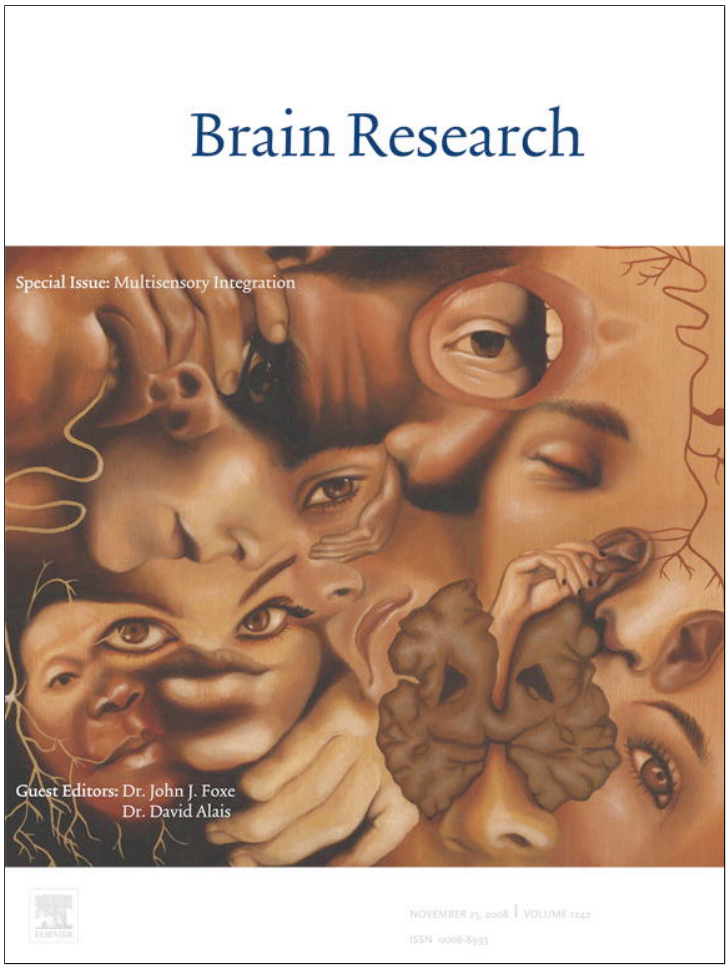

BRAIN RESEARCH 1242 (2008) 126–135

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Audio-visual integration of emotion expression

Olivier Collignon[a,b,*], *Simon Girard*[a], *Frederic Gosselin*[a], *Sylvain Roy*[a], *Dave Saint-Amour*[a,d], *Maryse Lassonde*[a,c], *Franco Lepore*[a,c]

[a]*Centre de Recherche en Neuropsychologie et Cognition (CERNEC), Université de Montréal, Montréal, Canada*
[b]*Neural Rehabilitation Engineering Laboratory, Centre for Research in Neurosciences (CRN), Université Catholique de Louvain, Brussels, Belgium*
[c]*Centre de Recherche CHU Sainte-Justine, Montréal, Canada*
[d]*Département d'Op htalmologie, Université de Montréal, Montréal, Canada*

ARTICLE INFO

Article history:
Accepted 10 April 2008
Available online 20 April 2008

Keywords:
Multisensory
Emotion
Disgust
Fear
Psychophysics

ABSTRACT

Regardless of the fact that emotions are usually recognized by combining facial and vocal expressions, the multisensory nature of affect perception has scarcely been investigated. In the present study, we show results of three experiments on multisensory perception of emotions using newly validated sets of dynamic visual and non-linguistic vocal clips of affect expressions. In Experiment 1, participants were required to categorise fear and disgust expressions displayed auditorily, visually, or using congruent or incongruent audio-visual stimuli. Results showed faster and more accurate categorisation in the bimodal congruent situation than in the unimodal conditions. In the incongruent situation, participant preferentially categorised the affective expression based on the visual modality, demonstrating a visual dominance in emotional processing. However, when the reliability of the visual stimuli was diminished, participants categorised incongruent bimodal stimuli preferentially via the auditory modality. These results demonstrate that visual dominance in affect perception does not occur in a rigid manner, but follows flexible situation-dependent rules. In Experiment 2, we requested the participants to pay attention to only one sensory modality at a time in order to test the putative mandatory nature of multisensory affective interactions. We observed that even if they were asked to ignore concurrent sensory information, the irrelevant information significantly affected the processing of the target. This observation was especially true when the target modality was less reliable. Altogether, these findings indicate that the perception of emotion expressions is a robust multisensory situation which follows rules that have been previously observed in other perceptual domains.

1. Introduction

Human beings must be able to understand the emotions of others in order to engage in successful social interactions. Affect perception, like speech perception, is a particular situation where the combination of information expressed from the face and the voice of the interlocutor optimises event identification. However, despite the fact that our ability to integrate these two sources in a unified percept could be a determinant for successful social behaviour, the

* *Corresponding author*. Université de Montréal, Département de Psychologie, CERNEC, 90 Vincent d'Indy, CP 6128, Succ. Centre-Ville, Montréal (Québec) H3C 3J7. Fax: +1 514 343 5787.
E-mail address: olivier.collignon@umontreal.ca (O. Collignon).
Abbreviations: RTs, Reaction Times; IE, Inverse efficiency

