions, reaction times to audiovisual stimuli were positively associated with activation in lateral and medial frontal parts of the in-

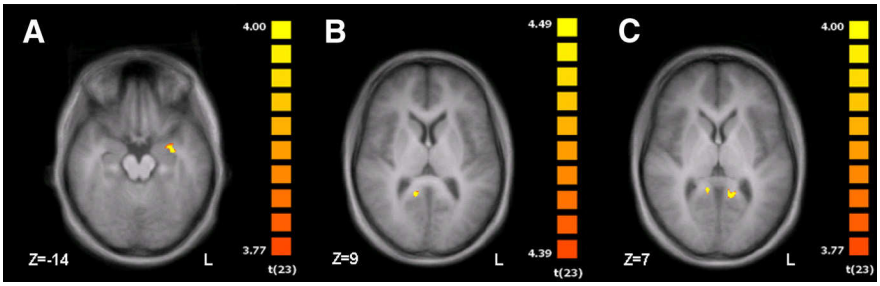

Figure 5. Conjunction analyses on multimodal integration. Left amygdala yielded stronger responses to bimodal compared with unimodal and stronger responses to congruent compared with incongruent trials (***A***: $p < 0.001$, uncorr.). Right (***B***) and left (***C***) vPCC integrated affective facial and vocal emotion independent from emotional category as confirmed by the conjunction (Congruent Neutral > Incongruent) ∩ (Congruent Angry > Incongruent) ∩ (Congruent Happy > Incongruent) (***B***: $p < 0.05$, FDR-corr.; ***C***: $p < 0.001$, uncorr.; for details, see Table 2).

