36 Multisensory Perception of Emotion, Its Time Course, and Its Neural Basis

BEATRICE DE GELDER, JEAN VROOMEN, AND GILLES POURTOIS

Introduction

Our senses provide information that often appears to arrive simultaneously from the same spatial location but via different modalities, such as when we *observe* a noisily bouncing ball, *hear* a laughing face, or *see* a burning fire and *smell* smoke. To the observer, spatial and temporal contiguity offers a strong incentive to draw together sensory cues as deriving from a single object or event. Cross-talk between the senses is probably adaptive. By reducing stimulus ambiguity and by insulating the organism from the effects of environmental noise, cross-talk between the senses improves performance. At the level of subjective experience, multisensory integration contributes to a sense of self and an intensified presence of the perceiver in his or her world. This aspect of multisensory integration is particularly relevant for multisensory perception of emotion, which is the focus of this chapter. Indeed, disorders of sensory integration have been associated with loss of the sense of self, as has been documented in schizophrenia (Bleuber, 1911; de Gelder, Vroomen, Annen, Masthof, & Hodiamont, 2003; de Gelder, Vroomen, & Hodiamont, 2003).

Audiovisual emotion perception: A new case of pairing based on event identity pairings

Facial expressions and emotional voice expressions are complex visual and auditory stimuli, whereas multisensory research has traditionally addressed very simple phenomena, such as the combined processing of a light flash and a sound beep. It was found that in such combinations, the presentation of a weak light flash enhanced localization of a weak auditory stimulus presented simultaneously. Many researchers in the field of multisensory perception have argued, either implicitly or explicitly, that focusing on simple stimuli is the safest route to understanding more complex stimuli. One well-known exception to this bias in favor of physically simple stimuli is audiovisual speech. Another is the audiovisual perception of emotion, which we present here.

Complex cases are inherently of greater interest because they concern situations that are more typical of the rich environment in which the brain operates. Perhaps of more importance, complex cases are also more likely to correspond to environmental situations that resemble constraints the brain faced in the course of evolution. The simplicity of a stimulus as defined in physical terms is not the same as simplicity as defined from an evolutionary point of view. For example, a square is physically a less complex stimulus than a face, yet the latter is evolutionarily more functional and thus "simpler" than the former. Of course, the overall goal is to apply to the study of complex cases methodological imperatives similar to those applied to the study of simple stimulus combinations in the past. Conversely, the investigation of more complex cases may illuminate important issues that await discussion for the simple cases. One such issue is the need to avoid interference from perceptual strategies of the observer; and this concern is equally relevant for the simple and the more complex cases.

We can approach the issue of constraints by asking what makes the more complex cases different from the better known, simpler ones. For this purpose we introduce a distinction between pairings based on *space-time coordinates* and those primarily based on *event identity* (Bertelson & de Gelder, 2003; de Gelder & Bertelson, 2003). The phenomenon of audiovisual pairing of emotion is one example of audiovisual phenomena where pairing seems to be based on event identity, similar to audiovisual speech pairings. When event identity is at stake, cross-talk between the senses is induced not so much by the requirement that the information arrives within the same space-time window, as is typically the case in laboratory experiments with simple stimulus pairings, but by the fact that each modality contributes to event identification. Recognition of event identity is thus an important ingredient of multisensory perception of complex stimulus pairings. Of course, dependence on time makes good functional sense, even when identity plays an important role. For example, synchrony plays a role in audiovisual speech integration (Bertelson, Vroomen, & de Gelder, 1997; Massaro & Egan, 1996; Munhall, Gribble, Sacco, & Ward, 1996). How temporal and spatial factors interact with contraints on pairing that have their basis in recognition of meaningful events is a topic for future research (de Gelder, 2000; de Gelder & Bertelson, 2003; Frissen & de Gelder, 2002).

A mention of the processes involved in recognition of event identity brings to the foreground higher cognitive processes, which play a more important role in complex event recognition. The multisensory perception of event identity depends to some extent on the perceiver's cognitive and emotional state. For instance, integration might depend on the viewer's beliefs about the likelihood that stimuli originated in a single object, or might even be related to the broader cognitive or motivational context in which the stimuli are presented. Such subjective biases would conflict with a major motivation for studying audiovisual integration, which is that it reflects truly perceptual, automatic, and mandatory processes that are not influenced by an observer's strategies or task settings (for recent discussions see Bertelson, 1999; Bertelson & de Gelder, 2003; Pylyshyn, 1999).

Fortunately, a wealth of recent empirical data supports the notion that stimuli that carry emotional information are perceived nonconsciously. Of course, this perceptual kernel can be integrated in later, more cognitive elaborations (LeDoux, 1996), and according to some definitions, emotions do indeed reflect higher cognitive states. This relation indicates that recognition of emotion has a perceptual basis that is insulated from subjective experience, just as, for example, the perception of color has a perceptual basis. For example, we now know that recognition of emotional stimuli proceeds in the absence of awareness (e.g., Morris, Ohman, & Dolan, 1999; Whalen et al., 1998) and, even more radically, in the absence of primary visual cortex (de Gelder, Pourtois, van Raamsdonk, Vroomen, & Weiskrantz, 2001; de Gelder, Vroomen, Pourtois, & Weiskrantz, 1999). The fact that facial expressions are perceived in a mandatory way is thus a good starting point for investigating whether presenting a facial expression together with an affective tone of voice will have an automatic and mandatory effect on perceptual processes in the auditory modality.

To summarize our discussion to this point, from the perspective of the brain's evolutionary history, auditory and visual expressions of emotion are simple stimuli that can be processed independently of subjective consciousness. Against this background it appears plausible that audiovisual pairing of emotional stimuli proceeds in an automatic and mandatory way even if identity plays an important role in this kind of multisensory event.

Before reviewing the research on this phenomenon, we wish to clarify the distinction between the multisensory perception of emotion, on the one hand, and on the other, similarities in the perception of emotion in visual and auditory modalities. As a first approach, we contrast the specific issues related to the perception of multisensory affect with studies that have investigated similarities between perceiving emotion in either the auditory or the visual modality.

Correspondences between perceiving emotion in faces and voices

The overwhelming majority or studies on human emotion recognition have used facial expressions (for a recent overview, see Adolphs, 2002; see also Adolphs, Damasio, Tranel, & Damasio, 1996), but only a few studies have studied how emotion in the voice is perceived (see Ross, 2000, for a review). Researchers have been interested in finding similarities between face and voice recognition and in acquiring empirical evidence for the existence of a common, abstract processing locus that would be shared by visual and auditory affective processes alike (Borod et al., 2000; Van Lancker & Canter, 1982). In support of this perspective, patients with visual impairments were tested for residual auditory abilities, and vice versa. For some time this research was also conducted within a framework of hemispheric differences, and evidence was adduced for right hemispheric involvement in the perception of facial and vocal expressions. More specific questions targeting individual emotions came to the fore as it became increasingly clear that different types (positive vs. negative) and different kinds of emotions (e.g., fear vs. happiness) are subserved by different subsystems of the brain (Adolphs et al., 1996).

It is fair to say that at present, there is no consensus on the existence of a dedicated functional and neuroanatomical locus where both facial and vocal expressions might be processed. A strong case was initially made for a role of the amygdala in recognizing facial as well as vocal expressions of fear, but so far studies have not yielded consistent results (Scott et al., 1997; but see Adolphs & Tranel, 1999; Anderson & Phelps, 1998).

Moreover, researchers have usually approached the issue of a common functional and possibly neuroanatomical basis for recognition of vocal and facial expressions by looking for correlations (Borod et al., 2000). But such data are not directly useful to understanding how, when both the face and the voice are present (as is often the case in natural conditions), the two information streams are actually integrated. Whether or not amodal or abstract representation systems exist that use representations that transcend either modality (but that can be shared by both, and thereby play a role in processing auditory as well as visual information) is at present an empirical question. It is also worth noting that the issue of a common basis reappears in a very different context, as we will see when discussing models of audiovisual integration.

