Brain and Cognition 69 (2009) 269–278

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Superior temporal activation in response to dynamic audio-visual emotional cues ☆

Diana L. Robins [a,*], Elinora Hunyadi [b], Robert T. Schultz [b]

[a] Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302-5010, USA
[b] Center for Autism Research, Children's Hospital of Philadelphia, 3535 Market Street, Suite 860, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Accepted 1 August 2008
Available online 21 September 2008

Keywords:
Audio-visual integration
fMRI
Prosody
Face perception
Emotion
Cross-modal

ABSTRACT

Perception of emotion is critical for successful social interaction, yet the neural mechanisms underlying the perception of dynamic, audio-visual emotional cues are poorly understood. Evidence from language and sensory paradigms suggests that the superior temporal sulcus and gyrus (STS/STG) play a key role in the integration of auditory and visual cues. Emotion perception research has focused on static facial cues; however, dynamic audio-visual (AV) cues mimic real-world social cues more accurately than static and/or unimodal stimuli. Novel dynamic AV stimuli were presented using a block design in two fMRI studies, comparing bimodal stimuli to unimodal conditions, and emotional to neutral stimuli. Results suggest that the bilateral superior temporal region plays distinct roles in the perception of emotion and in the integration of auditory and visual cues. Given the greater ecological validity of the stimuli developed for this study, this paradigm may be helpful in elucidating the deficits in emotion perception experienced by clinical populations.

1. Introduction

Emotion perception is a critical aspect of social interaction; in order to interact with others appropriately, it is essential to understand how social partners feel. Emotion processing is inherently multimodal (de Gelder & Vroomen, 2000), yet much of the cognitive neuroscience and neuroimaging literature on emotion perception uses artificial unimodal paradigms (e.g., static photos displaying emotional facial expressions). In order to understand the neural mechanisms that underlie emotional judgments during real-world social interactions, novel approaches need to be developed that realistically integrate emotional cues from affective prosody and facial expressions. Such paradigms will permit the study of brain activity during emotion perception in the context of social interchange.

A small number of recent studies have attempted to study the integration of affective prosody and emotional facial expression using neuroimaging techniques (Dolan, Morris, & de Gelder, 2001; Ethofer, Anders et al., 2006; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007; Pourtois, de Gelder, Bol, & Crommelinck, 2005). Most of these studies utilized stimuli consisting of a static photograph coupled with a short audio track; the exception is that Kreifelts and colleagues (2007) utilized dynamic video paired with a single spoken word. In studies that contrasted an audio-visual condition with unimodal conditions, increased activation during the audio-visual condition spanned from superior temporal gyrus to middle temporal gyrus (Kreifelts et al., 2007; Pourtois et al., 2005). When a congruent audio-visual condition was contrasted with incongruent affect presented in face and voice, similar to the paradigm developed by de Gelder and Vroomen (2000), the congruent condition was associated with greater left amygdala, right fusiform gyrus, left anterior cingulate, and left middle temporal gyrus (Dolan et al., 2001; Ethofer, Anders et al., 2006).

These findings are consistent with results from the more extensive literature examining the neural mechanisms of sensory integration, which utilizes paradigms such as the McGurk effect (McGurk & MacDonald, 1976), in which the mouth forms one phoneme which blends with an auditory percept of another phoneme, perception of two moving bars as colliding or passing through one another depending on the timing of an auditory burst (Bushara et al., 2003), lip reading (Calvert, Brammer, & Iverson, 1998; Sumby & Pollack, 1955) or reading written text (Frost, Repp, & Katz, 1989). These studies consistently demonstrate activation in the superior temporal cortex, most often in the posterior region (Beauchamp, Lee, Argall, & Martin, 2004; Bushara et al., 2003; Calvert & Campbell, 2003; Calvert, Campbell, & Brammer, 2000; Jones & Callan, 2003; Olson, Gatenby, & Gore, 2002; Saito et al., 2005; van Atteveldt, Formisano, Goebel, & Blomert, 2004; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003); in addition, other regions that demonstrate activation in some of the studies include the amygdala (Dolan et al., 2001), insula/claustrum (Bushara et al., 2003; Olson et al., 2002), the superior colliculi (Bushara et al., 2003; Calvert and Thesen, 2004).

☆ This study was carried out at the Yale University School of Medicine Child Study Center.