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.04.023

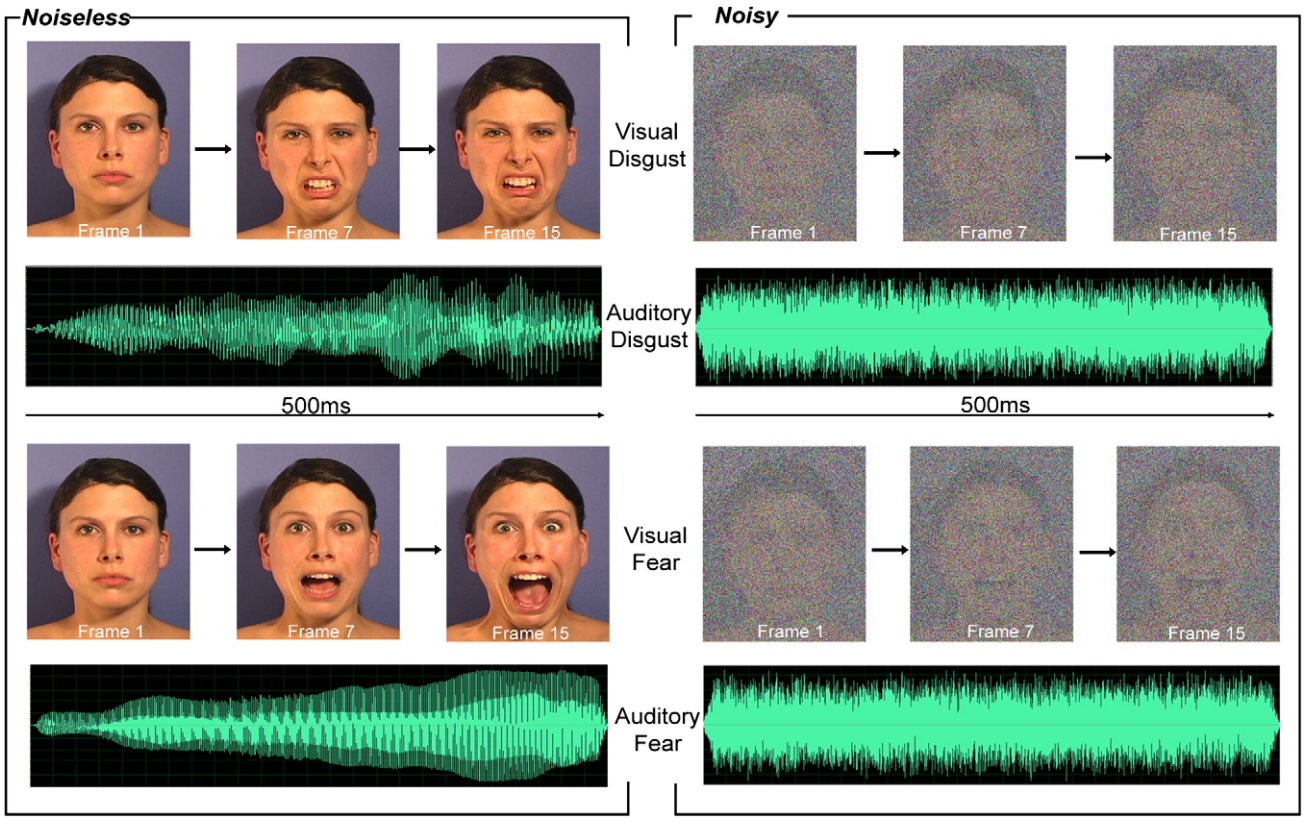

Fig. 1 – Schematic representation of the stimuli. In the three experiments, participants were required to discriminate between affective expressions of "fear" and "disgust". Stimuli consisted in video (from Simon et al., 2008) and non-linguistic vocal clips (from Belin et al., in press). Depending on the task, the clips were either displayed in noiseless condition or were presented with the addition of noise in order to decrease the reliability of the sensory information. These stimuli were either displayed alone and in bimodal congruent (the same expression in both modalities) or bimodal incongruent (different expressions in both modalities) combinations.

perception of affective states has typically been investigated using one modality at a time.

Recently, a few studies explored the multisensory nature of affective expressions (for review see Campanella and Belin, 2007). They indicated that congruency in information between facial expression and affective prosody facilitates behavioural reactions to emotional stimuli (Dolan et al., 2001; Massaro and Egan, 1996), and that information obtained via one sense can alter information processing in another (de Gelder and Vroomen, 2000; Ethofer et al., 2006a; Massaro and Egan, 1996). Such cross-modal biases occurred even when participants were instructed to base their judgement on just one of the modalities (de Gelder and Vroomen, 2000; Ethofer et al., 2006a), supporting the notion that processes underlying integration of facial and vocal affective information is automatic.

With only a few exceptions (de Gelder et al., 1999; Kreifelts et al., 2007), studies on bimodal perception of emotional expressions were conducted using static faces as stimuli. However, neuroimaging studies have revealed that the brain regions known to be implicated in the processing of facial affect–such as the posterior superior temporal sulcus (pSTS), the amygdala and the insula–respond more to dynamic than to static emotional expressions (e.g., Haxby et al., 2000; LaBar et al., 2003; Kilts et al., 2003). Also, most importantly, authors reported cases of neurologically affected individuals that were incapable of recognizing static facial expressions but could recognize dynamic expressions (Humphreys et al., 1993; Adolphs et al., 2003). Thus, it is more appropriate, in research dealing with the recognition of real-life facial expressions, to use dynamic stimuli because (1) dynamic facial expressions are encountered in everyday life and (2) dynamic and static facial expressions are processed differently. This issue is of particular interest in the investigation of audio-visual emotion processing, where the integration of dynamic prosody variation with still pictures results in very low ecologically relevant material. Although integration effects have undoubtedly been observed for voices paired with static faces (de Gelder and Vroomen, 2000), it is clear that such integrative processing would be much stronger when dynamic faces are used (Campanella and Belin, 2007; Ghazanfar et al., 2005; Schweinberger et al., 2007; Sugihara et al., 2006). For example, a recent study on person identification provides compelling evidence that the presentation of time-synchronized articulating faces influenced more strongly the identification of familiar voices than when accompanied by static faces (Schweinberger et al., 2007). Another clear illustration of this point comes from studies of audio-visual speech perception, and in particular the McGurck effect, where clips of faces in movement, but not still photograph, influence speech perception (McGurk and MacDonald, 1976; Campanella and Belin, 2007). Another limitation of the aforementioned studies

on bimodal emotion perception is that auditory affective material consisted of speech prosody (words, sentences) spoken with various emotional tones, with the possibility of affective tone of speech (emotional prosody) interacting with the affective value that may be carried by its semantic content (Scherer et al., 1984).

The present study thus attempts to asses the multisensory nature of the perception of affect expressions using ecologically relevant material that approximates real-life conditions of social communication. To do so, we used newly standardized and validated sets of dynamic visual (Simon et al., 2008) and nonverbal vocal (Belin et al., in press) clips of emotional expressions (Fig. 1). In Experiment 1, subjects were required to discriminate between fear and disgust affect expressions either displayed auditorily, visually or audio-visually, in a congruent (the same expressions in the two modalities) or incongruent way (different expressions in the two modalities). This method allows us to investigate whether the presentation of bimodal congruent stimuli improves the subject's performance and which modality dominates in a conflicting situation. Since we observed a visual dominance in the perception of multisensory affects, we also included a condition in which the reliability of the visual stimuli was decreased to challenge this dominance. To test if multisensory interaction in the processing of affective expression is a mandatory process, we conducted a second experiment with the same stimuli as those used in the first but with the explicit instruction to focus attention to only one sensory modality at a time. If multisensory interaction of affective information is an automatic process, it should take place even if the participant's attention is focused on only one modality (de Gelder and Vroomen, 2000; Massaro and Egan, 1996). Because the influence of a concurrent signal increases in situations where the reliability of a sensory channel is reduced (Ross et al., 2007), such as face perception in the dark or voice recognition in a noisy environment, the reliability of the visual and the auditory signals was manipulated.

The originality of this study resides in the use of highly ecological sets of stimuli in two experiments (the first with unconstrained and the second with constrained focus of attention) where the reliability of the sensory targets were individually challenged in order to shed light onto the mechanisms at play in the multisensory processing of affect expression.