Behavioral experiments measuring cross-modal bias between facial and vocal expressions: A first review and some methodological issues

In our first behavioral studies of intersensory perception of affect, we adapted a paradigm frequently used in studies of audiovisual speech (Massaro, 1998). We combined a facial expression with a short auditory vocal segment and instructed participants to attend to and categorize either the face or the voice, depending on the condition. These experiments provided clear evidence that an emotional voice expression that is irrelevant for the task at hand can influence the categorization of a facial expression presented simultaneously (de Gelder, Vroomen, & Teunisse, 1995). Specifically, participants categorizing a happy or fearful facial expression were systematically influenced by the expression of the voice (e.g., the face was judged as less fearful if the voice sounded happy) (Fig. 36.1). Massaro and Egan (1996) obtained similar results using a synthetic facial expression paired with a vocal expression. Subsequently we explored the situation in which participants were asked to ignore the face but had to rate the expression of the voice. A very similar cross-modal effect was observed for recognition of the emotional expression in the voice (de Gelder & Vroomen, 2000, Experiment 3). The effect of the face on voice recognition disappeared when facial images were presented upside down, adding further proof that the facial expression was the critical variable (de Gelder, Vroomen, & Bertelson, 1998). It is worth noting that these bias effects were obtained with stimulus pairs consisting of a static face and a vocal expression. With the exception of one behavioral study done in a cortically blind patient, in which we used short video clips with mismatched voice fragments (de Gelder, Pourtois, Vroomen, & Weiskrantz, 1999), all effects were obtained with static facial images. One might object that such pairs are not entirely ecological, because in naturalistic situations the voice one hears belongs to a moving face. But support for the use of static stimuli is provided by studies indicating that the perception of emotion in faces is linked to processes inducing imitation (Dimberg, Thunberg, & Elmehed, 2000) and experiencing of the emotion (Adolphs, Damasio, Tranel, Cooper, & Damasio, 2000).

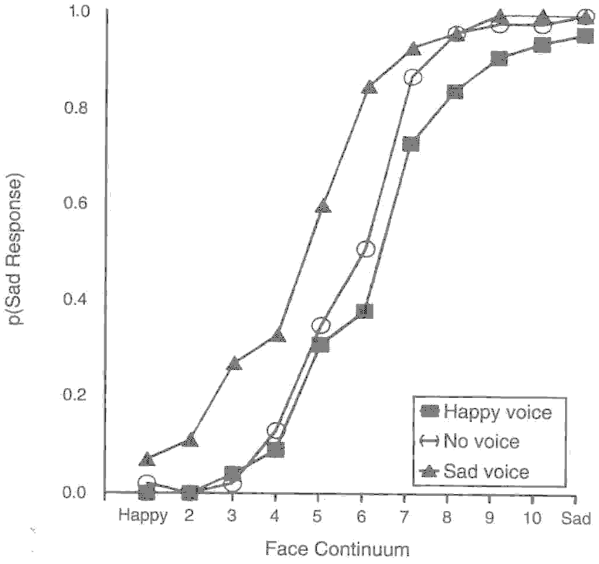

FIGURE 36.1 Behavioral results. A cross-modal bias effect from the voice to the face is indicated by a systematic response bias toward the concurrent tone of voice when participants were asked to judge the emotion in the face.

In the behavioral experiments, we used the paradigm of *cross-modal bias,* one of the few classical paradigms known to provide evidence for intersensory integration. (Other useful paradigms include the study of aftereffects and of intersensory fusion; for a discussion, see Bertelson, 1999; Bertelson & de Gelder, 2003.) Cross-modal bias measures the on-line effect of intersensory conflict on the task-relevant input modality; in other words, it measures how processing in the task-relevant modality is biased in the direction of the information presented in the other modality. As such, it is a measure of on-line conflict resolution. In contrast, studies of aftereffects investigate the consequences of intersensory conflict resolution by observing how perception in one modality is more or less permanently changed as a consequence of intersensory conflict. A major advantage of this approach is its robustness in the light of possible confounds. When measuring aftereffects one need be much less concerned with the possibility that the results are contaminated by the perceiver's perceptual bias.

Until now, aftereffects were exclusively measured in studies of spatial and temporal aspects of intersensory conflict. Only recently was the methodology applied successfully to the study of conflict at the level of stimulus identity for speech (Bertelson, Vroomen, & de Gelder, 2003). Because of its robustness, it is also a good tool for studying other situations of cross-talk between the senses due to presumed event identity effects. We are currently exploring the aftereffects of exposure to audiovisual conflict in affective pairs.

In cross-modal bias paradigms, the behavioral measures are accuracy and latency (see Massaro, 1999, for a full discussion). These measures can be influenced by factors such as attention, task demands, or subjective response bias. The situation of multisensory paradigms is quite similar to that used in studies of the redundant target effect (Miller, 1982, 1986). In both cases the experimenter measures the impact of a task-irrelevant secondary stimulus on the accuracy and latency of performance. By themselves, differences in either measure are not sufficient to demonstrate that multisensory integration has occurred and that performance primarily reflects the observer's experience of a single percept. Moreover, the typical effects on behavior of the presentation of a secondary stimulus as manifested in latency and accuracy are also observed when two stimuli are presented simultaneously and the secondary stimulus is presented in the same modality as the target. This is typically the case when a secondary visual stimulus is presented together with a primary visual stimulus. It is difficult to decide whether the observed behavioral effects reflect multisensory perceptual integration or instead are due to the influence of the secondary stimulus on how the task is performed. Unfortunately, such matters cannot be resolved by showing that the data fit a mathematical model, such as the FMLP (Massaro & Egan, 1996). More generally, such models do not distinguish well between effects that result from early integration and late decision-based ones (de Gelder & Vroomen, 2000). Behavioral methods need to be applied in concert with neurophysiological methods to make progress here.

A role for attention in audiovisual emotion perception

Cognitive theories of attention, like the one defended by Treisman (1996), predict that attention plays a critical role in combining isolated cues present in different modalities, thus making attention the prime candidate for bringing about intersensory integration. In other words, if attention plays a critical role, we would no longer observe a cross-modal bias when subjects are entirely focused on the task related to one modality and not paying attention to information in the other modality that is irrelevant to the task at hand. When this is indeed the case, it follows that for some intersensory processes, which one would then want to qualify as genuinely perceptual, attention does not play a critical role in bringing about integration or explaining its effects. In other words, if cross-modal integration of affective information is a truly automatic process, it should take place regardless of the demands of an additional task. Indeed, independence from demands on attentional capacity has long been one of the defining characteristics of "automatic" processes (see Shiffrin & Schneider, 1977). For example, we considered the role of either exogenous (reflexive, involuntary) or endogenous (voluntary) attention in ventriloquism. We found no evidence that either played a role in sound to visual location attraction (Bertelson, Vroomen, de Gelder, & Driver, 2000; Vroomen, Bertelson, & de Gelder, 2001).

To address whether attention plays a causal role in bringing about multisensory integration, one should really ask whether, if participants had to judge the voice, there would be a cross-modal effect from the facial to the vocal judgments if the face were not attended to. Recent research on attention has shown that irrelevant visual stimuli may be particularly hard to ignore under "low-load" attention conditions, yet they can be successfully ignored under higher-load conditions in which the specified task consumes more attentional capacity (e.g., Lavie, 1995, 2000). The facial expression in de Gelder and Vroomen's study (2000, Experiment 3), although irrelevant for the task, in which participants were required to judge emotion in the voice, might therefore have been unusually hard to ignore, due to the low-load attention of the task situation (the face was the only visual stimulus present, and the only task requirement was categorization of the voice). It is thus possible that the influence of the seen facial expression on judgments of the emotional tone of a heard voice would be eliminated under conditions of higher attentional load (e.g., with additional visual stimuli present and with a demanding additional task).

Attention as a possible binding factor was studied in a dual-task format in which we asked whether cross-modal integration of affective information would suffer when a demanding task had to be performed concurrently (Vroomen, Driver, & de Gelder, 2001). A positive result, meaning an effect of attentional load on the degree of integration taking place, would suggest that attentional resources are required for cross-modal integration of emotion to occur. If, on the other hand, a competing task did not influence performance, it is reasonably safe to assume that cross-modal interactions do not require attentional resources (Kahneman, 1973). Thus, we measured the influence of a visible static facial expression on judgments of the emotional tone of a voice (as in de Gelder & Vroomen, 2000) while varying attentional demands by presenting participants with an additional attention-capturing task.