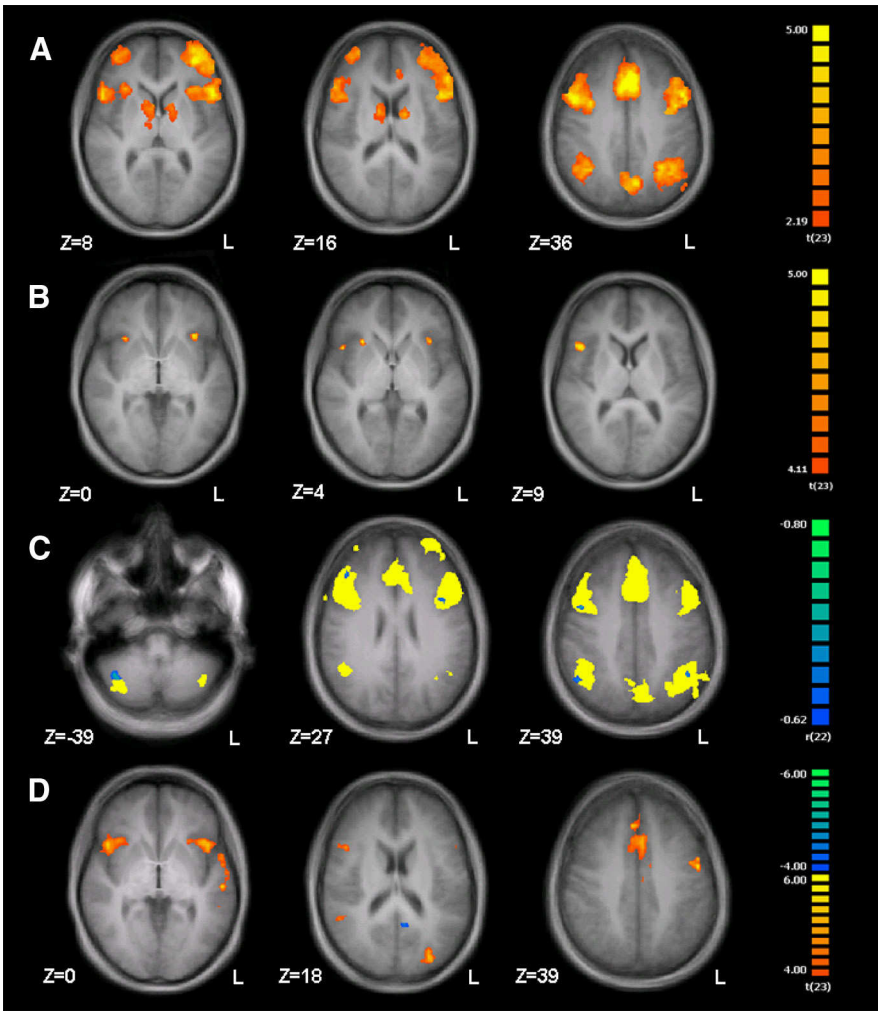

Figure 6. Functional exploration of the incongruency network. Activity in the incongruency network in incongruent compared with congruent trials (Fig. 4) was mainly independent from the decision for either facial or vocal emotion [***A***; (Decision for face > Congruent) ∩ (Decision for voice > Congruent); masked with $IC_{AV} > C_{AV}$; $p < 0.05$, FDR-corr.]. Differences between the decision for prosodic and face emotion were found in bilateral anterior insula and right frontal operculum (***B***; Decision for voice > Decision for face; masked with $IC_{AV} > C_{AV}$; $p < 0.05$, FDR-corr.; for details, see Table 2). The contrast ($C_{AV} > IC_{AV}$) showed a negative correlation with accuracy rates in the recognition of congruent emotions in parts of the incongruency network (***C***; $p < 0.05$, FDR-corr.; incongruency network shown in yellow). A parametric modulator based on trialwise reaction times showed for increased reaction times a stronger activation in frontal parts of the incongruency network along with a decreased activity in left vPCC (***D***; $p < 0.05$, FDR-corr.).

congruency network and negatively associated with activation in the left vPCC as part of the congruency network (Fig. 6*D*; $p < 0.05$, FDR-corr.). No significant difference emerged between congruent and incongruent AV emotions. For purely auditory and visual emotions, no association with RTs was observed. Posterior STS, pSTG, and thalamus were confirmed as AV integration areas independent from emotional congruency in the model with RTs as covariates. In a similar vein, amygdala and vPCC as well as the frontoparietal network with cerebellum and caudate nucleus remained significant in the contrast ($C_{AV} > IC_{AV}$) after controlling for reaction times.

Discussion

This is the first study to investigate the supramodal representation of emotional information with dynamic stimuli expressing facial and vocal emotions congruently and incongruently. In accordance with previously reported findings (Pourtois et al., 2000, 2005; Ethofer et al., 2006b; Kreifelts et al., 2007, 2010; Robins et al., 2009), audiovisual emotions evoked stronger responses in bilateral thalamus and superior temporal gyrus and sulcus compared with those expressed in one modality only. For congruent AV events, the vPCC and amygdala indicated functional additivity of bimodal emotional information.

Posterior superior temporal cortices have been suggested as key areas for the creation of crossmodal affective percepts (Pourtois et al., 2000, 2005; Ethofer et al., 2006b; Kreifelts et al., 2007, 2010; Robins et al., 2009). However, in our study, the responses in these areas were elicited by emotionally incongruent trials as well. Audiovisual integration studies reported enhanced activation in superior temporal regions to emotionally neutral combined speech and face stimuli (Campbell et al., 1998; Calvert et al., 2000; van Atteveldt et al., 2004; Miller and D'Esposito, 2005) as well as with sounds and images of tools and animals (Beauchamp et al., 2004; Fuhrmann Alpert et al., 2008), suggesting a more general role of these structures in multimodal perception. The same applies to the thalamus, which integrates audiovisual stimuli without emotional content (Bushara et al., 2003). Stronger responses at the posterior superior temporal sulcus have been observed to emotional compared with neutral stimuli (Kreifelts et al., 2007). However, in the same study, this supraadditive integration emerged for neutral stimuli as well. Increased responses to emotional stimuli may thus be just an unspecific influence of attention due to their higher salience. In summary, enhanced activation in superior temporal areas and thalamus cannot be attributed specifically to the supramodal representation of emotional information.

Extending the approach of previous studies, we included trials with emotions in both modalities that cannot be successfully integrated into a bimodal emotional percept, identifying brain regions that represent supramodal emotional aspects. A direct comparison of congruent with incongruent bimodal trials revealed a congruency-dependent activation in amygdala, vPCC, inferior temporal gyrus, superior temporal gyrus, and insula. Shared with the putative AV network, happy emotions yielded higher activity in the left amygdala, but only the vPCC responded to congruent facial and vocal expressions in all three emotion categories compared with incongruent stimuli.

