Selective Attention Modulates Early Human Evoked Potentials during Emotional Face–Voice Processing

Hao Tam Ho[1], Erich Schröger[2], and Sonja A. Kotz[3]

Abstract

■ Recent findings on multisensory integration suggest that selective attention influences cross-sensory interactions from an early processing stage. Yet, in the field of emotional face–voice integration, the hypothesis prevails that facial and vocal emotional information interacts preattentively. Using ERPs, we investigated the influence of selective attention on the perception of congruent versus incongruent combinations of neutral and angry facial and vocal expressions. Attention was manipulated via four tasks that directed participants to (i) the facial expression, (ii) the vocal expression, (iii) the emotional congruence between the face and the voice, and (iv) the synchrony between lip movement and speech onset. Our results revealed early interactions between facial and vocal emotional expressions, manifested as modulations of the auditory N1 and P2 amplitude by incongruent emotional face–voice combinations. Although audiovisual emotional interactions within the N1 time window were affected by the attentional manipulations, interactions within the P2 modulation showed no such attentional influence. Thus, we propose that the N1 and P2 are functionally dissociated in terms of emotional face–voice processing and discuss evidence in support of the notion that the N1 is associated with cross-sensory prediction, whereas the P2 relates to the derivation of an emotional percept. Essentially, our findings put the integration of facial and vocal emotional expressions into a new perspective—one that regards the integration process as a composite of multiple, possibly independent subprocesses, some of which are susceptible to attentional modulation, whereas others may be influenced by additional factors. ■

INTRODUCTION

Human communication benefits from a rich set of nonverbal cues that allow the immediate inference of a person's intentions and emotional states. This includes a wide range of facial expressions, gestures, postures, and emotional vocalizations that, in natural social interactions, are utilized concurrently. The multisensory characteristic of human communication entails that throughout a face-to-face interaction, various social, including emotional, information from vision and audition must be extracted simultaneously and combined to form a unified and coherent percept. The speed and ease with which audiovisual social information is integrated in everyday life point to an unconscious, effortless, and automatized process. As such, interactions between facial and vocal emotional expressions have been hypothesized to occur early and independently of attention (e.g., de Gelder & Vroomen, 2000). This study investigated this hypothesis using ERPs.

Early Interaction between Visual and Auditory Information

The notion that cross-sensory information interacts at an early processing stage, possibly within the sensory cortices and, thus, before attentional selection (see Driver, 2001, for a "selective review on selective attention"), is generally referred to as the "early integration hypothesis" (e.g., Koelewijn, Bronkhorst, & Theeuwes, 2010; Calvert & Thesen, 2004). It has received support from a number of behavioral and ERP findings. For instance, in their classic study, McGurk and MacDonald (1976) showed that when perceivers are faced with incongruent audiovisual speech (e.g., the sound /ba/ dubbed onto lip movements pronouncing /ga/), they tend to fuse the mismatched information into a new and more coherent percept (i.e., /da/). The illusion arises seemingly mandatorily and without conscious awareness, which has led to the conception that visual and auditory speech information interacts in a largely automatic manner. Early studies on the combined perception of emotional faces and voices point to similarly automatic and obligatory interactions between visual and auditory emotional information (de Gelder & Vroomen, 2000). Specifically, de Gelder and Vroomen (2000) reported that when ambiguous facial expressions (obtained by morphing sad and happy faces) are combined with, for example, a sad voice, the overall percept is more likely to be judged as "sad" rather than "happy" (Experiment 1). This influence of affective prosody on the perception of facial emotion prevailed, even when participants were instructed to ignore the voice (Experiment 2). In a similar fashion, happy and sad facial expression can influence the judgment of ambiguous

[1]Max Planck Institute for Human Cognitive and Brain Sciences, [2]University of Leipzig, [3]University of Manchester

Journal of Cognitive Neuroscience 27:4, pp. 798–818
doi:10.1162/jocn_a_00734

emotional prosody, despite instructions to ignore the face (Experiment 3). Following up these results, Vroomen, Driver, and de Gelder (2001) found that these cross-modal influences remain robust under high-load task conditions (participants performed concurrent tasks on simple visual or auditory distractor stimuli). In summary, evidence from audiovisual speech and emotion perception research suggests that cross-sensory interactions are unaffected by attentional modulations and task demand. At the same time, they show that incongruities between the visual and auditory information do not prevent sensory interactions (for further discussion, see de Gelder & Bertelson, 2003).

Adding to these findings, incongruent facial and vocal expressions have been shown to modulate early auditory evoked potentials, that is, the N1 (Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000) and P2 (Liu, Pinheiro, Zhao, et al., 2012; Balconi & Carrera, 2011; Pourtois, Debatisse, Despland, & de Gelder, 2002), which emerge ~100 and ~200 msec post-voice onset, respectively. These observations align with results from audiovisual speech perception studies that found interaction effects within similar time windows (Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005; Klucharev, Möttönen, & Sams, 2003). The general consensus is that such rapid effects reflect interactions between visual and auditory information at an early stimulus processing stage and possibly within the sensory-specific areas. This view is further corroborated by intracranial recordings and source localization of the N1 and P2 that point to neural generators within the auditory cortex (Besle, Bertrand, & Giard, 2009; Crowley & Colrain, 2004; Woods, 1995; Näätänen & Picton, 1987).

Effect of Task Demand on Multisensory Integration

In recent years, results have emerged that are inconsistent with the "early integration" hypothesis. For example, robust multisensory illusions, such as the McGurk effect, have been found to diminish when participants perform a concurrent, highly demanding task that requires them to pay close attention to either the visual or auditory modality (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Tiippana, Andersen, & Sams, 2004) and diverts their attention to the tactile modality (Alsius, Navarra, & Soto-Faraco, 2007). In addition, ERP results point to stronger audiovisual interactions, when attention is directed to the multisensory object, as compared with when attention is diverted from the multisensory object (Talsma, Doty, & Woldorff, 2007; Talsma & Woldorff, 2005). As these audiovisual interactions occurred within 50–100 msec post-stimulus onset, their modulation by attention implies that early stages of multisensory integration are not necessarily impervious to attentional influences. Corroborating these ERP findings, imaging results show that attention to congruent, as compared with incongruent, audiovisual speech information led to greater activation of early visual areas as well as the STS and superior colliculus than when attention was not directed to the audiovisual stimuli (Fairhall & Macaluso, 2009). The STS and superior colliculus are regions that have been strongly implicated in multisensory integration (for reviews, see, e.g., Stein & Stanford, 2008; Ghazanfar & Schroeder, 2006).

To consolidate these conflicting findings, current models of multisensory integration (e.g., Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010) have turned to Lavie's (1995, 2005) "perceptual load" theory, according to which perception proceeds in an automatic—in the sense of fast, unconscious, and obligatory—manner, yet is constrained by limited capacity. As a result, when limited resources are taken up, for example, by a high-load task, other processes may be compromised. With regard to multisensory integration, this means that information from different sensory modalities is integrated, by and large, independently of attention; however, multisensory interactions may weaken or fail altogether, when processing resources are exhausted (Talsma et al., 2010). In this context, Talsma and colleagues (2010) hypothesized that attention can be recruited to prioritize the multisensory process in question. We asked whether this model can be extended to the integration of facial and vocal emotional expressions.

Why Emotional Face–Voice Interactions Seem Robust against Task Demand

Although interactions between facial and vocal emotional expressions have been found to withstand high task demand (Vroomen et al., 2001), several factors were not key in previous studies. First, socially relevant multisensory information, such as audiovisual speech, exhibit a strong "unity effect," that is, the tendency to perceive different sensory information as coming from the same source (Tuomainen, Andersen, Tiippana, & Sams, 2005; for a review, see Navarra, Alsius, Soto-Faraco, & Spence, 2010). This tendency makes multisensory socially significant percepts particularly robust against interference in comparison with socially irrelevant stimuli, such as monkey vocalization and music playing (Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2008). Alternatively, it may not be the social significance of speech and emotional information per se that underlies this strong unity effect, but the overly familiarity of human perceivers with visual and auditory speech as coming from a common source. This notion appears to be corroborated by Lewald and Guski's (2003) findings, which suggest that a "unity assumption" (Welch & Warren, 1980) may also be formed in the course of a simple task, whereby participants have to judge the likelihood that two basic, low-level stimuli presented in the visual and auditory modality (with varying temporal and spatial disparities) have a common cause.

Like audiovisual speech, facial and vocal expressions induce a similarly strong "unity effect" that may be particularly difficult to impede. Furthermore, emotional expressions represent socially significant information and have been

shown to capture attention effectively (Vuilleumier, 2005; Compton, 2003). Consequently, it may require an especially resource-intensive task to impede emotional face–voice interactions as well as an equivalently relevant distractor (e.g., audiovisual speech) that can draw attention away from emotional expressions. However, Vroomen and colleagues (2001) used simple visual and auditory distractors (digits and sinusoidal tones, respectively) to test emotional face–voice interactions under "high" task demand. Such unimodal distractors have been found to affect multisensory processes to a much lesser extent than audiovisual distractors (Vatakis & Spence, 2006), possibly because unimodal distractors require less resources than multimodal distractors. This raises the question whether the absence of an effect of task demand in Vroomen and colleagues's (2001) study could be attributed to the use of unimodal distractors. Given Vatakis and Spence's (2006) finding, audiovisual distractors may be more appropriate to test how robust emotional face–voice interactions are against task demands.

Investigating the Influence of Attention on Emotional Face–Voice Perception

In essence, we argue that the effects of selective attention and task demand on the combined perception of emotional facial and vocal expressions have not been adequately investigated. As pointed out above, the unimodal distractors used to test task demand (see Vroomen et al., 2001) may not have placed sufficient processing demands on the perceptual system to disrupt emotional face–voice interactions. Furthermore, previous investigation of emotional face–voice perception in visual-only and auditory-only attention condition (de Gelder & Vroomen, 2000) used behavioral measures (i.e., forced choice and RT) that make it difficult to exactly determine how early the observed audiovisual emotional interactions occurred. At the same time, studies that employed a high temporal-resolution measure, such as ERP, did not manipulate or control for attention (Balconi & Carrera, 2011; Pourtois et al., 2000, 2002). Thus, it is unclear whether the early audiovisual emotional interactions reported in previous ERP studies arise in conditions in which, for example, one of the modalities need to be ignored or attention is directed to non-emotional (e.g., speech) audiovisual information.

To address these questions, we presented participants with videos of congruent and incongruent facial and vocal expressions and monitored their early brain responses using ERPs. The onset of the lip movement and speech sound in the videos could be synchronous or asynchronous. Participants performed four consecutive two-alternative forced-choice tasks that directed their attention to different aspects of the videos.

In the so-called attend-synchrony task, participants were asked to discriminate between synchronous and asynchronous audiovisual speech. As speech is socially significant, we assumed that this task would prioritize audiovisual speech processing causing resources to be drawn away from audiovisual emotion processing. According to the "early integration" hypothesis, audiovisual emotion processing should not be affected by this manipulation. In contrast, models of multisensory integration that are based on the "perceptual load" theory would predict a weakening or even failure of interaction between facial and vocal emotion expressions. This condition was compared with a second bimodal attention condition, the so-called attend-congruence task, in which participants had to determine whether facial and vocal expressions were congruent or incongruent. Thereby, attention was directed to the emotion conveyed by both the face and the voice. Here, emotional face–voice processing was expected to be prioritized, giving rise to strong interactions between facial and vocal expressions. Additionally, participants were given two unimodal attention tasks. In the attend-face and attend-voice condition, they had to judge the facial and vocal expression, respectively. These attentional manipulations were similar to that in de Gelder and Vroomen's (2000) study. We assumed that, to perform these two tasks successfully, participants would attempt to suppress emotional information in the irrelevant sensory modality to avoid potential interferences caused by incongruent information. If audiovisual emotional interaction is obligatory and occurs before attentional selection, as suggested by de Gelder and Vroomen's (2000) results, facial and vocal emotions should interact in spite of the suppression of affective information from the task-irrelevant modality.