A general concern in applying the dual-task method is whether tasks compete for the same pool of resources or whether there are multiple resource pools each of which deals separately with the various cognitive and perceptual aspects of the two tasks (Wickens, 1984). When tasks do not interfere, it may be that one of the tasks (or both) does not require any attentional resources (i.e., they are performed automatically), or it may be that they draw on different resource pools. To distinguish between these alternatives, we varied the nature of the additional task. If none of the different tasks interfered with the cross-modal interactions, this result would suggest that the cross-modal effect itself does not require attention. Participants judged whether a voice expressed happiness or fear while trying to ignore a concurrently presented static facial expression. As an additional task, participants were instructed to add two numbers together rapidly (Experiment 1), or to count the occurrences of a target digit in a rapid serial visual presentation (Experiment 2), or judge the pitch of a tone as high or low (Experiment 3). The face had an impact on judgments of the heard voice emotion in all experiments. This cross-modal effect was independent of whether or not subjects performed a demanding additional task. This result indicates that the integration of visual and auditory information about emotions is a mandatory process, unconstrained by attentional resources. It is also worth pointing out at this stage that in order to rule out response bias in later studies, we used an orthogonal task, for example sex classification, which did not require attending to the emotional content.

Recent neurophysiological techniques provide means of looking at intersensory fusion before its effects are manifest in behavior. Each of these methods has its limits. Theoretical conclusions that strive to be general and to transcend particular techniques, whether behavioral or neurophysiological, about multisensory perception require convergence from different methods. (For example, it is still difficult to relate a measure such as a gain in latency observed in behavioral studies and an increase in blood-oxygen-level-dependent [BOLD] signal measured on functional magnetic resonance imaging (fMRI), or the degree of cell firing and degree of BOLD signal.) Our first studies used electroencephalographic (EEG) recordings to acquire insight into the time course of integration. Our goal was to reduce the role of attention and of stimulus awareness through the use of indirect tasks. The focus on early effects and the selective study of clinical populations with specific deficits helped to clarify some central aspects of automatic multisensory perception of affect.

Electrophysiological studies of the time course of multisensory affect perception

To investigate the time course of face-voice integration of emotion, we exploited the high temporal resolution provided by event-related brain potentials (ERPs) and explored its neuroanatomical location using source localization models. Electrophysiological studies (either EEG or MEG) of multisensory perception have indicated large amplitude effects, sometimes consisting in an increase and at other times in a decrease of early exogenous components such as the auditory N1 or the visual P1 component (each generated around 100 ms for stimulus presentation in their respective modality) during presentation of multisensory stimuli (Foxe et al., 2000; Giard & Peronnet, 1999; Raij, Uutela, & Hari, 2000; Rockland & Ojima, 2003; Sams et al., 1991). Amplification of the neural signal in modality-specific cortex is thought to reflect an electrophysiological correlate of intersensory integration and has been observed when responses to multisensory presentations are compared with responses to the single-modality presentations individually. On the other hand, a decrease in amplitude is sometimes observed when the comparison focuses specifically on the contrast between congruent versus incongruent bimodal conditions (de Gelder, Pourtois, & Weiskrantz, 2002).

In our first study we used the phenomenon of *mismatch negativity* (MMN; Näätänen, 1992) as a means of tracing the time course of the combination of the affective tone of voice with information provided by the expression of the face (de Gelder, Böcker, Tuomainen, Hensen, & Vroomen, 1999). In the standard condition subjects were presented with concurrent voice and face stimuli with identical affective content (a fearful face paired with a fearful voice). On the anomalous trials the vocal expression was accompanied by a face with an incongruent expression. We reasoned that if the system was tuned to combine these inputs, as was suggested by our behavioral experiments, and if integration is reflected by an influence of the face on how the voice is processed, this would be apparent in some auditory ERP components. Our results indicated that when presentations of a voice-face pair with the same expression were followed by presentation of a pair in which the voice stimulus was the same but the facial expression was different, an early (170 ms) deviant response with

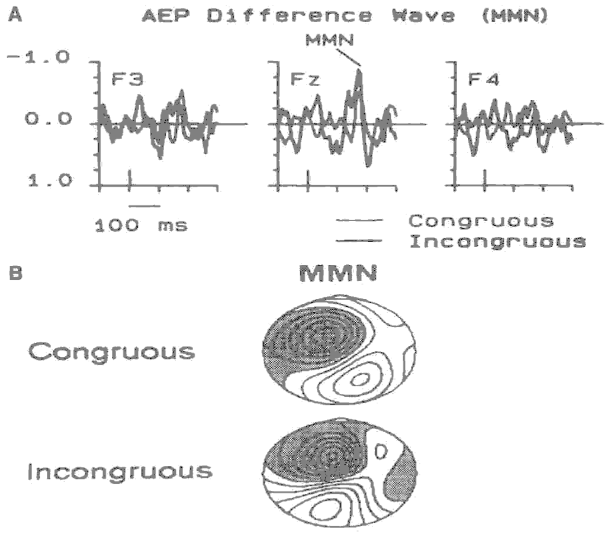

FIGURE 36.2 EEG results. (*A*) The grand averaged deviant − standard difference wave of the auditory brain potential for both congruent (thin line) and incongruent (thick line) face-voice pairs. Surplus negativity evoked by the deviant stimulus pair is plotted upward. The vertical bar on the *x*-axis indicates the onset of the auditory stimulus. Amplitude (μV) is plotted on the *y*-axis. (*B*) The isopotential map (0.1 μV between lines; shaded area negative) at 178 ms, showing the scalp distribution of the mismatch negativity (MMN) for voices with congruent and incongruent faces.

frontal topography was elicited (Fig. 36.2). This response strongly resembles the MMN, which is typically associated with a change (whether in intensity, duration or location) in a train of standard, repetitive auditory stimuli (Näätänen, 1992). Our results are consistent with previous EEG results (Surakka, Tenhunen-Eskelinen, Hietanen, & Sams, 1998) showing that pitch MMN may be influenced by the simultaneous presentation of positive nonfacial stimuli (colored light flashes).

Converging evidence for the integrated perception of facial expressions and spoken sentence fragments was provided in a follow-up EEG study (Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000) in which we measured the early effects of adding a facial expression to a sentence fragment spoken in an emotional tone of voice. Significant increases in the amplitude of P1 and auditory N200 were obtained for congruent face-voice pairs but not for incongruent ones, or for pairs in which the face was presented upside down. In a subsequent study (Pourtois, Debatisse, Despland, & de Gelder, 2002), we showed that congruent face-voice pairs elicited an earlier P2b component (the P2b component follows the auditory P2 exogenous component with a more posterior topography) than incongruent pairs (Fig. 36.3; Color Plate 17). This result suggests that the processing of affective prosody is delayed in the presence of an incongruent facial context. The temporal and topographical properties of the P2b component suggest that this component does not overlap with EEG components (e.g., the N2-P3 complex) known to be involved in cognitive processes at later decisional stages. Source localization carried out on the time window of the P2b component disclosed a single dipole solution in anterior cingulate cortex, an area selectively implicated in processing congruency or conflict between stimuli (MacLeod & MacDonald, 2000). The contribution of anterior cingulate cortex in dealing with perceptual and cognitive congruency has been shown in many previous brain imaging studies (Cabeza & Nyberg, 2000). The anterior cingulate is also one of the areas strongly associated with human motivational and emotion processes (Mesulam, 1998).

Neuroanatomy of audiovisual perception of emotion

There are as yet only few general theoretical suggestions in the literature concerning the neuroanatomical correlates of multisensory integration (e.g., Damasio, 1989; Ettlinger & Wilson, 1990; Mesulam, 1998). Recent studies have considered a variety of audiovisual situations, including arbitrarily associated pairs (Fuster, Bodner, & Kroger, 2000; Giard & Peronnet, 1999; Schröger & Widmann, 1998), naturalistic pairs, such as audiovisual speech pairs (Calvert et al., 1999; Raij et al., 2000), and audiovisual affect pairs (de Gelder, Böcker, et al., 1999; Dolan, Morris, & de Gelder, 2001; Pourtois et al., 2000).

Our first study directly addressing the audiovisual integration of emotion with brain imaging methods (fMRI) suggested that an important element of a mechanism for such cross-modal binding in the case of fearful face-voice pairs is be found in the amygdala (Dolan et al., 2001). In this study, subjects heard auditory fragments paired with either a congruent or an incongruent facial expression (happiness or fearfulness) and were instructed to judge the emotion from the face. When fearful faces were accompanied by short sentence fragments spoken in a fearful tone of voice, an increase in activation was observed in the amygdala (Fig. 36.4; Color Plate 18) and the fusiform gyrus. The increased amygdala activation suggests binding of face and voice expressions (Goulet & Murray, 2001), but further research is needed to investigate the underlying mechanism. Unlike in our behavioral studies, no advantage was observed for happy pairs. This could suggest that the rapid integration across modalities is not as automatic for happy expressions as it is for fear signals.

