Preattentive processing of audio-visual emotional signals

Föcker, J.; Gondan, Matthias; Röder, B.

Published in:
Acta Psychologica

DOI:
10.1016/j.actpsy.2011.02.004

Publication date:
2011

Document version
Early version, also known as pre-print

Citation for published version (APA):
Preattentive processing of audio-visual emotional signals

Julia Föcker a,b,⁎, Matthias Gondan c,d,1, Brigitte Röder a,2

a University of Hamburg, Biological Psychology and Neuropsychology, Von-Melle-Park 11, 20146 Hamburg, Germany
b Brain and Cognitive Sciences, University of Rochester, United States
c University of Regensburg, Department of Psychology, Universitätstr. 31, 93050 Regensburg, Germany
d Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 11 June 2010
Received in revised form 16 February 2011
Accepted 17 February 2011
Available online 12 March 2011

PsychNID classification:
2323
2326
2346

Keywords:
Multisensory
Emotion
Facial expression
Vocal expression
Response conflict
Attention

ABSTRACT

Previous research has shown that redundant information in faces and voices leads to faster emotional categorization compared to incongruent emotional information even when attending to only one modality. The aim of the present study was to test whether these crossmodal effects are predominantly due to a response conflict rather than interference at earlier, e.g. perceptual processing stages. In Experiment 1, participants had to categorize the valence and rate the intensity of happy, sad, angry and neutral unimodal or bimodal face–voice stimuli. They were asked to rate either the facial or vocal expression and ignore the emotion expressed in the other modality. Participants responded faster and more precisely to emotionally congruent bimodal stimuli were more efficiently processed than unimodal visual stimuli. To study the role of a possible response conflict, Experiment 2 used a modified paradigm in which emotional and response conflicts were disentangled. Incongruency effects were significant even in the absence of response conflicts. The results suggest that emotional signals available through different sensory channels are automatically combined prior to response selection.

⁎ Corresponding author at: University of Rochester, Brain and Cognitive Sciences, 404 Meliora Hall, Rochester, 14627, New York, United States. Tel.: +1 585 275 0695; fax: +1 585 442 9216.
E-mail addresses: jfoecker@bcs.rochester.edu (J. Föcker), gondan@imbi.uni-heidelberg.de (M. Gondan), brigitte.roeder@uni-hamburg.de (B. Röder).

1 Tel.: +49 6221 567784; fax: +49 6221 564195.
2 Tel.: +49 40 428383251; fax: +49 40 428386591.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Inferences about the emotional state of individuals can be drawn from their facial and vocal expressions (Darwin, 1872). Since these person-relevant features originate from different sensory channels, they have to be integrated to a coherent percept (de Gelder, Böcker, Tuomainen, Hensen, & Vroomen, 1999; Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000, 2002; Pourtois, Debatisse, Despland, & de Gelder, 2002). Recent studies have provided evidence of the idea that in addition to spatial and temporal stimulus features, the meaning of events is important for audio-visual integration as well (see Doehrmann & Naumer, 2008, for an overview). Redundant information in facial and vocal expressions, that is, the same emotional expression in both faces and voices, has been shown to lead to better emotional categorization performance compared to unimodal stimulation (i.e. either the face or the voice presented alone, Collignon et al., 2008; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007). In contrast, the categorization of emotional stimuli is impaired when the second channel of a bimodal stimulus provides incongruent information. These effects have been demonstrated both for static facial stimuli (de Gelder & Vroomen, 2000; de Gelder et al., 1999) and for dynamic audio-visual material (Collignon et al., 2008; Tanaka et al., 2010; Van den Stock, Grèzes, & de Gelder, 2008) and have been demonstrated in adults (Collignon et al., 2008; de Gelder & Vroomen, 2000), infants (Grossmann, Striano, & Friederici, 2006), blindsighted patients (de Gelder, Pourtois, & Weiskrantz, 2002; 2005), schizophrenics (de Jong, Hodiamont, Van den Stock, & de Gelder, 2009), adults with pervasive developmental disorder (Magnéée, de Gelder, van Engeland, & Kemner, 2008), and rhesus monkeys (Ghazanfar & Logothetis, 2003).

Similar to the automatic and pre-attentive binding of audio-visual speech signals (Bertelson, Vroomen, de Gelder, & Driver, 2000; Driver, 1996; but see Fairhall & Macaluso, 2009), audio-visual integration of emotions has been assumed to be an automatic process (Vroomen, Driver, & de Gelder, 2001). This hypothesis is supported by studies demonstrating that emotional binding is independent of whether or not participants attend to the auditory, visual or both sensory streams.
designed to test the extent of which this incongruency effect is due to
response keys (see Collignon et al., 2008), it cannot be excluded
that the perceptual level. To control for such possible response con-
icts rather than from a crossmodal interference at the perceptual level. To control for such possible response conficts in crossmodal interactions of emotional processing, a few studies
studies, auditory and visual stimuli were taken from different
stimulus sets, or they were presented from different spatial locations
(Collignon et al., 2008; de Gelder & Vroomen, 2000). Thus, the stimuli
were not temporally synchronized (see Table 1 for an overview) and
due to the lack of spatial congruency it is unlikely that the auditory and visual parts were ef-
ciently integrated (Stein & Stanford, 2008). Moreover, since different emotions were mapped on different spatial locations
auditory and visual parts were ef
ction at a level preceding decisional and motor processing stages.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Thirteen volunteers, recruited from the University of Hamburg
and from the local community, took part either for course credits or
a monetary compensation (7 Euro/h). One participant was excluded
from data analysis due to failure to follow task instructions.