The vPCC is involved in the processing of self-relevant emotional and nonemotional information as well as in self-reflection (Vogt et al., 2006). Via reciprocal connections of vPCC with subgenual ACC, the emotional information can gain access to the cingulate emotion subregions, helping to establish the personal relevance of sensory information coming into the cingulate gyrus (Vogt, 2005). Therefore, it seems a suitable candidate for supramodal representation of emotion information from different

modalities independent from low-level sensory features. As a whole, Mar (2011) describes the PCC and precuneus as regions supporting the inference on mental states of another and, therefore, integrating inputs from a wide variety of other brain regions that support memory, motor, and somatosensory processing. Our findings suggest that the PCC involvement in theory of mind abilities could represent convergence of multimodal emotion information in a social context.

Another important aspect of our study is a supraadditive response of the left amygdala to congruent happy stimuli both compared with unimodal and to incongruent runs. Left amygdala was reported in a number of studies on crossmodal effects in emotional processing, particularly in response to congruent fearful but not happy stimuli (Dolan et al., 2001) and to a shift in rating of facial expressions in presence of fearful prosody (Ethofer et al., 2006a). Kreifelts et al. (2010) reported left amygdala involvement in the integration of facial and vocal nonverbal social signals in videos expressing various emotions. Pourtois et al. (2005) described a lateralization of regions activated by audiovisual versus unimodal stimuli depending on the emotion expressed: convergence areas were situated mainly anteriorly in the left hemisphere for happy pairings and in the right hemisphere for fear pairings. This activation pattern is also in accordance with the valence theory which assigns positive emotions to the left and negative to the right hemisphere (for review, see Killgore and Yurgelun-Todd, 2007). The observed supraadditive effects of both multimodality and congruency for happy stimuli clearly document a supramodal representation of affective information in the amygdala; they may reflect a high self-relevance of congruent happy percepts in terms of social communication (Schmitz and Johnson, 2007) and, accordingly, be related to the regulation of vigilance (Yang et al., 2002).

The network for incongruent trials is likely to be the substrate of an enhanced neural processing effort in emotion decoding. Activity in the medial prefrontal cortex has been widely reported in studies using emotional tasks or stimuli, whereas activity in the adjacent ACC was more specifically related to emotional tasks with a cognitively demanding component (Phan et al., 2002; Mathiak and Weber, 2006). In addition, the well documented role of the ACC in cognitive conflict resolution (Kerns et al., 2004), which has also recently been demonstrated in the context of conflicting emotional stimuli (Müller et al., 2011), is likely to contribute to this part of the network. Increased activity in the intraparietal sulcus may reflect increased visual attention and working memory load (Pollmann and Maertens, 2005; Majerus et al., 2007). Activation of caudate nucleus along with medial and inferior frontal areas including the frontal operculum has been reported as a neural signature of prosodic classification (Kotz et al., 2003). In a similar vein, the neuroanatomical model of Ross (1981) assigned prosodic functions to right-hemispheric perisylvian cortices as a homolog to left-lateralized speech functions and has been confirmed by functional imaging studies (for review, see Wildgruber et al., 2006). Indeed, in our data, higher activity in right-accentuated inferior frontal areas predicted dominance of the prosodic over the visual perception, suggesting that the right Broca analog is a specific part of the network involved in the disambiguation of incongruent stimuli by evaluating prosodic cues. The notion of a supramodal emotion evaluation network is further supported by the negative correlation of parts of the latter with accuracy rates for audiovisual emotions, thus reflecting a higher processing workload in emotion decoding areas associated with lower emotion recognition skills. Accordingly, longer reaction times, indicating higher cognitive working load, were associated with stronger activation in frontal parts of the incongruency network, which lends further support to the assumption that emotional aspects of multisensory stimuli are not evaluated in early sensory regions, but later in higher processing areas (Chen et al., 2010).