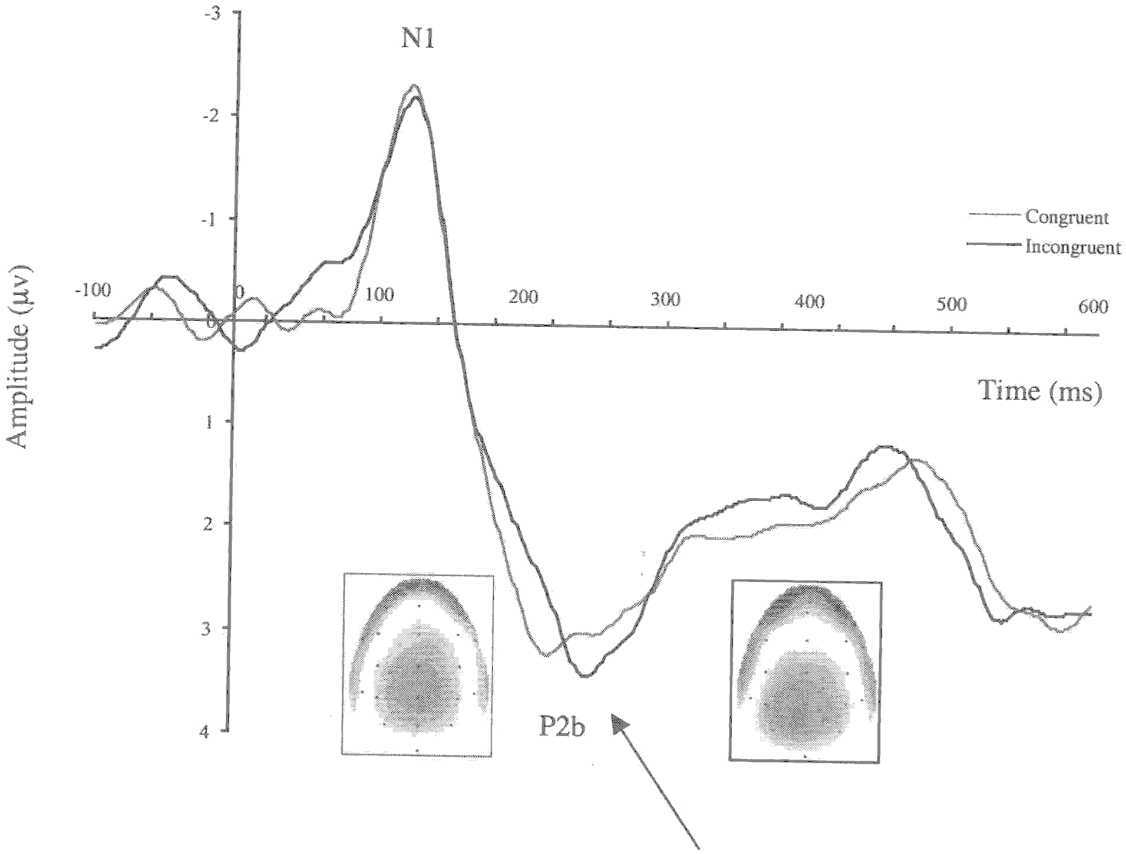

FIGURE 36.3 EEG results. Grand averaged auditory waveforms at the CPz electrode measured during the presentation of congruent and incongruent audiovisual stimulus pairs (and corresponding topographies at 224 ms and 242 ms in the congruent and incongruent condition, respectively). Auditory processing is delayed in time by around 220 ms when realized under an incongruent visual context. (See Color Plate 17.)

More generally, such a finding is consistent with increasing awareness of the neurobiological specificity of each emotion.

In a follow-up study (Pourtois, de Gelder, Bol, & Crommelinck, 2003) performed using $H_2^{15}O$ PET, we compared activations to unimodal stimuli and to bimodal pairs in order to find areas specifically involved in audiovisual integration. We also investigated whether activation in heteromodal areas would be accompanied by increased activation in modality-specific cortices (such as the primary auditory cortex or primary visual cortex). The latter phenomenon has been reported previously and can tentatively be viewed as a downstream consequence in modality-specific cortex of multisensory integration (see Calvert, Campbell, & Brammer, 2000; de Gelder, 2000; Dolan et al., 2001; Driver & Spence, 2000; Macaluso, Frith, & Driver, 2000). Such feedback or top-down modulations could be the correlate of the well-known cross-modal bias effects typically observed in behavioral studies of audiovisual perception (Bertelson, 1999; Bertelson & de Gelder, 2003). But some of these effects might in part depend on attention to the task-related modality. The fact that attentional demands can modulate the effects of multisensory integration is still entirely consistent with the notion that attention itself is not the basis of intersensory integration (for discussion, see Bertelson et al., 2000; de Gelder, 2000; McDonald, Teder-Sälejärvi, & Ward, 2001; Vroomen, Bertelson, et al., 2001, for a discussion). To avoid attentional modulation, we used a gender decision task that does not require attention to the emotion expressed, whether in the voice, the face, or both. Our main results suggest that the perception of audiovisual emotions activates a cortical region situated in the left middle temporal gyrus (MTG) and the left anterior fusiform gyrus (Fig. 36.5; Color Plate 19). The MTG has already been shown to be involved in multisensory integration (Streicher & Ettlinger, 1987) and has been described as a convergence region between multiple modalities (Damasio, 1989; Mesulam, 1998).

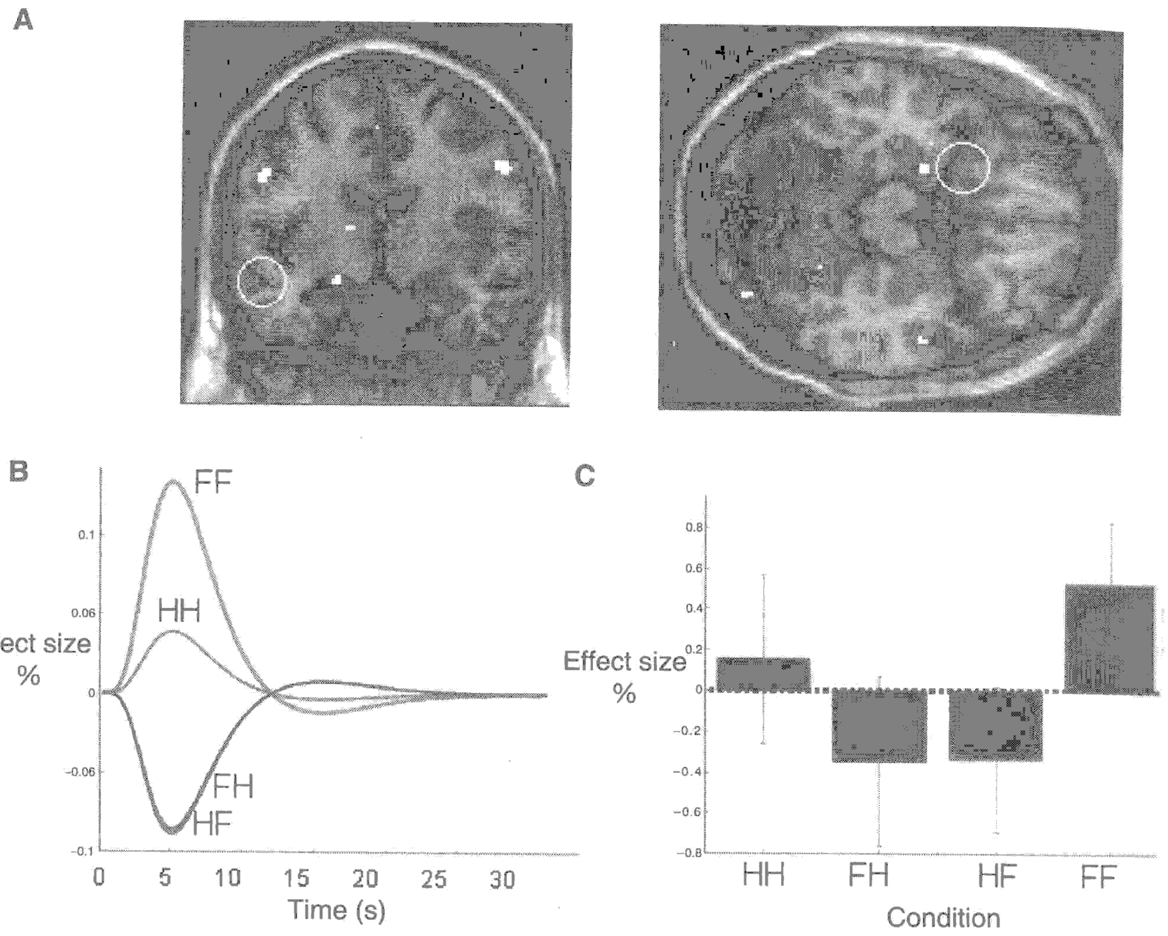

FIGURE 36.4 fMRI results. A statistical parametric map shows an enhanced response in the left amygdala in response to congruent fearful faces plus fearful voices. Condition: H, happy; F, fearful. (See Color Plate 18.)