37

Table 1
Overview of auditory and visual stimulus material used in previous studies.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Face stimuli/emotional pictures</th>
<th>Voice stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collignon et al. (2008)</td>
<td>Standardized set of dynamic color stimuli (Simon et al., 2008)</td>
<td>Montreal affective voices (Belin et al., 2008)</td>
</tr>
<tr>
<td>Dolan et al. (2001)</td>
<td>Static facial stimuli (Ekman, &amp; Friesen, 1976)</td>
<td>Recorded vocal stimuli</td>
</tr>
<tr>
<td>Ethofer et al. (2006)</td>
<td>Static facial stimuli (Ekman, &amp; Friesen, 1976)</td>
<td>Recorded vocal stimuli</td>
</tr>
<tr>
<td>Flom and Babrick (2007)</td>
<td>Videos</td>
<td>Recorded vocal stimuli</td>
</tr>
<tr>
<td>Ghazanfar and Logothetis (2003)</td>
<td>Videos of monkey faces</td>
<td>Monkey calls</td>
</tr>
<tr>
<td>Hagan et al. (2009)</td>
<td>Static facial stimuli (Ekman, &amp; Friesen, 1976)</td>
<td>Professional sound-effects from internet sites</td>
</tr>
<tr>
<td>Kreifelts et al. (2007)</td>
<td>Video sequences</td>
<td>and recorded vocal stimuli</td>
</tr>
<tr>
<td>Kreifelts et al. (2009, 2010)</td>
<td>Video sequences</td>
<td>Voices (two male and two female voices)</td>
</tr>
<tr>
<td>Magné et al. (2008)</td>
<td>Static facial stimuli (Ekman, &amp; Friesen, 1976)</td>
<td>Words spoken by three female and three male actors</td>
</tr>
<tr>
<td>Park et al. (2010)</td>
<td>Photographs (grey scaled and hair-masked)</td>
<td>Recorded vocal stimuli</td>
</tr>
<tr>
<td>Robins et al. (2009)</td>
<td>Dynamic facial stimuli</td>
<td>Recorded vocal stimuli</td>
</tr>
<tr>
<td>Spreckelmeyer et al. (2006)</td>
<td>International Affective Pictures (IAPS)</td>
<td>Voices of opera singers</td>
</tr>
<tr>
<td>Tanaka et al. (2010)</td>
<td>Video recordings of Japanese and Dutch speakers</td>
<td>Auditory recordings of Japanese and Dutch speakers</td>
</tr>
<tr>
<td>Van den Stock et al. (2007)</td>
<td>Photographs representing actors expressing different emotions</td>
<td>Spoken sentences</td>
</tr>
<tr>
<td>Van den Stock et al. (2008)</td>
<td>with their whole body but with the face blurred.</td>
<td>Recorded vocal stimuli (human and animal vocalizations)</td>
</tr>
<tr>
<td>Vroomen et al. (2001)</td>
<td>Static facial stimuli (Ekman, &amp; Friesen, 1976)</td>
<td>Recorded vocal stimuli</td>
</tr>
</tbody>
</table>
at the Hamburg University of Applied Sciences. In a second step all stimuli were digitized. The video signal was digitally sampled at 33 frames per second with a 24-bit resolution at 640×480 pixels size. Actors were filmed in front view; outer facial features including hair and ears were covered (Fig. 1). Videos were presented in gray scale with a black background. Lighting, background, height and distance were adjusted for all actors using Adobe Premiere Pro 2.0.

The accompanying utterances were recorded with a Neumann V87 Niere condenser microphone, equated for root mean square of 0.025; adjusted with the GOLDWAVE software and digitally sampled at 16 bit. During the presentation of unimodal auditory stimuli, a white fixation cross was presented in the middle of the screen (diameter 2.6 cm, visual angle 2.5°).

2.1.3. Stimulus selection and evaluation

For the purpose of this study, the 48 unimodal auditory stimuli were selected from a larger sample so that the mean duration of the unimodal auditory stimuli (761 ms; range: 500–1400 ms) did not differ between the four emotional conditions (main effect Emotional Prosody, \( F(3,44) = 0.18; P = 0.905 \)) (see Tables 2 and 3 for a more detailed stimulus description).

The auditory stimulus set was rated in a pilot study using a different participant sample (\( N = 32 \), mean age: 26 years; range: 20–48 years; 21 females, all right-handed) along the dimensions Valence, Arousal and Intensity for each of the emotions (angry, happy, sad and neutral). Participants first had to classify the auditory stimulus as expressing either an angry, happy, sad or neutral emotion. Moreover, they had to indicate whether they were spoken by a male or female speaker. Thirdly, each stimulus had to be rated along the dimensions Valence, Arousal and Intensity rating: not at all arousing = 0, maximally arousing = 10). Finally, they had to rate to what degree the emotions were represented in each of the stimuli (intensity rating: not at all = 0, to a very high degree = 10).

As intended, the valence, arousal, and intensity ratings of the final stimulus sample depended on Emotion (Intensity: \( F(3,93) = 10.59, P < 0.001 \); Arousal: \( F(3,93) = 24.86, P < 0.001 \); Valence: \( F(3,93) = 93.19, P < 0.001 \)). With regard to intensity ratings, the angry, sad and neutral voices revealed higher intensity ratings compared to happy voices (all \( P < 0.05 \)). Moreover, intensity ratings for sad voices were higher compared to neutral voices (\( P = 0.014 \)). According to the arousal ratings, all emotional voices (happy, angry, and sad) revealed higher arousal scores compared to neutral voices (all \( P < 0.05 \)). With regard to valence, angry voices revealed lower valence scores compared to happy, sad and neutral voices (all \( P < 0.05 \), Table 2). Moreover, valence scores for sad voices were lower compared to happy and neutral voices (all \( P < 0.05 \)).

For auditory stimuli, the corresponding facial stimuli were selected. Since previous studies have provided evidence of human facial expressions being much easier recognized than vocal prosody (Collignon et al., 2008; Hanley, Smith, & Hadfield, 1998; Joassin, Maurage, Bruyer, Crommelinck, & Campanella, 2004; Sheffert & Olson, 2004; von Kriegstein & Giraud, 2006) the corresponding facial expressions were selected by three new raters (two students, all female, mean age: 29 years; range: 23–39 years). Mean duration of unimodal visual stimuli was 902 ms (range: 670–1300 ms) and the duration did not significantly differ for the different emotions (\( F(3,44) = 0.89, P = 0.453 \); see Table 3 b). Visual stimuli had a longer duration than auditory stimuli because faces were visible for an average of 150 ms after the offset of the auditory stream (Fig. 1).

After selecting unimodal auditory and unimodal visual stimuli for the unimodal conditions, bimodal stimuli were created by combining a visual and an auditory stimulus. These stimuli pairings were emotionally congruent (e.g. happy face–happy voice) or emotionally incongruent (e.g. happy face–sad voice). In order to create bimodal stimuli, visual stimuli were combined with the auditory track of the same speaker articulating the same word. Each stimulus was recorded in three versions. To avoid that behavioral differences (e.g. the expected longer reaction times for emotionally incongruent

<table>
<thead>
<tr>
<th>Vocal emotional prosody</th>
<th>Valence Mean SE</th>
<th>Arousal Mean SE</th>
<th>Intensity Mean SE</th>
<th>Percent correct Mean SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad</td>
<td>3.44 0.3</td>
<td>5.16 0.4</td>
<td>8.7 0.2</td>
<td>97.66 0.67</td>
</tr>
<tr>
<td>Angry</td>
<td>2.26 0.2</td>
<td>5.48 0.4</td>
<td>8.3 0.2</td>
<td>90.63 1.8</td>
</tr>
<tr>
<td>Happy</td>
<td>6.26 0.2</td>
<td>5.38 0.3</td>
<td>7.2 0.2</td>
<td>80.73 2.06</td>
</tr>
<tr>
<td>Neutral</td>
<td>6.04 0.3</td>
<td>3.02 0.3</td>
<td>8 0.2</td>
<td>82.29 2.24</td>
</tr>
<tr>
<td>Mean</td>
<td>4.49 0.2</td>
<td>4.76 0.2</td>
<td>8.04 0.15</td>
<td>87.83 0.96</td>
</tr>
</tbody>
</table>

Note: Corresponding visual stimuli were selected depending on the prior selection of auditory stimuli and therefore were not rated.

