Integration of face and voice during emotion perception: is there anything gained for the perceptual system beyond stimulus modality redundancy?

Gilles Pourtois & Monica Dhar

Department of Experimental-Clinical and Health Psychology, Ghent University, Ghent, Belgium

Corresponding author:

Gilles Pourtois

Department of Experimental-Clinical and Health Psychology

Ghent University

Henri Dunantlaan 2

9000 Ghent

Belgium

Phone: +32 9 264 9144

Email: gilles.pourtois@ugent.be

Summary

In this chapter, we review empirical data and theoretical models which have been put forward in the affective science literature to account for the perception of emotions, when this process is simultaneously accomplished by sight and hearing. The visual component is provided by the face configuration that undergoes some geometric changes, which in turn lead to different and discrete emotion facial expressions. The auditory component is provided by the voice and its changes in pitch, duration and/or intensity leading to different affective tones of voice. Face-voice integration during emotion perception occurs when affective information conveyed by the two sensory modalities is integrated into a unified percept, or multisensory object. Although one may assume that the rapid and mandatory combination of multiple or complementary affective cues is adaptive (i.e. it likely reduces the effects of adverse factors like drifts or intrinsic noise), the central nervous system must however show some selectivity regarding which inputs from separate senses may eventually combine, as compared with merely redundant emotion signals. Indeed, not all spatial or temporal coincidences or co-occurrences lead to the perception of unified objects. Interestingly, results of behavioral studies confirm this conjecture, and indicate that the combination of emotional facial expressions with affective prosody leads to the creation of genuinely multisensory emotional objects, which show different properties compared to the combination of an emotional facial expression with another redundant or distracting emotional facial expression, or an emotion written word. Hence, the findings and models reviewed in this chapter suggest that some selectivity can be found in the way visual and auditory information is actually combined during emotion perception. The rapid and automatic paring of an emotional face with an affective voice might present a naturalistic situation in the sense that there is no need for mediation by higher-level cognitive, attentional or linguistic processes, which may be necessary for the efficient decoding of other stimulus categories or multisensory objects.

1 General Introduction

The perception of emotions in conspecifics stands out as one of the most important social skills in human cognition (Darwin, 1871, 1872; Frijda, 1989; A.R. Damasio, 1994). Yet, emotion perception is a complex phenomenon, as affective information is usually processed and conveyed concurrently by multiple sensory channels. Not only the muscles innervating the face swiftly change their configurations to convey or communicate a specific emotional expression, but also the tone of voice as well as the gait or body posture undergo compatible dynamic alterations to promote the efficient expression, communication, as well as decoding by conspecifics, of this emotion. Such emotional multisensory stimulations are the rule rather than the exception in natural environments (De Gelder & Bertelson, 2003), yet still very little is known about the actual brain mechanisms and cognitive processes underlying this remarkable perceptual ability, despite a clear recrudescence of empirical contributions in the field of Multisensory Integration since 10 years (Foxe & Molholm, 2009).

This scarcity is probably related to the fact that one of the dominant paradigms in emotion research is the cognitive approach (Fodor, 1983), which by definition seeks to decompose or break down complex mental functions (including emotion perception) into elementary or basic processes or principles, hence providing a strong analytical (as opposed to integrative) bias. Accordingly, emotion perception has mostly been studied in social or cognitive sciences by looking at mental processes or brain functions concerning a specific, isolated sensory modality (with a clear preference or advantage for vision, relative to the other senses), as opposed to systematic investigations exploring how multiple sensory cues are actually integrated during emotion perception. This observation does not imply that the cognitive approach makes the detailed study of multisensory integration almost impossible, but the modular and analytical perspective embraced by this dominant approach is sometimes hardly compatible with the fact that object-based multisensory perception is by essence a complex, permeable and non-encapsulated phenomenon, which does not necessarily obey to laws of organization that have been put forward primarily to account for modality-specific processes, like visual computations during object recognition or semantic processing during speech

perception for instance. Consistent with this conjecture, many studies have been designed to better characterize mechanisms of emotion perception when the affective information is conveyed predominantly by the face only (Ekman, 1992; Ekman & Friesen, 1976), or by the voice only (Banse & Scherer, 1996; Osullivan, Ekman, Friesen, & Scherer, 1985; K. R. Scherer, Banse, & Wallbott, 2001), but by comparison, few studies have been carried out to look at the nature and extent of perceptual effects when both channels (face and voice) are concurrently conveying important social or emotional signals, and they eventually interact with one another to yield a unified emotion percept (Massaro & Egan, 1996; De Gelder & Vroomen, 2000; Campanella & Belin, 2007). Likewise, studies looking at correlations between emotion face, affective voice and emotion word perception do exist and have been performed in the past (e.g. Borod et al., 1998; Borod et al., 2000), but these valuable studies suggesting the likely existence of amodal perceptual mechanisms during emotion perception do not directly address the question of how multiple sensory cues (e.g. face and voice) may be integrated during emotion perception, and what the resulting emotion percept may be. To address this complex question, another methodology and experimental approach, going beyond correlation methods, is required.

Presumably, an emotion is eventually perceived and experienced as such when these modality specific signals are combined and integrated together to yield a unified multimodal percept. What are the rules or constraints, if any, of this multisensory integration during emotion perception? Does multisensory integration of emotions represent another instance of stimulus redundancy (Marzi, Tassinari, Aglioti, & Lutzemberger, 1986; Miniussi, Girelli, & Marzi, 1998), or is there anything distinctive to this audio-visual integration process? The research presented in this chapter addresses these fundamental questions. Yet, the goal of this chapter is not to study emotions per se (see Vuilleumier & Pourtois, 2007; Brosch, Pourtois, & Sander, 2010 for recent reviews), but rather to shed light on mechanisms allowing the perceptual system to bind together affective information conveyed concurrently by multiple sensory channels. The focus is limited to visual and auditory information, and we do not consider other sensory inputs, like non-linguistic signals (e.g. vocalizations, see Morris, Scott, & Dolan, 1999;

Panksepp, 2005; Scott et al., 1997), emotional body language or gait (de Gelder, 2006), which also accompany expression of emotions.

The visual component is provided by the face configuration that undergoes some changes, which can eventually lead to different discrete emotional facial expressions (Mckelvie, 1995). Note however that unlike person identity information (Tanaka & Farah, 1993), it is still debated whether facial expression perception actually relies on the configuration of the face (as a whole), or instead the selective perceptual processing of some diagnostic faces parts, including the mouth and the eyes (see deGelder, Teunisse, & Benson, 1997; Smith, Cottrell, Gosselin, & Schyns, 2005; Adolphs et al., 2005). In any case, it can be argued that in some cases at least, affect-relevant information from the face is carried by the whole facial configuration (see deGelder et al., 1997), and this is the assumption adopted in our work. The auditory component is provided by the voice and its subtle changes in pitch, duration or intensity/loudness when articulating and producing speech sounds or fragments, which may lead to different discrete affective tones of voice. The respective contribution of variations in these psycho-acoustical parameters has been measured in natural or simulated affective speech (Cummings & Clements, 1995; Lieberman & Michaels, 1962; Williams & Stevens, 1972; K. Scherer, 1989). Many prosodic features contribute to the expression of vocal emotions, and it seems evident that the acoustic correlates are subject to large inter-individual differences (see Lieberman & Michaels, 1962). Despite the large inter-speaker variability, there is some general consensus that if prosodic features are ranked in terms of their respective contribution, then gross changes in pitch do contribute most to the transmission of emotions, duration is intermediate whereas loudness seems to be least important (Frick, 1985; I. R. Murray & Arnott, 1993), even if the simultaneous processing and integration of these different parameters is probably required to efficiently decode the emotion from the voice. Noteworthy, the manipulation of affective speech prosody can be done independently of the semantic content conveyed in the message and this is why the prosodic channel can be considered to be a separate channel. Hence, the primary goal is to explore the nature of the relationship unifying a given emotional face expression and a concurrent (either compatible or not)

affective tone of voice. We refer to this integration effect as “multisensory perception of emotion”, following standard practice (see de Gelder, Vroomen, & Pourtois, 2004).

This chapter is divided into two main and consecutive sections. In the first part, we review some classical empirical evidence and dominant theoretical frameworks that have been put forward in the cognitive sciences literature to account for multisensory perception in general, and multisensory perception of emotion more specifically. In the second part, we present new (unpublished) behavioral empirical data addressing the selectivity of multisensory perception of emotion. Several experiments were carried out to assess if the integration of face (emotional expression) and voice (affective tone of voice) during emotion perception may be somehow specific, relative to other forms of stimulus redundancy (e.g. two emotional faces shown simultaneously, relative to a single emotional face). The results of these experiments somehow enable to better demarcate the constraints on bimodal sensory inputs which have to be met to eventually yield genuine behavioral effects of multisensory perception of emotion.