* Corresponding author. Fax: +1 404 413 6207.
E-mail address: drobins@gsu.edu (D.L. Robins).

0278-2626/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandc.2008.08.007

Given the fairly small literature on the neural mechanisms of audio-visual emotion perception, it is helpful to examine the more extensive neuroimaging literature on unimodal emotion perception. A network of brain areas has been identified as playing a role in the perception of visual emotional cues from static facial expressions, relative to a variety of comparison conditions and utilizing a range of tasks requiring direct or indirect emotion processing. Some of these regions include the amygdala (Gur et al., 2002; Hariri, Bookheimer, & Mazziotta, 2000; Keightley et al., 2003), fusiform gyrus (Gur et al., 2002; Keightley et al., 2003; Kesler et al., 2001), insula (Keightley et al., 2003), and superior and middle temporal regions (Iidaka et al., 2001). A growing literature examining brain areas involved in the perception of affective prosody is not yet consistent for exact brain regions. However, most studies implicate frontal and temporal cortices, usually biased toward the right hemisphere (Buchanan et al., 2000; Mitchell, Elliott, Barry, Cruttenden, & Woodruff, 2003; Wildgruber et al., 2005).

The present two studies take the next step toward elucidating neural mechanisms of emotion perception by examining activation during the presentation of ecologically valid, dynamic, audio-visual stimuli in which the emotional prosody is consistent with the semantic content, improving on previous approaches such as the presentation of static photographs combined with brief auditory clips. Dynamic stimuli provide a better approximation of real-world social interactions than do static stimuli, because they require participants to monitor moment-to-moment changes in emotions expressed by others (Harwood, Hall, & Shinkfield, 1999; Sato, Kochiyama, Yoshikawa, Naito, & Matsumura, 2004). Thus, they permit a more comprehensive understanding of the neural pathways involved in emotion processing (de Gelder & Bertelson, 2003; de Gelder & Vroomen, 2000; Gepner, Deruelle, & Grynfeltt, 2001; Wildgruber et al., 2004; Wright et al., 2003).

Although this initial study involves neurotypical individuals, the stimuli developed for the current study may be useful in studies of atypical emotion perception, such as in autism spectrum disorders, anxiety disorders, mood disorders, and schizophrenia. Laboratory investigations are most useful when their findings can be generalized to real social interactions, and this is most likely to be possible when the stimuli most closely reflect the social demands and real-world experiences. Using autism as an example, the unimodal emotion perception literature indicates mixed results when examining deficits in emotion processing; some studies indicate that individuals with ASD have difficulty identifying facial expressions (e.g., Adolphs, Sears, & Piven, 2001; Celani, Battacchi, & Arcidiacono, 1999; Hobson, Ouston, & Lee, 1988; Yirmiya, Sigman, Kasari, & Mundy, 1992), and emotional prosody (Boucher, Lewis, & Collis, 2000; Fujiki, Spackman, Brinton, & Illig, 2008; Peppe, McCann, Gibbon, O'Hare, & Rutherford, 2007), compared to typically developing children, whereas others find no significant differences between individuals with ASD and controls (Gepner et al., 2001; Ozonoff, Pennington, & Rogers, 1990). However, there have been very few studies examining the integration of audio-visual emotion cues in autism (e.g., Haviland, Walker-Andrews, Huffman, Toci, & Alton, 1996; Loveland et al., 1995), and the prior studies use paradigms that are not comparable to naturalistic social interactions, such as preferential looking paradigms in which the participant sees two video displays and hears one audio track.

Furthermore, two aspects of the stimuli developed for the present studies warrant mention. First, in the dynamic audio-visual stimuli used in the current studies, the auditory segment lasts for the duration of the video clip, in contrast to previous studies (e.g., Kreifelts et al., 2007), in which the auditory stimulus was much shorter, and a portion of the bimodal condition was, in fact, video only. Second, in the current study the semantic content is emotionally ambiguous, meaning that the language naturally makes sense in multiple affective contexts (i.e., if the stimuli were shown in the context of a paragraph setting the emotional tone, the affective prosody would be consistent with the semantic content of the sentence). In previous studies, the semantic content of the auditory tracks was affectively neutral, meaning the words may not have seemed natural when spoken in emotional tone of voice (e.g., hearing the sentence, "The guest removed a room for Thursday" spoken with affective prosody; Ethofer, Anders et al., 2006), or participants were instructed to disregard the semantic content's emotional valence (e.g., hearing the word "pus" spoken in a happy voice; Kreifelts et al., 2007). Incongruence between semantic content and emotional tone could confound activation findings, by drawing the participant's attention to the incongruence or by affecting the way in which participants process the prosody. In effect, emotionally ambiguous semantic content reduces unwanted attention to the semantic content and allows the participant to focus on the affective prosody.