Fig. 1. Bimodal stimulus. Upper row: Emotional facial stimulus of a female actor articulating the German bisyllable “Lolo”. Lower row: A silent period of 50 ms was added at the beginning of the auditory stimulus. After voice offset, the video file continued for a mean duration of about 150 ms until the facial articulation finished.
Additionally, 6 deviants were presented in each condition (see Fig. 2): stimuli were presented, resulting in a total of 192 different stimuli. Of 60 cm a chin rest was used. To guarantee a constant viewing distance the sound track varied between 65 and 72 dB. Visual stimuli were placed on the left and right side of the computer monitor. The level of −

\[ r = 0.57, n = 48; \text{incongruent stimuli: } r = 0.74, n = 144; \ Z = -1.79, P = 0.07. \]

Auditory stimuli were presented with loudspeakers, which were placed on the left and right side of the computer monitor. The level of the sound track varied between 65 and 72 dB. Visual stimuli were presented in the center of the screen (width = 4° of visual angle; height = 9° of visual angle). To guarantee a constant viewing distance of 60 cm a chin rest was used.

Altogether, 48 unimodal auditory, 48 unimodal visual, 48 emotionally congruent bimodal and 48 emotionally incongruent bimodal stimuli were presented, resulting in a total of 192 different stimuli. Additionally, 6 deviants were presented in each condition (see Fig. 2): 6 unimodal auditory, 6 unimodal visual, 6 emotionally congruent bimodal and 6 emotionally incongruent bimodal deviant stimuli were created, thus 216 stimuli in total were used in Experiment 1.

Visual deviants consisted of a black dot (diameter = 6 mm, 0.6°) presented at one of the four possible locations of the face (forehead, nose, left or right cheek) for 100 ms within the last 130–330 ms of stimulus presentation. Auditory deviants consisted of one of the four tones (600 Hz, 900 Hz, 1200 Hz and 1500 Hz, duration: 100 ms) presented within the last 130–330 ms of the auditory stream. Participants had to detect dots presented at four different positions in the face and rare auditory tones presented with different frequencies.

For the emotionally incongruent bimodal stimuli, we combined each of the four emotional faces (happy, angry, neutral, and sad) with the three emotionally incongruent voice recordings (happy face–sad voice, etc. for a total of 12 incongruent combinations). Three different words spoken by four actors were used for these 12 different face–voice pairs (4 actors×3 words×12 different emotional face–voice pairings) resulting in 144 emotionally incongruent bimodal stimuli. Only a subset of 48 emotionally incongruent bimodal stimuli was used (Table 4).

Table 3
Mean and standard errors for the duration of a) unimodal auditory, b) unimodal visual and c) emotionally congruent bimodal and d) emotionally incongruent bimodal stimuli.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SE</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad</td>
<td>770 ms</td>
<td>51</td>
<td>893 ms</td>
<td>14</td>
</tr>
<tr>
<td>Angry</td>
<td>733 ms</td>
<td>69</td>
<td>945 ms</td>
<td>32</td>
</tr>
<tr>
<td>Happy</td>
<td>759 ms</td>
<td>41</td>
<td>870 ms</td>
<td>22</td>
</tr>
<tr>
<td>Neutral</td>
<td>781 ms</td>
<td>15</td>
<td>901 ms</td>
<td>53</td>
</tr>
<tr>
<td>Mean</td>
<td>761 ms</td>
<td>23</td>
<td>902 ms</td>
<td>18</td>
</tr>
<tr>
<td>c) Duration (emotionally congruent bimodal stimuli)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sad</td>
<td>948 ms</td>
<td>32</td>
<td>904 ms</td>
<td>24</td>
</tr>
<tr>
<td>Angry</td>
<td>995 ms</td>
<td>52</td>
<td>931 ms</td>
<td>32</td>
</tr>
<tr>
<td>Happy</td>
<td>923 ms</td>
<td>33</td>
<td>896 ms</td>
<td>29</td>
</tr>
<tr>
<td>Neutral</td>
<td>920 ms</td>
<td>48</td>
<td>884 ms</td>
<td>27</td>
</tr>
<tr>
<td>Mean</td>
<td>946 ms</td>
<td>21</td>
<td>904 ms</td>
<td>14</td>
</tr>
</tbody>
</table>

compared to emotionally congruent bimodal trials) were due to differences in synchrony of the combined auditory and visual stimuli, we always used a visual stimulus that did not originate from the same recording as the voice, even in congruent combinations.

Mean duration of the emotionally congruent bimodal stimuli was 946 ms (range: 734–1468 ms) and mean duration of emotionally incongruent bimodal stimuli was 904 ms (range: 633–1501 ms). Table 3c and d). Mean duration of emotionally congruent bimodal stimuli was longer than the emotionally incongruent bimodal stimuli and unimodal visual stimuli. Correlations between the duration of auditory and visual stimuli marginally differed between emotionally congruent and emotionally incongruent bimodal conditions (congruent: \( r = 0.57, n = 48 \); incongruent stimuli: \( r = 0.74, n = 144; Z = -1.79, P = 0.07 \)).

At the beginning of each block, participants were instructed to attend either to the faces (Attend Face) or voices (Attend Voice). Half of the participants began with the Attend Face condition, the other half with the Attend Voice condition. To prevent the participants from closing their eyes when they had to attend to vocal stimuli only, a secondary task was introduced. In the Attend Voice condition, participants had to detect a black dot (diameter = 6 mm, 0.6°) presented at one of 4 possible locations of the face (forehead; nose; left or right cheek) for 100 ms within the last 130–330 ms of stimulus presentation. In the Attend Face condition, participants had to detect one of four tones (600 Hz, 900 Hz, 1200 Hz and 1500 Hz, duration: 100 ms) presented within the last 130–330 ms of the auditory stream.

Each trial started with the presentation of a bimodal warning stimulus (duration = 500 ms; grey circle: diameter 2° of visual angle combined with multispeaker noise) (Fig. 3).