2. Results

2.1. *Experiment 1*

Correct discriminations (Fig. 2) were analysed by submitting Inverse Efficiency (IE) scores (see Data analysis section) to a 2 (Noises: Noisy or Noiseless) × 2 (Emotions: Fear or Disgust) × 3 (Stimuli: Visual, Auditory or Bimodal Congruent) repeated measures ANOVA. As expected, we obtained a main effect of the factor "Noises" ($F=16$, $p\leq.001$) showing better performance with noiseless than with noisy stimuli. Of great interest for the present study, we also obtained a main effect of the factor "Stimuli" ($F=8$, $p\leq.002$) demonstrating lower IE scores with bimodal stimuli compared to visual ($p\leq.01$) and auditory ($p\leq.005$) stimuli alone. The ANOVA also revealed a significant interaction between the factors "Noises" and "Stimuli" ($F=22$, $p\leq$10E-5). Post-hoc analyses revealed that bimodal superiority compared to auditory stimuli was present in both conditions of noise but that superiority of bimodal over visual stimuli was only present with noisy stimuli ($p\leq$10E-6). Finally, we also obtained a significant interaction between the factors "Emotions" and "Stimuli" ($F=5$, $p\leq.01$) showing that unimodal auditory fear stimuli were less easily recognized than unimodal auditory disgust stimuli.

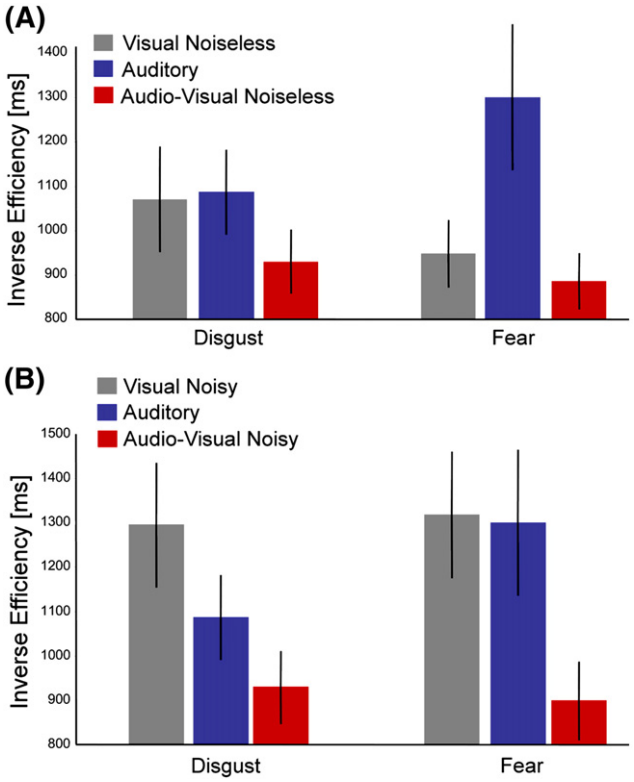

Fig. 2 – Mean IE scores and standard errors obtained in Experiment 1 for unimodal stimuli (blue for auditory, grey for visual) and congruent audio-visual stimuli (red) for both emotion expressions. The figure displays the results obtained with noiseless visual stimuli (panel A) for noisy visual stimuli (panel B). IE scores are obtained by dividing RTs by correct response rates, thus eliminating any potential speed/accuracy tradeoff effects in the data; the lower the score, the more efficient the performance (Spence et al., 2001; Roder et al., 2007). The best performance was obtained in the bimodal conditions for both emotion expressions, especially with noisy visual stimuli.

To further test the presence of multisensory gain in reaction times (RTs) data, we investigated if the redundancy gain obtained for RTs in the bimodal conditions exceeded the statistical facilitation predicted by probability summation using Miller's race model of inequality (Miller, 1982) (see Data analysis section for details). In all cases, we observed violation of the race model prediction over the fastest quantiles of the reaction time distribution, supporting interaction accounts for faster RTs in bimodal than in unimodal conditions of presentation (Fig. 3).

Because there are no "correct" responses with incongruent bimodal stimuli, a tendency to respond either "fear" or "disgust"