2 Object-based multisensory perception

2.1 Introduction

We restrict our review to what is usually referred to as object-based multisensory perception (Lehmann & Murray, 2005). As it turns out, multisensory perception of emotion can be considered as an instance of multisensory object recognition, and is similar to many other cases of object perception where convergent information about the same object is presented through different sensory modalities. As a first approximation, the kinds of audio-visual objects that have been mostly studied appear to be of two categories: simple/arbitrary audio-visual pairings vs. complex/natural pairings (see Pourtois & de Gelder, 2002). A common example of simple audio-visual pairings is the combination of light flashes with tone bursts, or the combination of specific tone frequencies with simple geometric figures (Stein & Meredith, 1993; Giard & Peronnet, 1999; Fuster, Bodner, & Kroger, 2000; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). Such pairings are obviously arbitrary and usually the subject is trained intensively to associate them, and later perceive them as paired in the context of the

experiment. The situation is quite different with more complex audio-visual pairs consisting of speech sounds and lip movements, or facial emotional expressions and affective tones of voice (De Gelder & Bertelson, 2003; Campanella & Belin, 2007). These complex pairings are natural, as they do not require any training for the perceiver to treat these pairs as such in the laboratory. In fact, in the course of studying these pairings naturally associated (e.g. an emotional facial expression combined with an affective tone of voice), the experimenter may even create conditions allowing pulling them apart and dissociating them (see Mcgurk & Macdonald, 1976; De Gelder & Vroomen, 2000). This is often done in order to obtain incongruent pairs and compare them with the more natural situation of congruence. Natural and arbitrary pairs thus seems to pull the researchers in opposite directions, to some extent, and it is plausible to argue that the underlying multisensory integration processes may be different depending on whether the audio-visual pairs are natural, or rather arbitrary (see Pourtois & de Gelder, 2002 for evidence).

Object-based multisensory perception is widespread in daily environments. However, there are only a few multisensory objects that have been studied in depth so far in cognitive sciences. Space perception, language perception and the perception of temporal events are three domains of human cognition where multisensory research has brought valuable insight. In the domain of space perception, many multisensory or crossmodal effects have been shown previously that all reflect our ability to integrate spatial information when this information is concurrently provided by the visual and auditory (or proprioceptive or tactile) modality (Driver & Spence, 1998a, 1998b, 2000). For example, the distance between spatially disparate auditory and visual stimuli tends to be underestimated with temporally coincident presentations, a phenomenon known as the ventriloquist effect/illusion (Bermant & Welch, 1976; Bertelson, 1999). Visual capture is another instance found in the spatial domain (Hay, Pick, & Ikeda, 1965). It involves a spatial localization situation in which the visual information is in conflict with that of another modality, namely proprioceptive information, and perceived location is determined predominantly by visual information. Likewise, when speech sounds (syllables) are presented simultaneously with incongruent lip movements, subjects report a percept that neither belongs to the visual modality, nor to the auditory one, but

that represents either a fusion or combination between the two inputs (Mcgurk & Macdonald, 1976). These results indicate that the visual and auditory components of syllables do combine and this combination translates as a new speech percept. Natural speech perception therefore provides a compelling case of multisensory integration (Dodd & Campbell, 1987). A third compelling instance or illusion of object-based multisensory integration is found in the temporal domain and may be seen, to some degree, as a symmetric case to that observed with the ventriloquist illusion. Here a visual illusion is induced by sound (Shams, Kamitani, & Shimojo, 2000). When a single flash of light is accompanied by multiple auditory beeps, the single flash is perceived as multiple flashes. This phenomenon is partly consistent with previous behavioral results that showed that sound can alter the visually perceived direction of motion (Sekuler, Sekuler, & Lau, 1997). Altogether, these effects suggest that visual perception is malleable by signals from other sensory modalities, such as auditory perception is malleable by signals from other sensory modalities. More generally, the dominance of one modality over the other does not seem therefore to be fixed or absolute, but instead may depend upon the context in which crossmodal effects take place. For space perception, the visual modality dominates over the auditory, and this situation is reversed during the perception of discrete temporal events (for which the auditory domain takes the lead on visual cues).

Traditionally, two sets of constraints have been envisaged in the literature (Bertelson, 1999). The first, referred to as structural factors, primarily concerns the spatial and temporal properties of the sensory inputs. The other set, often discussed as cognitive factors, is related to a whole set of higher-level, semantic or attention-related factors, including the subject's knowledge of, and familiarity with the multisensory situation (Talsma et al., 2010). Structural factors are the ones that have attracted by far the most attention from researchers in the field of multisensory integration (see G.A. Calvert, Spence, & Stein, 2004). By comparison, the role of cognitive factors is still under-investigated, although more recent work has started to explore the links between selective attention brain mechanisms and multisensory integration brain processes (see Talsma et al., 2010). However, from a conceptual viewpoint, it seems plausible to argue that some additional cognitive or object-based constraints on multisensory perception

actually take place, to prevent the organism to register many invalid and spurious incidences of multimodality, as solely defined based on the spatial and temporal coincidences of the visual and auditory inputs. Yet, there are only a few studies that have addressed this question, and tested to which extent object-based constraints may influence mechanisms of multisensory perception (see Pourtois & de Gelder, 2002; De Gelder & Bertelson, 2003).

Object-based multisensory perception is a complex issue, since beyond the spatial and temporal determinants of the input, the nature of the object to perceive may vary a lot from one condition (or encounter) to another. In this context, one may consider emotions just as one class of perceptual objects, besides other categories like speech (i.e. speech sounds presented simultaneously with lipread information/lip movements, see Mcgurk & Macdonald, 1976; G. A. Calvert et al., 1997) or space (i.e. although spatial localization is determined predominantly by visual cues, the presentation of concurrent spatial auditory or tactile cues strongly bias and influence visual spatial localization abilities, see Stein & Meredith, 1993; Driver & Spence, 1998a; Bertelson, 1999), as reviewed here above. Several objects or dimensions are actually susceptible to being perceived by multiple sensory channels at the same time, and therefore a central (still unanswered) question concerns the existence of general principles that would govern multisensory perception. Structural factors, such as temporal and spatial coincidence (see Stein & Meredith, 1993), may be envisaged as such. On the other hand and contrary to this view, one might postulate that each domain or object of perception (e.g. emotion, speech, space) actually possesses its own organization principles and that the overlap between these domains is fairly limited. Presumably, multisensory perception of emotion most likely shares some invariance in the basic perceptual mechanisms of audio-visual integration with these other domains (speech and space perception), while some specificity may well be present, although this question still remains open.

2.2 Multisensory perception: behavioral effects and cognitive models

In behavioral research on audio-visual integration, a few classical models have been proposed (Miller, 1982, 1986; Massaro, 1998; Bertelson, 1999; Dodd & Campbell, 1987;

De Gelder & Bertelson, 2003). The behavioral measures on the basis of which audio-visual integration is inferred are predominantly accuracy and response latency. When participants respond better and faster to the bimodal (audio-visual) stimulus than to either the visual only or auditory only stimulus, there is evidence that the response is presumably based on multisensory integration (Giard & Peronnet, 1999; Talsma et al., 2010). However, this evidence is inevitably indirect, and other accounts, like for example a race model (Raab, 1962) that do not assume integration of the two separate modality inputs can in principle still explain the same pattern of behavioral results. Although multisensory integration intuitively refers to the notion that the brain combines different input modalities, it is actually a theoretical notion advanced in order to account for a wide range of (behavioral) observations showing that there are bi-directional interaction effects between different sensory modalities. Traditionally, faster RTs for bimodal stimulus pairs than unimodal stimuli is compatible with the Redundant Signal Effect (RSE, see Miller, 1982, 1986). If an RSE is obtained for (congruent) audio-visual stimulus pairs, it does not necessarily mean that audio-visual integration (or neural interaction) occurs (Miller, 1986), however. Firstly, RSEs are also obtained with redundant stimuli presented in the same modality. The RSE is therefore not specific to multisensory perception, and is also found in spatial summation experiments in which a redundant simple visual stimulus (e.g. the simultaneous and synchronous presentation of the same simple visual stimulus at two separate spatial positions, usually on each side in the visual field to allow callosal interhemispheric transfer) is detected faster than a non-redundant visual stimulus, an effect classically referred to as Redundant Target Effect (RTE, see Marzi et al., 1986; Miniussi et al., 1998; M. M. Murray, Foxe, Higgins, Javitt, & Schroeder, 2001; de Gelder, Pourtois, van Raamsdonk, Vroomen, & Weiskrantz, 2001; Savazzi & Marzi, 2002, 2004, 2008; Turatto, Mazza, Savazzi, & Marzi, 2004). Secondly, faster RTs for (congruent) bimodal stimulus pairs (relative to unimodal stimuli) could be explained by a horse race model that does not imply interaction between sensory modalities (Raab, 1962), as briefly explained here above. In this perspective, each stimulus of a pair independently competes for response initiation and the faster of the two mediates the response. Thus, “simple” probability (or statistical) summation could yield the RSE. Indeed, the likelihood of either of two stimuli

yielding a fast RT is higher than that from one (unisensory) stimulus alone. On the other hand, RSE could also be explained by a co-activation model that implies that the two modalities are integrated together and interact prior to motor response initiation (Miller, 1982). In order to distinguish between these two opposite accounts (race model vs. co-activation model), Miller (1982) proposed to analyze RTs using cumulative probability functions and to test for what he called the inequality assumption. The inequality places an upper limit on the cumulative probability of RTs at a given latency for a stimulus pair. For any latency, *t*, the race model holds when the cumulative probability value is less than or equal to the sum of the cumulative probabilities from each of the single stimulus minus an expression of their joint probability. Hence, based on a formal analysis of RT distribution (and the violation of the inequality assumption, or not, see Miller, 1982, 1986), it is possible to establish whether a simple statistical facilitation/summation, or instead a co-activation (integration) between the two modalities (or sensory inputs) occurs during the processing of bimodal stimulus pairs (see Molholm et al., 2002 for an example).