After a variable interstimulus interval (600–700 ms, uniform distribution), either a unimodal or a bimodal stimulus was presented. Participants had to categorize the emotion of the attended modality as quickly and accurately as possible as neutral, angry, happy, or sad. They responded with one of the four fingers of their right hand (index, middle, ring and little finger) by pressing one of the four buttons of the computer keyboard. Three seconds after the offset of the target

2.1.4. Procedure

The experiment was run in a sound attenuated and dimly lit room at the University of Hamburg.

At the beginning of each block, participants were instructed to attend either to the faces (Attend Face) or voices (Attend Voice). Half of the participants began with the Attend Face condition, the other half with the Attend Voice condition. To prevent the participants from closing their eyes when they had to attend to vocal stimuli only, a secondary task was introduced. In the Attend Voice condition, participants had to detect a black dot (diameter = 6 mm, 0.6°) presented at one of 4 possible locations of the face (forehead; nose; left or right cheek) for 100 ms within the last 130–330 ms of stimulus presentation. In the Attend Face condition, participants had to detect one of four tones (600 Hz, 900 Hz, 1200 Hz and 1500 Hz, duration: 100 ms) presented within the last 130–330 ms of the auditory stream.

Each trial started with the presentation of a bimodal warning stimulus (duration = 500 ms; grey circle: diameter 2° of visual angle combined with multispeaker noise) (Fig. 3).

After a variable interstimulus interval (600–700 ms, uniform distribution), either a unimodal or a bimodal stimulus was presented. Participants had to categorize the emotion of the attended modality as quickly and accurately as possible as neutral, angry, happy, or sad. They responded with one of the four fingers of their right hand (index, middle, ring and little finger) by pressing one of the four buttons of the computer keyboard. Three seconds after the offset of the target

![Fig. 2](image-url) Distribution of stimuli (standards and deviants) in the unimodal, congruent bimodal and incongruent bimodal condition separately for the Attend Face and Attend Voice condition.
stimulus, the same stimulus was presented a second time. With an ISI of 50 ms, a five-point rating scale was presented. Participants had to rate the intensity of emotional expression of the attended modality (1 = low and 5 = high). As soon as the participant had responded, the next trial started with the warning stimulus. The assignment of the emotion to one of the four buttons was balanced across participants. Two practice blocks were used to familiarize the participants with the task. For deviants, participants had to additionally press the space bar following the emotional classification. After pressing the space bar, the next trial started immediately.

Stimuli were presented in a random sequence in 12 blocks of 27 stimuli each. In the Attend Face condition, only unimodal visual and bimodal stimuli were presented, whereas in the Attend Voice condition, only unimodal auditory stimuli and bimodal stimuli were used (see Fig. 2). Identical bimodal stimuli were used for both the Attend Face and the Attend Voice condition.

### 2.1.5. Data analysis

Mean reaction times, percent correct, inverse efficiency (IE) scores and mean intensity scores to standards were determined for each experimental condition and participant (Townsend & Ashby, 1978, see below). Incorrectly classified stimuli or trials with reaction times below 200 ms or exceeding 3950 ms (stimulus duration plus interstimulus interval) were disregarded. Since IE scores integrate both response times and percentage correct, we based our conclusions of Experiment 1 mainly on IEs. In order to allow us to compare results from Experiment 1 and Experiment 2, we analyzed both mean reaction times and percentage correct, as well.

Separate analyses were calculated for the mean response times observed under the two attentional conditions (Attend Face and Attend Voice), for two reasons: First, within these conditions, statistical analyses were cut down to two one-way ANOVAs with a single factorial manipulation of three levels (namely, Stimulus Type: unimodal, emotionally congruent bimodal, and emotionally incongruent bimodal). In this case, significance levels of the three possible pairwise comparisons do not need to be corrected if the overall analysis is significant (closed testing method, see e.g. p. 138 in Hsu, 1996). Secondly, the main effect of Attention would simply reflect that response times to emotional facial stimuli differ from response times for emotional voices. This main effect was not part of our research questions. We expect qualitatively similar effects of Stimulus Type within the two levels of Attention, that is, that the same pairwise comparisons are significant. Thus, results of the condition-specific tests confirm each other and were, thus, reported without correction for multiplicity.

Participant-specific mean reaction times were submitted to a one-way repeated measurement ANOVA with the factor Stimulus Type (unimodal, emotionally congruent bimodal, and emotionally incongruent bimodal). The Huynh and Feldt (1976) correction for violations of the sphericity assumption was employed. The conventional significance level of $\alpha = 5\%$ was used. Conditional on a significant one-way-analysis, posthoc pairwise comparisons were performed using two-tailed $t$-tests for paired samples (in line with the directional nature of our expectations, we additionally report the results of one-tailed tests at those places where conclusions of one-tailed tests and two-tailed tests differ).

To account for a possible speed–accuracy trade-off, inverse efficiency scores were calculated by dividing mean response times by the proportion of correct trials (Townsend & Ashby, 1978). Similar to the mean reaction times, lower IE values indicate better performance. If the percentage of correct responses does not take extremely low values, participant-specific differences of IE scores can be assumed to be asymptotically normally distributed under the null-hypothesis, such that parametric ANOVAs and paired $t$-tests can be used to demonstrate performance differences between the experimental conditions. The IE scores were, thus, analyzed in the same way as the mean reaction times. The corresponding analysis was performed for the intensity ratings (for correctly classified stimuli only).

For the percentage of correct responses, convergence to normality is supposed to be slow, and variance inhomogeneity is to be expected for different conditions. Therefore, we used non-parametric Friedman tests for demonstrating differences between the three levels of Stimulus Type. For the reasons stated above, this analysis was run twice, once for the Attend Face and once for the Attend Voice condition.

### Table 4

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Words</th>
<th>Actor</th>
<th>Stimulus Type</th>
<th>Stimulus</th>
<th>ISI (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>Voice</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1. Angry</td>
<td>Happy</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2. Angry</td>
<td>Sad</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3. Angry</td>
<td>Neutral</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>4. Happy</td>
<td>Sad</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>5. Happy</td>
<td>Neutral</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>6. Happy</td>
<td>Angry</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>7. Sad</td>
<td>Angry</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>8. Sad</td>
<td>Happy</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9. Sad</td>
<td>Neutral</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>10. Neutral</td>
<td>Happy</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>11. Neutral</td>
<td>Sad</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>12. Neutral</td>
<td>Angry</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
Fig. 4. Experiment 1: Reaction times (ms), percent correct (%), inverse efficiency scores (ms) and mean perceived emotional intensity scores for unimodal, emotionally congruent bimodal and emotionally incongruent bimodal stimuli (averaged across emotions). Horizontal lines indicate significant differences between conditions ($P < 0.05$).
Conditional on a significant Friedman test, pairwise comparisons were calculated using Wilcoxon tests for the three levels of Stimulus Type. Exploratory analyses of the influence of Emotion are reported in descriptive form only.