Participants in the present studies viewed short movies blocked by modality (audio, video, and audio–video) and/or emotion (angry, fearful, happy, and neutral), as well as unimodally presented facial and auditory emotional cues while undergoing fMRI scanning. Activation or enhancement of activation to the AV emotional stimuli was contrasted with activation during unimodal conditions; effects of emotion were also investigated. The neural substrates underlying perception of emotion in different modalities were examined with fMRI using region-of-interest (ROI) and whole-brain analyses.

A priori regions-of-interest (ROIs) for AV integration (Study 1a and Study 2) are based on the literature examining AV integration of emotion perception as well as the more extensive literature of AV integration in language and sensory paradigms. It is hypothesized that the AV condition will demonstrate increased activation in the superior temporal sulcus (STS), fusiform gyrus, cingulate gyrus, insula, superior colliculi, and amygdala, relative to the unimodal conditions.

A priori ROIs for the emotion contrasts, regardless of modality (Study 1b and Study 2), are based on the unimodal emotion perception literature as well as the two studies that used AV emotional conditions. Based on the findings from the literature on unimodal emotion perception as well as the more recent exploration of multimodal and dynamic emotion perception, it is hypothesizes that amygdala, fusiform gyrus, superior and middle temporal cortex, and insula will demonstrate increased activation during perception of emotional stimuli relative to neutral stimuli.

2. Materials and methods

2.1. Participants

2.1.1. Study 1

Ten individuals (mean age 22.3 ± 4.6, range 18–33 years) were recruited for this study. Participants included three males (mean age = 27.3, *SD* = 5.5, age range = 22–33 years) and seven females (mean age = 20.4, *SD* = 2.1, age range = 18–23 years).

2.1.2. Study 2

Five individuals (mean age 20.6 ± 1.8, range 18–23) who had not participated in Study 1 were recruited for Study 2. Participants included three males (mean age = 20.7, *SD* = 2.5, age range = 18–23) and two females (mean age = 20.5, *SD* = 0.7, age range = 20–21).

For both studies, each participant gave written informed consent, which was approved by the Human Investigation Committee at the Yale University School of Medicine, and has therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. All participants reported no his-

tory of neurological or significant psychiatric illness or injury. Participants were excluded based on self-report of metal in their body. Each participant received $50 compensation.

2.2. Stimuli

Novel audio–visual (AV) stimuli were developed at the Yale Child Study Center (Robins & Schultz, 2004; see http://www2.gsu.edu/~wwwpsy/faculty/robins.htm for sample stimuli). Two professional actors (one female and one male) delivered 10 sentences in four emotions (angry, fearful, happy, and neutral). Sentence content was emotionally ambiguous (see Table 1 for sentences); each was feasibly delivered in all four emotional tones. For example, "The door is open" can be expressed as *angry* if someone left a door open and rain flooded the hallway, *fearful* if the speaker fears an intruder opened the door, *happy* if the speaker is welcoming someone, and *neutral* as a statement of fact. Video recording captured both emotional intonation and facial expression. Stimuli were separated into audio-only (A) and video-only (V) format, as well as kept intact (audio–video; AV), totaling 240 stimuli (80 A, 80 V, and 80 AV). Visual stimuli included just the actor's head with no sound accompanying the video display. During presentation of audio stimuli, the screen was black. Stimuli averaged 1.5 s (±0.3; range = 0.8–2.1 s) in duration. Both fMRI studies utilized stimuli from this set of 240 videos.

Behavioral piloting (n = 16; 8 males, 8 females) with A and V stimuli used a forced choice format; participants viewed audio-only clips and video-only clips, and labeled the emotion as angry, fearful, happy, or neutral. Multiple takes of each unimodal stimulus were screened for a minimum accuracy of 80%. Three rounds of pilot testing were conducted until the 80% criteria was reached on one audio and one video clip for each sentence performed by each actor (total: 80 audio, 80 video).