Besides these important technical considerations related to the definition or qualification of multisensory perception effects, in fact, very few (computational) models have been developed in the literature to account for these multisensory behavioral effects. Notably, the Fuzzy Logical Model of Perception (FLMP, see Massaro, 1987; Massaro, 1998) represents such a valuable attempt. The key assumption behind the model of Massaro is that sensory information is always processed the same way, whatever the domain of application. In this perspective, audio-visual integration is just one instance of perception besides other cases and the underlying mechanisms responsible for audio-visual perception are similar to the mechanisms involved in other domains of cognition or perception. To validate his model, Massaro has provided data on bimodal speech perception from children, elderly, hearing impaired or bimodal emotion perception that all fit the FLMP (see also Massaro & Egan, 1996). This is an apparent strength of this computational model: this model is able to describe a wide range of human performance patterns during audio-visual perception. However, a possible downside is that the FLMP remains only descriptive, because this model does not implement any pre-conception about the nature of the components it seeks to describe (see Burnham,

1999). In the FLMP, four sequential stages of processing are postulated. The first step is feature evaluation, which is assumed to be carried out independently and separately for each modality source. The second stage is an integration of the features available after the first stage. This is of course the stage of interest, with regards to mechanisms of multisensory integration. Integration is achieved through a multiplicative combination of the response strengths of components of information input. Then, the result of this integration is matched against a prototype stored in memory during the assessment stage. Finally, a response is selected based on the most consistent prototype, given the visual and auditory cues. The proposal of a first evaluation stage carried out separately for each modality source is debated, and does not agree with independent evidence from neuroimaging or neurophysiology work showing reliable crossmodal effects not only in multimodal or heteromodal brain regions (A. R. Damasio, 1989; Mesulam, 1998; Pourtois, de Gelder, Bol, & Crommelinck, 2005; Ethofer, Pourtois, & Wildgruber, 2006), but also (and already) in unisensory or modality-specific cortices (see G. A. Calvert et al., 1999; Macaluso, Frith, & Driver, 2000; G. A. Calvert, 2001). Moreover, this independence of the auditory and visual components during audio-visual perception has been called into question, at least for the case of speech perception. An alternative account is the possibility of intermodal cues (see Campbell, Dodd, & Burnham, 1998). Another controversial property of the first evaluation step is related to the nature of the representations that drive this process. Indeed, in this model (Massaro, 1998), the algorithm of perception tags each feature with a continuous value and this characteristic runs against several empirical data that showed a categorical perception function during speech perception (see Liberman, Harris, Hoffman, & Griffith, 1957). Despite these critiques or limitations, the FLMP undoubtedly provides one of the few valuable computational models aimed at describing the critical computations involved during the perception and later integration of visual and auditory cues when presented simultaneously (object-based multisensory integration).

2.3 Multisensory perception effects revealed using the crossmodal paradigm

Several methods have been used to disclose, at the behavioral level, evidence of object-based multisensory perception. A classical method that we have used in this

work, is referred to as the crossmodal paradigm (see Bertelson, 1999). This specific paradigm is actually part of a larger sets of methods, used to indirectly measure the impact of stimulus processing in one sensory modality onto another. Other indirect methods include the use of after-effects (see Held, 1965), intersensory fusions (Mcgurk & Macdonald, 1976) or staircases (Bertelson, 1999).

The crossmodal paradigm is reminiscent of older studies on intermodal discrepancy following prismatic adaptation (Hay et al., 1965), on audio-visual space perception (Bermant & Welch, 1976), and has been used in audio-visual speech studies (Driver, 1996; Massaro, 1987, 1998) and in crossmodal attention studies (Driver & Spence, 1998a). In this paradigm, the systematic influence of one modality on the other is assessed using a strict methodology that requires a narrowing of the subject's attentional resources to one modality only, during stimulus processing. Then, a relative "automatic" crossmodal bias effect from the unattended modality to the attended modality can be measured. Therefore, the impact of one modality on the other is measured indirectly in the crossmodal paradigm. This procedure offers a double methodological advantage. Firstly, it has been shown to be more sensitive than direct measures and this procedure is better suited than other methodologies to capture genuine perceptual, as opposed to post-perceptual effects (see Bertelson, 1999). Secondly, it allows to manipulate the level or amount of congruence between the two modalities, unlike other contrasting methods capitalizing solely on the direct comparison between unidmodal and multimodal stimulus conditions (see Giard & Peronnet, 1999; Molholm et al., 2002). Hence, the experimenter may use this powerful method and set up in the laboratory artificial conditions in which the level of congruence between the two sensory modalities systematically or parametrically varies (see De Gelder & Vroomen, 2000). This method enables quantification of the actual crossmodal impact from one modality onto the other (e.g. from vision to audition, or vice versa) during the perception of bimodal stimulus pairs, including the combination of an emotional facial expression with an affective tone of voice (see Pourtois et al., 2005; Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000; Pourtois, Debatisse, Despland, & de Gelder, 2002), as reviewed in the next section.

3 Multisensory perception of emotion

3.1 Possible functions of multisensory perception of emotion

The fact that emotional information concurrently presented in different sensory modalities is integrated is likely to occur for reasons that go far beyond a simple back-up function allowing the system to overcome a given sensory loss and to rely on the spared/redundant modality to continue to operate. At least three distinct arguments or points can be evoked to support the functionality and need for integration during multisensory perception of emotion.

A first support for functionality comes from several older developmental studies (see Lewkowicz, 2000) that have clearly shown that very young infants look longer at face stimuli accompanied by voices (see Haith, Bergman, & Moore, 1977). Five to seven-month-old infants also look longer at a face that carries the same expression as the voice than at a face carrying a different expression (Walker & Grolnick, 1983). These results suggest that the recognition of affective expressions may be first multimodal, before a differentiation occurs between the face and the voice (Walker-Andrews, 1997). There would be an ontogenetic priority in favor of multisensory perception. Furthermore, these results suggest a possible modular organization for audio-visual perception of emotion, which is not consistent with a simple back-up function.

The second element is that each sense (here the visual and the auditory channel) actually provides a qualitatively distinct subjective impression of the environment, including emotion perception. Although referring to the same event (e.g. an angry affective state), the emotion conveyed by ear (voice of wrath) and by eye (a furious facial expression) is not simply redundant but both senses complement each other given the specificity and specialization of each sense (de Gelder, Vroomen, & Pourtois, 1999). This argument is therefore about the sensory specificity and complementary of multisensory perception of emotion.

A third defense for the importance of the function of multisensory perception of emotion is the optimization. Indeed, multisensory perception of emotion consists of enhancing detection and discrimination of emotions, as well as speed responsiveness to these

highly relevant biological stimuli (Sander, Grafman, & Zalla, 2003). The fact that the perception of emotions is by nature multimodal and audio-visual, allows the perceptual system to disambiguate the actual functional meaning of the emotional input using a stable amodal or supramodal representation (see Farah, Wong, Monheit, & Morrow, 1989; Borod et al., 2000). There are large inter-individual differences between human beings (as well as animals) in the ability to express and perceive different emotions. Moreover, humans have numerous ways to express and perceive the same emotion. As a consequence, the rapid and automatic combination of different channels of communication probably acts as an optimizer or catalyzer to rapidly perceive and efficiently recognize a given emotional state or object. From an evolutionary perspective (see also A.R. Damasio, 1994), integration of multiple affective inputs across different sensory modalities makes adaptive sense, given the enhanced biological significance of emotional stimuli. It also makes sense, given the fact that combining different sources of information (face and voice) usually leads to more accurate and faster judgments, as well as more appropriate behaviors, as stressed in the second section of this chapter here below. However, this compensatory function may not be specific to multisensory perception of emotions, and could also explain other multisensory perceptual phenomena, like crossmodal spatial mechanisms at stake during the ventriloquist illusion for instance (Bertelson, 1999).

3.2 Behavioral evidence for multisensory perception of emotion

In their seminal study, de Gelder and Vroomen (2000) performed a series of elegant behavioral experiments looking at crossmodal effects from the voice to the face, and vice versa, during emotion perception (see also Massaro & Egan, 1996). These authors used an experimental situation in which varying degrees of (in)congruence were created between emotional facial expression and affective tone of voice. Two contrasting emotions (happy vs. sad) in the voice were manipulated. The same sentence (with a neutral semantic content) was uttered by a semi-professional actor, either with a happy or sad tone of voice. Using a standard morphing technique (see Etcoff & Magee, 1992; Beale & Keil, 1995; deGelder et al., 1997), a visual continuum of varying emotional facial expression with 11 steps starting from one emotion at one extreme (happy) and

going to another emotion (sad) at the other extreme was created. In the two first experiments, de Gelder and Vroomen (2000) combined the 11 faces with the two auditory conditions and compared these pairings to the condition where the morphed faces were presented alone (no accompanying sound). The task of the participant was to judge the emotion (Experiment 1) or to judge the emotion conveyed by the face (happy vs. sad) while ignoring the concurrent voice (Experiment 2). Results clearly showed that the identification of the emotion in the face was categorical (see Etcoff & Magee, 1992; Beale & Keil, 1995; deGelder et al., 1997), but more importantly was systematically biased in the direction of the simultaneously presented affective tone of voice. More specifically, this effect consisted in the fact that the likelihood to give a sad response when judging the emotional face (along the continuum) was reduced if the face was paired with a happy voice, regardless of the amount of sadness perceived in the face (i.e. general lateral shift of the psychometric response function). Moreover, RT results also showed that congruent bimodal stimulus pairs were judged faster than either incongruent stimulus pairs or single-modality stimuli (i.e. faces only).