Possible Functions of the N1 and P2 in Emotional Face–Voice Processing

On the basis of previous results, we expected interactions between facial and vocal emotion to be manifested as modulations of the N1 and P2 amplitude (Balconi & Carrera, 2011; Pourtois et al., 2000, 2002). Although these two components are often taken to be linked to one another (Crowley & Colrain, 2004), divergent functions have been associated with the N1 and P2, which suggests that they are, in fact, independent of one another. For example, findings from audiovisual speech perception research have linked the N1 to a cross-sensory anticipatory process, whereby facial cues, such as muscle movements, that naturally precede the voice, give rise to predictions about the auditory stimulus (e.g., Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). In contrast, the P2 has been associated with the detection of emotional significance in emotion perception research (e.g., Kotz & Paulmann, 2011; Sauter & Eimer, 2010). However, it should be noted that, in their original model of emotional voice processing, Schirmer and Kotz (2006) related the P2 to a more complex process, whereby vocal (and possibly also facial) cues are integrated to derive an emotional gestalt. Essentially, the N1 is thought to reflect a top–down process, whereas the P2 may be driven

bottom–up. As such, the two components should show differential modulations of selective attention.

METHODS

Participants

Thirty-two individuals (16 women), all right-handed and with normal hearing and normal or corrected-to-normal vision, participated in the experiment. The data of one participant were excluded from further analysis because of excessive artifact contamination. The mean age of the remaining 31 participants (16 women) was 25.32 years (SD = 3.52 years). All participants received payment for their time.

Stimulus Material

A schematic depiction of a stimulus example can be found in Figure 1A. The stimulus material consisted of videos recorded with SONY HDR-HC7 camcorder (Sony, Tokyo, Japan) in HDV1080i quality. The sound was additionally recorded with a Zoom Handy Recorder H4, which offered a better quality than the camera's built-in microphone. The video was subsequently overlaid with the separately recorded sound in Final Cut Pro 7 (Apple, Inc., Cupertino, CA) using the onset of the original camera sound for alignment. A 24-year-old semi-professional actress uttered two nonlexical interjections, "ah" and "oh," with either an angry or a neutral facial expression and congruent emotional prosody 20–30 times. Fifteen videos per interjection and emotion were selected to create an equal number of congruent and incongruent stimuli. For incongruent stimuli, neutral facial expressions were overlaid with an angry voice and angry facial expressions with a neutral voice. The onset of the original voice was used to align the incongruent voice with the lip movement. Additionally, a set of 24 asynchronous videos were created. Asynchrony was achieved by shifting the voice onset 300–500 msec before the first lip movement. The asynchronous videos were presented as catch trials in the experiment and afterwards excluded from the analysis. The sound in all videos was normalized to 0 dB (default setting) using root mean square in Final Cut Pro. In the ERP experiment, the volume was adjusted to a comfortable level for all participants. Congruent and incongruent videos were 1.1–2.2 sec in length. Before the ERP experiment, the videos were rated in terms of valence and arousal (Bradley & Lang, 1994) by 32 different participants (16 women) and tested in an emotion identification experiment with 40 different participants (20 women; see supplementary material for details). The rating results are plotted in Figure 4.

Design and Tasks

The experiment consisted of four blocks, each comprised 144 videos (60 congruent, 60 incongruent, and 24 asynchronous) presented in random order. Participants performed a different two-alternative forced-choice task in each block. In the attend-face task, they indicated whether the facial expression was angry. In the attend-voice task,

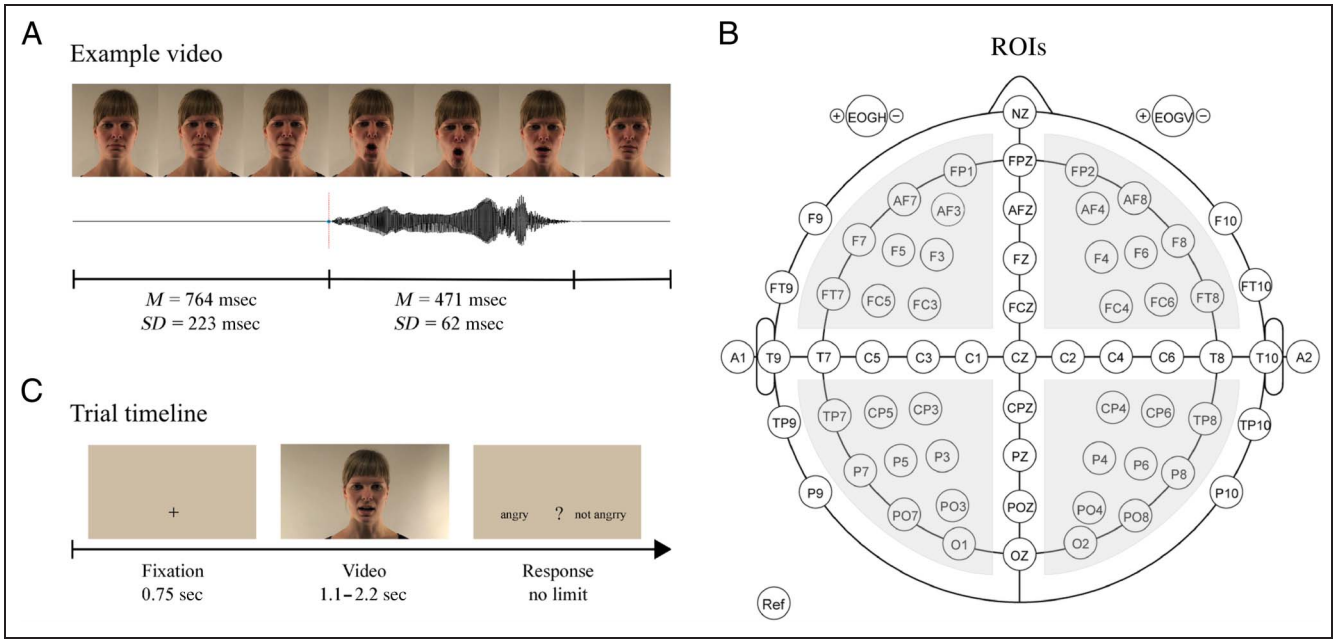

Figure 1. (A) Selected frames from an example video depicting angry facial movements with its congruent angry waveform; facial movements preceded the voice in all videos by 764 msec on average (SD = 223 msec); the mean duration of the voice was 471 msec (SD = 62 msec). (B) Electrodes grouped into four ROIs; midline and central electrodes were analyzed separately; the outmost electrodes were excluded from the analysis. (C) Schematic depiction of a trial's time course. All trials started with a fixation cross lasting for 0.75 sec, followed by a video played in full length. At the end of each video, participants had to make a choice.

they judged whether the emotion expressed by the voice was anger. In the attend-congruence task, participants discriminated between congruent and incongruent facial and vocal expressions. Finally, in the attend-synchrony task, they determined whether the onset of the voice was synchronous with the lip movement. All participants performed the attend-synchrony task first and the attend-congruence task last. The order of the attend-face and attend-voice task was counterbalanced across participants.

Procedure and Trials

Participants were seated approximately 1 m from a computer screen inside a sound attenuated booth. The sound was delivered via SONY MDR-XD100 headphones. Participants were given written and oral instructions at the start of each block. To familiarize themselves with the task, they completed a short practice trial before each block. As illustrated in Figure 1C, each trial in all experimental conditions began with a fixation cross ("+") on a blank computer screen for 750 msec. Subsequently, a video was presented and played in full length. A question mark appeared immediately at the end of the video with the answer choices displayed to the left and the right of the question (e.g., "angry ? not angry" or "not congruent ? congruent"). For half of the participants, the not-choice was always on the right-hand side, and for the other half, it was always on the left-hand side. Participants were instructed to wait for the question mark to give their response. Delaying participants' answer was necessary to avoid electrophysiological activities related to motor response. Participants had as much time as needed to respond.

EEG Recording and Preprocessing

The EEG was recorded from 63 equidistantly positioned (10–20 system) scalp electrodes (Ag/AgCl) built into an elastic cap (Easy-Cap, Falk Minow Services, Herrsching-Breitbrunn, Germany) using the software BrainVision Recorder (Brain Products GmbH, Munich, Germany). The sampling rate was 500 Hz. Average reference was used during recording. The data were re-referenced to the nose electrode offline. Two bipolar eye electrodes were used to monitor vertical and horizontal eye movement. The impedance was kept below 5 kΩ. Long epochs of 3.5 sec (−0.5 to 3 sec from video onset) were extracted using the EEGLAB 8.0.3.5b toolbox (Delorme & Makeig, 2004) in MATLAB 7.7.0 (The Mathworks, Natick, MA). The epochs were submitted to a Windowed Sinc FIR filter with Blackman window from the *firfilt* plugin in EEGLAB (Widmann & Schröger, 2012). The cutoff frequency was set to 2 Hz (as recommended by Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002) to remove slow potentials (see, e.g., van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011) that could result in baseline offsets. The filter order (defined as filter length minus 1) was 2750, which was estimated using a transition bandwidth of 1 Hz. The transition bandwidth of a windowed sinc FIR filter is a function of the filter order and the window type (Widmann & Schröger, 2012). In the case of the Blackman windowed sinc FIR filter, the transition bandwidth can be estimated as 5.5/filter order × sampling rate. Subsequently, independent component analysis (ICA) was conducted using the *runica* algorithm. Independent component analysis-based artifact identification methods were employed to identify ocular and muscle artifacts (Mognon & Jovicich, 2011; Winkler, Haufe, & Tangermann, 2011). The results of the artifact detection procedures were visually inspected before any component was manually removed. Shorter epochs of 1.5 sec (−0.5 to 1.0 sec from voice onset) were extracted and submitted to an automatic artifact rejection procedure whereby epochs that contain abnormally distributed activity (3 standard deviations from mean kurtosis) were discarded. Trials that led to incorrect responses were also excluded from the analysis. Baseline correction was applied using a prestimulus interval of 500 msec (−500 to 0 msec from voice onset). Finally, the averaged data were low pass filtered with a FIR filter (Windowed Sinc, Blackman window) from the *firfilt* plugin. The cutoff frequency was 30 Hz (Filter order: 276).