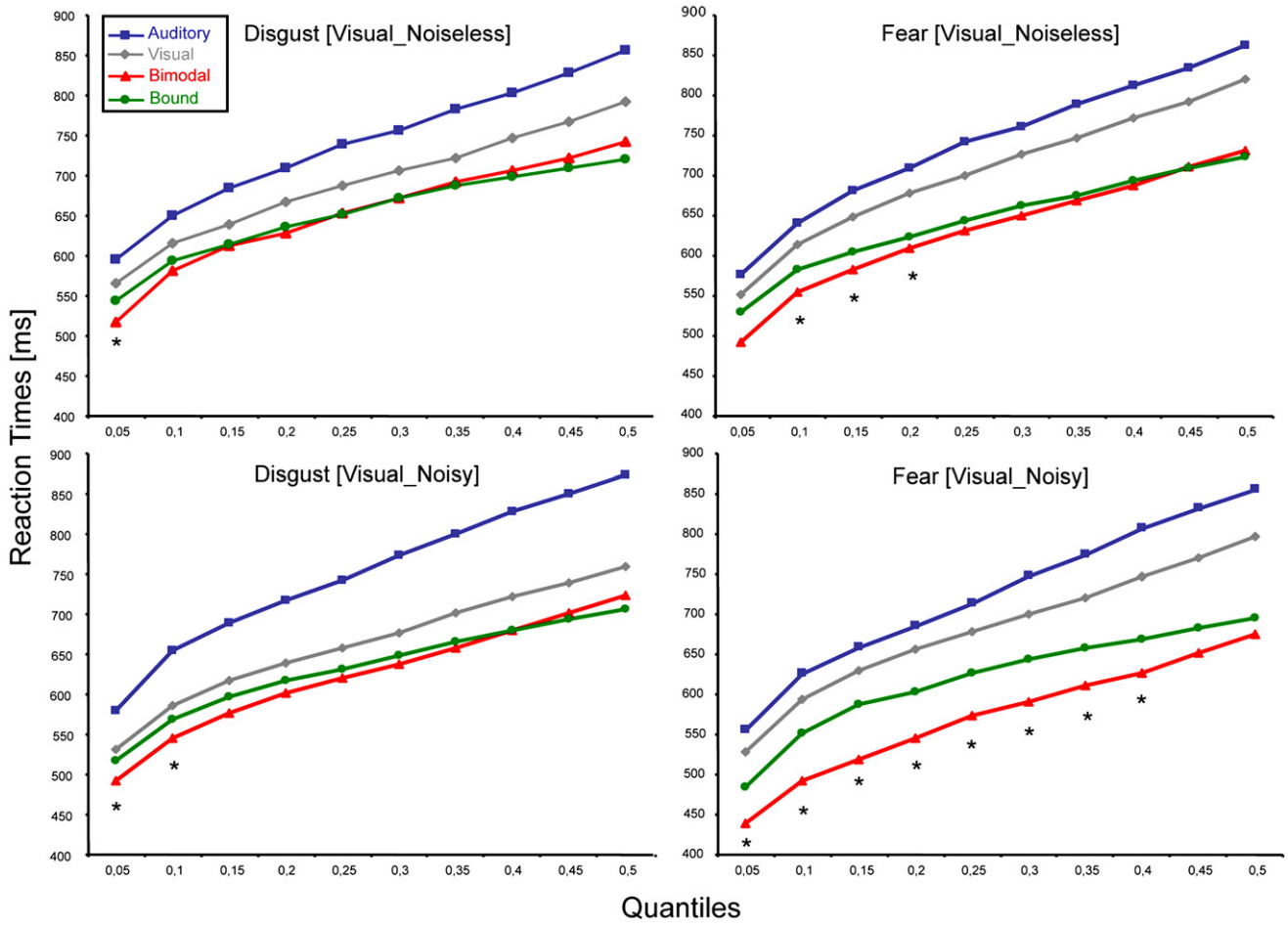

Fig. 3 – Redundancy gain analysis and test for violation of the race model inequality in Experiment 1 (Miller, 1982; Ulrich et al., 2007). The scatter plots illustrate the cumulative probability distributions of RTs (only the first half of the quantiles are displayed) with congruent bimodal stimuli (red triangles) and their unisensory counterparts (blue squares for auditory, grey lozenges for visual), as well as with the race model bound (green dots) computed from the unisensory distributions. RTs were obtained either with noiseless visual stimuli (panel A and B) or with noisy visual stimuli (panel C and D). Bimodal values inferior to the bound indicate violation of the race model and the asterisks refer to statistical significance. In all conditions–but especially for noisy visual stimuli–the race model inequality is significantly violated over the fastest quantiles of the reaction time distribution, supporting interaction accounts.

was estimated by subtracting the proportion of "fear" responses from the proportion of "disgust" responses (pDisgust – pFear) in the four incongruent conditions of stimulation (noiseless fearful face/disgust voice; noisy fearful face/disgust voice; noiseless disgust face/fearful voice; noisy disgust face/fearful voice). The index, which varies between −1 (subject always responded "fear") to 1 (subject always responded "disgust"), was analysed by means of a two way 2×2 ANOVA with "Visual Emotion" (fear or disgust) and "Noises" (noiseless or noisy) as within-subject factors. We observed a highly significant interaction effect ($F=25$, $p\leq.0002$) showing that the index was more positive with "normal visual disgust/auditory fear" stimuli than with "noisy visual disgust/auditory fear" stimuli ($p\leq.01$) and that the index was significantly more negative with "normal visual fear/auditory disgust" stimuli than with "noisy visual fear/auditory disgust" stimuli ($p\leq.003$). In other words, it indicates that with noiseless visual stimuli, the participants oriented their responses toward the visual modality whereas with noisy visual stimuli, the participants have a tendency to categorise the stimuli in the affect class expressed in the auditory modality (Fig. 4).

The fact that emotions in both modalities are equally well recognized in noiseless condition of presentation does not exclude a difference between visual and auditory stimuli in perceived emotional intensity. It is thus possible that emotions expressed via the visual modality are perceived as more intense, which may possibly underlie the visual dominance observed in the incongruent condition with noiseless stimuli. To test this assumption, 28 naïve subjects (18 females; mean age 28, S.D. 7; 2 left-handed) rated each facial and vocal expressions with respect to the intensity of disgust and fear on a 100-point visual-analogue scale. Participants were instructed to "rate the intensity of each of the two emotions in the clip from 0 = not at all to 100 = the most intense possible". Each clip was therefore rated on the two emotional categories (see Table 1). We carried out a 2 (Emotion: Disgust and fear)×2 (Modality: Auditory, Visual) ANOVA on the recorded data. Results showed a significant effect of the factor "Emotion" ($F=58$, $p\leq$10E-6) revealing

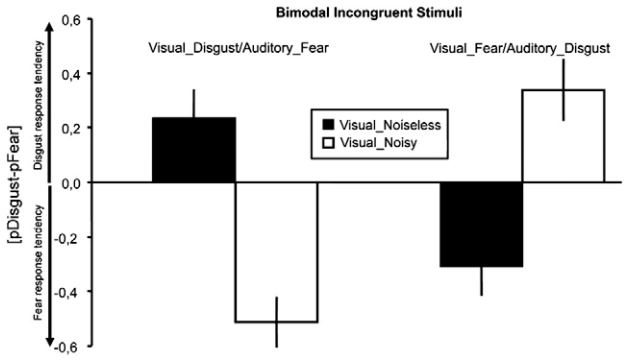

Fig. 4 – Bias to respond either "fear" or "disgust" in incongruent bimodal conditions was estimated by subtracting the proportion of "fear" responses from the proportion of "disgust" responses (pDisgust–pFear) in Experiment 1. Participants tend to report the emotion expressed in the visual modality with intact stimuli (black histograms). However, they categorise more readily the stimuli as the affect expressed in the auditory modality with noisy visual stimuli (white histograms).

that "Fear" expressions were rated significantly more intense than "Disgust" expressions. We also observed a significant effect of the factor "Modality" showing that emotions expressed in the visual modality were rated significantly more intense than emotions expressed in the auditory modality ($F=32$, $p\leq 5E\text{-}6$). This last result thus confirms the visual dominance observed in our tasks and may explain why in incongruent situation with noiseless visual stimuli, participants choose to orient their response toward the visual modality (Fig. 4).

2.2. *Experiment 2*

IE scores obtained in Experiment 2 (Fig. 5) were analysed with a 2 (Attention: attend visual or attend auditory) × 2 (Noises: noisy or noiseless) × 2 (Emotions: fear or disgust) × 3 (Stimuli: unimodal, congruent bimodal, incongruent bimodal) ANOVA. We first observed a significant main effect of the factor "Attention" ($F=5$, $p\leq .04$) showing better performance when participants attended the visual modality than the auditory one. As expected, we also obtained a significant main effect of the factor "Noises" ($F=46$, $p\leq 10E\text{-}5$) demonstrating the disrupting role of the addition of noise to the stimuli. Importantly, we obtained a powerful main effect of the factor "Stimuli" ($F=43$, $p\leq 10E\text{-}7$), demonstrating better performance (lower IE scores) with bimodal congruent stimuli compared to unimodal ones ($p\leq .02$) and worse performance with bimodal incongruent stimuli compared to unimodal ($p\leq 10E\text{-}5$) or congruent bimodal ones ($p\leq 10E\text{-}7$).

The ANOVA also showed a significant interaction between the factor "Attention" and "Emotions" ($F=10$, $p\leq .05$), which resulted from inferior performances when participants discriminated fear stimuli in the auditory modality than in the visual one. An interaction between the factors "Attention" and "Stimuli" was also observed ($F=6$, $p\leq .006$). This effect indicates that performance decreased significantly more in the bimodal incongruent condition of stimulation when the subjects attended to the auditory modality than to the visual modality ($p\leq .0005$). Interestingly, we also obtained an interaction between the factors "Noises" and "Stimuli" ($F=8$, $p\leq .001$) showing that the irrelevant modality especially influenced the noisy targets.