Activation of the fusiform gyrus is consistent with the results of our previous fMRI study (Dolan et al., 2001) of recognition of facial expression paired with tones of voice. Interestingly, our results showed that within the left MTG (BA 21), there was no difference between visual and auditory levels of activation, but instead there was a significant increase for the audiovisual condition compared with the unimodal conditions. Of note, activation in the left MTG did not correspond to an increase in regional cerebral blood flow (rCBF) in regions that are modality-specific. Moreover, activations were also observed separately for the two emotions when visual and auditory stimuli were presented concurrently. "Happy" audiovisual trials activate different frontal and prefrontal regions (BA 8, 9, 10, and 46) that are all lateralized in the left hemisphere, while audiovisual "fear" activates the superior temporal gyrus in the right hemisphere, confirming strong hemispheric asymmetries in the processing of positively (pleasant) versus negatively (unpleasant) valenced stimuli (see Davidson & Irwin, 1999, for a review).

An intriguing possibility is that presentation in one modality activates areas typically associated with stimulation in the other modality. For example, using fMRI we investigated auditory sadness and observed strong and specific orbitofrontal activity. Moreover, in line with the possibility just raised, among the observed foci there was strong activation of the left fusiform gyrus, an area typically devoted to face processing, following presentation of sad voices (Malik, de Gelder, & Breiter, in prep.). A similar effect was oberved by Calvert and collaborators studying audiovisual speech. They found activation of auditory cortex following presentation of silent speech movements (Calvert et al., 1997). For the case of emotion, such effects also make sense when one takes into account that sensorimotor cortex plays a role in the perception of emotional expressions of the face (Adolphs, 2002).

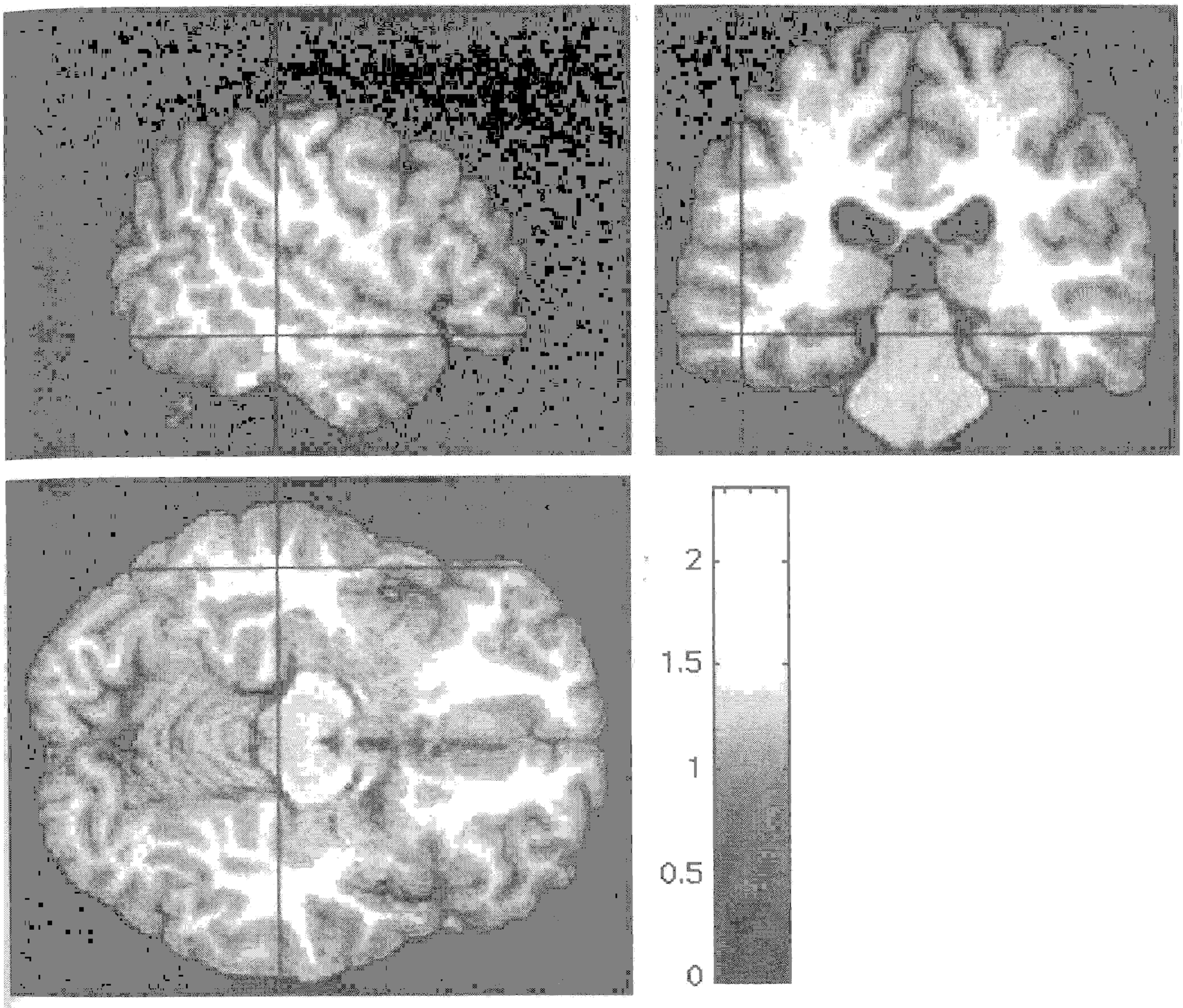

FIGURE 36.5 PET results. Coronal, axial, and sagittal PET sections showing significant activation of a multisensory region in the left middle temporal gyrus ($-52x$, $-30y$, $-12z$) in eight normal subjects during audiovisual trials (happy and fearful emotions) compared with unimodal trials (Visual + Auditory). (See Color Plate 19.)

Much more research is needed before we can begin to understand what is common to the different instances of audiovisual integration. A small number of candidate areas for audiovisual integration have been reported so far, but it is difficult to generalize from the available evidence. Comparisons are also complicated by differences in paradigms, in the choice of baseline, and in the use of different control conditions in each of these studies. For example, arbitrary audiovisual pairs have been used as a control condition for audiovisual speech pairs (Raij et al., 2000; Sams et al., 1991), but other studies have used meaningless grimaces (Calvert et al., 1997). In other studies the sum of unimodal activation in each modality has been used as a baseline (Calvert et al., 2000). Also, different analytic strategies are available, such as strategies that contrast unimodal versus multimodal situations or bimodal congruent situations with bimodal incongruent ones.

Selectivity in audiovisual affect pairing

How selective is the pairing mechanism underlying pairings in which event identity plays a critical role? So far we have mentioned studies of audiovisual affect in which the visual stimuli consisted of facial expressions. Such pairings are based on congruence in stimulus identity—that is, emotional meaning—across the two input modalities. However, other visual stimuli, such as objects and pictures of visual scenes, also carry affective information and are equally conspicuous in the daily environment.

Similarly, there are other sources of auditory affect information besides affective prosody, the most obvious candidates being word meaning and nonverbal auditory signals. If semantic relationship were the only determinant of identity-based pairings, either of those visual inputs should combine with either of those two alternative auditory messages. Selectivity is an important issue for identity pairings, and learning more about it should reveal important insights into the biological basis of multisensory perception. On the other hand, we do not know to what extent the familiar boundaries of our own species are the limits of our biological endowment. For example, for human observers, the facial expressions of higher apes might bind more easily with vocalizations than visual scenes do because vocalizations share more biologically relevant properties with human faces (de Gelder, van Ommeren, & Frissen, 2003).