ERP Analysis

Statistical analyses were conducted on epochs time-locked to the voice onset. The epochs spanned −500 to 1000 msec. In most videos, the voice started 400 msec after the face, except for two congruent (one neutral) and two incongruent videos. Those videos were included in the experiment but subsequently excluded from the data analysis, as the preceding face in those four videos likely evoked visual ERP components that may overlap with the auditory ERP components evoked by the following voice. Additionally, we excluded the peripheral scalp electrodes that are typically very noisy (see Figure 1B). The remaining scalp electrodes were grouped into four ROIs: left anterior, right anterior, left posterior, and right posterior (depicted in Figure 1B), the midline (anterior: FPz, AFz, Fz, FCz; posterior: FPz, AFz, Fz, FCz) and central electrodes (left: C1, C3, C5, T7; right: C2, C4, C6, T8). The analyses of the midline and central electrodes were conducted separately from the four ROIs. By means of visual inspection, two time windows were selected that encompass the N1 (70–140 msec) and the P2 amplitude (150–230 msec) from voice onset (see Figure 3). Mean amplitudes were computed for each time window and group of electrodes. Subsequently, the values were submitted to a repeated-measures ANOVA with the factors: Task (4 levels: attend-synchrony, attend-face, attend-voice, and attend-congruence), Voice (2 levels: neutral voice, angry voice), and Congruence (2 levels: congruent face, incongruent face), LR (2 levels: left hemisphere, right hemisphere), and PA (2 levels: posterior electrodes, anterior electrodes) for the analysis of the four ROIs. The analysis of the central electrodes did not include the factor PA, and the

Table 1. Participants' Performance in the Four Task and Stimulus Conditions

		Attend-synchrony		*Attend-congruence*		*Attend-face*		*Attend-voice*	
Voice	*Face*	*M (%)*	*SE (%)*	*M (%)*	*SE (%)*	*M (%)*	*SE (%)*	*M (%)*	*SE (%)*
Angry	Congruent	95.07	1.07	96.77	0.67	98.3	0.5	97.94	0.47
	Incongruent	87.19	3.2	97.04	0.61	98.3	0.61	95.34	1.27
Neutral	Congruent	88.08	1.72	95.7	0.95	98.3	0.48	98.21	0.5
	Incongruent	87.54	2.24	95.79	0.92	98.21	0.46	95.88	0.89

The table shows the mean hit rates (*M*) and standard errors (*SE*) in percent (%). Note that, although participants performed above chance (mean accuracy > 80%), the hits rates in the attend-synchrony task is particularly low relative to the hit rates in the attend-face task, for example.

factor LR was excluded from the analysis of the midline electrodes. The analyses were conducted using the programming language R (Version 3.0.2) for statistical computing (R Core Team, 2013) in conjunction with the software RStudio (Version 0.98.490; RStudio, 2013). Repeated-measures ANOVAs were conducted using the package *ez* (Version 4.2-2; Lawrence, 2013) that estimates effect sizes using generalized eta squared (η^2_G) (Cousineau, 2005). *p* values obtained in pairwise comparisons were adjusted using the Holm–Bonferroni method (Holm, 1979).

RESULTS

Task Performance

As can be seen in Table 1, participants performed all four tasks above chance; the mean accuracy was greater than 85%. For the statistical analysis, d' (a measure of sensitivity or discriminability; see Macmillan & Creelman, 2005) was computed and entered into a repeated-measures ANOVA with the factors Task (4 levels: attend-synchrony, attend-face, attend-voice, attend-congruence), Congruence (2 levels: congruent, incongruent), and Voice (2 levels: neutral, angry). Mean d' values for all task and stimulus conditions are graphically depicted in Figure 3 (right). The statistical results are listed in Table 2. Of interest is the significant three-way interaction between Task, Congruence, and Voice, $F(3, 90) = 5.24$, $p_{GG} < .01$, $\eta^2_G = .02$. This interaction was further analyzed in four repeated-measures ANOVAs with the factors Congruence and Voice conducted at each level of Task. The analysis revealed a significant main effect of Congruence in the attend-voice condition, $F(1, 30) = 10.47$, $p < .01$, $\eta^2_G = .07$. As can be seen in Figure 3 (right), task performance in the attend-voice task was worse when facial and vocal expressions were incongruent. In the attend-synchrony condition, the analysis yielded a significant interaction between Congruence and Voice, $F(1, 30) = 8.65$, $p < .01$, $\eta^2_G = .03$. Mean d' values indicated that performance in the attend-synchrony task was generally worse than in the other conditions (see Figure 3). Pairwise comparisons conducted on d' scores confirmed this observation (all ps < .001). However, when both the face and voice expressed anger, discrimination improved notably in the attend-synchrony condition (see Figure 3). Paired-sample *t* tests with d' also indicated that discrimination was significantly better when facial and vocal expressions were congruent than incongruent in the angry voice condition, $t(30) = 2.88$, $p < .01$, (Cohen's) $d = .53$. No such improvement was found in the neutral voice condition, $t(30) = -0.12$, $p = .91$, $d = -.02$.

Table 2. Statistical Results of Participants' Task Performance

Effect	*DFn*	*DFd*	*F*	*p*	*ges*	*W*	p_W	*GG*	p_{GG}	*p < .05*
Task	3	90	20.64	.00	0.21	0.10	.00	0.44	.00	*
Congruence	1	30	6.87	.01	0.01					*
Voice	1	30	8.46	.01	0.01					*
Task × Congruence	3	90	4.76	.00	0.02	0.36	.00	0.62	.01	*
Task × Voice	3	90	7.18	.00	0.02	0.65	.03	0.82	.00	*
Congruence × Voice	1	30	1.87	.18	0.00					
Task × Congruence × Voice	3	90	5.24	.00	0.02	0.59	.01	0.77	.01	*

Participants' task performance was statistically analyzed using d' (a measure of sensitivity or discriminability). Three factors were entered into the repeated-measures ANOVA: Task (4 levels), Congruence (congruent, incongruent), and Voice (angry, neutral). Where Mauchly's test (*W*) indicated that the Sphericity assumption was violated, the Greenhouse–Geisser correction (*GG*) was applied. Effect size was estimated using generalized eta-squared (η^2_G). The asterisks (*) indicate effects that yielded a significance level of $p < .05$.

ERPs

The grand averages are shown in Figures 2–4. Figure 2 gives a general overview of the N1 and P2 results. Figure 3 depicts the auditory evoked potentials to congruent and incongruent stimuli in each task collapsed across the two voice conditions, whereas Figure 4 shows the responses to congruent and incongruent stimuli in each voice condition

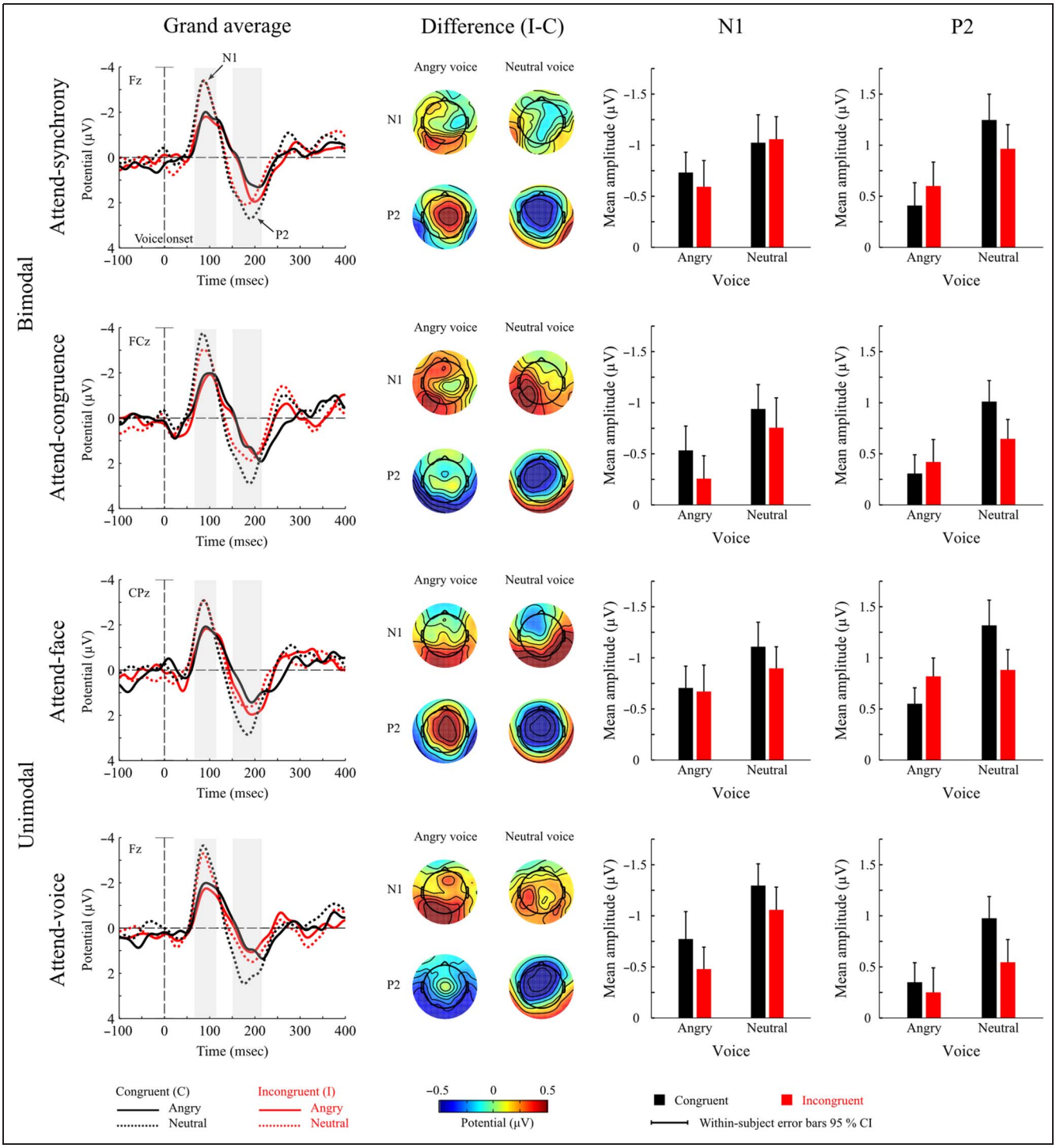

Figure 2. Left: Grand-averaged ERPs extracted from voice onset. The highlighted areas (in gray) depict the approximate time window of the N1 (70–140 msec) and P2 (150–230 msec). Mid: Topographies of the difference between congruent and incongruent stimuli for angry and neutral voice computed by subtracting the N1 and P2 mean amplitudes in the congruent condition from that in the incongruent condition. Right: The bars depict mean amplitudes to congruent (black) and incongruent (red) angry and neutral voice within the two time windows of interest. For the N1 time window, mean amplitudes were collapsed over all ROI electrodes and averaged across participants. For the P2 time window, mean amplitudes were collapsed over anterior ROI electrodes, where the Congruence and Voice showed a significant interaction ($p < .05$). Error bars represent within-subject 95% confidence intervals (Cousineau, 2005).