We also examined the possibility that our results are at least partly due to carry-over effects (see e.g., Wylie et al., 2004). For example, the auditory influence observed when subjects were instructed to attend only to the visual modality may have occurred only in vision-only blocks that immediately followed audition-only blocks. To test this hypothesis, we examined whether the influence of the irrelevant modality was also present in the first half of testing, when participants had not already been instructed to attend to the irrelevant modality. We thus carried out another 2 (Modality attended first: Auditory, Visual) × 2 (Attention: Auditory, Visual) × 2 (Emotions: fear or disgust) × 3 (Stimuli: unimodal, congruent bimodal, incongruent bimodal) ANOVAs on IE scores. Noisy stimuli were not included in the analyses since, as explained in Experimental procedures, they always followed noiseless stimuli. We obtained a main effect of the factor "Attention", showing better results when the visual modality was attended compared to the auditory modality ($F=6$, $p\leq .027$). As expected, we also obtained a main effect of the factor "Stimuli" ($F=52$, $p\leq 10E\text{-}7$), indicating better results in the bimodal congruent condition than in the bimodal incongruent condition. However, no interaction with the factor "Modality attended first" was found to be significant. This clearly demonstrates that our results are independent of carry-over effects and thus support the idea that multisensory interactions in the processing of affect expressions are automatic.

3. Discussion

In the present study, participants were required to discriminate between "fear" and "disgust" emotion expressions displayed either auditorily, visually, or audio-visually via short dynamic facial and non-linguistic vocal clips. Our results provide compelling evidence for the multisensory nature of emotion processing and extend further our comprehension of the mechanisms at play in the integration of audio-visual expression of affect.

In Experiment 1, when participants were instructed to process emotional information in both modalities, they showed improved performance (lower IE scores) with congruent bimodal stimuli compared to either unimodal condition (Fig. 2). Moreover, we observed that RTs in congruent bimodal conditions exceeded the race model estimation (Miller, 1982) over the fastest quantiles of the reaction time distribution, providing evidence that information from the visual and auditory sensory modalities truly interacted to produce the RT facilitation. This evidence reinforces the notion of an intrinsic multisensory nature of affective expression recognition. It is worth noting that these effects are greater in the noisy conditions, as predicted by the "inverse effectiveness" principle—which states that the result of multisensory integration is inversely proportional to the effectiveness of the relevant stimuli (Stein and Meredith, 1993). This result highlight the substantial advantage of relying on multisensory emotion processing in everyday life situations

Table 1 – Intensity rating by expression category

Expression perceived	Target expression			
	Visual_Fear	Visual_Disgust	Auditory_Fear	Auditory_Disgust
Fear	80.3 (14.9)	1.4 (5.9)	70.2 (15.3)	1.2 (2.4)
Disgust	0.7 (2.1)	61.8 (20.1)	1.2 (3.3)	55.5 (22.1)

Naïve subjects rated the intensity of each facial and vocal expression of disgust and fear on a 100-point visual-analogue scale, with 0 = "not at all" to 100 = "most intense".

where the reliability of one of the sensory channels can be reduced, like in darkness or in a noisy environment.

We also used a condition where participants faced incongruent audio-visual stimuli (different affect expressions in both modalities). When incongruent bimodal pairs contained noiseless visual stimuli, the participants oriented their responses more often toward the visual modality. This result suggests some kind of visual dominance or "visual capture" in the perception of affective expression. This may be related to the fact that the perceived intensity of our emotional clips delivered visually were judged as more intense than the ones delivered via the auditory modality (see Table 1). However, when participants were presented with audio-visual stimuli composed of noisy visual stimuli, the participants categorised more often the affect expressed in the auditory modality (Fig. 4). These results demonstrate that visual dominance in affect perception does not occur in a rigid, hardwired manner, but follows flexible situation-dependent rules that allow information to be combined with maximal efficacy (Ernst and Bulthoff, 2004). Our results of audio-visual integration in affective expression recognition thus obey a perceptual framework where the degree of relative uncertainty in different sensory domains dictates whether the overall perceptual output is derived from one modality rather than another (Ernst and Bulthoff, 2004). This process of giving more weight to the less ambiguous modality would clearly offer ecological benefits since decoding emotion expression is often associated with multiple sources of sensory information with their

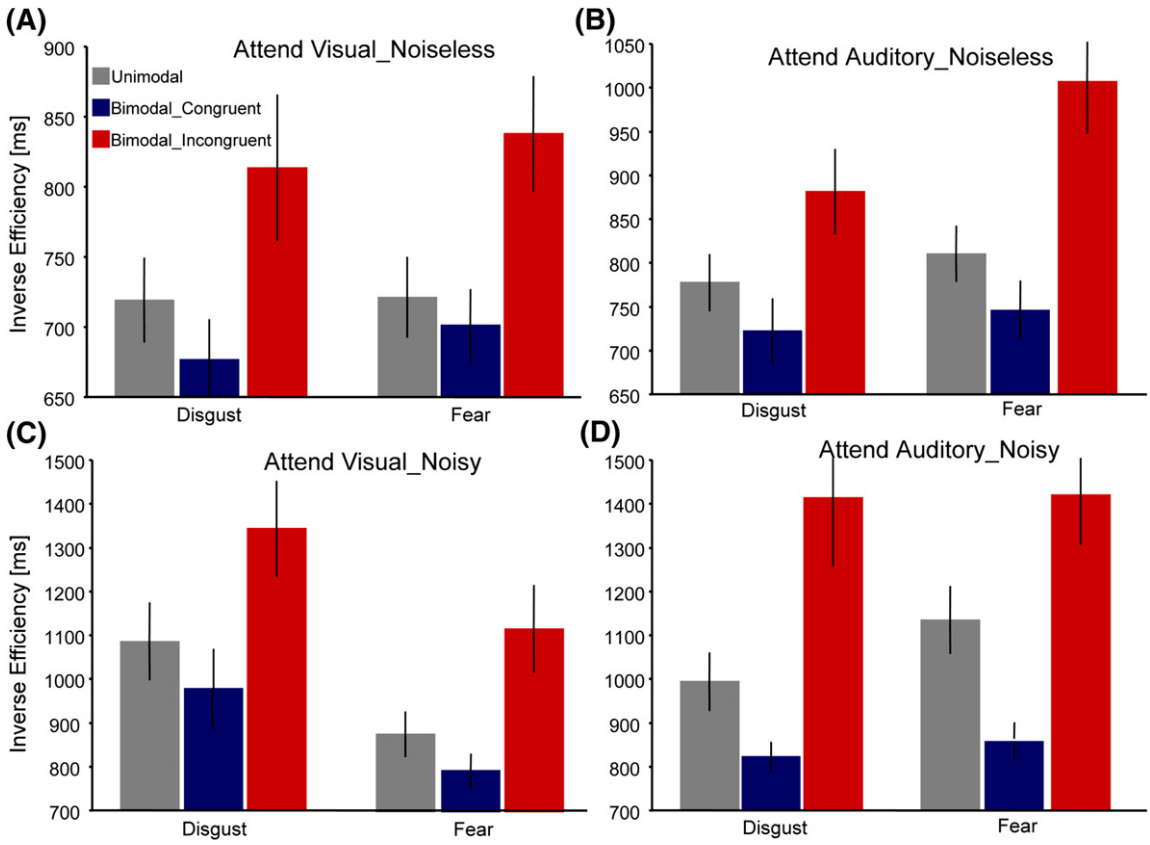

Fig. 5 – Mean IE scores and standard errors obtained in Experiment 2 for unimodal stimuli (grey), congruent bimodal stimuli (blue) and incongruent bimodal stimuli (red) for both emotion expressions. Panels A and C display performance when participants were instructed to discriminate the emotion expressed visually either with (A) or without noise (C). Panels B and D display performance when participants were instructed to discriminate the emotion expressed auditorily either with (D) or without noise (B). Congruent bimodal stimuli were better recognized than unimodal stimuli, and unimodal stimuli were better recognized than incongruent bimodal stimuli—this was especially true for noisy target stimuli.