The matter of selectivity has recently been investigated in a number of ways. In one ERP study, the possible biological basis of identity pairings was explored by contrasting two kinds of auditory components, prosodic and semantic, for the same visual stimulus. Our goal was to find indicators for the difference between the two kinds of auditory stimuli combined with the same visual stimulus. We contrasted the impact of affective prosody (Prosodic condition) versus word meaning (Semantic condition) of a spoken word on the way a facial expression was processed. To test for the presence of specific audiovisual responses in the EEG at the level of the scalp, we compared ERPs to audiovisual trials (AV) with brain responses for visual plus auditory stimuli trials (A + V) (Barth, Goldberg, Brett, & Di, 1995). Subjects performed a gender discrimination task chosen because it was unrelated to the effects studied. In both conditions, ERPs for AV trials were higher in amplitude than ERPs for A + V trials. This amplification effect was manifested for early peaks with a central topography, such as N1 (at 110 ms), P2 (at 200 ms), and N2 (at 250 ms). However, our results indicated that the time course of responses face-voice pairings differs from that of responses to face-word pairings. The important finding was that the amplification effect observed for AV trials occurred earlier for face-voice pairings than for face-word parings. The amplification effect associated with AV presentations and manifested at the level of the scalp was already observed at around 110 ms in the Prosodic condition, whereas this effect was at its maximum later (around 200 ms) in the Semantic condition. These results suggest different nonoverlapping time courses for affective prosodic pairing versus semantic word pairing.

We subsequently addressed the same issue using a different technique, single-pulse transcranial magnetic stimulation (TMS) (Pourtois & de Gelder, 2002). Two types of stimulus pairs were compared, one consisting of arbitrary paired stimuli in which the pairing was learned, and the other consisting of natural pairings as described above. Participants were trained on the two types of pairs to ensure that the same level of performance was obtained for both. Our hypothesis was that TMS would interfere with cross-modal bias obtained with meaningless shape-tone pairs (Learned condition) but not with the cross-modal bias effect of voice-face pairs (Natural condition). Single-pulse TMS applied over the left posterior parietal cortex at 50, 100, 150, and 200 ms disrupted integration at 150 ms and later, but only for the learned pairs. Our results suggest that content specificity as manipulated here could be an important determinant of audiovisual integration (Fig. 36.6; Color Plate 20). Such a position is consistent with some recent results indicating domain specific sites of intersensory integration. Content is likely to represent an important constraint on audiovisual integration.

A different way of investigating selectivity is by looking at possible contrasts between pairings that are presented under conditions of normal stimulus awareness and outside the scope of visual consciousness. This approach might be particularly sensitive to the contrast between facial expressions and emotional scenes, given the special evolutionary status of faces. We comment on this research line in the next section, in a discussion on consciousness.

Qualitative differences between conscious and nonconscious audiovisual perception

An important aspect of audiovisual integration is the role of stimulus awareness. This is also a dimension that has so far rarely been considered but that appears important in light of findings about the unconscious processing of facial expressions (de Gelder, Vroomen, et al., 1999; Morris et al., 1999; Whalen et al., 1998). If we can obtain evidence that unseen stimuli, or at least stimuli the observer is not aware of, still exert a cross-modal bias, then the case for an automatic, mandatory perceptual phenomenon is even stronger. By the same token, the requirement that audiovisual bias should be studied in situations that are minimally transparent to the observer is met equally well when observers are unaware of the second element of the stimulus pair. Patients with visual agnosia that includes an inability to recognize facial expressions pose a unique opportunity for investigating this issue. We studied a patient with visual agnosia and severe facial recognition problems due to bilateral occipitotemporal damage (Bartolomeo et al., 1998; de Gelder, Pourtois, Vroomen, & Bachoud-Levi, 2000; Peterson,

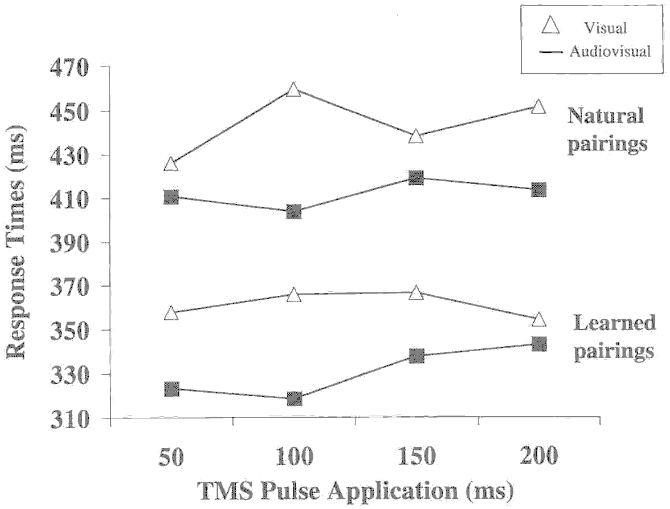

FIGURE 36.6 Results of the TMS experiment. Response times are plotted as a function of the SOA between stimulus presentation and pulse deliverance (a single pulse was delivered 50, 100, 150, or 200 ms after stimulus presentation) when TMS is applied over the left posterior parietal cortex. In the Learned condition (burst tone paired with geometrical figure), there was significant interaction between Modality × SOA, indicating that audiovisual trials are not faster than visual trials at 200 ms. In the Natural condition (tone of voice paired with facial expression), the interaction Modality × SOA is not significant. (See Color Plate 20.)

de Gelder, Rapcsak, Gerhardstein, & Bachoud-Levi, 2000). Her recognition of facial expressions is almost completely lost, although recognition of emotions in voices is intact. This combination allowed us to look at spared covert recognition of facial expressions, as we could use a cross-modal bias paradigm. With this indirect testing method we found clear evidence of covert recognition, as her recognition of emotions in the voice was systematically affected by the facial expression that accompanied the voice fragment she was rating (de Gelder et al., 2000).

Patients with hemianopia but who have retained some residual visual abilities (see Weiskrantz, 1986, 1997) offer an even more radical opportunity to study audiovisual integration under conditions in which there was no awareness of the stimuli presented. Moreover, such cases offer a window onto the neuroanatomy of nonconscious visual processes and the role of striate cortex in visual awareness. Phenomenologically, these patients manifest the same pattern as patients with visual agnosia, because in both cases, conscious recognition of the facial expression is impaired. In the hemianopic patient G. Y. we found behavioral and electrophysiological evidence for a cross-modal bias of unseen facial expressions on processing of the emotion in the voice (de Gelder, Vroomen, & Pourtois, 2001). More recently we looked at a possible interaction between awareness and type of audiovisual pairing (de Gelder et al., 2002). For this purpose we designed two types of pairs, each with a different visual component. One type of pair consisted of facial expression/voice pairs and the other type consisted of emotional scene/voice pairs. In this study, the face/voice pairs figured as the natural pairings and the scene/voice pairs as the semantic ones. Intersensory integration was studied in two hemianopic patients with a complete unilateral lesion of the primary visual cortex. ERPs were measured in these two patients, and we compared the pattern obtained in the intact hemisphere, with patients conscious of the visual stimuli, with that obtained in the blind hemisphere, where there was no visual awareness. We explored the hypothesis that unlike natural pairings, semantic pairings might require conscious perception and mediation by intact visual cortex (possibly based on feedback to primary cortex from higher cognitive processes). Our results indicate that adding visual affective information to the voice results in an amplitude increase of auditory-evoked potentials, an effect that obtains for both natural and semantic pairings in the intact field but is limited to the natural pairings in the blind field (Fig. 36.7; Color Plate 21). These results are in line with previous studies that have provided evidence in favor of qualitatively different processing systems for conscious and nonconscious perception (LeDoux, 1996; Weiskrantz, 1997).