Another question addressed in this study was whether this crossmodal bias effect during emotion perception could also be obtained from the face to the voice, or only from the voice to the face as reviewed here above. Were these crossmodal bias effects during emotion perception bidirectional and symmetric? In a third experiment, de Gelder and Vroomen (2000) directly addressed this question and created for this purpose a symmetric situation where the crossmodal impact from the face to the voice during emotion perception could be measured and assessed. They created a symmetric experiment to Experiment 2 in which a 7-steps voice continuum between two extreme emotions (fear vs. happy) was made up. Like was the case for the visual continuum used in the first two experiments, a vocal continuum was created using a computer-assisted auditory morphing procedure, by manipulating in a parametric fashion the physical distance between several features of the voices. This sophisticated procedure essentially works out on a modeling and subsequent parametric modulation of the fundamental frequency (F0) of the two original auditory fragments (fear voice and happy voice, see Vroomen, Collier, & Mozziconacci, 1993). Changing simultaneously and parametrically the duration, pitch range and pitch register of the two original utterances

allowed to create several discrete steps along a vocal continuum, progressively going from one emotion (happy) to the other (fear). These seven voice fragments were then combined with two facial expressions (fearful vs. happy) to yield 14 stimulus pairs with varying levels of emotion (in)congruence. In this experiment, participants were instructed to judge the emotion conveyed by the voice, while ignoring (though attending to) the face information. Results showed a systematic bias of emotional voice identification by the concurrent facial expression, as well as a RT facilitation for congruent bimodal stimulus pairs, relative to incongruent bimodal pairs. Altogether, these results suggest bidirectional (from face to voice and vice versa) crossmodal bias effects during emotion perception.

Although certainly convincing, these behavioral results (De Gelder & Vroomen, 2000) are also compatible with other explanations that do not postulate any access to the emotional content of the face or the voice in order to trigger the crossmodal bias effect during emotion perception. For example, one may speculate that the crossmodal bias effect from the face to the voice during emotion perception described here above may not be specific to the affective content of the face, but instead may be obtained with any other visual stimuli that have an affective content. Hence, an important additional evidence would be to show that the actual (covert) processing of the affective information from the emotional face is crucial in order to obtain a reliable crossmodal bias effect (from the face to the voice) during emotion perception. This question was addressed in a different study and the results basically confirmed this hypothesis (De Gelder, Vroomen, & Bertelson, 1998). Presenting emotional faces upside down disrupts the perceptual processing of the emotional facial expression (Mckelvie, 1995; deGelder et al., 1997). Based on this face inversion effect for emotional facial expressions, de Gelder et al. (1998) surmised that the crossmodal bias effect from the face to the voice would be strongly attenuated with the presentation of inverted, relative to upright emotional facial expressions. Results of this study confirmed this prediction, as when subjects were asked to identify the emotional tone of voice in the presence of inverted emotional facial expressions, the crossmodal bias effect from the face to the voice basically disappeared, whereas it was still well present when these faces were presented upright (see also De Gelder & Vroomen, 2000). These behavioral results

suggest that the crossmodal bias effect (from the face to the voice) during emotion perception is a function of the expression conveyed by the face. More generally, these findings add support to the notion that crossmodal affective biases are to some extent automatic and perceptual in nature, and they cannot be easily reduced to some post-perceptual voluntary adjustments.

3.3 A role for attention in multisensory perception of emotion?

These findings indicate that these crossmodal effects during emotion perception are likely to be perceptual, mandatory and automatic, as opposed to post-perceptual (e.g. response bias) or influenced by attention, subjective beliefs or decision processes (see Bertelson, 1999). Indeed, the fact that this systematic crossmodal influence was observed even when participants were instructed to voluntarily ignore one of the two sensory modalities seems to indicate that multisensory integration of affective information takes place "automatically" to some extent, regardless of attentional factors. This property (i.e. independence from demands on attentional capacity) has long been one of the defining characteristics of "automatic" processes (see Kahneman, 1973; Schneider & Shiffrin, 1977, but see Moors & De Houwer, 2006 for a more recent and refined theoretical account of "automaticity"). To some degree, this observation is also consistent with the notion that the integration between an emotional facial expression and an affective tone of voice is occurring at a pre-attentive level (see also Driver, 1996), as demonstrated using other investigation techniques, like the recording of event-related brain potentials in healthy adult participants, which suggest early perceptual effects within modality-specific cortices during multisensory perception of emotion (see Pourtois et al., 2000; Pourtois et al., 2002). This audio-visual integration of emotional signals could take place during an early perceptual stage of stimulus processing, before (selective) attention comes into play (Talsma et al., 2010).

On the other hand, strict perceptual properties associated with multisensory perception of emotion may appear unlikely, given the fact that recognition of emotion stands as a particularly content-rich process, and seems more akin to higher-level cognition than perception. In this context, the (multisensory) perception of emotion should make a poor candidate for qualifying as a case of perception-based audio-visual integration. Yet,

there is a wealth of recent empirical studies (including in brain-imaging) that have brought support to the notion that emotion perception and recognition (at least for some specific emotions like fear or anger) is a true perceptual process, or at least has a hard perceptual core (see Calder, Lawrence, & Young, 2001; Pourtois, Grandjean, Sander, & Vuilleumier, 2004; Vuilleumier, 2005; Phelps, Ling, & Carrasco, 2006; Bocanegra & Zeelenberg, 2009a, 2009b; see also chapter by Kreifelts, Wildgruber and Ethofer in this book). Hence, multisensory perception of emotion is likely to rely on genuine perceptual mechanisms, which allow a pre-attentive binding of affective information simultaneously conveyed by multiple sensory cues (visual and auditory). Indirect evidence obtained in specific brain-damaged patients also lent support to this conclusion (i.e. multisensory perception of emotion is a perceptual process). In one of these neuropsychology studies, two patients with selective striate cortex damages but unaware low-level residual visual abilities ("Blindsight", see Weiskrantz, 1986) were nevertheless shown to benefit partly from the presentation of an emotional visual stimulus (either a face or a scene) in their blind visual field during multisensory integration of emotion (i.e. this visual stimulus for which they therefore remained unaware had nonetheless a reliable crossmodal influence on the processing of a concurrent affective tone of voice, see de Gelder, Pourtois, & Weiskrantz, 2002). These results speak for multisensory integration mechanisms occurring without attention, possibly even without (visual) stimulus awareness (see also de Gelder, Pourtois, Vroomen, & Bachoud-Levi, 2000).

As suggested here above, the observation that crossmodal influences (from the face to the voice and vice versa) during emotion perception are observed although the participants were instructed to ignore one of the two modalities (i.e. crossmodal bias effect) can be taken as evidence, at least partly (see Schneider & Shiffrin, 1977) that multisensory perception of emotion is automatic and does not depend upon the availability of attentional resources. However, even if the instructions are to a ignore stimulus in one modality and attend to the concurrent emotional stimulus in the other modality, it may well be that it is actually difficult to ignore this former stimulus and modality (e.g. because of an intrinsic "sensory" dominance for example). Indeed, research on attention has clearly shown that irrelevant/unattended visual stimuli may be particularly hard to ignore under low-load conditions, when the specified task consumes

more attentional capacity (Lavie, 1995, 2005). Hence, one may speculate that in the case of crossmodal influence from the face (which has to be ignored) to the voice, it may be hard to ignore the face, a highly biologically relevant visual stimulus, despite the instructions, due to the low-load nature of the experimental situation (i.e. the emotional facial expression was the only visual stimulus present and the task required an identification of the affective tone of voice, see De Gelder & Vroomen, 2000). Interestingly, this issue was actually addressed in a different study (Vroomen, Driver, & de Gelder, 2001). The authors used a dual-task paradigm asking participants either to add two digits presented visually together (i.e. high load) while judging the affective tone of voice (presented simultaneously with a congruent or incongruent emotional facial expression), or simply detect zeros in a rapidly presented sequence of digits shown visually (i.e. low load) while performing the same task. Moreover, in a third condition, they also gave a secondary auditory task to participants, consisting of deciding whether a tone was high or low (i.e. low load) while judging the emotion from the voice. This experimental design allowed the authors to test whether the crossmodal bias effect from the face to the voice was (load) attention dependent or not, i.e. would be reduced by either a secondary auditory task or by a secondary visual task, the latter which could be either easy or more difficult. Results showed that the crossmodal bias effect was actually independent of whether or not subjects performed a demanding additional (distracting) task. In all three cases, the visible static emotional face had a reliable impact on judgments of the heard emotional voices. Moreover, the systematic influence of the seen emotional facial expression on judgments of the emotional tone of the heard voice was not eliminated under conditions of high load (see Vroomen et al., 2001). These behavioral results therefore confirmed that multisensory perception of emotion is automatic to some extent, since it arises regardless of the attentional demands imposed by an additional task (see Lavie, 2005).