reliability varying from place to place. For example, although facial expression could provide the most reliable information to interpret an emotion in daylight face-to-face interaction, at night, vocal information is often more useful.

In Experiment 2, we explicitly requested the participants to focus their attention to only one sensory modality at a time while completely disregarding the irrelevant modality (Fig. 5). This procedure was used to investigate if multisensory interactions in the processing of emotion expressions originated in an automatic, mandatory perceptual process rather than in a post-perceptual judgement (i.e. a decision taken by the perceiver after both inputs were processed) (de Gelder and Vroomen, 2000). The underlying idea is that if audio-visual integration operates in an automatic fashion, multisensory influence should occur even if the participants only focus their attention toward one single modality. Results clearly demonstrated performance increase when the non-target modality was emotionally congruent with the attended modality and a decrease in performance when the concurrent modality was incongruent with the expression displayed in the target sensory channel. These effects attest to the automaticity of multisensory interactions in the perception of emotion expressions. This situation could be related to a kind of "emotional Stroop" where the automatic nature of integration in bimodal emotion induces an inability for participants to focus on only one modality, even if instructed to do so. Other evidence for this hypothesis comes from studies showing evidence of the limited role of awareness in bimodal perception, by demonstrating nonconscious influence of a non-recognized facial expression presented in the blind hemifield of hemianopic patients in the categorisation of affective tone in voice (de Gelder et al., 2005, 2002). The automaticity hypothesis is further corroborated by electrophysiological findings showing that face/voice emotion interactions take place at early perceptual stages rather than during late decisional stages (de Gelder et al., 1999; Ethofer et al., 2006b; Pourtois et al., 2000) and occur irrespective of attentional resources deployment (Vroomen et al., 2001). Moreover, we also observed that the influence of the irrelevant modality was especially strong when delivered with noisy sensory targets. This is again in accord with the "inverse effectiveness" principle (Stein and Meredith, 1993) and could be linked to the proposition mentioned above that bimodal emotional information is integrated by attributing less weight to uncertain sensory information. In the present situation, when the attended modality was less reliable, participants automatically attributed more weight to the irrelevant sensory modality in their processing of bimodal emotional expression.

More generally, we observed better global performance when participants attended to the visual modality compared to the auditory one. This result lends support to the idea that, at least for the case of "fear" or "disgust" discrimination in the present study, the visual modality transmits the most salient information. This could also be related to the fact that the decrease of performance obtained in the bimodal incongruent condition of stimulation is greater when the subjects attend to the auditory modality than to the visual one (Fig. 5). In other words, this reflects a more disruptive power of the irrelevant visual expressions on auditory judgement than irrelevant auditory expression on visual judgement. This result is certainly in line with the enhanced salience of the visual stimuli which induces more disruption than the auditory ones when in concurrence with the sensory target. Such observations should be linked to our observation that in an unconstrained situation (Experiment 1), without noise, participants preferentially categorise incongruent audio-visual situation toward the emotion expressed in the visual modality.

Recent neuroimaging studies have suggested that the combination of affect expressions as observed in the present study is likely to be implemented in specific anatomical convergence zones, probably between the visual and auditory modalities such as in the amygdala and the middle temporal gyrus in the left hemisphere (Ethofer et al., 2006b). Because the use of dynamic visual stimuli and non-linguistic vocal expressions is a typical situation of our environment and thus possesses a higher ecological value, we believe that such material may provide great insight into the neural network involved in audio-visual coupling of emotion (Kreifelts et al., 2007). Also, the brevity of our clips (500 ms) makes them suitable for comparison between electrophysiological and neuroimaging techniques. Moreover, the use of dynamic visual and non-verbal vocal emotional stimuli may provide more ecologically valid understandings of emotion processing disorders postulated in some psychopathologies (autism, schizophrenia). In particular, future studies should address the possibility that specific deficits in emotional integrative processes could exist in the absence of any abnormality of unimodal face or voice processing (Delbeuck et al., 2007; de Gelder et al., 2003).

4. Experimental procedures

4.1. *Participants*

Sixteen paid volunteers participated in Experiment 1 (8 females; mean age 26, S.D. 9; all right-handed). The same number of subjects participated in Experiment 2 (8 females; mean age 25, S.D. 10; all right-handed with the exception of 1 left-handed female). Four subjects took part in the two experiments. All participants were without any recorded history of neurological or psychiatric problems, reported normal hearing and normal or corrected-to-normal vision and did not use psychotropic medication at the time of testing. The study was approved by the local ethics committee and all subjects gave their written informed consent prior to inclusion in the study.

4.2. *Stimuli*

We decided to focus on "Fear" and "Disgust" emotion expressions because from an evolutionary perspective, both emotions share a common goal which is to alarm the observer in situation of direct threat and thus may be more important for survival than other basic emotion such as happiness. Indeed, in the multisensory domain, Dolan et al. suggested that the rapid integration across modalities is not as automatic for happy expressions as it is for fear signals (Dolan et al., 2001). Furthermore, despite the fact that both emotions belong to the category of 'negative affect', disgust and fear expressions can be clearly distinguished from one another (Belin et al., in press;

Ekman and Friesen, 1976; Simon et al., 2008) and serve as a model to study the existence of separate neural substrates underlying the processing of individual emotion expressions (Calder et al., 2001).

The visual stimuli came from a standardized set of 64 dynamic color stimuli of prototypical facial expressions of pain, the so-called basic emotions (Happiness, Anger, Disgust, Fear, Surprise, Sadness) and neutrality (Simon et al., 2008). For the purpose of this study, we selected the 3 male and the 3 female actors who produced the most unambiguous facial expressions of "fear" and "disgust" emotions. The facial expressions were "prototypical" insofar as they possessed the key features (identified using the Facial Action Coding System: FACS) identified by Ekman and Friesen (1976) as being representative of everyday facial expressions. The same male and female actors portrayed the two emotions. The use of several different stimuli (3 males/3 females) for each discrete emotion avoided potential confounds with actor's identity or sex in the "fear–disgust" discrimination tasks. The original video clips which lasted 1 s were cut with Adobe Premiere 6.5 (Adobe Systems Inc.) to obtain 500 ms-clips. All the clips started with a neutral frame before initiation of the facial movement. These new clips were then resized (image size: 350×430 pixels, frame rate=29.97 frames/s) in Adobe After-Effects 7.0 (Adobe Systems Inc.). A recognition rate of approximately 85% was obtained in the unimodal visual conditions for both emotions, attesting that the stimuli were reliably discriminated by our volunteers.