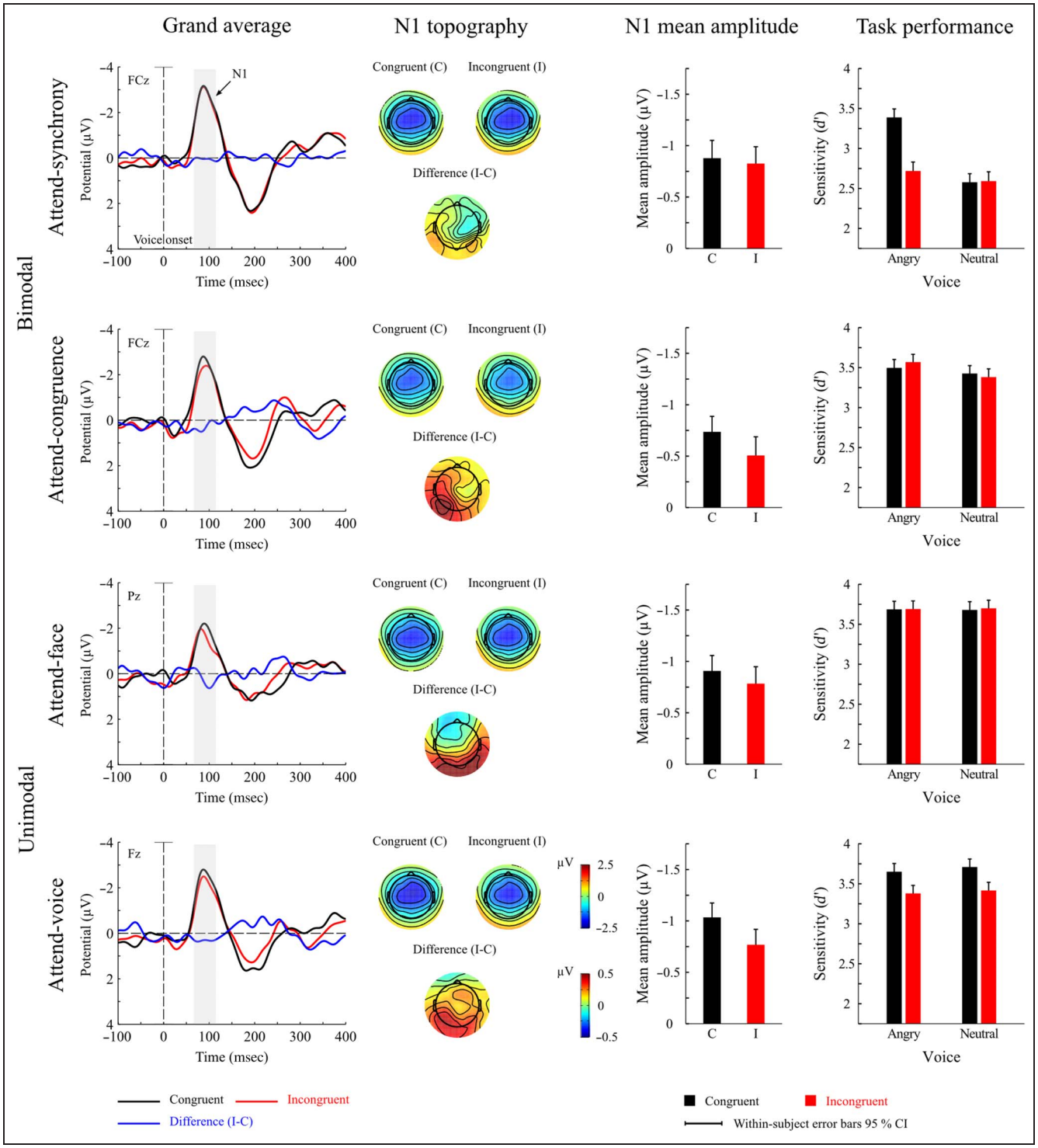

Figure 3. Left: Grand-averaged ERPs extracted from voice onset and collapsed across the two Voice conditions. The highlighted area (in gray) depicts the approximate time window of the N1 (70–140 msec). Note the reduced N1 amplitude to incongruent stimuli in the attend-congruence, attend-face, and attend-voice task. For comparison, congruent and incongruent stimuli in the attend-synchrony task show virtually no difference in the same time window. Mid: The upper topographies depict the scalp distribution of the N1 response to congruent and incongruent stimuli across angry and neutral voice. The lower difference topographies were computed by subtracting the congruent from the incongruent condition. Right: The bar graphs on the left reflect the N1 mean amplitudes to congruent (black) and incongruent (red) condition; the values were computed by averaging the ROI electrodes. The bar graphs on the right summarize participants' performance in the four discrimination tasks. As can be seen, participants' mean discriminability or sensitivity (d') is particularly low in the attend-synchrony task. Note also the lower d' value to incongruent stimuli in the attend-voice task.

collapsed across the four task conditions. The results of the repeated-measures ANOVA for the N1 and P2 time window are listed in Tables 3 and 4, respectively. In both the N1 and P2 time window, Voice showed a significant main effect, that is, $p < .05$ (see also Figure 2). This was expected, as angry voice and neutral voice have different acoustic characteristics. For the same reason, we unpacked significant interactions, such as Congruence × Voice, only by Voice. In this example, we compared congruent and incongruent conditions within angry and neutral voice, but not angry and neutral voice within congruent or incongruent condition. Finally, because of length restrictions, we will only discuss effects of interest.

N1

The analysis of the N1 time window yielded a significant main effect of Congruence in the ROIs, $F(1, 30) = 9.04$, $p < .05$, $\eta^2_G = .013$, and at the midline electrodes, $F(1, 30) = 4.19$, $p < .05$, $\eta^2_G = .003$. In addition, Congruence showed a significant interaction with PA at the same electrode sites, ROI: $F(1, 30) = 13.58$, $p < .01$, $\eta^2_G = .003$; midline: $F(1, 30) = 5.04$, $p < .05$, $\eta^2_G = .001$. Planned comparisons with paired-sample t tests indicated that the Congruence effect was significant only at posterior electrodes, ROI: $t(30) = -3.72$, $p < .01$, $d = -.68$; midline: $t(30) = -2.50$, $p < .05$, $d = -.46$. From Figure 3, it can be seen that the N1 amplitude in the incongruent stimulus condition is reduced in comparison with the congruent condition. This is noticeable in all but the attend-synchrony task condition. Mean amplitudes computed over the posterior ROIs pointed to a smaller N1 response when facial and vocal expressions were incongruent, $M = -0.44$, $SE = 0.09$, than when they were congruent, $M = -0.72$, $SE = 0.11$. Similar observations were made at midline electrodes, incongruent: $M = -0.76$, $SE = 0.11$; congruent: $M = -0.96$, $SE = 0.12$.

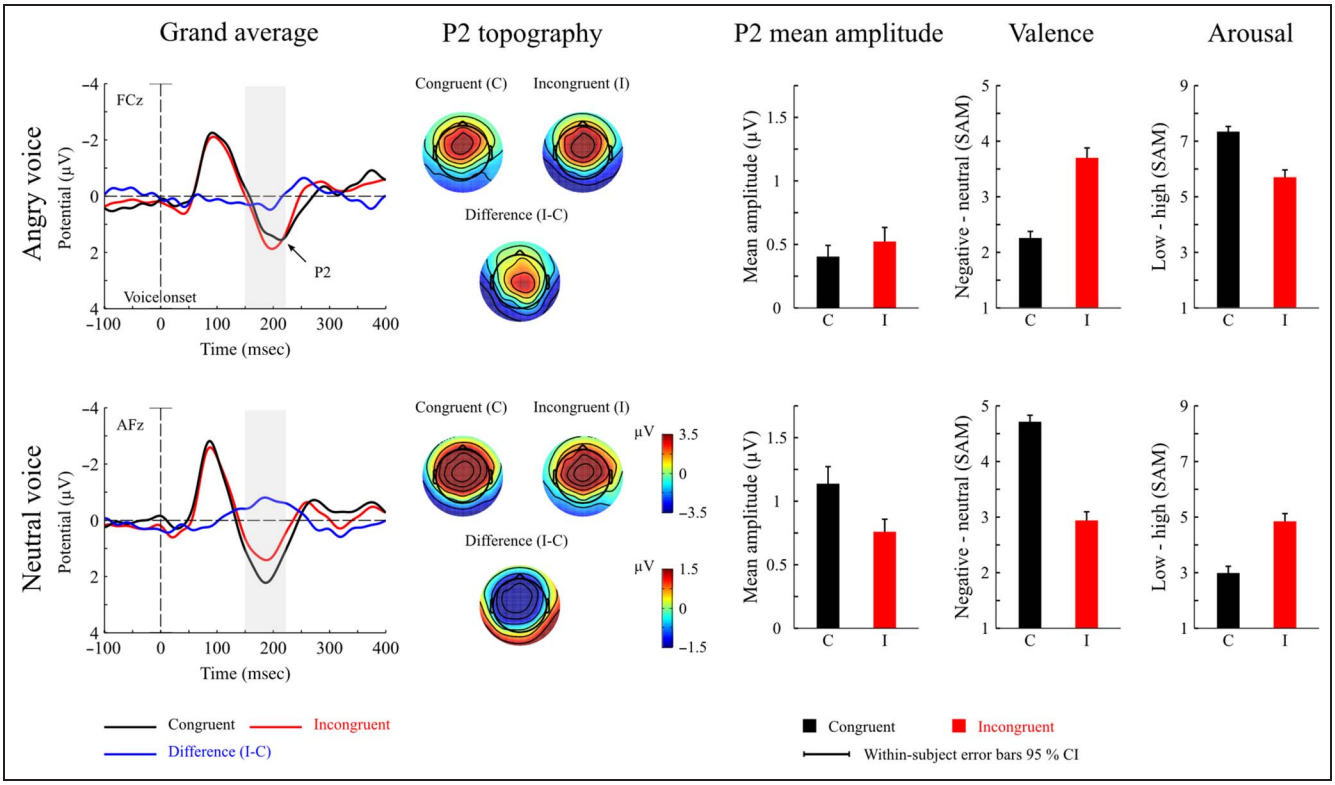

Figure 4. Left: Grand-averaged ERPs to congruent and incongruent angry and neutral voice; the ERPs were collapsed across the four task conditions. The highlighted area (in gray) depicts the approximate time window of the P2 (150–230 msec). Note the greater P2 amplitude to incongruent than congruent stimuli in the angry voice condition. This congruence effect reverses in the neutral voice condition, where the P2 response to incongruent stimuli is reduced compared with congruent stimuli. Mid: The upper topographies depict the P2 response to congruent and incongruent angry and neutral voice collapsed across all task conditions. The lower difference topographies were computed by subtracting the congruent from the incongruent condition. Right: The left bar graphs reflect the P2 mean amplitudes to congruent and incongruent stimuli collapsed across the task conditions; the values were computed by averaging only anterior ROI electrodes. The bar graphs in the middle and on the right summarize the results from a separate valence and arousal rating study with 32 different participants (16 female). (See supplementary material for a detailed summary of the rating study.) On the 9-point SAM scale (y axis), a valence score (left) of 1 means that participants perceived both facial and vocal emotional expressions as *very negative*, 5 = *neutral*, 9 = *very positive*; an arousal score (right) of 1 means that participants perceived the same expressions as *very low in arousal*, 5 = *medium*, 9 = *very high* (Bradley & Lang, 1994). Note that when the angry voice was combined with an incongruent neutral face, the overall percept was rated as less negative and lower in arousal than when the angry voice was combined with a congruent angry face. Both valence and arousal effect reversed in the case of neutral voice. Correspondingly, we observe a decrease in the P2 amplitude to the incongruent angry voice and an increase in the P2 amplitude to the incongruent neutral voice.