Movie presentation was programmed using PsyScope (Cohen, MacWhineey, Flatt, & Provost, 1993), and presented to participants using a Macintosh G3 laptop computer. Stimuli were projected onto a screen placed at the foot of the scanner, in a dimly lit room.

2.3. Procedure

After screening to ensure MRI safety, participants were placed in the scanner and provided with a button box. Each participant's head was stabilized using foam cushions and a piece of tape placed across the forehead. Participants viewed stimuli using a pair of mirrors placed on the head coil. View was centered and focused for each participant before scanning commenced.

Participants in both studies were told that stimulus presentation would vary by modality and emotion. Their task was to watch and listen attentively to each stimulus, and to press the button at the end of every stimulus, as quickly as possible. The aim of the button press was to ensure that participants attended to the stimuli; all participants successfully completed this task.

Table 1
Ten ambiguous sentences in AV stimuli

Sentence	Number of words	Characters per word	Reading level
Clouds are in the sky	5	3.4	<.1
I didn't expect you	4	4.0	.7
It might happen	3	4.3	1.3
It's across the street	4	4.8	.7
It's dark already	3	5.0	5.2
Look in the box	4	3.0	<.1
Put it down	3	3.0	<.1
The dog is barking	4	3.8	.7
The door is open	4	3.2	.7
Turn off the television	4	5.0	6.6

2.3.1. Study 1

Study 1 entailed a block design, with six runs in total, divided into two parts (designated as Study 1a and Study 1b); each part consisted of three runs. Runs for Study 1a and 1b were alternated within each participant, and were counterbalanced across participants. In Study 1a, stimuli were blocked by sensory modality (A, V, and AV), with emotion type mixed within each block (run time = 5 min, 33 s; see Fig. 1). This design allowed for investigation of AV integration relative to each unimodal condition (A and V). Study 1a contained 12 blocks per run (mean = 13.8 s, SD = 0.42, range 13.5–14.7 s), with four blocks per condition; block order was pseudorandomly sequenced. During Study 1b, stimuli were blocked by emotion (angry, happy, fearful, and neutral), with sensory modality intermixed within each block. This design measured neural activation to emotional stimuli relative to neutral stimuli. Study 1b contained 16 blocks per run (mean = 13.9, SD = 0.32, range 13.5–14.8 s), with four presentations of each emotion in pseudorandom design (run time = 7 min, 30 s).

For all runs, each block consisted of five stimuli presented with a 1000 ms ISI. Between movie condition blocks was a 10-s interval in which static pictures of houses were presented. The house condition was not used as part of any contrast in the present study.

2.3.2. Study 2

Study 2 also used a block design, with six runs. However, in contrast to Study 1, which presented blocks that mixed emotions in the modality runs, and blocks that mixed modalities during the emotion runs, Study 2 stimuli were limited to fearful and neutral stimuli blocked by emotion and modality. Fear was chosen as the single emotion to present based on the large literature on fear perception, particularly relating to amygdala activation, compared to relatively fewer prior studies on anger and happiness perception. This design allowed for the analysis of the unique contributions of modality and emotionality effects during the passive viewing task. Each run contained six different types of blocks: A fear (AF), V fear (VF), AV fear (AVF), A neutral (AN), V neutral (VN), and AV neutral (AVN). Runs consisted of 12 blocks (mean = 13.9 s, SD = 0.14 s, range 13.2–14.8 s), with two blocks of each condition presented in pseudorandom order. Rest between blocks consisted of 9 s of blank screen followed by 1 s of the fixation point. Run time was 5 min, 43 s.

2.4. Magnetic resonance imaging data acquisition and analysis

Functional magnetic resonance imaging (fMRI) was conducted on a 3.0 Tesla Siemens Trio scanner at the Yale University School of Medicine Magnetic Resonance Research Center, with a standard birdcage head coil. Following localizer scans, 2D anatomical scans