4 Multisensory perception of emotion vs. emotion stimulus redundancy

4.1 Introduction

The behavioral evidence reviewed so far is consistent with the notion that the perceptual system integrates emotional information from the face and the voice,

probably at an early/pre-attentive stage following stimulus onset (see Pourtois et al., 2000). Indeed, one may speculate that the ability to combine multiple inputs from different sensory sources is advantageous for an organism in the case of affect perception, as it appears to be the case for other forms of multisensory perception (e.g. speech or space perception). But reasonably, in the absence of any limits on which inputs make "good" pairings, such a theoretical advantage would be quickly lost, to some extent. Yet, still little is known about possible constraints on affective pairings and on the possible role that such constraints may play when turning to the case of multisensory perception of emotion. This question is therefore somehow related to the selectivity of the crossmodal bias effect during emotion perception, as reviewed here above (see also de Gelder et al., 2002). Two extremes, but equally plausible alternatives, may be suggested in this respect. Either the crossmodal bias between the voice and the face during emotion perception actually reflects the existence of a general mechanism for affect perception whereby the perceptual system (blindly) samples and merges all sources of affect information available at a given moment (time) and position (space) (Borod et al., 2000). Or alternatively, the crossmodal bias effect during emotion perception is narrowly restricted to the combination of an emotional facial expression with an effective tone of voice, and this specific audio-visual situation requires selective (perceptual) mechanisms (see Pourtois & de Gelder, 2002; Pourtois et al., 2005). For example, we do not know if task irrelevant stimuli presented in the periphery, like for example an additional emotional facial expression or a written emotion word, would not have a comparable influence on emotional ratings of a central emotional facial expression, just like the affective prosody in a concurrent voice does (see De Gelder & Vroomen, 2000). If the crossmodal bias effect would present some of the same characteristics as for example the interference effect observed in Stroop-like tasks (see MacLeod, 1991), it would slow-down rather than speed up the response to the central target emotional face stimulus. Hence, unconstrained integration (i.e. occurring regardless of the object or content) would probably expose the organism to vicarious influences away from the main task at hand.

In a series of behavioral studies, we actually addressed this question of selectivity and compared the crossmodal bias effect (from the voice to the face) during emotion

perception to other cases of pairings or emotion stimulus redundancy (while keeping the sensory modality – vision - constant). Either a central emotion expression was paired with a congruent or incongruent affective tone of voice (audio-visual integration of emotion), or instead with another/distracting congruent or incongruent emotional facial expression (visual redundancy, see Marzi et al., 1986; Miniussi et al., 1998). Likewise, we also looked at the pairing of the central emotional facial expression with a distracting written emotion word, shown at the same unattended spatial location. We predicted that the crossmodal bias effect would be qualitatively different, relative to these two other instances of emotion redundancy ("intramodal" bias). More specifically, we surmised a facilitation for congruent bimodal face-voice parings during emotion perception (see De Gelder & Vroomen, 2000), whereas the presentation of an additional emotional facial expression or written emotion word would primarily slow down perceptual decision during incongruent pairings, consistent with an interference effect (see MacLeod, 1991). Such an asymmetric outcome would indicate that multisensory perception of emotion cannot simply be assimilated to a case of emotion stimulus redundancy, and that there is more to gain for the perceptual system in the simultaneous presentation of a face and a voice during emotion perception, than the mere juxtaposition of multiple/redundant emotional stimuli within the same sensory modality (here with a focus on vision and emotion face stimuli).

4.2 Methods

Thirty-one adult participants (mean age: 20) were instructed to discriminate the emotion expression (angry vs. sad) of a central target face. This central target face was presented either alone, or accompanied by an affective distracter. This distracter could be either auditory (an angry or sad tone of voice), or visual (the written name of an emotion word or another face). Following standard practice, (see Driver, 1996; Bertelson, 1999), we manipulated the emotional congruence between the target face and the affective information presented concurrently and to be ignored. We used a within-subject design, the same procedure and stimulus duration of the central target face across the different conditions. Notably, the effect of the auditory distracter on emotion face perception was studied in a separate block than the effect of the visual

distracters (either a written emotion name or an emotional facial expression). Note that only emotions with a negative valence, i.e. angry and sad, were used, in order to avoid possible confounds in the interpretation between the role of (in)congruence between affective content of target and distracter, and actual valence of the emotion displayed (positive vs. negative).

The target face (5 cm width x 6.5 cm height) consisted of the static black and white photograph of one out of six actors, posing either a sad or an angry emotional facial expression (see Ekman & Friesen, 1976) and was briefly presented in the center of a 17-inch screen for 150 ms. Auditory stimuli were 12 different spoken words always with the same neutral content (/plane/) pronounced by semi-professional actors, either with a sad or angry affective tone of voice (see Pourtois et al., 2000; Pourtois et al., 2002; Pourtois et al., 2005 for additional details regarding these previously validated auditory stimuli). Mean duration of the auditory fragments was 348 ms. Based on the emotion content of the face and the voice, congruent and incongruent audio-visual pairs were created. The spoken distracter was always presented at such a time that its offset coincided with that of the central face stimulus (duration of 150 ms). Face distracters were identical to the targets. All combinations involved two pictures of the same actor, displaying the same emotion (thus twice the same picture) on congruent trials, or different emotions on incongruent trials. The emotional face distracter was presented in full synchrony with the target face, 5 cm above it (distance from the screen was 60 cm). Emotion written words were two adjectives (/ANGRY/ vs. /SAD/, in French) printed in Times police 24 (3 cm width x 1 cm height). Like for the distracting face, the distracting word was presented synchronously with the central emotional face, 5 cm above it. Congruent and incongruent trials were created based on the (mis)match between the emotion displayed by the central face and that conveyed by the written word briefly presented in the upper visual field.

The experimental session included control trials during which the central target face was presented alone (no distracter) and trials during which it was accompanied by a distracter, in random order. All trials started with the presentation of a fixation cross in the center of the screen for 250 ms, followed by a 600 ms blank screen, and then the

presentation of the central target face for 150 ms. Such a short stimulus presentation for the emotional face presumably reduced peripheral eye explorations (e.g. vertical saccades back and forth between the two positions in the visual field). At the offset of the central target face, the screen went blank again for 1200 ms (allowing registration of the actual response made by the participant), before the next trial started. Participants were instructed to perform a two-alternative forced choice task about the actual emotional expression (sad vs. angry) of the central emotional face, and to respond as accurately and as fast as possible by pressing one of two keys of a response pad with their dominant hand. They were explicitly told to base their response only on the emotional target face, and to ignore either the auditory or visual distracter (either a face or a written word). The testing included two main blocks. In one block, control trials (n=60) were intermixed with audio-visual trials (n=120, 60 per level of congruence). In the other, control trials (n=60) were intermixed with visual-redundant trials (n=240, 60 per type of distracter and per level of congruence). The order of the two blocks was counterbalanced across participants. Within each block and across participants, trial order was randomized.

4.3 Results

For the audio-visual condition/block, a repeated measures analysis of variance (ANOVA) with two factors (Emotion of the central face and Trial type) was computed on mean error rates (mean error rate was 15%). This analysis disclosed a significant main effect of trial type ([$F(2,60)=6.15$, $p<.005$], with no significant modulation by emotion. Post-hoc comparisons (based on paired t-tests) showed that congruent trials produced on average less discrimination errors than control [$t30=3.37$, $p<.005$] and incongruent [$t30=2.52$, $p=.014$] trials. The difference between control and incongruent trials was not significant [$t30<1$]. The statistical analysis carried out on mean RTs (for correct discriminations only) basically revealed a similar outcome, indicated by a clear RT facilitation for congruent audio-visual pairs, relative to either control trials (i.e. emotional face alone) or incongruent audio-visual trials (see Figure 1A). This analysis revealed a significant effect of trial type [$F(2,60)=12.3$, $p<.001$], with no reliable modulation by the emotion content of the face. Post-hoc comparisons confirmed faster perceptual

decisions for congruent audio-visual trials, relative to control trials [t30=2.47, p=.016] or incongruent audio-visual trials [t30=2.49, p=.016]. There was therefore no speed-accuracy trade-off, as participants responded faster and made less errors for congruent audio-visual trials compared to the other conditions (emotional face alone or emotional face combined with an incongruent affective tone of voice). More importantly, these behavioral results confirmed a systematic and significant crossmodal bias effect from the to be ignored voice to the attended face (see De Gelder & Vroomen, 2000), indicated by a facilitation of the ratings/perceptual judgments of the central emotional face when it was accompanied by a congruent affective tone of voice (Figure 1A). Central to the present investigation is the question whether a similar facilitatory perceptual effect could be observed when the same central emotional face is no longer combined with an affective tone of voice, but instead with another "distracting" emotional face or a written emotion word.

Results obtained for the other block (visual-redundant trials) show a very different outcome (see Figure 1B). First, the 2 (Emotion) x 2 (Congruence) x 2 (trial type: written word vs. face) ANOVA performed on mean error rates did not yield any significant effect. By comparison, the ANOVA performed on mean RTs disclosed a significant effect of Congruence [F(1,30)=25.85, p<.001], as well as a significant interaction between Emotion and Congruence [F(1,30)=9.78, p<.005]. However, this significant effect of congruence clearly indicated slower RTs with incongruent trials than either control or congruent trials (Figure 1B), and this effect turned out to be larger for sad faces than angry faces. Moreover, this significant interference effect was found to be the same, regardless of the nature of the affective distracter, either an emotional face or a written emotion word (Figure 1B).