Limitations

Although well controlled, our stimuli contained the confounding aspects of task difficulty and ambiguity. There was no correct response to incongruent emotional stimuli, and accordingly the judgment of the dominant emotion was more difficult in the incongruency condition. As a consequence, the neural correlates of task difficulty cannot be fully disentangled from those of emotional congruency. To partially resolve this issue, we removed error trials before the analysis and controlled for accuracy rates in emotion recognition. However, we would like to emphasize that the aspects of difficulty and ambiguity are inherent to the task. Unambiguous emotional congruency facilitates emotion recognition, which is the major benefit of multimodal emotions; as such, congruent and incongruent trials are by definition characterized by differences in difficulty levels.

As another limitation, we cannot fully exclude the probability that sometimes not both modalities were attended in bimodal trials. Although we controlled for individual modality preferences and dominant modalities in incongruent trials, there is no possibility to assess modality preferences for congruent stimuli. However, the higher recognition rates for congruent bimodal emotions at least suggest that subjects attended both modalities when bimodal stimuli were presented, regardless of emotional congruency.

Conclusion

Our findings indicate that facial and vocal emotions do not merge on the perceptual level of audiovisual integration, but in later processing stages in the limbic system. The vPCC may be a central structure for representing congruency of auditory and visual emotions independent from emotional category.

References

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190–1192.

Bushara KO, Hanakawa T, Immisch I, Toma K, Kansaku K, Hallett M (2003) Neural correlates of crossmodal binding. Nat Neurosci 6:190–195.

Calvert GA, Thesen T (2004) Multisensory integration: methodological approaches and emerging principles in the human brain. J Physiol Paris 98:191–205.

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649–657.

Campbell R, Dodd B, Burnham D (1998) Hearing by eye II: advances in the psychology of speech-reading and audio-visual speech. Hove, UK: Psychology Press.

Chen YH, Edgar JC, Holroyd T, Dammers J, Thönnessen H, Roberts TP, Mathiak K (2010) Neuromagnetic oscillations to emotional faces and prosody. Eur J Neurosci 31:1818–1827.

de Gelder B, Vroomen J (2000) The perception of emotions by ear and by eye. Cogn Emot 14:289–311.

de Gelder B, Böcker KB, Tuomainen J, Hensen M, Vroomen J (1999) The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. Neurosci Lett 260:133–136.

Dolan RJ, Morris JS, de Gelder B (2001) Crossmodal binding of fear in voice and face. Proc Natl Acad Sci U S A 98:10006–10010.

Dyck M, Winbeck M, Leiberg S, Chen Y, Gur RC, Mathiak K (2008) Recognition profile of emotions in natural and virtual faces. PLoS One 3:e3628.

Dyck M, Winbeck M, Leiberg S, Chen Y, Mathiak K (2010) Virtual faces as a tool to study emotion recognition deficits in schizophrenia. Psychiatry Res 179:247–252.

Ethofer T, Anders S, Erb M, Droll C, Royen L, Saur R, Reiterer S, Grodd W, Wildgruber D (2006a) Impact of voice on emotional judgment of faces: an event-related fMRI study. Hum Brain Mapp 27:707–714.

Ethofer T, Pourtois G, Wildgruber D (2006b) Investigating audiovisual integration of emotional signals in the human brain. Prog Brain Res 156:345–361.

Fuhrmann Alpert G, Hein G, Tsai N, Naumer MJ, Knight RT (2008) Temporal characteristics of audiovisual information processing. J Neurosci 28:5344–5349.

Ghazanfar AA, Logothetis NK (2003) Facial expressions linked to monkey calls. Nature 423:937–938.

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.

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.

Kerns JG, Cohen JD, MacDonald AW 3rd, Cho RY, Stenger VA, Carter CS (2004) Anterior cingulate conflict monitoring and adjustments in control. Science 303:1023–1026.

Killgore WD, Yurgelun-Todd DA (2007) The right hemisphere and valence hypotheses: could they both be right (and sometimes left)? Soc Cogn Affect Neurosci 2:240–250.

King AJ, Palmer AR (1985) Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. Exp Brain Res 60:492–500.

Kotz SA, Meyer M, Alter K, Besson M, von Cramon DY, Friederici AD (2003) On the lateralization of emotional prosody: an event-related functional MR investigation. Brain Lang 86:366–376.

Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D (2007) Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. Neuroimage 37:1445–1456.

Kreifelts B, Ethofer T, Huberle E, Grodd W, Wildgruber D (2010) Association of trait emotional intelligence and individual fMRI activation patterns during the perception of social signals from voice and face. Hum Brain Mapp 31:979–991.