2.2. Results

Fig. 4 shows the mean IE scores, response times, percentage of correct responses and mean perceived emotional intensity for the different conditions used in the experiment.

Descriptive statistics (IE scores, response times and percent correct, and mean intensity rating) for each emotion and experimental condition are reported in Table 5.

2.2.1. IE scores

For both Attend Face and Attend Voice conditions, there was a main effect of Stimulus Type (Attend Face: F(2,22) = 11.38, P = 0.004 and Attend Voice: F(2,22) = 28.88, P<0.001) revealing higher performance to emotionally congruent bimodal trials compared to emotionally incongruent bimodal trials (Attend Face: t(11) = −3.88, P = 0.003 and Attend Voice: t(11) = −5.28, P = 0.001) (Fig. 4).

Higher performance in emotionally congruent bimodal trials compared to unimodal trials was only observed in Attend Face (t(11) = −1.34, P = 0.206). Moreover, higher performance to unimodal compared to emotionally incongruent bimodal trials was observed for both attentional conditions (Attend Face: t(11) = −2.81, P = 0.017 and Attend Voice: t(11) = −6.12, P<0.001).

2.2.2. Mean reaction times

Separate ANOVAs for the Attend Face and the Attend Voice condition were calculated to test the reliability of incongruency effects under both attention conditions. For both conditions, there was a main effect of Stimulus Type (Attend Face: F(2,22) = 6.17, P = 0.014 and Attend Voice: F(2,22) = 25.28, P<0.001) revealing shorter mean reaction times for emotionally congruent bimodal stimuli compared to emotionally incongruent bimodal stimuli (Attend Face: t(11) = −3.21, P = 0.008 and Attend Voice: t(11) = −7.19, P<0.001) (Fig. 4).

Moreover, shorter reaction times for unimodal compared to emotionally incongruent bimodal trials were observed both in the Attend Face and the Attend Voice condition (Attend Face: t(11) = −2.26, P = 0.045 and Attend Voice: t(11) = −6.94, P<0.001) (Fig. 4). The difference between emotionally congruent bimodal and unimodal trials was not significant (Attend Face: t(11) = −0.57, P = 0.575 and Attend Voice: t(11) = −0.003, P = 0.998).

2.2.3. Percentage of correct responses

The proportion of correct responses was generally higher in congruent than in incongruent stimuli for both, the Attend Face and the Attend Voice condition (Attend Face: χ²(2) = 20.59, P<0.001 and Attend Voice: χ²(2) = 14.00, P<0.001). Condition specific Wilcoxon tests for paired samples were significant (Attend Face: Z = 3.06, P = 0.002 and Attend Voice: Z = 2.90, P = 0.004). Moreover, higher rates of correct classifications were observed in the unimodal compared to the emotionally incongruent bimodal trials (Attend Face: Z = 2.37, P = 0.018 and Attend Voice: Z = 2.82, P = 0.005) and in the emotionally congruent bimodal compared to the unimodal trials (Attend Face: Z = −2.98, P = 0.003 and Attend Voice: Z = −1.65, two-tailed P = 0.099, one-tailed P of exact Wilcoxon test = 0.055) (see Fig. 4).

2.2.4. Perceived emotional intensity

The main effect of Stimulus Type was significant for both the Attend Face and the Attend Voice condition (Attend Face: F(2,22) = 8.50, P = 0.006 and Attend Voice: F(2,22) = 8.23, P = 0.005).

The emotionally congruent bimodal condition was perceived as more intense compared to the emotionally incongruent bimodal condition in both Attend Face and Attend Voice condition (Attend Face: t(11) = 3.20, P = 0.008 and Attend Voice: t(11) = 3.26, P = 0.008). This was true for the unimodal condition as well which was perceived as more intensive as compared to the emotionally incongruent bimodal condition in both attention conditions (Attend Face: t(11) = 2.98, P = 0.012 and Attend Voice: t(11) = 2.87, P = 0.015). No significant differences were observed between intensity scores to emotionally congruent bimodal and unimodal stimuli (Attend Face: t(11) = −0.85, P = 0.412 and Attend Voice: t(11) = 0.24, P = 0.812) (see Fig. 4).

2.2.5. Deviants

No significant main effect of Stimulus Type was observed for correct deviant classifications (Attend Face: χ²(2) = 3.29, P = 0.192 and Attend Voice: χ²(2) = 2.80, P = 0.272).

2.3. Discussion

In Experiment 1 we investigated whether the presentation of an emotional expression in an attended modality modulates the processing of an emotional expression in an unattended modality. Facial and vocal stimuli were presented either unimodally or bimodally. Participants had to categorize the emotional expression of the face or vocal stimuli were presented either unimodally or bimodally.

The expressed valence in the auditory and visual part was either the same (congruent) or differed (incongruent). Participants were...
able to successfully discriminate between the four different facial expressions and vocal prosodies (percentage of correct responses to unimodal presentations was above 80%). By contrast categorization performance of bimodal stimuli was clearly influenced by the unattended modality in both the Attend Face and the Attend Voice condition. Despite the instruction to ignore the second modality input, performance and intensity ratings were higher in emotionally congruent than in emotionally incongruent combinations of faces and voices. These findings are in accordance with the view that emotional expressions in both faces and voices are automatically extracted and combined (e.g. Öhman et al., 2001; Vroomen et al., 2001).

In addition, incongruent emotional information resulted in behavioral decrements relative to unimodal performance irrespective of which modality was attended. This result substantiates the view that emotional expressions of different modalities are automatically extracted and combined. Thus, meaning as a supramodal feature seems to play a critical role in multisensory interactions (Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004).

For IE scores and percentage of correct classifications crossmodal facilitation of emotionally congruent bimodal compared to unimodal stimuli was observed. In particular, this multisensory gain was significant in the Attend Face condition. It might be speculated that the lack of multisensory gains in the Attend Voice condition was due to a ceiling effect in the voice only condition. Participants performed 89% correct in this condition. As expected from the inverse effectiveness law (Stein & Stanford, 2008), multisensory gains are mostly expected for unisensory stimuli with a low processing efficiency. However, to confirm this conclusion, reaction times for unattended unisensory stimuli would be needed, which were not available in the present study.

One might argue that the incongruency effect emerged from lower synchrony of auditory and visual stimuli when combined to emotionally incongruent bimodal stimuli compared to emotionally congruent bimodal stimuli. We consider this explanation as unlikely since we used the auditory track and video recordings of different actors both for the emotionally congruent and for the emotionally incongruent bimodal condition. Moreover, we calculated the correlation between auditory and visual stimuli separately for congruent and incongruent stimuli. Though these correlations marginally differ, the correlation is slightly higher, if at all, for emotionally incongruent bimodal stimuli rendering the account unlikely that synchrony differences between the emotionally congruent and emotionally incongruent bimodal stimuli were the main reason for longer reaction times in the incongruent than the congruent condition.