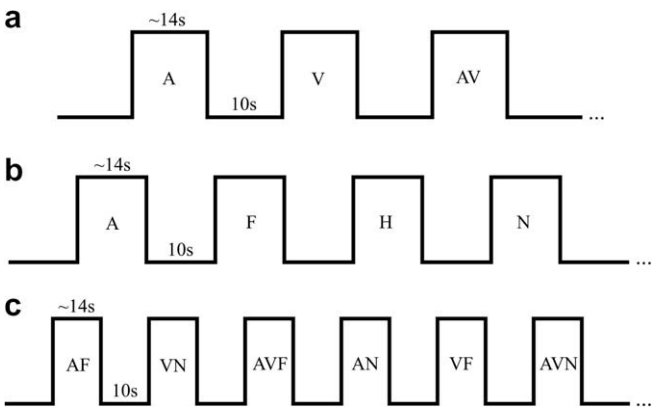

Fig. 1. Block design for Study 1a (a), Study 1b (b), and Study 2 (c).

were acquired for in plane coregistration with the EPI functional data (T1 flash, axial oblique plane through the AC-PC, 34 slices, 4 mm^3 isotropic voxels, no gap TR = 300, TE = 2.47, flip angle = 60°) with full cortex coverage, first slice prescribed "one slice above vertex" (top of brain). Next, six functional runs were acquired in the axial AC/PC plane, using a gradient echo, single-shot echoplanar sequence (TR = 1950, TE = 25, flip angle = 60°, 34 slices, 4 mm^3 isotropic voxels, no gap). The last scan acquired was the 3D MPRAGE 1 mm^3 anatomical image, also for functional localization (176 slices, 1 mm^3 isotropic voxels, TR = 2530, TE = 3.66, flip angle = 7°).

All image preprocessing for Study 1 and Study 2 was performed using Brain Voyager 2000 (Goebel, 2000). Data analyses were performed using Brain Voyager 2000 (Goebel, 2000) and Brain Voyager QX (Goebel, 2004); the only difference between the studies is that Study 1 used random effects analysis, and Study 2 used fixed effects analysis due to its small sample size. Preprocessing included intrasession alignment, motion correction, 7 mm FWHM Gaussian spatial smoothing, and linear trend removal. The first four volumes of each run were discarded. The functional image was coregistered to the 3D anatomical image, and the 3D image was then transformed into standard Talairach space by applying piece-wise linear transformation. The Talairach and coregistration transformations were applied to the functional data in order to interpolate it into standard 3D 3 mm^3 space. Non-brain space and cerebellum were masked, and white matter was ignored. All images are shown using radiological convention (left = right). Parametric maps were obtained using a general linear model (GLM) with multiple conditions.

Analyses examined specific task contrasts using the *t* statistic. For both Study 1 and Study 2, the threshold for analysis of a priori regions-of-interest was $p < .01$; for whole-brain analyses, a higher threshold of $p < .001$ was used to account for multiple comparisons. Specific post hoc analyses examining the contrast of individual emotions relative to neutral stimuli (e.g., Angry > Neutral) indicated widespread activation at the threshold of $p < .001$. In order to identify foci of activations and reduce the incidence of activated regions running together, a threshold of $p < .0001$ was used to describe these results. Conjunction analyses were also used, which combine *t*-tests to indicate regions that meet the threshold for each *t*-test included the conjunction. For example, rather than average the two unimodal conditions (A and V) in comparison to the cross-modal condition (AV), the more stringent conjunction analysis only identifies those regions significantly more active for both AV versus A *and* AV versus V, using the notation (AV > A)∩(AV > V), demonstrating preferential activation to cross-modal stimuli.

Peak activations were determined by examining *t* scores in the entire activated region; Tables 2–4 provide cluster size in mm^3, and Talairach coordinates of peak voxel. Figures show cross-hairs on the peak voxel, and event-related time course averaging was performed using the peak voxel. All functional data is presented on a composite 3D anatomical image, consisting of the average brain of all participants in each study.

3. Results

3.1. Study 1a

The posterior superior temporal sulcus (pSTS) demonstrated activation significantly above baseline in all three modality conditions (A, V, and AV). In addition, a conjunction analysis comparing activation maps for (AV > A)∩(AV > V) indicated increased bilateral pSTS to AV stimuli relative to both types of unimodal stimuli ($p < .01$; see Table 2 and Fig. 2). However, no preferential activation to AV was seen in other targeted ROIs (e.g., amygdala, superior colliculi) even when a lower threshold of $p < .05$ was used.