4.4 Discussion

Based on previous results and findings (see De Gelder & Vroomen, 2000; Pourtois et al., 2005), we predicted that a gain in accuracy and response latencies (RT) would be observed when an emotional facial expression had to be judged as part of a multisensory emotion object (i.e. audio-visual pairing). Results of this study confirmed this prediction. However, when the exact same emotional target face stimuli were rated

in the presence of a concurrent distracting emotional face or visual emotion word, there was also a reliable influence from this latter unattended visual stimulus on the ratings of the central face, but from a different nature (compare Figure 1A to Figure 1B). Unlike what we found for face-voice pairs during emotion perception, when emotional visual distracters are congruent with the central emotional face targets, they do not facilitate or enhance the perceptual processing of these targets. Instead, these visual emotional distracters have a negative impact on response latencies when they carry an incongruent emotional meaning, relative to the central visual target. Taken together, these behavioral results somehow suggest a special status for face-voice combinations during emotion perception, relative to the mere redundancy of emotion information within the same (visual) sensory modality. As reviewed here above, it has been argued that facilitation or enhancement of responses/decisions to a target (i.e. perceptual "benefit") presented together with a congruent distracter may be indicative of perceptual integration between the two inputs (see Miller, 1982; Stein & Meredith, 1993; Massaro, 1998; Marzi et al., 1986; Miniussi et al., 1998). In contrast, a response "cost" associated with the presence of incongruent distracters inevitably evokes response competition (or response bias) phenomena, such as typically observed in the well-known Stroop effect (MacLeod, 1991).

More generally, these new behavioral results suggest a general mechanism for within-modality perception of affect (as the effect was the same for the unattended additional emotional face and the emotional written word), and are therefore consistent with the computational model of perception proposed by Massaro (1998). However, the new critical result is that in this condition (emotion stimulus redundancy within the same sensory modality), congruent trials are processed the same way as control trials (face alone, see Figure 1B), and therefore this interference effect is substantially different from the response facilitation observed with audio-visual pairings during emotion perception (see Figure 1A). In other words, emotion congruence across the concurrent inputs (face + face or face + word) does not lead to a gain in response latencies, but incongruence leads to a processing cost, whose origin is likely to be found at the level of the selection of the motor response (i.e. post-perceptual effect) and which may be dependent upon the availability of attentional resources (see Talsma et al., 2010).

Indeed, these findings may be compatible with a relatively late response competition view between visual stimuli, such as postulated previously in other interference situations (see MacLeod, 1991). These results therefore suggest a dissociation between the combinatory processes unifying an emotional face with a concurrent affective tone of voice (see Pourtois et al., 2005), and those underlying the perception of multiple or redundant visual emotional stimuli. Whereas the former may depend on perceptual, possibly even pre-attentive mechanisms (see also de Gelder et al., 2002; de Gelder et al., 2004), the latter may reflect another class of integration processes which do not rely so much on perceptual mechanisms, and which in turn may be influenced by higher order cognitive processes, including decision making and selective attention (Talsma et al., 2010).

A potential objection is that the response facilitation found with audio-visual pairs during emotion perception may be somehow a consequence of the auditory component in the pairing, rather than the affective congruence of the central emotion face stimulus with this emotional auditory distracter. For example, with distracters that are in the same modality as the central target, they may not produce any facilitatory effect due to an attentional bottleneck phenomenon (see Pashler, 1994; Marois & Ivanoff, 2005). This alternative interpretation is unlikely though, because prior behavioral studies have shown that the crossmodal bias effect during emotion perception was not altered by a secondary task, even if this latter actually required extra processing load in either the visual or auditory modality (see Driver, 1996; Vroomen et al., 2001; Lavie, 2005). Hence, a putative limitation of processing resources does not seem to be a critical factor hampering multisensory perception of emotion (see also Pourtois et al., 2005; Pourtois et al., 2000).

5 General Conclusions

Sensory modalities are traditionally characterized by the type of physical stimulation they are more sensitive to, light for vision, sound for hearing, skin pressure for touch, molecules in the air for smell, etc (Mesulam, 1998). This approach to the study of perception does not do justice to natural beginnings of perception. Indeed, in most of the cases, the organism is confronted with different sensory inputs often taking place at

the same time and place, and the perceiver reports objects with multiple sensory attributes. Hence, in many natural situations, different senses receive more or less simultaneously correlated information about the same object or event. The sensory specificity does not correspond either to what usually happens at the other extreme of the perception process, namely the perceiver's intuition that after perceiving or recognizing an object or event, or after remembering or imagining it, different sensory modalities are intimately linked with one another. In line with this notion of sensory specificity, there seems to be a sort of general consensus in the field about the assumption that information from primary and modality specific cortices is combined in heteromodal areas of the brain (see Stein & Meredith, 1993; Mesulam, 1998; G. A. Calvert, 2001; Ethofer et al., 2006), an integration process that eventually yields multisensory-determined objects.

In this chapter, we have reviewed evidence from behavioral studies and cognitive models showing that the perception of emotion can be qualified as an object-based multisensory phenomenon. Emotional facial expression and affective tone of voice do combine during emotion perception to eventually yield strong perceptual effects, which can be distinguished from the mere redundancy of emotional signals within a given sensory modality (here with a focus on the visual modality). This multisensory phenomenon is likely to be an early perceptual, maybe pre-attentive integration effect, which does not resemble behavioral manifestations of emotion stimulus redundancy, for which a clear-cut post-perceptual cost, rather than a perceptual benefit was observed in our study. On the other hand, this observation does not contradict the notion that selective attention mechanisms can boost or alter multisensory perception effects (see Talsma et al., 2010), but under certain laboratory circumstances at least, the combination of an emotional facial expression with a concurrent affective tone of voice can take place irrespective of the fact that stimulus content in one modality is directly ignored (unattended), or attentional load is reliably increased (see Vroomen et al., 2001).

As we have argued throughout this chapter, this audio-visual integration during emotion perception rests on an argument of adaptiveness. The combination of complementary

affective information conveyed by multiple channels is probably adaptive, because it can reduce the effects of potentially adverse factors like drifts or intrinsic noise on perceptual performance. But in the absence of any limitations, multisensory perception would be rather inefficient to deal with these natural variations. The new behavioral results presented in this chapter are in line with this assumption and previous studies (see De Gelder & Vroomen, 2000), as they show that the combination of an emotional facial expression with an affective tone of voice does not reflect a general (amodal) perceptual effect (see Massaro, 1998; Massaro & Egan, 1996; Borod et al., 2000), but instead it may serve a specific optimization function for the organism, aimed at binding selectively visual and non-lexical (psycho-acoustic) auditory cues during emotion perception, as these two cues usually convey simultaneously and naturistically critical and converging emotional information about the actual mental state, and possible intentions or action tendencies of peers or conspecifics (Frijda, 1989). For this reason, is the combination of multisensory inputs during emotion perception probably a key perceptual process relying on specific cognitive processes and neural systems (see Pourtois et al., 2005), likely sharing similarities with other multisensory objects, including space perception (see Bertelson, 1999), even though this conjecture remains largely speculative at this stage and it would need some direct confirmation at the empirical level as well.

Acknowledgements

Writing and elaboration of this chapter was made possible thanks to the financial support provided by the European Research Council (Starting Grant #200758) and Ghent University (BOF Grant #05Z01708) to GP. The behavioral results presented in this chapter were already partly introduced and discussed in the unpublished thesis manuscript written up and submitted by Gilles Pourtois (Tilburg University, April 2002).

References

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*(7021), 68-72.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614-636.

Beale, J. M., & Keil, F. C. (1995). Categorical Effects in the Perception of Faces. *Cognition, 57*(3), 217-239.

Bermant, R. I., & Welch, R. B. (1976). Effect of Degree of Separation of Visual-Auditory Stimulus and Eye Position Upon Spatial Interaction of Vision and Audition. *Perceptual and Motor Skills, 43*(2), 487-493.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersleben, T. Bachman & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347-369). Amsterdam: Elsevier Science.

Bocanegra, B. R., & Zeelenberg, R. (2009a). Dissociating Emotion-Induced Blindness and Hypervision. *Emotion, 9*(6), 865-873.

Bocanegra, B. R., & Zeelenberg, R. (2009b). Emotion Improves and Impairs Early Vision. *Psychological Science, 20*(6), 707-713.

Borod, J. C., Cicero, B. A., Obler, L. K., Welkowitz, J., Erhan, H. M., Santschi, C., et al. (1998). Right hemisphere emotional perception: Evidence across multiple channels. *Neuropsychology, 12*(3), 446-458.

Borod, J. C., Pick, L. H., Hall, S., Sliwinski, M., Madigan, N., Obler, L. K., et al. (2000). Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. *Cognition & Emotion, 14*(2), 193-211.

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition & Emotion, 24*(3), 377-400.

Burnham, D. (1999). Perceiving talking faces: from speech perception to a behavioral principle (reviewed by D. Burnham). *Trends in cognitive sciences, 3*, 487-488.

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of fear and loathing. *Nature reviews, 2*(5), 352-363.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral cortex (New York, N.Y., 11*(12), 1110-1123.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*(12), 2619-2623.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science (New York, N.Y, 276*(5312), 593-596.

Calvert, G. A., Spence, C., & Stein, B. E. (2004). *The Handbook of Multisensory Processes*. Cambridge: MIT Press.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in cognitive sciences, 11*(12), 535-543.

Campbell, R., Dodd, B., & Burnham, D. (1998). *Hearing by Eye II: Advances in the psychology of speechreading and audio-visual speech*. Hove, UK: Psychology Press.

Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *Journal of the Acoustical Society of America, 98*, 88-98.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: a system-level proposal for the neural substrates of recall and recognition. *Cognition, 33*, 25-62.

Damasio, A. R. (1994). *Descartes 'Error. Emotion, Reason and the Human Brain*. New-York: Putman Books.