Table 3. Statistical Results for the N1 (70–140 msec) Time Window

Effect	*DFn*	*DFd*	*F*	*p*	η^2_G	*W*	p_W	*GG*	p_{GG}	$p < .05$
ROI										
Task	3	90	3.22	.03	.01	0.93	.85	0.96	.03	*
Congruence	1	30	9.04	.01	.01					*
Voice	1	30	25.66	.00	.05					*
LR	1	30	1.42	.24	.00					
PA	1	30	23.23	.00	.06					*
Task × Congruence	3	90	0.77	.51	.00	0.80	.26	0.88	.50	
Task × Voice	3	90	0.60	.62	.00	0.92	.79	0.94	.61	
Congruence × Voice	1	30	0.10	.75	.00					
Task × LR	3	90	0.24	.87	.00	0.71	.08	0.81	.83	
Congruence × LR	1	30	1.73	.20	.00					
Voice × LR	1	30	4.34	.05	.00					
Task × PA	3	90	1.08	.36	.00	0.85	.44	0.89	.36	
Congruence × PA	1	30	13.58	.00	.00					*
Voice × PA	1	30	0.55	.47	.00					
LR × PA	1	30	4.76	.04	.00					*
Task × Congruence × Voice	3	90	0.49	.69	.00	0.80	.28	0.86	.66	
Task × Congruence × LR	3	90	2.88	.04	.00	0.74	.12	0.85	.05	*
Task × Voice × LR	3	90	0.32	.81	.00	0.92	.79	0.95	.80	
Congruence × Voice × LR	1	30	0.25	.62	.00					
Task × Congruence × PA	3	90	1.59	.20	.00	0.80	.27	0.89	.20	
Task × Voice × PA	3	90	2.82	.04	.00	0.91	.74	0.94	.05	*
Congruence × Voice × PA	1	30	0.30	.59	.00					
Task × LR × PA	3	90	2.96	.04	.00	0.89	.63	0.93	.04	*
Congruence × LR × PA	1	30	0.61	.44	.00					
Voice × LR × PA	1	30	0.79	.38	.00					
Task × Congruence × Voice × LR	3	90	0.36	.78	.00	0.79	.25	0.86	.75	
Task × Congruence × Voice × PA	3	90	0.87	.46	.00	0.91	.75	0.94	.46	
Task × Congruence × LR × PA	3	90	2.71	.05	.00	0.76	.15	0.87	.06	
Task × Voice × LR × PA	3	90	2.74	.05	.00	0.66	.04	0.78	.06	
Congruence × Voice × LR × PA	1	30	0.03	.87	.00					
Task × Congruence × Voice × LR × PA	3	90	0.57	.64	.00	0.49	.00	0.75	.59	
Midline										
Task	3	90	2.54	.06	.01	0.91	.76	0.94	.07	
Congruence	1	30	4.19	.05	.00					
Voice	1	30	17.58	.00	.03					*
PA	1	30	24.54	.00	.06					*
Task × Congruence	3	90	0.79	.50	.00	0.77	.19	0.86	.49	

Table 3. (*continued*)

Effect	*DFn*	*DFd*	*F*	*p*	η^2_G	*W*	p_W	*GG*	p_{GG}	$p < .05$
Task × Voice	3	90	0.54	.65	.00	0.96	.96	0.97	.65	
Congruence × Voice	1	30	1.17	.29	.00					
Task × PA	3	90	1.56	.21	.00	0.83	.38	0.90	.21	
Congruence × PA	1	30	5.04	.03	.00					*
Voice × PA	1	30	0.25	.62	.00					
Task × Congruence × Voice	3	90	0.13	.94	.00	0.86	.49	0.90	.93	
Task × Congruence × PA	3	90	1.71	.17	.00	0.90	.70	0.94	.17	
Task × Voice × PA	3	90	1.30	.28	.00	0.94	.89	0.96	.28	
Congruence × Voice × PA	1	30	0.43	.52	.00					
Task × Congruence × Voice × PA	3	90	0.88	.45	.00	0.90	.68	0.94	.45	
Central										
Task	3	90	4.50	.01	.02	0.91	.72	0.94	.01	*
Congruence	1	30	3.05	.09	.00					
Voice	1	30	27.62	.00	.06					*
LR	1	30	0.69	.41	.00					
Task × Congruence	3	90	0.75	.53	.00	0.85	.45	0.91	.52	
Task × Voice	3	90	1.06	.37	.00	0.87	.57	0.92	.37	
Congruence × Voice	1	30	0.21	.65	.00					
Task × LR	3	90	0.21	.89	.00	0.86	.52	0.91	.87	
Congruence × LR	1	30	2.67	.11	.00					
Voice × LR	1	30	3.83	.06	.00					
Task × Congruence × Voice	3	90	0.11	.95	.00	0.78	.21	0.85	.93	
Task × Congruence × LR	3	90	3.89	.01	.00	0.74	.13	0.84	.02	*
Task × Voice × LR	3	90	0.39	.76	.00	0.91	.76	0.94	.75	
Congruence × Voice × LR	1	30	0.01	.94	.00					
Task × Congruence × Voice × LR	3	90	0.72	.54	.00	0.87	.56	0.91	.53	

Repeated-measures ANOVAs were conducted on mean amplitudes computed across the four ROIs, midline, and central electrodes. The factors entered into the ANOVAs were Task (4 levels), Congruence (congruent, incongruent), Voice (angry, neutral), LR (left, right), and PA (posterior, anterior). Where Mauchly's test (*W*) indicated that the Sphericity assumption was violated, the Greenhouse–Geisser correction (*GG*) was applied. Effect size was estimated using generalized eta-squared (η^2_G). The asterisks (*) indicate effects that yielded a significance level of $p < .05$.

The results further included a significant interaction between Task, Congruence, and LR in the ROIs, $F(3, 90) = 2.88$, $p < .05$, $\eta^2_G = .001$, and at the central electrodes, $F(3, 90) = 3.89$, $p < .05$, $\eta^2_G = .001$. This interaction was unpacked by Task in separate repeated-measures ANOVAs. The Congruence × LR interaction remained significant in the attend-face task, $F(1, 30) = 4.56$, $p < .05$, $\eta^2_G = .002$. Mean amplitudes computed over the right ROIs pointed to a reduced N1 response to incongruent as compared with congruent facial and vocal expressions (see Figure 3). Yet, paired-sample *t* tests showed that this Congruence effect was merely marginally significant, $t(30) = -1.71$, $p = .097$, $d = -.31$. As in the attend-face task, incongruent emotional face–voice combinations also showed a smaller N1 amplitude than congruent combinations in the attend-voice and attend-congruence task (see Figure 3). Although this Congruence effect was significant in the attend-voice condition, $F(1, 30) = 7.68$, $p < .05$, $\eta^2_G = .05$, it only reached near significance in the attend-congruence condition, $F(1, 30) = 3.84$, $p = .06$, $\eta^2_G = .04$. No significant results were obtained for the attend-synchrony task. Indeed, as Figure 3 illustrates, the N1 responses to incongruent

Table 4. Statistical Results for the P2 (150–230 msec) Time Window

Effect	*DFn*	*DFd*	*F*	*p*	η^2_G	*W*	p_W	*GG*	p_{GG}	$p < .05$
ROI										
Task	3	90	8.46	.00	.03	0.98	.99	0.99	.00	*
Congruence	1	30	1.62	.21	.00					
Voice	1	30	25.07	.00	.05					*
LR	1	30	2.31	.14	.00					
PA	1	30	60.04	.00	.14					*
Task × Congruence	3	90	0.81	.49	.00	0.92	.81	0.95	.49	
Task × Voice	3	90	1.18	.32	.00	0.86	.50	0.91	.32	
Congruence × Voice	1	30	5.15	.03	.01					*
Task × LR	3	90	0.81	.49	.00	0.92	.77	0.94	.48	
Congruence × LR	1	30	12.52	.00	.00					*
Voice × LR	1	30	1.32	.26	.00					
Task × PA	3	90	3.55	.02	.00	0.83	.36	0.90	.02	*
Congruence × PA	1	30	3.28	.08	.00					
Voice × PA	1	30	0.00	.97	.00					
LR × PA	1	30	17.11	.00	.00					*
Task × Congruence × Voice	3	90	0.77	.51	.00	0.97	.96	0.98	.51	
Task × Congruence × LR	3	90	0.52	.67	.00	0.95	.93	0.97	.66	
Task × Voice × LR	3	90	0.33	.81	.00	0.84	.40	0.90	.78	
Congruence × Voice × LR	1	30	0.85	.36	.00					
Task × Congruence × PA	3	90	0.03	.99	.00	0.91	.76	0.95	.99	
Task × Voice × PA	3	90	1.16	.33	.00	0.81	.30	0.89	.33	
Congruence × Voice × PA	1	30	9.57	.00	.00					*
Task × LR × PA	3	90	2.27	.09	.00	0.85	.45	0.90	.09	
Congruence × LR × PA	1	30	5.30	.03	.00					*
Voice × LR × PA	1	30	5.59	.02	.00					*
Task × Congruence × Voice × LR	3	90	0.56	.64	.00	0.56	.00	0.75	.59	
Task × Congruence × Voice × PA	3	90	0.32	.81	.00	0.84	.42	0.91	.79	
Task × Congruence × LR × PA	3	90	0.37	.77	.00	0.92	.77	0.94	.76	
Task × Voice × LR × PA	3	90	0.42	.74	.00	0.87	.54	0.92	.72	
Congruence × Voice × LR × PA	1	30	0.02	.89	.00					
Task × Congruence × Voice × LR × PA	3	90	1.68	.18	.00	0.81	.30	0.89	.18	
Midline										
Task	3	90	7.86	.00	.03	0.98	.99	0.99	.00	*
Congruence	1	30	1.56	.22	.00					
Voice	1	30	40.80	.00	.07					*
PA	1	30	62.25	.00	.13					*
Task × Congruence	3	90	0.73	.54	.00	0.89	.66	0.92	.53	

Table 4. (*continued*)

Effect	*DFn*	*DFd*	*F*	*p*	η^2_G	*W*	p_W	*GG*	p_{GG}	$p < .05$
Task × Voice	3	90	1.57	.20	.00	0.81	.29	0.87	.21	
Congruence × Voice	1	30	17.84	.00	.02					*
Task × PA	3	90	2.65	.05	.00	0.76	.17	0.85	.06	
Congruence × PA	1	30	4.74	.04	.00					*
Voice × PA	1	30	0.37	.55	.00					
Task × Congruence × Voice	3	90	1.15	.33	.00	0.96	.95	0.97	.33	
Task × Congruence × PA	3	90	0.20	.89	.00	0.94	.88	0.96	.89	
Task × Voice × PA	3	90	0.33	.81	.00	0.89	.65	0.94	.79	
Congruence × Voice × PA	1	30	6.29	.02	.00					*
Task × Congruence × Voice × PA	3	90	0.95	.42	.00	0.87	.56	0.92	.41	
Central										
Task	3	90	7.55	.00	.04	0.96	.95	0.98	.00	*
Congruence	1	30	1.17	.29	.00					
Voice	1	30	42.08	.00	.12					*
LR	1	30	0.05	.83	.00					
Task × Congruence	3	90	0.71	.55	.00	0.90	.70	0.94	.54	
Task × Voice	3	90	1.54	.21	.01	0.90	.70	0.94	.21	
Congruence × Voice	1	30	18.03	.00	.02					*
Task × LR	3	90	0.63	.60	.00	0.82	.34	0.90	.58	
Congruence × LR	1	30	14.74	.00	.00					*
Voice × LR	1	30	2.55	.12	.00					
Task × Congruence × Voice	3	90	0.66	.58	.00	0.91	.76	0.94	.57	
Task × Congruence × LR	3	90	0.29	.83	.00	0.98	.99	0.99	.83	
Task × Voice × LR	3	90	0.24	.87	.00	0.76	.17	0.88	.85	
Congruence × Voice × LR	1	30	1.57	.22	.00					
Task × Congruence × Voice × LR	3	90	0.82	.49	.00	0.53	.00	0.74	.46	

Repeated-measures ANOVAs were conducted on mean amplitudes computed across the four ROIs, midline, and central electrodes. The factors entered into the ANOVA were Task (4 levels), Congruence (congruent, incongruent), Voice (angry, neutral), LR (left, right), and PA (posterior, anterior). Where Mauchly's test (*W*) indicated that the sphericity assumption was violated, the Greenhouse–Geisser correction (*GG*) was applied. Effect size was estimated using generalized eta-squared (η^2_G). The asterisks (*) indicate effects that yielded a significance level of $p < .05$.

and congruent stimuli show virtually no difference. At the central electrodes, the Congruence × LR interaction was found to be significant in both the attend-synchrony, $F(1, 30) = 6.06, p < .05, \eta^2_G = .003$, and attend-face condition, $F(1, 30) = 4.98, p < .05, \eta^2_G = .002$. Yet, subsequent *t* tests yielded no significant results in the attend-face condition. Only in the attend-synchrony condition a significant hemispheric difference was found within the incongruent stimulus condition, $t(30) = 2.09, p < .05, d = .38$. Again, in the attend-congruence task, the Congruence × LR interaction was merely marginally significant, $F(1, 30) = 3.72, p = .06, \eta^2_G = .005$. The Congruence effect in the attend-voice condition was also only marginally significant, $F(1, 30) = 3.42, p = .07, \eta^2_G = .02$.