The auditory stimuli came from the "Montreal affective voices", a standardized set of emotional vocal expressions designed for research on auditory affective processing with the avoidance of potential confound from linguistic content (Belin et al., in press). The set consisted of 70 short, non-linguistic interjections (the vowel /a/) expressing basic emotions (anger, disgust, fear, happiness, sadness, and surprise) plus a neutral expression recorded in ten different actors. The set is freely available at ftp://132.204.126.245:21021. Among this set, we selected "Fear" and "Disgust" vocalizations portrayed by the 3 male and the 3 female actors producing the stimuli with the highest level of distinctiveness. Again, each actor portrayed both emotions. The selected affective interjections were then edited in short meaningful segments of 500 ms (rise/fall time 10 ms) and normalized peak value (90%) using Adobe Audition 2.0 (Adobe Systems Inc.). A recognition rate of approximately 85% for "disgust" and 80% for "fear" was obtained in the unimodal auditory conditions, attesting that the stimuli were reliably discriminated by our volunteers.

The bimodal stimuli were obtained by simultaneously presenting visual and auditory clips. The matching could either be "congruent", with audio and video tracks portraying the same emotion (i.e., Fearful Face/Fearful Voice), or "incongruent", with audio and video tracks portraying different emotions (i.e., fearful Face/Disgust voice). Each actor in the visual clips was assigned with a specific "voice" for the two emotions throughout the experiment, either in the congruent or incongruent conditions.

4.3. Procedure

The participants sat in a silent and darkened room. Stimuli were displayed using Presentation software (Neurobehavioral Systems, Inc.) running on a Dell XPS laptop computer with Windows XP operating system. Behavioural responses and reaction times were recorded using Logview, a custom-made software specifically designed to analyse behavioural data obtained with Presentation. Visual stimuli (width=10° of visual angle; height=12.5° of visual angle) were presented in the centre of the screen over a constant grey background. The viewing distance was maintained constant at 60 cm by using a chinrest. Auditory stimuli were presented binaurally trough headphones (Philips HJ030) at a self-adjusted comfortable level.

4.3.1. *Experiment 1*

The participants were required to discriminate fear and disgust emotion expression stimuli presented only auditorily, only visually, or audio-visually. Audio-visual stimuli could be either incongruent (different expression in the two modalities) or congruent (the same expression in the two modalities). The participants were required to respond as quickly and as accurately as possible in a forced-choice discrimination paradigm, by pressing the appropriate keyboard keys. The response keys were counterbalanced across subjects. The subjects were instructed to identify the portrayed emotion as either "fear" or "disgust" based on their initial reaction to the stimuli, even if they perceived a conflict between the senses. This paradigm was used to investigate eventual modality-dominance in conflicting situations. The participants were presented with a total of 480 stimuli (2 [Emotions]×4 [Conditions]×60 [stimuli; 10 by each actor (6)]) randomly interleaved in 5 separate blocks of 96 stimuli. Since preliminary results demonstrated a visual dominance in the discrimination of incongruent audio-visual stimuli, we decreased the reliability of the visual target to diminish the visual dominance in the incongruent condition. To do so, we adjusted the visual signal-to-noise ratio of the video clips to lower the accurate discrimination of the stimuli presented only visually to a level of approximately 70% correct. The adjustment was carried out individually using the QUEST procedure (Watson and Pelli, 1983) implemented in the Psychtoolbox (Brainard, 1997; Pelli, 1997) for Matlab (The MathWorks, Inc.). During the adjustment phase, the participants viewed 60 video clips (30 "fear" and "30" disgust) to which white Gaussian noise was added in each of the three color channels. The participants were then presented with a total of 360 "noisy" visual stimuli (2 [Emotions]×3 [Conditions: Visual alone, audio-visual congruent, audio-visual incongruent]×60 [stimuli; 10 by each actor (6)]) randomly interleaved in 5 separate blocks of 72 stimuli. The noisy visual stimuli were always presented after the noiseless visual stimuli because pre-tests showed that performance increased rapidly over the first 100 trials or so before reaching an asymptote and we wanted to make sure that the noise adjustment was performed after the asymptote was attained.

Each stimulus presentation was followed by a 2000 ms grey background (the response period), then a central cross appeared for 500 to 1500 ms (random duration) prior to the next stimulus (Mean ISI 3000 ms; range 2500–3500 ms). Trials to which participants did not respond were considered as omissions and were discarded. Breaks were encouraged between blocks to maintain a high concentration level and prevent mental fatigue.

4.3.2. *Experiment 2*

In experiment 2, participants were also submitted to a "fear" vs "disgust" discrimination task but were explicitly requested to focus their attention to only one sensory modality at a time while completely disregarding the irrelevant modality. We used this procedure to test if multisensory interaction in the processing of affective expression is a mandatory process which takes place even if the subject's attention is focused on a single modality (de Gelder and Vroomen, 2000). In the visual condition, participants saw stimuli displayed only visually and audio-visually; audio-visual stimuli were either congruent or incongruent. In the auditory condition, participants were presented stimuli only auditorily and audio-visually; audio-visual stimuli were either congruent or incongruent. In both conditions, we presented noiseless stimuli in a first condition and noisy stimuli in a second condition. The visual and auditory signals were degraded to test if multisensory influence is strengthened at a low level of certainty, as predicted by the "inverse effectiveness" principle (Stein and Meredith, 1993). The noisy stimuli were obtained following the method described above (for auditory stimuli, unidimensional white Gaussian noise was added to the signal) with a discrimination rate of stimuli presented only in one modality approximately set to 70% correct.

Presentation of noisy stimuli always followed the presentation of noiseless stimuli conditions to ensure that performance had stabilized to its asymptotic level. The order of the auditory and visual conditions was counterbalanced across subjects. In each of the four conditions, a total of 360 stimuli (2 [Emotions] × 3 [Conditions] × 60 [stimuli; 10 by each actor (6)]) were delivered in five separate blocks of 72 stimuli. The stimulus-delivery procedure was the same as in Experiment 1.