With the possibility of different systems for conscious and nonconscious processes, some novel and intriguing possibilities arise. First, it is in principle possible that conflicts between the two systems could arise when two different stimuli are simultaneously presented in the

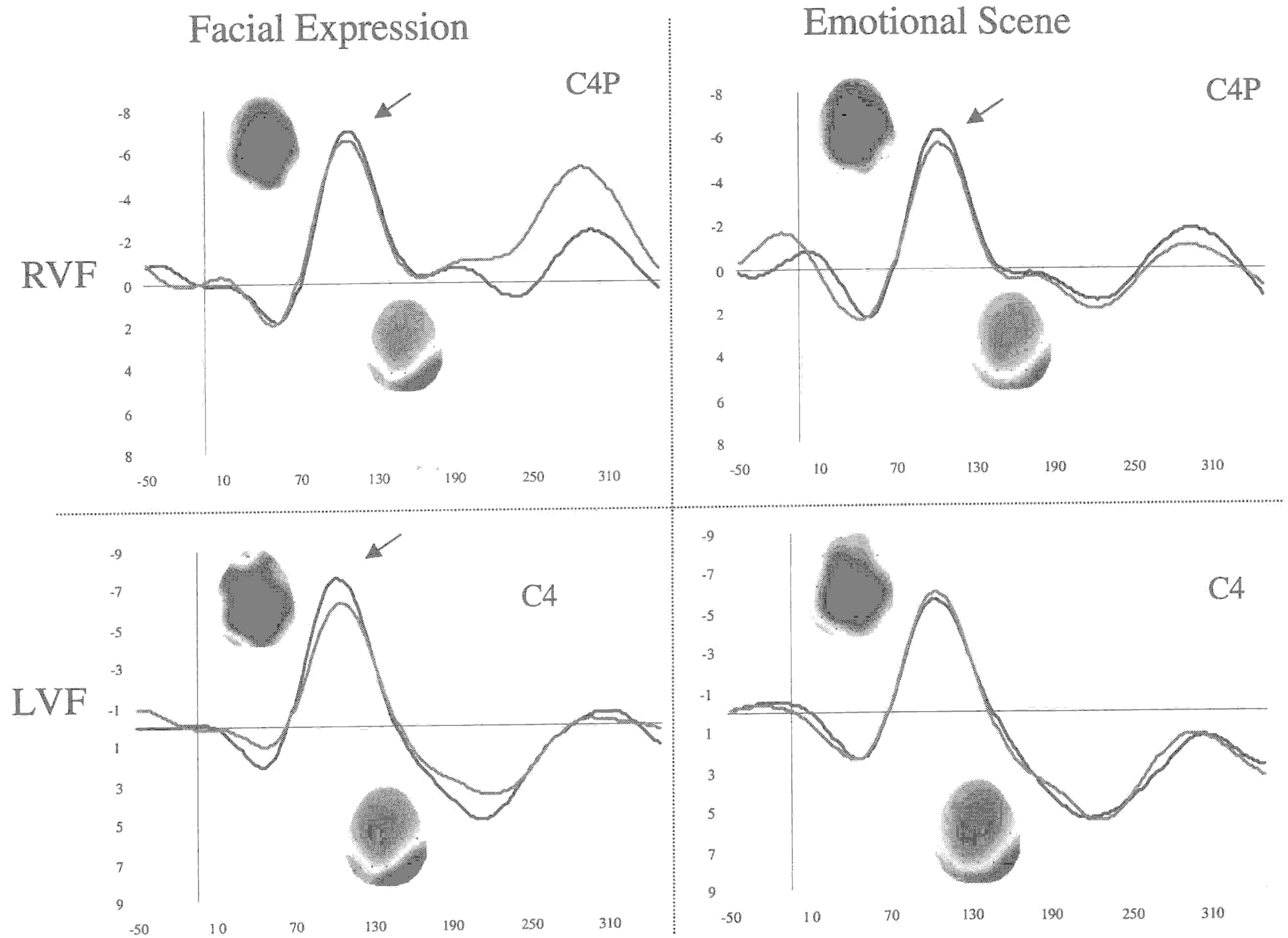

FIGURE 36.7 EEG results in a hemianopic patient with a complete unilateral lesion of the primary visual cortex. Shown are grand averaged auditory waveforms and corresponding topographies (horizontal axis) obtained at central electrodes in each visual condition (congruent pairs in black, incongruent pairs in red) and for each visual hemifield (left/blind vs. right/intact). Congruent pairs elicited a higher auditory N1 component than incongruent pairs for visual presentations (facial expressions or emotional scenes) in the intact field but only for facial expressions in the blind field. For each topographic map (N1 and P2 components), the time interval is 20 ms and

blind field and in the intact field. Second, different integration effects might obtain for conscious and nonconscious presentation of visual stimuli, and integration effects might obtain for the former but not for the latter. We began to explore such situations by taking advantage of the blindsight phenomenon. Participants were presented with visual stimuli (either faces or scenes) and auditory stimuli. Only amygdala responses to faces (and not scenes) are enhanced by congruent voices (in either the intact or the blind field). There are enhanced fusiform responses to faces (but not to scenes) with congruent voices (as in de Gelder et al., 2002). When looking at blind versus intact differences there are activations of superior colliculus (SC) for blind fearful faces (but not for blind fearful scenes). Fusiform responses are enhanced more by congruent voice-face pairings in the intact field than in the blind field. But the SC activation in response to fearful faces is enhanced by congruency between voices and faces, and the enhancement is greater for presentation in the blind field than for presentation in the intact field (de Gelder, Morris, & Dolan, in prep.). These results indicate that awareness of the stimulus plays an important role in the pairings obtained. This aspect warrants more attention in other areas of multisensory research.

Models for multisensory perception of affect

So far, our studies have explored the perceptual dimension of multisensory affect perception mainly by trying to rule out postperceptual biases. If we take audiovisual speech as the closest example at hand, it is fair to say that at present, a number of different models of integration can be envisaged (Summerfield, 1987), without any compelling arguments in favor of any of them. More research is needed before some theoretical models offered for audiovisual speech can be completely ruled out for the case of audiovisual emotion. For example, we have distinguished between the traditional problem of finding parallels between facial and vocal expression deficits. Although these are two different issues, we do not rule out that emotional face- and voice-recognition processes might have a lot in common. Nor can one rule out at present a second theoretical possibility, which assumes recoding of one of the input representations into the format of the other (e.g., visual representations recoded into auditory ones). Likewise, we must still consider that both sensory representations are recoded into a supramodal abstract representation system (Farah, Wong, Monheit, & Morrow, 1989). A complicating factor that is specific to the case of audiovisual affect is that the integration mechanism might be sensitive to the specific affective content, such that the mechanism for fear integration might actually be different from that for sad or happy pairings. For example, we observed an increase in amygdala activation for fearful voice-face combinations but not for happy ones (Dolan et al., 2001). Converging evidence from different methods is needed to make some progress in understanding these complex issues.

One reason why we became interested in the issue of selectivity of pairings is that it might provide cues as to the underpinnings of emotional perception. Would a special status of voice-face pairs indicate that there exists a specific functional underpinning for this kind of pair? According to this view, scene-voice pairs, scene-word pairings, or face-word pairs would have different characteristics than face-voice pairs. One possible interpretation, although a rather speculative one, for the special status of the voice prosody-face pairings is that they are mediated by action schemes (de Gelder & Bertelson, 2003). The analogy that comes to mind here is that of articulatory representation or phonetic gestures, in the sense of abstract representations that figure in the description of auditorily transmitted as well as visually transmitted speech. Ultimately, a notion of motor schemes could be developed along similar lines as was done for the motor theory of speech perception (Liberman & Mattingly, 1985). This action framework suggests an interesting interpretation for the special status of some kinds of audiovisual pairings. Some stimulus combinations make pairings that are always congruent when they occur in naturalistic circumstances. They are so tightly linked that a special effort is needed in the laboratory to separate them and to rearrange the pairings into incongruent combinations for the purpose of experiments. We have used the notion of biologically determined (de Gelder & Bertelson, 2003) or of naturalistic pairings (Pourtois & de Gelder, 2002) to characterize that type of pairing. Which pairings are naturalistic is obviously a significant issue for future empirical research. Evolutionary arguments by themselves are too general and too open-ended to provide specific constraints on the action repertoire of an organism.

Conclusions

Our overview of audiovisual emotion perception has brought together different topics of interest in this new field. Various themes have emerged in the course of this discussion. Some themes are methodological. An example is the need to base conclusions on converging lines of evidence obtained with different methods, such as behavioral and brain imaging studies, because these methods use different metrics. Comparing data is hard, but with progress in multimodal imaging methods, such comparisons should become easier. Other themes that have been addressed in this chapter are more theoretical. We asked whether attention is the glue of audiovisual emotion perception, and we have described some data pointing to a negative answer. We made a beginning by distinguishing different types of audiovisual pairs as we contrasted arbitrary and natural pairs. In the category of natural emotion pairs, and in particular with respect to the auditory component, the difference between semantic and prosodic components turned out to be important. We questioned whether awareness was an important factor in audiovisual integration of emotions, and we found that in fact it was, at least in some cases. Moreover, awareness needs to be considered a critical variable for understanding the kinds of pairings obtained under different circumstances. All of these themes are important for future research on audiovisual emotion, yet none is specific to this domain. Instead, we are dealing with a phenomenon that is basic to our understanding of multisensory integration and important for the study of emotion.

REFERENCES

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews, 1,* 21–61.