Majerus S, Bastin C, Poncelet M, Van der Linden M, Salmon E, Collette F, Maquet P (2007) Short-term memory and the left intraparietal sulcus: focus of attention? Further evidence from a face short-term memory paradigm. Neuroimage 35:353–367.

Mar RA (2011) The neural bases of social cognition and story comprehension. Annu Rev Psychol 10:103–134.

Massaro DW, Egan PB (1996) Perceiving affect from the voice and the face. Psychon Bull Rev 3:215–221.

Mathiak K, Weber R (2006) Toward brain correlates of natural behavior: fMRI during violent video games. Hum Brain Mapp 27:948–956.

Miller LM, D'Esposito M (2005) Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J Neurosci 25:5884–5893.

Müller VI, Habel U, Derntl B, Schneider F, Zilles K, Turetsky BI, Eickhoff SB (2011) Incongruence effects in crossmodal emotional integration. Neuroimage 54:2257–2266.

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660.

Phan KL, Wager T, Taylor SF, Liberzon I (2002) Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage 16:331–348.

Pollmann S, Maertens M (2005) Shift of activity from attention to motor-related brain areas during visual learning. Nat Neurosci 8:1494–1496.

Pourtois G, de Gelder B, Vroomen J, Rossion B, Crommelinck M (2000) The time-course of intermodal binding between seeing and hearing affective information. Neuroreport 11:1329–1333.

Pourtois G, Debatisse D, Despland PA, de Gelder B (2002) Facial expressions modulate the time course of long latency auditory brain potentials. Brain Res Cogn Brain Res 14:99–105.

Pourtois G, de Gelder B, Bol A, Crommelinck M (2005) Perception of facial expressions and voices and of their combination in the human brain. Cortex 41:49–59.

Remedios R, Logothetis NK, Kayser C (2009) Monkey drumming reveals common networks for perceiving vocal and nonvocal communication sounds. Proc Natl Acad Sci U S A 106:18010–18015.

Robins DL, Hunyadi E, Schultz RT (2009) Superior temporal activation in response to dynamic audio-visual emotional cues. Brain Cogn 69:269–278.

Ross ED (1981) The aprosodias: functional-anatomic organization of the affective components of language in the right hemisphere. Arch Neurol 38:561–569.

Sander D, Grafman J, Zalla T (2003) The human amygdala: an evolved system for relevance detection. Rev Neurosci 14:303–316.

Schmitz TW, Johnson SC (2007) Relevance to self: a brief review and framework of neural systems underlying appraisal. Neurosci Biobehav Rev 31:585–596.

Stein BE, Meredith MA (1993) Merging of the senses. Cambridge, MA: MIT.

Stein BE, Stanford TR (2008) Multisensory integration: current issues from the perspective of the single neuron. Nat Rev Neurosci 9:255–266.

Stein BE, Wallace MT (1996) Comparisons of cross-modality integration in midbrain and cortex. Prog Brain Res 112:289–299.

Sugihara T, Diltz MD, Averbeck BB, Romanski LM (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. J Neurosci 26:11138–11147.

Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain. New York: Thieme Medical Publishers.

Thönnessen H, Boers F, Dammers J, Chen YH, Norra C, Mathiak K (2010) Early sensory encoding of affective prosody: neuromagnetic tomography of emotional category changes. Neuroimage 50:250–259.

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

Vogt BA (2005) Pain and emotion interactions in subregions of the cingulate gyrus. Nat Rev Neurosci 6:533–544.

Vogt BA, Vogt L, Laureys S (2006) Cytology and functionally correlated circuits of human posterior cingulate areas. Neuroimage 29:452–466.

Wallraven C, Breidt M, Cunningham DW, Bülthoff H (2008) Evaluating the perceptual realism of animated facial expressions. ACM TAP 4:1–20.

Wildgruber D, Ackermann H, Kreifelts B, Ethofer T (2006) Cerebral processing of linguistic and emotional prosody: fMRI studies. Prog Brain Res 156:249–268.

Yang TT, Menon V, Eliez S, Blasey C, White CD, Reid AJ, Gotlib IH, Reiss AL (2002) Amygdalar activation associated with positive and negative facial expressions. Neuroreport 13:1737–1741.