Our results suggest that multisensory interactions of emotional experience take place, at least to some extent, outside the focus of attention (i.e. crossmodal incongruency effects occurred even though one sensory channel is not task relevant). This result contrasts with a report of Talsma, Doty, and Woldorff (2007). However, they used different kinds of stimulus material such as tones and visual gratings. By contrast our paradigm used more complex visual and auditory materials (such as emotional human faces and voices). These seemingly contradictory findings may be due to the fact that emotional human faces and voices might capture attention automatically in contrast to abstract stimulus material used by Talsma and colleagues.

Given that Experiment 1 suggests that participants were able to discriminate between single emotions and that crossmodal interferences were observed for both the Attend Face and the Attend Voice task, we conducted a second experiment to test whether or not the incongruency effect mainly arises from a response conflict. To this aim, two emotions were mapped on one button, thereby disentangling perceptual decisions from processes of response selection.

## 3. Experiment 2

In the second Experiment, two emotions rather than one were assigned to a response button. Thus, in the emotionally incongruent bimodal condition the emotional expression of the primary and secondary modality were mapped either on the same or on different response keys (Fig. 5). Incongruency effects observed in the first condition can unequivocally be interpreted as an interference effect between auditory and visual emotional information arising at processing stages prior to response selection.

### 3.1. Method

#### 3.1.1. Participants

Twenty four participants took part in this experiment. Twelve participants (mean age: 24 years; range: 20–29 years; SD = 2.6; 11 female; all right handed) were run with the button-emotion assignment response key I: angry–neutral, response key II: happy–sad, the other twelve participants (mean age: 25 years; range: 18–33 years; SD = 4.4; 9 female, 11 right handed) with the button-

![Fig. 5. Experiment 2. Trials of the emotionally incongruent bimodal condition with the emotion of the secondary modality mapped (a) either on the same (happy–angry) or (b) on a different key from the emotion of the primary modality (happy–sad). (a) Participants had to attend to the face. A happy face was combined with an angry voice. Both task-relevant (happy face) and task-irrelevant (angry voice) emotions were mapped on the left button. (b) A happy face combined with a sad voice was presented. Task-relevant (happy face) and task-irrelevant emotions (sad voice) are mapped on two different buttons.](image-url)
emotion assignment response key I: angry–happy, response key II: neutral–sad.

3.1.2. Procedure

Two different assignments of vocal and facial expressions were mapped on two buttons of the computer keyboard (assignment 1: angry–neutral (button 1) versus happy–sad (button 2); assignment 2: angry–happy (button 1) versus neutral–sad (button 2)) (see button assignment 2 in Fig. 5). This particular expression–response key assignment resulted in one condition with response conflict (emotionally incongruent bimodal condition with response conflict), whereas the other conditions were free of response conflicts (unimodal, congruent bimodal and incongruent condition without response conflict) (see Fig. 5a and b).

The experimental design and the stimulus material were the same as described in Experiment 1. In different blocks, participants had to attend either to the faces or to the voices and they had to categorize as fast and as accurately as possible the emotional valence of the attended face or voice, respectively, while ignoring possible input from the second modality. After categorizing the emotional expression, they had to rate its intensity on a five point rating scale. The mapping of emotions to the right index and middle finger was counterbalanced across the participants. For each condition (unimodal, emotionally congruent bimodal, emotionally incongruent bimodal condition with response conflict, emotionally incongruent bimodal without response conflict), 48 stimuli were presented for both Attend Face and Attend Voice. Additionally, six deviants were presented in each condition (see Experiment 1). Thus, a total of 432 trials were presented. Attend Face and Attend Voice condition altered every second block.

3.1.3. Data analysis

Mean reaction times, percent correct, IE scores, and mean emotional intensity ratings were calculated separately for each condition (unimodal, emotionally congruent bimodal, emotionally incongruent bimodal without response conflict, and emotionally incongruent bimodal with response conflict) and button assignment (version 1 versus version 2). It has to be noted that it is not always possible to distinguish between correct responses to the primary modality and possible “incorrect” responses to the secondary modality since two emotions were mapped on one button. Thus, confirmatory analyses are based on the mean reaction times only; supplementary analyses of percent correct, IE scores, and mean intensity scores are only descriptive in nature. The perceived emotional intensity was analyzed, as in Experiment 1. As we did not expect different rating intensities for the two incongruent conditions, ratings were averaged across both conditions.

Participant-specific mean reaction times and IE scores were submitted to a one-way ANOVA with the repeated measurement factor Stimulus Type (emotionally congruent bimodal and emotionally incongruent bimodal with and without response conflict), with the Huynh and Feldt (1976) correction for violations of the sphericity assumption. Because the main hypothesis primarily concerns the three bimodal conditions (emotionally congruent bimodal, emotionally incongruent bimodal without conflict, and emotionally incongruent bimodal with conflict) we report only an exploratory analysis of the responses to unimodal stimuli. As for Experiment 1, separate analyses were carried out for the two attentional conditions (Attend Face and Attend Voice). Conditional on significant main effects of Stimulus type, pairwise comparisons were made for the three levels of Stimulus Type using two-tailed t-tests for dependent samples. For percent correct, we used non-parametric Friedman tests for demonstrating differences between the three levels of Stimulus Type. This analysis was again run twice, once for Attend Face and once for Attend Voice. Conditional on a significant Friedman test, pairwise comparisons were made using Wilcoxon tests. In Table 7 we report results of post-hoc comparisons for unimodal versus emotionally congruent bimodal and emotionally incongruent bimodal trials for reaction times, IE scores, percentage of correct responses, and mean intensity scores.

### Table 6
Mean RT (ms), correct percent (%), IE scores (ms) and mean intensity scores of Experiment 2. Standard errors are reported in parentheses (SE). Mean intensity scores are averaged across the two incongruent conditions.