Adolphs, R., Damasio, H., Tranel, D., Cooper, G., & Damasio, A. D. (2000). A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *Journal of Neuroscience, 20,* 2683–2690.

Adolphs, R., Damasio, H., Tranel, D., & Damasio, A. R. (1996). Cortical systems for the recognition of emotion in facial expressions. *Journal of Neuroscience, 16,* 7678–7687.

Adolphs, R., & Tranel, D. (1999). Preferences for visual stimuli following amygdala damage. *Journal of Cognitive Neuroscience, 6,* 610–616.

Anderson, A. K., & Phelps, E. A. (1998). Intact recognition of vocal expressions of fear following bilateral lesions of the human amygdala. *Neuroreport, 9,* 3607–3613.

Barth, D. S., Goldberg, N., Brett, B., & Di, S. (1995). The spatiotemporal organization of auditory, visual, and auditory-visual evoked potentials in rat cortex. *Brain Research, 678,* 177–190.

Bartolomeo, P., Bachoud-Levi, A. C., de Gelder, B., Denes, G., Dalla Barba, G., Brugieres, P., et al. (1998). Multiple-domain dissociation between impaired visual perception and preserved mental imagery in a patient with bilateral extrastriate lesions. *Neuropsychologia, 36,* 239–249.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersleben, T. Bachmann, & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347–362). Amsterdam: Elsevier.

Bertelson, P., & de Gelder, B. (2003). The psychology of multisensory perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention.* Oxford, England: Oxford University Press.

Bertelson, P., Vroomen, J., & de Gelder, B. (1997). Auditory-visual interaction in voice localisation and b-modal speech recognition: The effects of desynchronization. In C. Benoit & R. Campbell (Eds.), *Proceedings of the Workshop on Audiovisual Speech Processing: Cognitive and Computational Approaches* (pp. 97–100). Rhodes.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14,* 592–597.

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62,* 321–332.

Bleuler, E. (1911). Dementia Praecox oder Gruppe der Schizophrenien [Dementia praecox or the group of schizophrenias]. In: Aschaffenburg Handbuch der Psychiatrie. Leipzig: Deuticke.

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 and Emotion, 14,* 193–211.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition: II. An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12,* 1–47.

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, 276,* 593–596.

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,* 2619–2623.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

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

Davidson, R. J., & Irwin, W. (1999). The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences, 3,* 11–21.

de Gelder, B. (2000). More to seeing than meets the eye. *Science, 289,* 1148–1149.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception, and ecological validity. *Trends in Cognitive Sciences, 10,* 460–467.

de Gelder, B., Böcker, 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. *Neuroscience Letter, 260,* 133–136.

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

de Gelder, B., Pourtois, G. R. C., Vroomen, J. H. M., & 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,* 425–444.

de Gelder, B., Pourtois, G. R. C., & 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, 99,* 4121–4126.

de Gelder, B., van Ommeren, B., & Frissen, I. (2003). Feelings are not specious: Recognition of facial expressions and vocalizations of chimpanzees (*Pan troglodytes*) by humans. Presented at the Annual Meeting of the Cognitive Neuroscience Society.

de Gelder, B., & Vroomen, J. (2000). The perception of emotion by ear and by eye. *Cognition & Emotion, 14,* 289–311.

de Gelder, B., Vroomen, J. H. M., Annen, L., Masthof, E., & Hodiamont, P. P. G. (2003). Audiovisual integration in schizophrenia. *Schizophrenia Research, 59,* 211–218.

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., & Hodiamont, P. (2003). *Deficits in multisensory perception of affect in schizophrenia.* Manuscript submitted for publication.

de Gelder, B., Vroomen, J., & Pourtois, G. (2001). Covert affective cognition and affective blindsight. In B. de Gelder, E. de Haan, & C. A. Heywood (Eds.), *Out of mind: Varieties of unconscious processing.* Oxford, England: Oxford University Press.

de Gelder, B., Vroomen, J., Pourtois, G., & Weiskrantz, L. (1999). Non-conscious recognition of affect in the absence of striate cortex. *Neuroreport, 10,* 3759–3763.

de Gelder, B., Vroomen, J., & Teunisse, J.-P. (1995). Hearing smiles and seeing cries: The bimodal perception of emotions. *Bulletin of the Psychonomic Society,* 30.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science, 11,* 86–89.

Dolan, R., Morris, J., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences, USA, 98,* 10006–10010.

Driver, J., & Spence, C. (2000). Multisensory perception: Beyond modularity and convergence. *Current Biology, 10,* 731–735.

Ettlinger, G., & Wilson, W. A. (1990). Cross-modal performance: Behavioural processes, phylogenetic considerations and neural mechanisms. *Behavioral Brain Research, 40,* 169–192.

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,* 461–470.

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cognitive Brain Research, 10,* 77–83.

Frissen, I., & de Gelder, B. (2002). Visual bias on sound location modulated by content-based processes. *Journal of the Acoustical Society of America, 112,* 2244.

Fuster, J. M., Bodner, M., & Kroger, J. K. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature, 405,* 347–351.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multisensory object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11,* 473–490.

Goulet, S., & Murray, E. A. (2001). Neural substrates of crossmodal association memory in monkeys: The amygdala versus the anterior rhinal cortex. *Behav Neuroscience, 115,* 271–284.

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,* 451–468.

Lavie, N. (2000). Selective attention and cognitive control. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 175–194). Cambridge, MA: MIT Press.

LeDoux, J. E. (1996). *The emotional brain.* New York: Simon & Schuster.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science, 289,* 1206–1208.

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences, 4,* 383–391.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioural principle.* Cambridge, MA: MIT Press.

Massaro, D. W. (1999). Speechreading: Illusion or window into pattern recognition? *Trends in Cognitive Sciences, 3,* 310–317.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review, 3,* 215–221.

McDonald, J. J., Teder-Sälejärvi, W. A., & Ward, L. M. (2001). Multisensory integration and crossmodal attention effects in the human brain. *Science, 292,* 1791.

Mesulam, M. M. (1998). From sensation to cognition. *Brain, 121,* 1013–1052.

Miller, J. O. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14,* 247–279.

Miller, J. O. (1986). Time course of coactivation in bimodal divided attention. *Perception & Psychophysics, 40,* 331–343.

Morris, J. S., Ohman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences, USA, 96,* 1680–1685.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351–362.

Näätänen, R. (1992). *Attention and brain function.* Hillsdale, NJ: Erlbaum.

Peterson, M. A., de Gelder, B., Rapcsak, S. Z., Gerhardstein, P. C., & Bachoud-Levi, A. C. (2000). Object memory effects on figure assignment: Conscious object recognition is not necessary or sufficient. *Vision Research, 40,* 1549–1567.

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

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

Pourtois, G., de Gelder, B., Bol, A., & Crommelinck, M. (2003). *Convergence of visual and auditory affective information in human multisensory cortex.* Manuscript submitted for publication.

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.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341–365.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron, 28,* 617–625.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology, 50,* 19–26.

Ross, E. D. (2000). Affective prosody and the aprosodias. In M. M. Mesulam (Ed.), *Principle of behavioral and cognitive neurology* (pp. 316–331). Oxford, England: Oxford University Press.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Letters, 127,* 141–145.

Schröger, E., & Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. *Psychophysiology, 35,* 755–759.

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,* 254–257.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: 2. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190.

Streicher, M., & Ettlinger, G. (1987). Cross-modal recognition of familiar and unfamiliar objects by the monkey: The effects of ablation of polysensory neocortex or of the amygdaloid complex. *Behavioural Brain Research, 23,* 95–107.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). Hillsdale, NJ: Erlbaum.

Surakka, V., Tenhunen-Eskelinen, M., Hietanen, J. K., & Sams, M. (1998). Modulation of human auditory information processing by emotional visual stimuli. *Cognitive Brain Research, 7,* 159–163.

Treisman, A. (1996). The binding problem. *Current Opinions in Neurobiology, 6,* 171–178.

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition, 1,* 185–195.

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63,* 651–659.

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independant of attentional resources? *Cognitive and Affective Neurosciences, 1,* 382–387.

Weiskrantz, L. (1986). *Blindsight: A case study and implications.* Oxford, England: Oxford University Press.

Weiskrantz, L. (1997). *Consciousness lost and found.* Oxford, England: Oxford University Press.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience, 18,* 411–418.

Wickens, D. D. (1984). Processing resources in attention. In R. Pararsuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). Orlando, FL: Academic Press.