Darwin, C. (1871). *The descent of Man*. London: John Murray.

Darwin, C. (1872). *The Expression of Emotions in Man and Animals*. London: John Murray.

de Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature reviews, 7*(3), 242-249.

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in cognitive sciences, 7*(10), 460-467.

de Gelder, B., Pourtois, G., van Raamsdonk, M., Vroomen, J., & Weiskrantz, L. (2001). Unseen stimuli modulate conscious visual experience: evidence from inter-hemispheric summation. *Neuroreport, 12*(2), 385-391.

de Gelder, B., Pourtois, G., Vroomen, J., & Bachoud-Levi, A. C. (2000). Covert processing of faces in prosopagnosia is restricted to facial expressions: evidence from cross-modal bias. *Brain and cognition, 44*(3), 425-444.

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. *Proceedings of the National Academy of Sciences of the United States of America, 99*(6), 4121-4126.

De Gelder, B., & Vroomen, J. (2000). Perceiving Emotions by Ear and by Eye. *Cognition & Emotion, 14*(289-311).

De Gelder, B., Vroomen, J., & Bertelson, P. (1998). Upright but not inverted faces modify the perception of emotion in the voice. *Current Psychology of Cognition, 17*, 1021-1031.

de Gelder, B., Vroomen, J., & Pourtois, G. (1999). Seeing cries and hearing smiles. Crossmodal perception of emotional expressions. In G. Aschersleben, T. Bachmann & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 425-438). Amsterdam: Elsevier.

de Gelder, B., Vroomen, J., & Pourtois, G. (2004). Multisensory Perception of Emotion, Its Time Course, and Its Neural Basis. . In G. Calvert, C. Spence & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 581-597). Cambridge, MA: MIT Press.

deGelder, B., Teunisse, J. P., & Benson, P. J. (1997). Categorical perception of facial expressions: Categories and their internal structure. *Cognition & Emotion, 11*(1), 1-23.

Dodd, B., & Campbell, R. (1987). *Hearing by eye: the psychology of lip-reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381*(6577), 66-68.

Driver, J., & Spence, C. (1998a). Cross-modal links in spatial attention. *Philosophical transactions of the Royal Society of London, 353*(1373), 1319-1331.

Driver, J., & Spence, C. (1998b). Crossmodal attention. *Current opinion in neurobiology, 8*(2), 245-253.

Driver, J., & Spence, C. (2000). Multisensory perception: beyond modularity and convergence. *Current biology, 10*(20), R731-735.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*, 169-200.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo-Alto: Consulting Psychologists Press.
Etcoff, N. L., & Magee, J. J. (1992). Categorical Perception of Facial Expressions. *Cognition, 44*(3), 227-240.
Ethofer, T., Pourtois, G., & Wildgruber, D. (2006). Investigating audiovisual integration of emotional signals in the human brain. *Progress in Brain Research, 156*, 345-361.
Farah, M. J., Wong, A. B., Monheit, M. A., & Morrow, L. A. (1989). Parietal Lobe Mechanisms of Spatial Attention - Modality-Specific or Supramodal. *Neuropsychologia, 27*(4), 461-470.
Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.
Foxe, J. J., & Molholm, S. (2009). Ten years at the Multisensory Forum: musings on the evolution of a field. *Brain topography, 21*(3-4), 149-154.
Frick, R. W. (1985). Communicating Emotion - the Role of Prosodic Features. *Psychological Bulletin, 97*(3), 412-429.
Frijda, N. (1989). *The Emotions*. Cambridge: Cambridge University Press.
Fuster, J. M., Bodner, M., & Kroger, J. K. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature, 405*(6784), 347-351.
Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience, 11*(5), 473-490.
Haith, M. M., Bergman, T., & Moore, M. J. (1977). Eye contact and face scanning in early infancy. *Science, 198*, 853-855.
Hay, J. C., Pick, H. L., & Ikeda, K. (1965). Visual Capture Produced by Prism Spectacles. *Psychonomic Science, 2*(8), 215-216.
Held, R. (1965). Plasticity in sensory-motor systems. *Scientific American, 213*, 84-94.
Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.
Lavie, N. (1995). Perceptual Load as a Necessary Condition for Selective Attention. *Journal of Experimental Psychology-Human Perception and Performance, 21*(3), 451-468.
Lavie, N. (2005). Distracted and confused?: selective attention under load. *Trends in cognitive sciences, 9*(2), 75-82.
Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Brain research, 24*(2), 326-334.
Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*(2), 281-308.
Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The Discrimination of Speech Sounds within and across Phoneme Boundaries. *Journal of Experimental Psychology, 54*(5), 358-368.
Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to emotional content of speech. *Journal of the Acoustical Society of America, 34*, 922-927.
Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science (New York, N.Y, 289*(5482), 1206-1208.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin, 109*, 163-203.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in cognitive sciences, 9*(6), 296-305.
Marzi, C. A., Tassinari, G., Aglioti, S., & Lutzemberger, L. (1986). Spatial summation across the vertical meridian in hemianopics: a test of blindsight. *Neuropsychologia, 24*(6), 749-758.
Massaro, D. W. (1987). *Speech perception by ear and by eye: a paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Massaro, D. W. (1998). *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge: MIT Press.
Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review, 3*, 215-221.
Mcgurk, H., & Macdonald, J. (1976). Hearing Lips and Seeing Voices. *Nature, 264*(5588), 746-748.
Mckelvie, S. J. (1995). Emotional Expression in Upside-down Faces - Evidence for Configurational and Componential Processing. *British Journal of Social Psychology, 34*, 325-334.
Mesulam, M. M. (1998). From sensation to cognition. *Brain, 121*, 1013-1052.
Miller, J. (1982). Divided Attention - Evidence for Co-Activation with Redundant Signals. *Cognitive Psychology, 14*(2), 247-279.
Miller, J. (1986). Timecourse of Coactivation in Bimodal Divided Attention. *Perception & Psychophysics, 40*(5), 331-343.
Miniussi, C., Girelli, M., & Marzi, C. A. (1998). Neural site of the redundant target effect electrophysiological evidence. *Journal of cognitive neuroscience, 10*(2), 216-230.
Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain research, 14*(1), 115-128.
Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*, 297-326.
Morris, J. S., Scott, S. K., & Dolan, R. J. (1999). Saying it with feeling: neural responses to emotional vocalizations. *Neuropsychologia, 37*(10), 1155-1163.
Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, 93*, 1097-1108.
Murray, M. M., Foxe, J. J., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2001). Visuo-spatial neural response interactions in early cortical processing during a simple reaction time task: a high-density electrical mapping study. *Neuropsychologia, 39*(8), 828-844.
Osullivan, M., Ekman, P., Friesen, W., & Scherer, K. (1985). What You Say and How You Say It - the Contribution of Speech Content and Voice Quality to Judgments of Others. *Journal of Personality and Social Psychology, 48*(1), 54-62.
Panksepp, J. (2005). Psychology. Beyond a joke: from animal laughter to human joy? *Science (New York, N.Y, 308*(5718), 62-63.
Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin, 116*(2), 220-244.

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science, 17*(4), 292-299.

Pourtois, G., & de Gelder, B. (2002). Semantic factors influence multisensory pairing: a transcranial magnetic stimulation study. *Neuroreport, 13*(12), 1567-1573.

Pourtois, G., de Gelder, B., Bol, A., & Crommelinck, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex; a journal devoted to the study of the nervous system and behavior, 41*(1), 49-59.

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*(6), 1329-1333.

Pourtois, G., Debatisse, D., Despland, P. A., & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Brain research, 14*(1), 99-105.

Pourtois, G., Grandjean, D., Sander, D., & Vuilleumier, P. (2004). Electrophysiological correlates of rapid spatial orienting towards fearful faces. *Cerebral cortex (New York, N.Y., 14*(6), 619-633.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New-York Academy of Sciences, 24*.

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences, 14*(4), 303-316.

Savazzi, S., & Marzi, C. A. (2002). Speeding up reaction time with invisible stimuli. *Current biology, 12*(5), 403-407.

Savazzi, S., & Marzi, C. A. (2004). The superior colliculus subserves interhemispheric neural summation in both normals and patients with a total section or agenesis of the corpus callosum. *Neuropsychologia, 42*(12), 1608-1618.

Savazzi, S., & Marzi, C. A. (2008). Does the redundant signal effect occur at an early visual stage? *Experimental brain research. Experimentelle Hirnforschung, 184*(2), 275-281.

Scherer, K. (1989). Vocal measurement of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: theory, research, and experience* (Vol. 4, pp. 233-259). San-Diego, CA: Academic Press.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and Automatic Human Information-Processing .1. Detection, Search, and Attention. *Psychological Review, 84*(1), 1-66.

Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnson, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature, 385*(6613), 254-257.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*(6614), 308.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408*(6814), 788.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological science, 16*(3), 184-189.
Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge: Bradford Books.
Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in cognitive sciences, 14*(9), 400-410.
Tanaka, J. W., & Farah, M. J. (1993). Parts and Wholes in Face Recognition. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 46*(2), 225-245.
Turatto, M., Mazza, V., Savazzi, S., & Marzi, C. A. (2004). The role of the magnocellular and parvocellular systems in the redundant target effect. *Experimental brain research. Experimentelle Hirnforschung, 158*(2), 141-150.
Vroomen, J., Collier, R., & Mozziconacci, S. (1993). Duration and intonation in emotional speech. *Proceedings of the Third European Conference on Speech Communication and Technology*, 577-580.
Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, affective & behavioral neuroscience, 1*(4), 382-387.
Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences, 9*(12), 585-594.
Vuilleumier, P., & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia, 45*(1), 174-194.
Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: differentiation of multimodal information. *Psychological Bulletin, 121*, 437-456.
Walker, A., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant behavior and development, 6*, 491-498.
Weiskrantz, L. (1986). *Blindsight. A case study and implications*. Oxford: Oxford University Press.
Williams, C. E., & Stevens, K. N. (1972). Emotions and Speech - Some Acoustical Correlates. *Journal of the Acoustical Society of America, 52*(4), 1238-&.