P2

From Figure 3 (right bar graphs), it can also be seen that the P2 amplitude was modulated by emotional face–voice congruence comparable to the N1 results. Correspondingly, the analysis of the P2 time window yielded a significant interaction between Congruence and LR across all ROIs, $F(1, 30) = 12.52, p < .001, \eta^2_G = .01$, and at

central electrodes, $F(1, 30) = 14.74, p < .01, \eta^2_G = .002$. Paired-sample t tests revealed that the Congruence effect was only significant in the left hemisphere, ROI: $t(30) = 2.38, p < .05, d = .43$; central: $t(30) = 2.20, p < .05, d = .39$. In contrast to the N1 results, the P2 congruence effect appears to vary little across the four tasks (see Figure 3). This was confirmed by statistical analysis; no significant interaction between Task and Congruence was observed (all ps $> .05$). Instead, we see that the direction of the Congruence effect differs between Neutral and Angry voice. A look at the mean amplitudes suggests (see also Figure 4) that Incongruent angry voice led to an increase in the P2 amplitude whereas Incongruent neutral voice gave rise to a reduced P2 response. In accordance with this observation, the statistical analysis yielded significant interactions between Congruence and Voice at all sites, ROI: $F(1, 30) = 5.15, p < .05, \eta^2_G = .005$; midline: $F(1, 30) = 17.84, p < .01, \eta^2_G = .02$; central: $F(1, 30) = 18.03, p < .01, \eta^2_G = .02$. Additionally, Congruence and Voice showed significant interactions with PA in the ROIs, $F(1, 30) = 9.57, p < .01, \eta^2_G = .003$, and at the midline, $F(1, 30) = 6.29, p < .05, \eta^2_G = .002$. Subsequent analyses revealed that the Congruence × Voice interaction was significant only in the anterior ROIs, $F(1, 30) = 27.13, p < .001, \eta^2_G = .06$. At the midline, the interaction was significant at both the anterior, $F(1, 30) = 37.91, p < .001, \eta^2_G = .09$, and posterior electrodes, $F(1, 30) = 4.25, p < .001, \eta^2_G = .01$. Paired-sample t tests indicated that the difference in P2 amplitude between Congruent and Incongruent neutral voice was significant in the anterior ROIs, $t(30) = 4.24, p < .001, d = .76$, and at anterior midline electrodes, $t(30) = 5.04, p < .001, d = .91$. With regard to Angry voice, the analysis revealed a significant congruence effect only at anterior midline electrodes, $t(30) = -2.72, p < .05, d = -.49$.

DISCUSSION

To summarize the above results, we found congruent and incongruent combinations of angry and neutral facial and vocal emotional expressions to modulate the auditory N1 and P2 amplitudes. The N1 and P2 were elicited approximately 70–140 msec and 150–230 msec after voice onset, respectively. Such rapid congruence effects have been interpreted as reflecting interactions between facial and vocal emotional information at an early, possibly pre-attentive processing stage (Balconi & Carrera, 2011; Pourtois et al., 2000, 2002). Here, we show that audiovisual emotional interactions within the N1, but not P2, were modulated by attention. More specifically, incongruent emotional face–voice combinations led to a reduced N1 amplitude, when compared with congruent combinations. This congruence effect was most robust in the attend-voice condition, yet weakened in the attend-face and attend-congruence condition. The effect disappeared altogether in the attend-synchrony condition. These observations strongly suggest that selective attention can influence early interactions between facial and vocal emotional information. Below, we will discuss three factors that may have been involved in the weakening of the N1 congruence effect, namely modality dominance, task-induced expectancy, and task demand, and propose that the N1 relates to a cross-sensory predictive process that is induced by anticipatory facial movements, which typically precede the voice (Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005).

In the P2 time window, incongruent and congruent emotional face–voice combinations also led to different ERP responses. However, the direction of the congruence effect differed depending on the preceding facial emotional expression. More precisely, when an angry face preceded a neutral voice, the P2 amplitude was reduced in comparison with when the neutral voice was paired with a congruent, that is, neutral facial expression. The effect reversed, that is, the P2 amplitude increased, when a neutral face preceded an angry voice as compared with when both the facial and vocal expressions were angry. Furthermore, we found no significant interaction with Task, suggesting that the P2 congruence effect was unaffected by our attentional manipulations. These results clearly contrast with that of the N1. Hence, we propose that the two components are functionally dissociated in emotional face–voice perception. We will discuss evidence in support of the notion that the P2 relates to a process, whereby visual and auditory emotional information extracted from the facial and/or vocal expression is integrated to derive an emotional percept (e.g., Kotz & Paulmann, 2011; Schirmer & Kotz, 2006). Finally, we will assess in how far our findings align with current models of multisensory integration (e.g., Talsma et al., 2010).

Interactions between Facial and Vocal Emotions within the N1 Time Window

Modulation of the N1 amplitude by emotional face–voice congruence has been reported previously (Pourtois et al., 2000). Our study extends this finding in showing that this congruence effect can be influenced by selective attention. In two unimodal tasks, that is, the attend-face and attend-voice, we asked participants to discriminate between neutral and angry facial and vocal expressions, respectively. Similar manipulations of attention were employed in a behavioral study by de Gelder and Vroomen (2000), who found facial emotional expressions to influence the judgment of vocal emotional expressions, and vice versa, despite instructions to ignore the task-irrelevant modality. Therefore, the authors concluded that facial and vocal emotional information interacts regardless of whether perceivers attend to only the face or voice.

Visual inspection of the N1 response elicited in this study suggests that facial and vocal emotional expressions interacted in both the attend-face and attend-voice condition, that is, whether attention was directed to only

the face or the voice (see Figure 3). In both task conditions, incongruent emotional face–voice combinations led to a reduced N1 amplitude as compared with congruent combinations. However, statistical analyses indicated that the N1 reduction in the incongruent condition was significant for the attend-voice but not attend-face condition. In essence, our results suggest that audiovisual emotional interaction within the N1 time window can occur independently of visual attention, yet interaction between facial and vocal emotional information weakens when the auditory modality is ignored.

A similar dichotomy between the two unimodal attention conditions is displayed in the behavioral results. We expected incongruent emotional cues from the task-irrelevant modality to distract perceivers, giving rise to erroneous responses. In accordance with this prediction, the discriminability (see Figure 2) decreased significantly in the attend-voice condition when facial expressions failed to match the vocal expressions. However, no such decline was observed in the attend-face condition, suggesting that participants successfully prevented incongruent vocal emotional information from interfering and hence interacting with facial emotional information. This suppression may account for the weakening of the congruence effect in the attend-face task. In contrast, facial emotional information may be more difficult to ignore, which could have resulted in the robust congruence effect observed in the attend-voice condition.

The Role of Modality Dominance in the Suppression of Emotional Information

In line with the above results, Talsma and colleagues (2007) also found that low-level visual and auditory information (i.e., gratings and sinusoidal tones, respectively) interacted early (i.e., around 50 msec after audiovisual stimulus onset) when attention was focused on both modalities; yet, when participants needed to ignore one of the modalities, audiovisual interactions emerged later in time and task performance declined. Talsma and colleagues (2007) speculated that suppression of task-irrelevant information in the unimodal attention condition could have prevented early audiovisual interactions. This suppression likely led to processing costs, which explains the decline in task performance.

Related to Talsma and colleagues' (2007) account, our results suggest that modality plays a crucial role in the suppression of task-irrelevant cues. In the context of emotional face–voice perception, facial emotional information may be more difficult to ignore than vocal emotional information. This dichotomy can be explained in terms of the "modality appropriateness" hypothesis, according to which the modality that provides the more accurate or appropriate information tends to dominate, that is, to have greater influence on the overall percept (e.g., Spence & Squire, 2003; Welch & Warren, 1980). In a separate emotion identification study that employed the same video stimuli as the present experiment (see supplementary material for details), we found that facial expressions were recognized with high accuracy, whereas vocal expressions were often confused. This observation is similar to that reported in previous emotion identification studies (e.g., Paulmann & Pell, 2011; Tanaka et al., 2010; Collignon et al., 2008). On the assumption that the higher reliability of facial emotional information makes the face the more dominant modality in emotion perception, the present N1 results suggest that the more dominant a modality is, the more difficult it is to ignore the modality.

In summary, employing more basic, low-level visual and auditory stimuli, Talsma and colleagues (2007) were able to show that suppression of information from either modality can obstruct early audiovisual interactions. Extending these findings, our use of facial and vocal emotional expressions in this study revealed that ignoring a modality can be difficult during audiovisual perception, if it is the more dominant one. It should be noted that such visual dominance is not observed exclusively in emotional face–voice perception, nor is it always the visual modality that dominates (for a review, see, e.g., Spence & Squire, 2003). Therefore, we would be cautious to interpret this difference in findings as to reflect a difference between social and nonsocial audiovisual stimuli.

Audiovisual Interactions in the N1 Time Window Depend on Sufficient Processing Resources

Our interpretation of the behavioral results diverts somewhat from Talsma and colleagues' (2007) in that we take the decline in task performance in the current study as an indication that perceivers failed to suppress facial emotional information, whereas Talsma and colleagues (2007) interpreted the lower accuracy rate in their study as a reflection of successful suppression. As the authors pointed out in their discussion, the suppression of task-irrelevant information in the unimodal task condition likely resulted in processing costs. Therefore, it is possible that the absence of an audiovisual interaction effect in Talsma and colleagues' (2007) study arose from a shortage in processing resources rather than from the suppression itself. Such an account would be in line with models of multisensory integration that view interactions between information from different sensory modalities as obligatory and automatic processes that are, however, constrained by limited capacity (e.g., Talsma et al., 2010).