<table>
<thead>
<tr>
<th>Stimulus Type</th>
<th>RT (ms)</th>
<th>IE scores (ms)</th>
<th>Correct (SE)</th>
<th>Intensity (rating)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attend Face</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unimodal</td>
<td>1251 (55)</td>
<td>1405 (66)</td>
<td>89 (1)</td>
<td>3.61 (0.08)</td>
</tr>
<tr>
<td>Bimodal congruent</td>
<td>1201 (56)</td>
<td>1307 (67)</td>
<td>92 (0.6)</td>
<td>3.59 (0.09)</td>
</tr>
<tr>
<td>Incongruent no conflict</td>
<td>1290 (57)</td>
<td>1433 (71)</td>
<td>90 (1.2)</td>
<td>3.59 (0.10)</td>
</tr>
<tr>
<td>Incongruent with conflict</td>
<td>1315 (62)</td>
<td>1661 (108)</td>
<td>81 (2.03)</td>
<td></td>
</tr>
<tr>
<td><strong>Attend Voice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unimodal</td>
<td>1302 (53)</td>
<td>1461 (62)</td>
<td>89 (1.1)</td>
<td>3.51 (0.09)</td>
</tr>
<tr>
<td>Bimodal congruent</td>
<td>1291 (53)</td>
<td>1383 (63)</td>
<td>93 (0.8)</td>
<td>3.50 (0.11)</td>
</tr>
<tr>
<td>Incongruent no conflict</td>
<td>1387 (59)</td>
<td>1580 (61)</td>
<td>87 (1.2)</td>
<td>3.20 (0.09)</td>
</tr>
<tr>
<td>Incongruent with conflict</td>
<td>1443 (68)</td>
<td>1939 (105)</td>
<td>76 (2.4)</td>
<td></td>
</tr>
</tbody>
</table>

3.2. Results

Table 6 shows the mean reaction times, IE scores, percentage of correct responses and mean intensity scores for the different conditions used in Experiment 2.

3.2.1. Mean reaction times

Separate analyses for the Attend Face condition revealed shorter reaction times to emotionally congruent bimodal stimuli compared to the emotionally incongruent bimodal stimuli (with and without response conflict; main effect of Stimulus Type: $F(2,46) = 19.21$, $P<0.001$; emotionally congruent bimodal versus emotionally incongruent bimodal without conflict: $t(23) = -5.12$, $P<0.001$; emotionally incongruent bimodal with conflict versus emotionally congruent bimodal and emotionally incongruent bimodal without conflict: $t(23) = -3.28$, $P=0.001$).

### Table 7
Pairwise t-tests (mean RT and IE scores) and Wilcoxon tests (percent correct) for comparing unimodal with emotionally congruent bimodal, emotionally incongruent bimodal trials with and without a response conflict.

<table>
<thead>
<tr>
<th>Stimulus Type</th>
<th>RT (ms)</th>
<th>IE scores (ms)</th>
<th>Percent correct (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attend Face</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unimodal vs. congruent</td>
<td>3.22</td>
<td>0.004</td>
<td>4.63</td>
</tr>
<tr>
<td>Unimodal vs. incongruent with response conflict</td>
<td>-3.3</td>
<td>0.002</td>
<td>-4.38</td>
</tr>
<tr>
<td>Unimodal vs. incongruent without response conflict</td>
<td>-2.87</td>
<td>0.008</td>
<td>-1.65</td>
</tr>
<tr>
<td><strong>Attend Voice</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unimodal vs. congruent</td>
<td>0.76</td>
<td>0.453</td>
<td>3.86</td>
</tr>
<tr>
<td>Unimodal vs. incongruent without response conflict</td>
<td>-3.61</td>
<td>0.001</td>
<td>-3.45</td>
</tr>
<tr>
<td>Unimodal vs. incongruent with response conflict</td>
<td>-5.47</td>
<td>&lt;0.001</td>
<td>-6.98</td>
</tr>
</tbody>
</table>

The result pattern including unimodal, emotionally congruent bimodal and emotionally incongruent bimodal trials with and without a response conflict in Experiment 1, with highest performance observed in the congruent condition, intermediate performance in the unimodal condition, and lowest performance in the incongruent condition with response conflict (see Tables 6 and 7, Fig. 6).
emotionally congruent bimodal versus emotionally incongruent bimodal with conflict: $t(23) = -6.67, P < 0.001$ (Fig. 6b).

The mean response times observed in the two emotionally incongruent bimodal conditions (emotionally incongruent bimodal condition with response conflict and without response conflict) did not significantly differ from each other (emotionally incongruent bimodal without a response conflict versus emotionally incongruent bimodal with a response conflict: $t(23) = -1.09, P = 0.285$).

In Attend Voice, shorter reaction times were observed for the emotionally congruent bimodal condition compared to the emotionally incongruent bimodal condition with and without response conflict (Attend Voice: main effect of Stimulus Type: $F(2,46) = 20.52, P < 0.001$; emotionally congruent bimodal versus emotionally incongruent bimodal without a response conflict: $t(23) = -3.95, P < 0.001$; emotionally congruent bimodal versus emotionally incongruent bimodal with a response conflict: $t(23) = -5.99, P < 0.001$). Reaction times to emotionally incongruent bimodal stimuli without response conflict were shorter compared to emotionally incongruent bimodal stimuli with a response conflict ($t(23) = -2.51, P = 0.019$).5

3.2.2. IE scores

The main effect of Stimulus Type (emotionally congruent bimodal, emotionally incongruent bimodal without response conflict, and emotionally incongruent bimodal with response conflict) was significant for both attention conditions (Attend Face: $F(2,46) = 24.08, P < 0.001$ and Attend Voice: $F(2,46) = 45.76, P < 0.001$).

Participants reacted more efficiently to congruent bimodal compared to incongruent conditions (Attend Face: congruent versus incongruent without conflict: $t(23) = -5.61, P < 0.001$; congruent versus incongruent with conflict: $t(23) = -6.27, P < 0.001$ and Attend Voice: congruent versus incongruent without conflict: $t(23) = -6.18, P < 0.001$; congruent versus incongruent with response conflict: $t(23) = -8.29, P < 0.001$).

Within incongruent stimuli, performance on trials without response conflict was better than trials with conflict in both attention conditions (Attend Face: $t(23) = -4.22, P < 0.001$ and Attend Voice: $t(23) = -5.12, P < 0.001$).

3.2.3. Percent correct

The main effect of Stimulus Type was significant for the Attend Face and the Attend Voice condition (Attend Face: $\chi^2(2) = 24.08, P < 0.001$ and Attend Voice: $\chi^2(2) = 36.06, P < 0.001$). Condition specific Wilcoxon tests for paired samples were significant when comparing congruent bimodal stimuli with incongruent bimodal stimuli with conflict for both, Attend Face and Attend Voice conditions (Attend Face: $Z = 3.98, P < 0.001$ and Attend Voice: $Z = 4.10, P < 0.001$). The difference between congruent stimuli and conflict-free incongruent stimuli was significant in Attend Voice only ($Z = 3.62, P < 0.001$; Attend Face: $Z = 1.52, P = 0.128$). A higher number of correct classifications were observed in incongruent trials without response conflict compared to emotionally incongruent trials with response conflict (Attend Face: $Z = 3.90, P < 0.001$ and Attend Voice: $Z = 3.99, P < 0.001$).