Figure Caption

Figure 1. (A) Mean RTs (± 1 S.E.M) obtained during the block containing control trials (emotional face only) intermixed with either congruent or incongruent audio-visual trials (emotional face + affective tone of voice). Results show a facilitation of perceptual decisions for the central emotional face stimulus, when this face stimulus was paired with a congruent (though unattended) affective tone of voice. (B) Mean RTs (± 1 S.E.M) obtained during the block containing control trials (emotional face only) intermixed with either congruent or incongruent within-modality redundant trials. The visual distracter was either an emotion written word, or another, secondary emotional facial expression

(always shown at a fixed location in the upper visual field). Results show an interference effect created by the presentation of an incongruent (though unattended) affective distracter, as opposed to a benefit in perceptual processing when the exact same emotional face was paired with a congruent affective tone of voice.

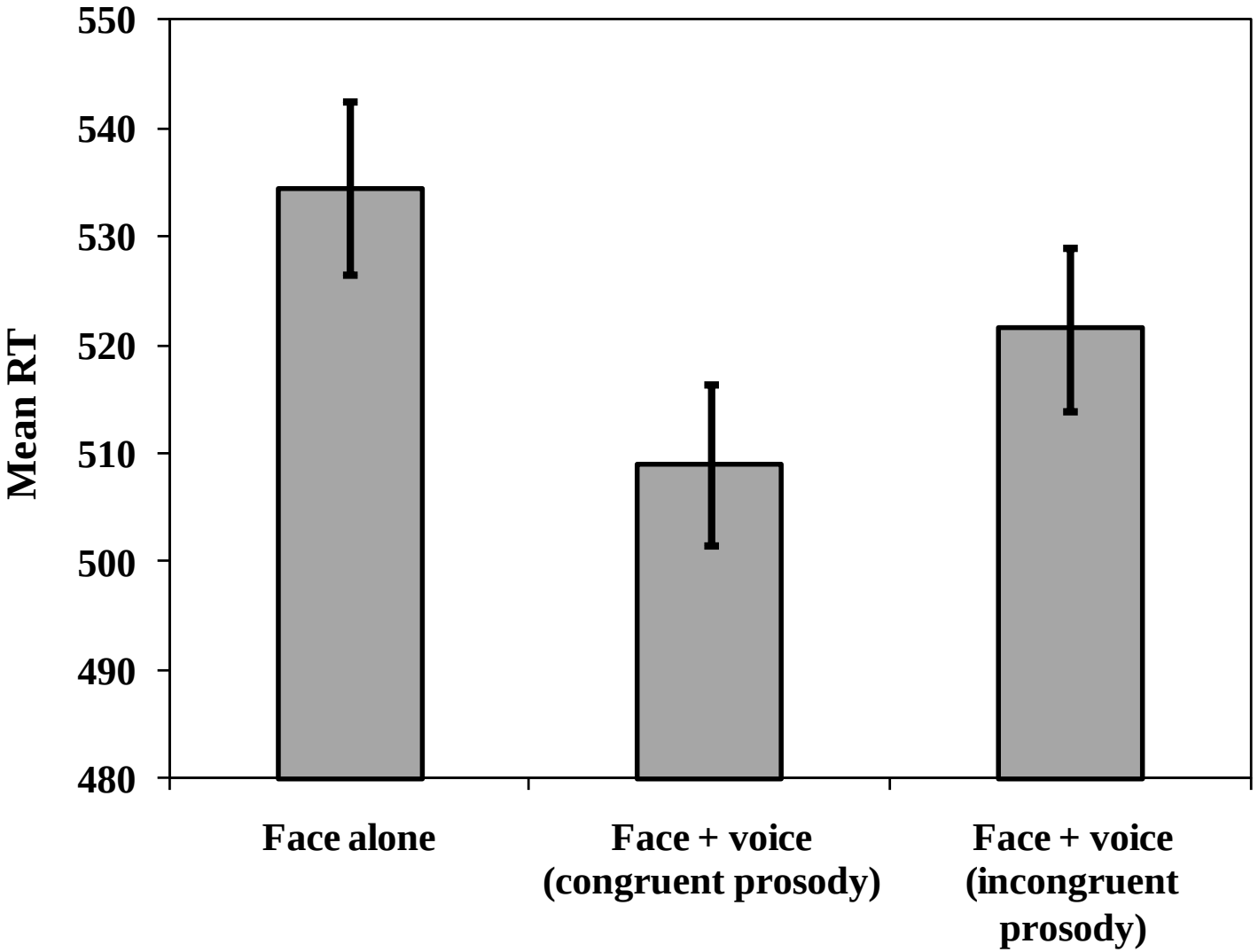
550
540
530
520
510
500
490
480
Mean RT
Face alone
Face + voice (congruent prosody)
Face + voice (incongruent prosody)

Figure 1A

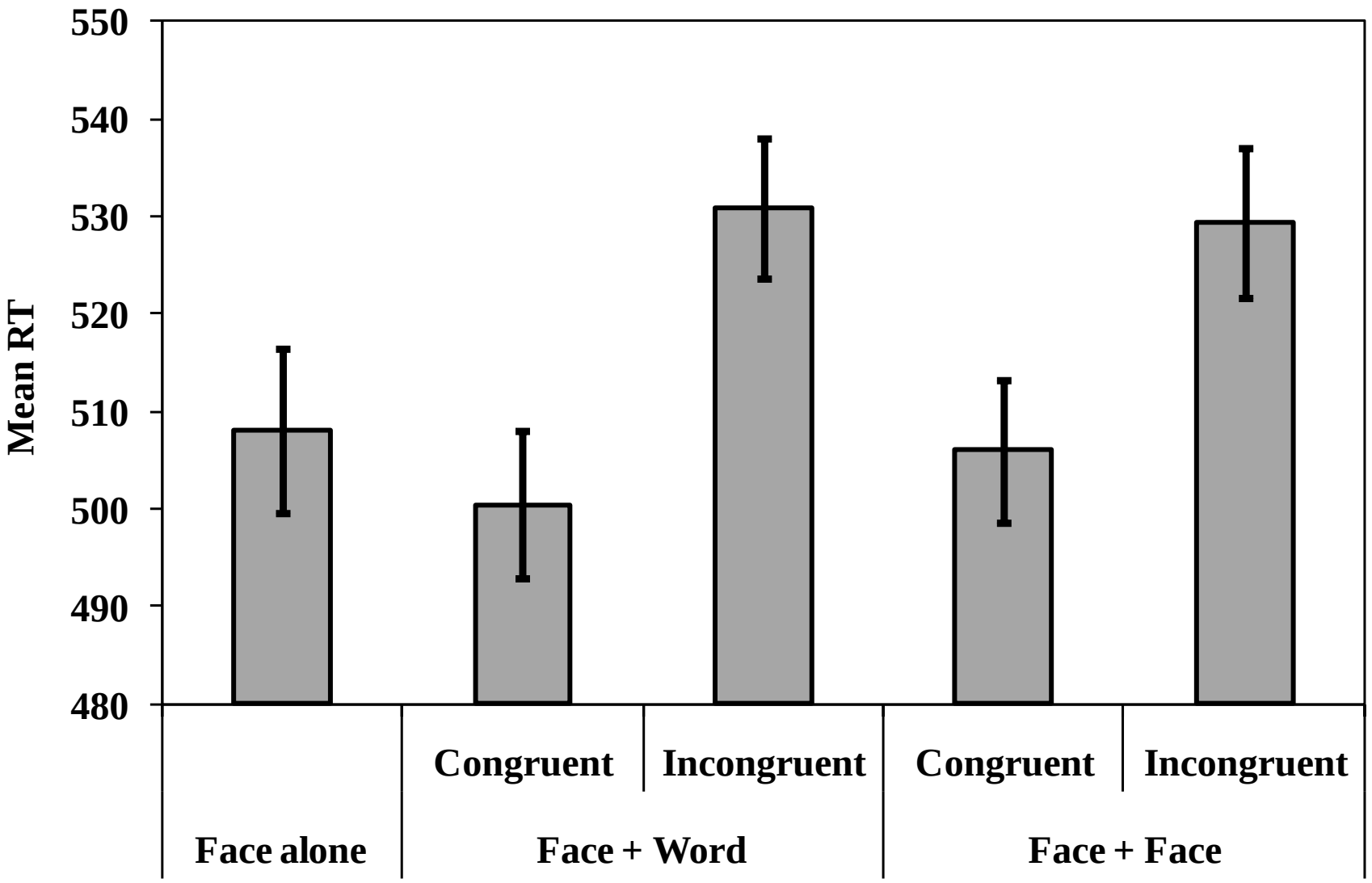
550
540
530
520
510
500
490
480
Mean RT
Congruent
Incongruent
Congruent
Incongruent
Face alone
Face + Word
Face + Face

Figure 1B