In accordance with these models, recent studies on the McGurk effect showed that the audiovisual illusion disappeared when participants performed a concurrent demanding task (Alsius et al., 2005, 2007). In a similar vein, we observed that the N1 congruence effect disappeared completely when participants were required to discriminate between synchronous and asynchronous audiovisual speech. Complementarily, task performance

deteriorated significantly in comparison with the other conditions, pointing to high processing costs in the attend-synchrony task. This, in turn, could have undermined interactions between facial and vocal emotional expressions within the N1 time window. Our results contrast with that of Vroomen and colleagues (2001), who failed to find an effect of task demand on early emotional face–voice interaction. Their study differs from ours in that the authors employed unimodal distractors, which have no social relevance (i.e., digits and sinusoidal tones). As previously noted, unimodal distractors may not take up enough resources to impede audiovisual emotional interactions which, in turn, would explain the absent effect of task demand on emotional face–voice interaction in Vroomen and colleagues' (2001) study.

Task-induced Expectancy Attenuates N1 Congruence Effect

On the above account, one would expect visual and auditory emotional information to interact most strongly in the N1 time window when attention is directed to both the facial and vocal expressions as in the attend-congruence task. Although the grand averages pointed to a reduction in the N1 amplitude to incongruent as compared with congruent face–voice combinations, this difference failed to reach statistical significance. It appears that audiovisual interactions within the N1 time window weaken in the attend-congruence condition in relation to the attend-voice condition. A possible explanation could be that participants were prepared to hear a potentially incongruent emotional prosody at the onset of each video, which resulted in an attenuation of the N1 effect. This would add to the above observation that audiovisual emotional interactions can be influenced by top–down control (attend-face condition) and is consistent with the notion that N1 modulation in audiovisual perception is related to a cross-sensory anticipatory process (e.g., Vroomen & Stekelenburg, 2010; Stekelenburg & Vroomen, 2007).

This notion has been advocated most prominently in studies on audiovisual speech perception (e.g., van Wassenhove et al., 2005). In natural audiovisual speech, facial muscle movements typically precede the voice. Perceivers may use these visual cues to make predictions about the following auditory stimulus, which, in turn, can ease processing facilitating behavioral response. Indeed, suppression of the N1 amplitude to audiovisual, as compared with auditory-only, speech stimuli often comes with higher accuracy and faster RTs (e.g., van Wassenhove et al., 2005; Besle, Fort, Delpuech, & Giard, 2004; Klucharev et al., 2003). However, under certain circumstances, facial cues may lead to incorrect predictions. In this case, the prediction needs to be corrected, which may require a reanalysis of the input, giving rise to processing costs (e.g., Jakobs et al., 2009). Therefore, if perceivers are aware of potential mismatches between the visual and auditory stimulus, they are likely to make use of this information to prevent such prediction errors from occurring.

Cross-sensory Anticipation and N1 Suppression

Our results are at odds with Stekelenburg and Vroomen's (2007) finding that the N1 was not differentially modulated by congruent versus incongruent audiovisual (non-McGurk) speech information. This has led the authors to conclude (i) that the N1 is not sensitive to the content of the audiovisual input and, further, (ii) that its amplitude reduction is largely related to anticipatory visual motion that cues perceivers about the onset of the auditory stimulus (see also Vroomen & Stekelenburg, 2010). Given the present findings, we argue that the N1 is sensitive to the emotional content of the audiovisual input. Furthermore, we propose that the absence of a congruence effect in Stekelenburg and Vroomen's (2007) study could be due to the subtlety of audiovisually mismatched speech or the ambiguity of visual speech information (i.e., lip movement), which may not allow reliable predictions of the auditory speech content (for a review on lip reading, see, e.g., Summerfield, 1992). As Vroomen and Stekelenburg (2010) showed in their follow-up study, the N1 amplitude was not modulated when visual information could not reliably predict the auditory stimulus, suggesting that perceivers cease to use anticipatory visual cues when these prove to be unreliable.

Compared with lip movements, facial expressions likely lead to stronger predictions of the vocal emotional content. In this case, a violation of the prediction may be more pronounced, giving rise to a reduction in the N1 amplitude. At the same time, Stekelenburg and Vroomen (2007), among others, have interpreted a reduced N1 amplitude in audiovisual integration as reflecting confirmed prediction. Findings from local field potential studies with macaque monkeys show that the perception of both incongruent and congruent emotional face–voice combinations, as compared with that of vocal expressions alone, leads to suppressive effects within the auditory cortex (Kayser, Logothetis, & Panzeri, 2010). In particular, Kayser and colleagues (2010) found suppression of auditory activity induced by incongruent emotional face–voice combinations to be greater than that induced by the congruent face–voice combinations. As mentioned in the introduction, the human auditory cortex has been considered a possible neural source of the audiovisually modulated N1 (e.g., Besle et al., 2009; Woods, 1995). In the macaque, suppression of auditory activity has been shown to arise only when the auditory stimulus lags more than 200 msec behind the visual stimulus (Ghazanfar, Maier, Hoffman, & Logothetis, 2005), adding to the notion that, in audiovisual integration, auditory suppressive effects may relate to a cross-sensory anticipatory process.

Arguably, "anticipation" is a vague term and, in the context of audiovisual speech and emotion perception,

"motor resonance" may serve as a deeper explanation for the N1 modulations. According to the so-called "motor-theory of social cognition," perceivers understand another's actions, emotions, and thoughts via simulations, that is, by simulating the other's actions or imagining themselves being in the other's situation; the discovery of the so-called "mirror neurons" provided strong support for this theory (for reviews, see, e.g., Gallese, Keysers, & Rizzolatti, 2004; Gallese & Goldman, 1998; for opposing arguments, see, e.g., Jacob & Jeannerod, 2005). Although we do not want to downplay the role of simulations in social cognition, we are sceptical that motor resonance is what underlies the N1 modulations in audiovisual perception. As Vroomen and Stekelenburg (2010) have shown, the N1 was also modulated by anticipatory motion of a nonsocial visual stimulus (i.e., two colliding discs) that allowed perceivers to reliably predict the onset of a nonsocial auditory stimulus (i.e., a sinusoidal tone). To understand the exact mechanism that subserves this cross-sensory anticipatory process, we contend that more observations are necessary.

Functional Dissociation between N1 and P2 in Emotional Face–Voice Processing

Similar to the N1, the P2 amplitude showed modulations by emotional face–voice congruence. However, these modulations varied little across the task conditions, as opposed to the N1 modulation. Indeed, the statistical analysis yielded no significant interactions between Task and Congruence, suggesting that the P2 modulation was resilient to our attentional manipulations. In agreement with Vroomen and Stekelenburg (2010), we propose that the N1 and P2 reflect dissociated processes in the integration of multisensory information. This contrasts with the more traditional view that regards the N1 and P2 as one complex ERP response (Crowley & Colrain, 2004). However, we found no evidence that modulation of the P2 was related to that of the N1. For instance, emotional face–voice interactions within the P2 time window emerged despite the absence of a N1 congruence effect in the attend-synchrony. Similar observations were made by Talsma and colleagues (2007). As previously discussed, they found earlier audiovisual interactions to be susceptible to manipulations of attention, yet later interaction effects appeared to be unaffected by attentional influences.

Furthermore, we found emotional face–voice congruence to modulate the P2 amplitude differently than the N1 amplitude. As summarized above, incongruent emotional face–voice combinations only led a reduced P2 amplitude when an angry facial expression was combined with a neutral prosody. In contrast, when a neutral face was combined with an angry vocal expression, the P2 amplitude was increased. Such a reversal of the congruence effect was not observed in the N1 time window. To put it simply:

(i) Angry face + neutral voice < neutral face + neutral voice.
(ii) Neutral face + angry voice > angry face + angry voice.

Looking at (i) and (ii), it becomes clear that the P2 was reduced whenever the facial expression depicted anger or the P2 was enlarged whenever the facial expression was neutral. With the present experimental paradigm, it is not possible to determine whether it is the angry or neutral facial expression that induces the above P2 modulations. It may be also the case that the P2 is reduced by angry facial expressions as well as enlarged by neutral facial expressions. Future studies could clarify this question, for instance, by including a voice-only condition that could serve as a baseline. Taken together, these observed differences between the N1 and P2 strongly point to two dissociated processes.

Derivation of an Emotional "Gestalt" ~200 msec Post-stimulus Onset

Our P2 findings contrast with that of Balconi and Carrera (2011), who found only an effect of emotional face–voice congruence but no interaction between congruence and vocal emotion. As a result, the authors proposed that the P2 is "a cognitive marker of cross-modal integration" irrespective of the emotional content of the audiovisual stimuli. However, as elucidated above, our results suggest that modulation of the auditory P2 in emotional face–voice perception is not driven by congruence per se but shows influences of facial emotion. This observation aligns with a number of studies that found the P2 amplitude to be modulated by emotionally significant voices (Schirmer, Chen, Ching, Tan, & Hong, 2013; Garrido-Vásquez et al., 2012; Liu, Pinheiro, Deng, et al., 2012; Liu, Pinheiro, Zhao, et al., 2012; Paulmann, Seifert, & Kotz, 2010; Sauter & Eimer, 2010; Paulmann, Pell, & Kotz, 2008) and faces (Paulmann & Pell, 2009; Holmes, Kiss, & Eimer, 2006; Holmes, Vuilleumier, & Eimer, 2003). In two recent studies, modulation of the P2 was found to correlate with the level of perceived valence (Schirmer et al., 2013) and arousal of the emotional stimuli (Paulmann, Bleichner, & Kotz, 2013), adding to the notion that the component is involved in emotion processing.

In a similar vein, we found a correspondence between the P2 modulations observed in the present ERP study and valence and arousal ratings collected in a separate behavioral study. We presented the same videos of congruent and incongruent facial and vocal expressions to a different group of participants who were asked to rate the stimuli in terms of valence and arousal (see supplementary material for details). As can be seen in Figure 4, the overall audiovisual percept was rated more negative

and aroused when an angry face was combined with a neutral voice, as compared with when both facial and vocal expressions were neutral. The same incongruent stimuli led to a smaller P2 amplitude than the congruent stimuli. Conversely, when a neutral face was combined with an angry voice, the overall audiovisual percept was rated less negative and aroused than when both facial and vocal expressions were angry. In this case, the incongruent stimuli gave rise to a larger P2 amplitude than the congruent stimuli.

Furthermore, our finding that the P2 modulations in this study was unaffected by our attentional manipulations is also in line with observations from emotional voice and face perception research. As discussed in Kotz and Paulmann's (2011) review, modulation of the P2 by facial or vocal emotional information arises irrespective of whether attention is directed explicitly or implicitly to the emotional expression. This, in turn, implies that the underlying process is independent of attentional influences. Such findings have led to the notion that the P2 reflects "the rapid detection of emotional significance" (Sauter & Eimer, 2010). Given the parallels between emotional face and voice perception in terms of P2 modulation, Sauter and Eimer (2010) proposed that this process possibly involves "supramodal brain mechanisms" (see also Kotz & Paulmann, 2011)). Our results may be regarded as evidence for supramodal brain mechanisms in the modulation of the P2 (see also Liu, Pinheiro, Zhao, et al., 2012; Balconi & Carrera, 2011). At the same time, they suggest that the process underlying the P2 in emotional face–voice perception is somewhat more complex than "detection of emotional significance."