3.2.4. Perceived emotional intensity

The main effect of Stimulus Type (unimodal, emotionally congruent bimodal and emotionally incongruent bimodal) was significant for both conditions (Attend Face: $F(2,46) = 17.03, P < 0.001$ and Attend Voice: $F(2,46) = 33.96, P < 0.001$). Post hoc t-tests revealed higher intensity ratings for congruent compared to incongruent stimuli (Attend Face: $t(23) = 5.22, P < 0.001$ and Attend Voice: $t(23) = 6.45, P < 0.001$).

---

5 Mean reaction times were unaffected by button assignment. There was neither a significant main effect of the group factor Button Assignment nor an interaction of Button Assignment with any of the other factors (Button Assignment: $F(1,122) = 1.04, P = 0.310$; Button Assignment x Attention: $F(1,122) = 0.77, P = 0.388$; Button Assignment x Stimulus Type: $F(3,66) = 1.03, P = 0.379$; Button Assignment x Stimulus Type x Attention: $F(3,66) = 0.08, P = 0.565$).
and the expressions depend on attention and whether they take place prior to a response selection process. Moreover, the present data suggest that incongruency effects might have been overestimated at least for some conditions in previous studies (Collignon et al., 2008; de Gelder & Vroomen, 2000).

3.2.5. Deviants
Deviant detection was nearly perfect: The mean detection rate for dots was 95% (SE = 1.4) and the mean detection rate for tones was 99% (SE = 0.4).

No significant main effect of Stimulus Type was observed for correct deviant classifications (Attend Face: $\chi^2(2) = 4.00, P = 0.135$ and Attend Voice: $\chi^2(2) = 2.40, P = 0.301$).

3.3. Discussion of Experiment 2

In order to conclude that auditory and visual emotional cues are automatically extracted and combined at a processing step preceding response selection, it was necessary to disentangle a possible contribution of a response conflict from interference at pre-response selection stages. In the second experiment, two emotions were mapped on either two different or one single response button, allowing us to estimate the contribution of a response conflict to the observed crossmodal incongruency effects. Results show that the incongruency effect was significant even after controlling for a response conflict. Our results, thus, strongly suggest that emotional information simultaneously provided by both the auditory and visual channel is matched prior to the selection of the response.

Significant differences for mean reaction times between the emotionally incongruent bimodal condition with response conflict and the incongruent condition without response conflict were only observed in the Attend Voice condition, but not in the Attend Face condition (Fig. 6). By contrast and importantly when eliminating response conflict, the incongruency effect was reliable for both the Attend Voice and the Attend Face condition. These results suggest that at least a portion of the incongruency effect precedes the response selection process. Moreover, the present data suggest that incongruency effects might have been overestimated at least for some conditions in previous studies (Collignon et al., 2008; de Gelder & Vroomen, 2000).

4. General discussion

The present study used crossmodal incongruency to investigate whether audio-visual interactions of vocal and facial emotional expressions depend on attention and whether they take place prior to response selection. In contrast to previous studies (e.g. de Gelder et al., 1999; Grossmann et al., 2006; Hagan et al., 2009; Magnée et al., 2008), task-relevant and task-irrelevant modalities originated from the same stimulus set, that is, vocal and visual recordings came from the same actor. Thus, in contrast to the bimodal stimulation used previously, our audio-visual stimuli were better synchronized and, thus, more likely to be integrated (see Stein & Stanford, 2008).

Similar audio-visual incongruency effects were observed irrespective of the task-relevant modality. Emotional signals provided by different sensory input channels seem, thus, to be integrated even if one sensory channel is not task relevant and, in principle, could be ignored. Since the same incongruency effects were observed after controlling for a response conflict (see Experiment 2), it seems justified to conclude that emotional audio-visual interaction takes place prior to a response selection.

We are nevertheless unable to exactly identify the processing stage of crossmodal interactions based on our behavioral data. There are, however, several recent electrophysiological and imaging studies suggesting early perceptual processing of audio-visual emotional information, independent of attention (de Gelder et al., 1999; Pourtois et al., 2000, 2002; Speerkelmeyer, Kutus, Urbach, Altenmüller, & Münte, 2006; Vuilleumier, Armony, Driver, & Dolan, 2001). For example, early emotional audio-visual integration effects have been reported for the time range of the N1 and P2, that is event-related potentials that have been associated with perceptual processing (N1: de Gelder et al., 1999; Pourtois et al., 2000; P2: Pourtois et al., 2002; Speerkelmeyer et al., 2006).

It might be argued that the relatively long response latencies observed in the present study are incompatible with an automatic or fast integration process. However, it has to be noted that our stimuli had a duration varying from 500 to 1501 ms. Moreover, we used a categorization task rather than a simple detection task. Thus, reaction times above 1000 ms can be expected. There are studies reporting that the integration of crossmodal information pertaining to objects depends on attention (e.g., Talsma et al., 2007). It might be speculated that whether or not attention is necessary or modulates crossmodal binding depends on the stimulus type rather than overall load (as suggested by Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). Emotional signals are of high relevance, particularly in a social context. For example, emotional signals are processed with high preference and efficiency (LeDoux, 1996; Vuilleumier et al., 2001). A rapid and early integration of emotional signals across modalities might further enhance the processing of such signals, increasing the efficiency of social interactions.

Bimodal facilitation effects were only observed for the Attend Face condition. We speculate that this finding may be linked to the fact that participants were slightly better able to categorize the auditory than the visual stimuli. This advantage might be due to the high relevance of prosody for language comprehension (Thönnessen et al., 2010). To recognize a particular prosody, various acoustic features, such as pitch, pitch variations, syllable duration and voice quality need to be combined. However, other studies have reported that visual expressions are judged with higher accuracy than emotional prosody in many everyday situations resulting in a higher reliance on visual as compared to auditory cues in emotional processing (e.g., Collignon et al., 2008). Nevertheless, as shown by the same authors, the dominance of one or the other modality depends on the reliability of the single modality input. Given the visually clearly recognizable facial expressions used in the present study, our results are consistent with the often observed visual dominance in emotional categorization.

5. Conclusion

Crossmodal interactions during the processing of auditory and visual emotional signals occur prior to response selection and seem to occur, at least partially, independent of attention. Thus, crossmodal integration of emotional signals provided via different modality channels seems to facilitate attentional orienting rather than to depend on attention.

Acknowledgements

We thank Anna Best, Karin Deazle, Daniela Heddaeus, Henrike Janssen and Liang Zhang for creating the stimulus material and Nils Skotara for programming support. We thank Corinna Klinge for providing us the rating program and Dagmar Tödter for participant recruitment. This study was supported by the Federal Ministry of Education and Research (G01GW0561).

References