4.4. *Data analysis*

To take both response speed and accuracy into account, Inverse Efficiency (IE) scores were derived by dividing response times (150–1500 ms post-stimulus) by correct response rates separately for each condition (thus, a higher value indicates worse performance). IE scores, which constitute a standard approach to combine RT and accuracy measures of performance (Townsend and Ashby, 1978, 1983), can be considered as "corrected reaction times" that discount possible criterion shift or speed/accuracy tradeoffs (Spence et al., 2001; Roder et al., 2007). IE scores were submitted to repeated measures analysis of variance (ANOVARM). Based on significant *F*-values, Bonferroni post-hoc analyses were performed when appropriate. Accuracy and RTs data are illustrated in three supplementary figures (Figs. 6, 7 and 8). Analysis of "redundancy gain" and violation of the race model inequality (Miller, 1982) were tested on RTs of Experiments 1 and 2. In the race model, faster RTs obtained in bimodal situations are produced because the two unimodal stimuli set up a race for the control of response and the faster process wins, that is, there is no need to postulate neural interaction between the two stimuli. However, if RTs obtained in the bimodal condition are better than the predictions of the race model, this provides evidence that information from the visual and auditory sensory modalities interacted to produce the RT facilitation. Analyses of violation of the race model inequality were carried out using the RMITest software which implements the algorithm described in Ulrich, Miller and Schröter (2007). The algorithm estimates the cumulative probability distributions of RT in the two unimodal conditions and the bimodal condition, and tests whether redundant-targets (the bimodal condition) RTs are significantly faster than would be predicted by a race model (with *t*-tests).

The individual variability of the main results in both experiments is presented as supporting material in supplementary Figs. 9, 10, and 11.

Acknowledgments

We thank Stephane Denis for his help with the experimental setup. OC is a postdoctoral researcher at the Belgian National Funds for Scientific Research (F.R.S.-FNRS). This work was supported by the FRSQ Rehabilitation network (REPAR to OC), the Canada Research Chair Program (ML, FL) and the Natural Sciences and Engineering Research Council of Canada (ML, FL, FG).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.brainres.2008.04.023.

REFERENCES

Adolphs, R., Tranel, D., Damasio, A.R., 2003. Dissociable neural systems for recognizing emotions. Brain Cogn. 52, 61–69.

Belin, P., Fillion-Bilodeau, S., Gosselin, F., in press. The "Montreal Affective Voices": a validated set of nonverbal affect bursts for research on auditory affective processing. Behav Res Methods.

Brainard, D.H., 1997. The psychophysics toolbox. Spatial Vision 10, 433–436.

Calder, A.J., Lawrence, A.D., Young, A.W., 2001. Neuropsychology of fear and loathing. Nat. Rev., Neurosci. 2, 352–363.

Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends Cogn. Sci. 11, 535–543.

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

de Gelder, B., Bocker, K.B., 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.

de Gelder, B., Pourtois, G., Weiskrantz, L., 2002. Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures. Proc. Natl. Acad. Sci. U. S. A 99, 4121–4126.

de Gelder, B., Vroomen, J., Annen, L., Masthof, E., Hodiamont, P., 2003. Audio-visual integration in schizophrenia. Schizophr. Res. 59, 211–218.

de Gelder, B., Morris, J.S., Dolan, R.J., 2005. Unconscious fear influences emotional awareness of faces and voices. Proc. Natl. Acad. Sci. U. S. A 102, 18682–18687.

Delbeuck, X., Collette, F., Van der Linden, M., 2007. Is Alzheimer's disease a disconnection syndrome? Evidence from a crossmodal audio-visual illusory experiment Neuropsychologia 45, 3315–3323.

Dolan, R.J., Morris, J.S., de Gelder, B., 2001. Crossmodal binding of fear in voice and face. Proc. Natl. Acad. Sci. U. S. A 98, 10006–10010.
Ekman, P., Friesen, W.V., 1976. Pictures of facial affect. Consulting Psychologist Press, Palo Alto (CA).
Ernst, M.O., Bulthoff, H.H., 2004. Merging the senses into a robust percept. Trends Cogn Sci. 8, 162–169.
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.
Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K., 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J. Neurosci. 25, 5004–5012.
Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cogn. Sci. 4, 223–233.
Humphreys, G.W., Donnelly, N., Riddoch, M.J., 1993. Expression is computed separately from facial identity, and it is computed separately for moving and static faces: neuropsychological evidence. Neuropsychologia 31, 173–181.
Kilts, C.D., Egan, G., Gideon, D.A., Ely, T.D., Hoffman, J.M., 2003. Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. Neuroimage 18, 156–168.
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.
LaBar, K.S., Crupain, M.J., Voyvodic, J.T., McCarthy, G., 2003. Dynamic perception of facial affect and identity in the human brain. Cereb. Cortex 13, 1023–1033.
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.
Massaro, D.W., Egan, P.B., 1996. Perceiving affect from the voice and the face. Psychon. Bull. Rev. 3, 215–221.
Miller, J., 1982. Divided attention: evidence for coactivation with redundant signals. Cognit. Psychol. 14, 247–279.
Pelli, D.G., 1997. The videotoolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision 10, 437–442.
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.
Roder, B., Kusmierek, A., Spence, C., Schicke, T., 2007. Developmental vision determines the reference frame for the multisensory control of action. Proc. Natl. Acad. Sci. U. S. A 104, 4753–4758.
Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., Foxe, J.J., 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex 17, 1147–1153.
Scherer, K., Ladd, D., Silverman, K., 1984. Vocal cues to speaker affect: testing two models. The J. Acoust. Soc. Am. 76, 1346–1356.
Schweinberger, S.R., Robertson, D., Kaufmann, J.M., 2007. Hearing facial identities. Q. J. Exp. Psychol. 60, 1446–1456.
Simon, D., Craig, K.D., Gosselin, F., Belin, P., Rainville, P., 2008. Recognition and discrimination of prototypical dynamic expressions of pain and emotions. Pain 135, 55–64.
Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT, Cambridge (MA).
Spence, C., Shore, D.I., Gazzaniga, M.S., Soto-Faraco, S., Kingstone, A., 2001. Failure to remap visuotactile space across the midline in the split-brain. Can. J. Exp. Psychol. 55, 133–140.
Sugihara, T., Diltz, M.D., Averbeck, B.B., Romanski, L.M., 2006. Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. J. Neurosci. 26, 11138–11147.
Townsend, J.T., Ashby, F.G., 1978. Methods of modeling capacity in simple processing systems. In: Castellan, N.J, Restle, F. (Eds.), Cognitive Theory, vol. 3. Erlbaum, Hillsdale, NJ, pp. 199–239.
Townsend, J.T., Ashby, F.G., 1983. Stochastic Modelling of Elementary Psychological Processes. Cambridge University Press, (NY).
Ulrich, R., Miller, J., Schroter, H., 2007. Testing the race model inequality: an algorithm and computer programs. Behav. Res. Methods 39, 291–302.
Vroomen, J., Driver, J., de Gelder, 2001. Is cross-modal integration of emotional expressions independent of attentional resources? Cogn. Affect. Behav. Neurosci. 1, 382–387.
Watson, A.B., Pelli, D.G., 1983. QUEST: a Bayesian adaptive psychometric method. Percept. Psychophys. 33, 113–120.
Wylie, G.R., Javitt, D.C., Foxe, J.J., 2004. Don't think of a white bear: an fMRI investigation of the effects of sequential instructional sets on cortical activity in a task-switching paradigm. Hum. Brain Mapp. 21 (4), 279–297.