In their original model of emotional voice processing, Schirmer and Kotz (2006) linked the P2 to a process, whereby "emotionally significant acoustic information [is integrated] to derive an [auditory] emotional 'gestalt.'" Following Schirmer and Kotz's (2006) proposal, this process takes place within the anterior STS—a brain region that is associated with a multitude of different functions (for a review, see, e.g., Hein & Knight, 2008). According to Hein and Knight (2008), the anterior portion of the STS is activated in speech and face processing as well as audiovisual integration, which makes it a good candidate for audiovisual emotional integration (see also Kreifelts, Ethofer, Shiozawa, Grodd, & Wildgruber, 2009). However, whether the anterior STS is the site at which the P2 is generated is unclear. Attempts to localize the neural generators of the P2 point to different areas of the temporal lobe; what seems likely is that the P2 is generated from several neural sources (Crowley & Colrain, 2004). Whether the anterior STS is one of these generators is an exciting question that may be explored in future studies.

Does the P2 Reflect a General Stimulus Classification Process?

On a final note, we would like to point out that other factors, which are not necessarily emotion specific, have been found to influence the auditory P2. Results by Pinheiro and colleagues (2011), for instance, showed interactions between emotional prosody and lexicality within the P2 time window. Specifically, a larger P2 amplitude was elicited by (emotionally spoken) sentences as compared with sentences without semantic content ("pure prosody" sentences). Modulation of the P2 has been also observed in non-emotional audiovisual context (Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005; Klucharev et al., 2003). Of note is Stekelenburg and Vroomen's (2007) finding that the P2 is differentially modulated by congruent and incongruent audiovisual speech stimuli (that do not elicit a McGurk effect). In an early study, García-Larrea, Lukaszewicz, and Mauguière (1992) proposed that the P2 could denote a general stimulus classification process. On this account, the modulations of the P2 observed in this study may be only indirectly linked to the emotional content of the stimuli, as they simply reflect the classification of the four combinations of facial and vocal emotional expressions. Therefore, to clarify the process underlying the P2 in multisensory perception, we suggest that future studies investigate at least two questions. First, what other factors, besides emotion, modulate the component and, second, do these factors influence the P2 in a similarly way as facial and vocal emotional expressions?

Emotional Face–Voice Perception as a Composite of Multiple Subprocesses

Taken together, our ERP findings reveal that emotional face–voice perception is subserved by at least two processes denoted by the auditory N1 and P2. These processes appear to be independent of one another in the sense that, whatever factor (e.g., attention) affects the N1, its modulation has no obvious consequence on the P2. This finding diverts somewhat from Talsma and colleagues' (2010) model of multisensory integration, which regards the integration process as a sequence of multiple processing steps, each of which is triggered by the preceding one. In this view, attention is primarily needed when one of the processes cannot be carried out effectively, because the available resources are insufficient. In contrast, we consider multisensory integration as a composite of multiple subprocesses that do not necessarily take place sequentially, but can occur, more or less, in parallel. Some of these subprocesses, such as the one underlying the N1, may be influenced by various factors, including attention, expectancy, and task demand, whereas others, such as the one subserving the P2, may show modulation of additional factors that could be explored in future research. Finally, we have discussed findings which suggest that the processes underlying the N1 and P2, discussed above, may not be specific to emotional face–voice perception. This raises the possibility that these processes are also involved in other types of human communication which, as discussed at the beginning, is in its natural form multisensory; this includes audiovisual speech perception, but also

audiovisual perception of music and dance performance, and so forth.

Conclusion

This study set out to investigate the hypothesis that interactions between facial and vocal emotional expressions occur early and independently of attention. Early interactions between facial and vocal emotional information were, indeed, found within the auditory N1 and P2 time window. However, inconsistent with the above hypothesis, interactions within the earlier N1 time window were susceptible to manipulations of attention. In contrast, audiovisual emotional interactions within the later P2 time window appeared to be unaffected by attentional modulation. This difference between the N1 and P2 suggests that modulations of these two components reflect two independent processes subserving the combined perception of facial and vocal emotional expressions. We discussed two possible functions underlying the N1 and P2, namely cross-sensory anticipation and derivation of an emotional percept. Related to these two processes, we identified a number of open questions that may need to be addressed in future research, in particular, whether the processes associated with the N1 and P2 are emotion specific. In essence, our findings challenge the widely held "early integration" view of emotional face–voice perception. They show that although, in everyday life, we integrate facial and vocal emotional expressions in a seemingly automatic manner—that is fast, unconscious, and apparently without much effort—under certain task conditions, interactions between facial and vocal emotional information may weaken or fail altogether.

Acknowledgments

This research was conducted as part of the DFG (Deutsche Forschungsgesellschaft) graduate program "Function of Attention in Cognitive Processes." The authors would like to thank Andreas Widmann, Maren Grigutsch, and Burkhard Maess for advice on data analysis and Claudia Teickner for help with data collection.

Reprint requests should be sent to Hao Tam Ho, Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstr. 1A, 04103 Leipzig, Germany, or via e-mail: htho@cbs.mpg.de.

REFERENCES

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15,* 839–843.

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research, 183,* 399–404.

Balconi, M., & Carrera, A. (2011). Cross-modal integration of emotional face and voice in congruous and incongruous pairs: The P2 ERP effect. *Journal of Cognitive Psychology, 23,* 132–139.

Besle, J., Bertrand, O., & Giard, M.-H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research, 258,* 143–151.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20,* 2225–2234.

Bradley, M., & Lang, P. J. (1994). Measuring emotion: The self-assessment semantic differential manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25,* 49–59.

Calvert, G., & Thesen, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris, 98,* 191–205.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., et al. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242,* 126–135.

Compton, R. (2003). The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and Cognitive Neuroscience Reviews, 2,* 115–129.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1,* 42–45.

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology, 115,* 732–744.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7,* 460–467.

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

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134,* 9–21.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology, 92,* 53–78.

Fairhall, S. L., & Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *European Journal of Neuroscience, 29,* 1247–1257.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory. *Trends in Cognitive Sciences, 2,* 493–501.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8,* 396–403.

García-Larrea, L., Lukaszewicz, A. C., & Mauguière, F. (1992). Revisiting the oddball paradigm. Non-target vs neutral stimuli and the evaluation of ERP attentional effects. *Neuropsychologia, 30,* 723–741.

Garrido-Vásquez, P., Pell, M. D., Paulmann, S., Strecker, K., Schwarz, J., & Kotz, S. A. (2012). An ERP study of vocal emotion processing in asymmetric Parkinson's disease. *Social Cognitive and Affective Neuroscience, 8,* 918–927.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25,* 5004–5012.

Ghazanfar, A. A., & Schroeder, C. C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10,* 278–285.

Hein, G., & Knight, R. (2008). Superior temporal sulcus—It's my area: Or is it? *Journal of Cognitive Neuroscience, 20,* 2125–2136.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.
Holmes, A., Kiss, M., & Eimer, M. (2006). Attention modulates the processing of emotional expression triggered by foveal faces. *Neuroscience Letters, 394,* 48–52.
Holmes, A., Vuilleumier, P., & Eimer, M. (2003). The processing of emotional facial expression is gated by spatial attention: Evidence from event-related brain potentials. *Brain Research, 16,* 174–184.
Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences, 9,* 21–25.
Jakobs, O., Wang, L. E., Dafotakis, M., Grefkes, C., Zilles, K., & Eickhoff, S. B. (2009). Effects of timing and movement uncertainty implicate the temporo-parietal junction in the prediction of forthcoming motor actions. *Neuroimage, 47,* 667–677.
Kayser, C., Logothetis, N. K., & Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Current Biology, 20,* 19–24.
Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research, 18,* 65–75.
Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134,* 372–384.
Kotz, S., & Paulmann, S. (2011). Emotion, language, and the brain. *Language and Linguistics Compass, 3,* 108–125.
Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009). Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia, 47,* 3059–3066.
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. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences, 9,* 75–82.
Lawrence, M. A. (2013). *ez:* Easy analysis and visualization of factorial experiments (Version 4.2-2). Retrieved from cran.r-project.org/package=ez.
Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research, 16,* 468–478.
Liu, T., Pinheiro, A., Deng, G., Nestor, P. G., McCarley, R. W., & Niznikiewicz, M. A. (2012). Electrophysiological insights into processing nonverbal emotional vocalizations. *NeuroReport, 23,* 108–112.
Liu, T., Pinheiro, A., Zhao, Z., Nestor, P. G., McCarley, R. W., & Niznikiewicz, M. A. (2012). Emotional cues during simultaneous face and voice processing: Electrophysiological insights. *PLoS One, 7,* e31001.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Hove, UK: Psychology Press.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 260,* 746–748.
Mognon, A., & Jovicich, J. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology, 48,* 1–12.
Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology, 24,* 375–425.
Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion, 11,* 4–11.
Paulmann, S., Bleichner, M., & Kotz, S. A. (2013). Valence, arousal, and task effects in emotional prosody processing. *Frontiers in Psychology, 4,* 345.
Paulmann, S., & Pell, M. D. (2009). Facial expression decoding as a function of emotional meaning status: ERP evidence. *NeuroReport, 20,* 1603–1608.
Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion, 35,* 192–201.
Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104,* 262–269.
Paulmann, S., Seifert, S., & Kotz, S. A. (2010). Orbito-frontal lesions cause impairment during late but not early emotional prosodic processing. *Social Neuroscience, 5,* 59–75.
Pinheiro, A. P., Galdo-Álvarez, S., Rauber, A., Sampaio, A., Niznikiewicz, M., & Gonçalves, O. F. (2011). Abnormal processing of emotional prosody in Williams syndrome: An event-related potentials study. *Research in Developmental Disabilities, 32,* 133–147.
Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., & Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *NeuroReport, 11,* 1329–1333.
Pourtois, G., Debatisse, D., Despland, P.-A., & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive Brain Research, 14,* 99–105.
R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from www.r-project.org/
RStudio. (2013). RStudio: Integrated development environment for R (Version 0.98.490). Boston, MA. Retrieved from www.rstudio.org/.
Sauter, D. A., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience, 22,* 474–481.
Schirmer, A., Chen, C.-B., Ching, A., Tan, L., & Hong, R. Y. (2013). Vocal emotions influence verbal memory: Neural correlates and interindividual differences. *Cognitive, Affective & Behavioral Neuroscience, 13,* 80–93.
Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10,* 24–30.
Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology, 13,* R519–R521.
Stein, B. B. E., & Stanford, T. T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9,* 255–266.
Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19,* 1964–1973.
Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335,* 71–78.
Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex, 17,* 679–690.
Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14,* 400–410.
Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience, 17,* 1098–1114.

Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder, B. (2010). I feel your voice. Cultural differences in the multisensory perception of emotion. *Psychological Science, 21,* 1259–1262.

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Brain Research, 14,* 106–114.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16,* 457–472.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96,* B13–B22.

van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *Neuroimage, 55,* 1208–1218.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 1181–1186.

Vatakis, A., Ghazanfar, A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8,* 1–11.

Vatakis, A., & Spence, C. (2006). Temporal order judgments for audiovisual targets embedded in unimodal and bimodal distractor streams. *Neuroscience Letters, 408,* 5–9.

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127,* 12–23.

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

Vroomen, J., & Stekelenburg, J. J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22,* 1583–1596.

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences, 9,* 585–594.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Widmann, A., & Schröger, E. (2012). Filter effects and filter artifacts in the analysis of electrophysiological data. *Frontiers in Psychology, 3,* 233.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions, 7,* 30.

Woods, D. (1995). The component structure of the N1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology-Supplements Only, 44,* 102–109.