3.2. Study 1b

The effect of emotion relative to neutral stimuli was highly significant; therefore, the more stringent threshold of $p < .001$ was used for all tables and figures to facilitate the identification of unique activations. Comparison of an average of all emotional stimuli to neutral stimuli indicated greater activation (all $p < .001$) to emotional stimuli in bilateral anterior superior temporal gyrus (aSTG) and bilateral fusiform gyrus (FG; see Table 3 and Fig. 3a–d). A conjunction analysis of each individual emotion compared to neutral ([A > N]∩[F > N]∩[H > N]) demonstrated significant activation only in right aSTG and left FG (see Table 3 and Fig. 3e–f). To further explain the difference between these two analyses, consider that the first emotion analysis addresses whether preferential activation is found for the average of all emotions compared to no emotion (neutral). The second emotion analysis addresses whether preferential activation is found for every emotion relative to no emotion. It is important to note that activations in these analyses were overlapping, and peaks in right aSTG and left FG are within 4 mm of one another.

Effects also were seen for individual emotions compared to no emotion (see Table 3). Angry > Neutral (see Table 3) was a highly significant contrast; therefore, activations are presented at an even higher threshold of $p < .0001$. Angry blocks led to greater activation than Neutral blocks in right anterior STG, and a region spanning the left STG and the lateral fissure, $p < .0001$. Fearful > Neutral demonstrated significant activations in the left superior frontal sulcus ($p < .0001$). Happy > Neutral showed no significant findings

Table 2
Study 1a: Activations regarding AV integration

Brain region	Cluster size (mm^3)	Talairach			*t* value
		x	*y*	*z*	
(AV > A)∩(AV > V)					
Left pSTS	282	−66	−31	4	5.68[a]
Right pSTS	49	42	−34	10	4.10[a]
Left pSTS	12	−45	−37	10	3.66[a]

A priori ROI analysis on group averaged maps.
[a] $p < .01$.

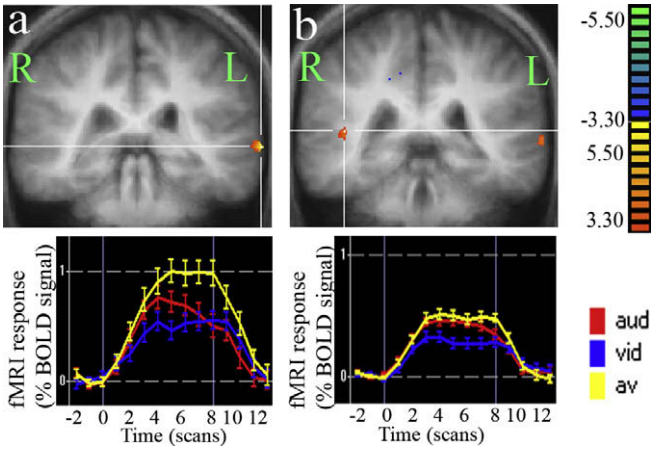

Fig. 2. Increased activation in bilateral STS (a: −66, −31, 4; b: 42, −34, 10) for (AV > A)∩(AV > V); $t(9) = 3.3$, $p < .01$. Cross-hairs on brain images represent the voxel of peak activation. Time course shown below each coronal image represents activation in the peak voxel. On the *x*-axis, 0 indicates block onset, and 8 indicates block offset.

Table 3
Study 1b: Activations regarding emotion

Brain region	Cluster size (mm^3)	Talairach			*t* value
		x	*y*	*z*	
Emotion > neutral					
Left aSTG	194	−63	−10	1	6.69[b]
Right aSTG	190	63	−13	4	5.84[b]
Right fusiform gyrus	138	36	−37	−14	6.14[b]
Right aSTG	126	54	2	−5	6.08[b]
Left fusiform gyrus	27	−42	−46	−17	5.62[b]
(Angry > Neutral)∩(Fearful > Neutral)∩(Happy > Neutral)					
Left FG	28	−39	−46	−17	4.38[a]
Right anterior STG	27	58	2	−2	4.27[a]
Angry > neutral					
Left aSTG/lateral fissure	726	−57	−13	10	10.70[c]
Right aSTG	203	57	2	1	9.38[c]
Right pSTG	180	66	−34	13	9.70[c]
Right aSTG	48	48	17	−14	9.07[c]
Right aSTG	7	54	11	−5	7.76[c]
Fearful > neutral					
Left superior frontal sulcus	14	−12	26	46	8.92[c]

Whole-brain analysis on group averaged maps.
[a] $p < .01$.
[b] $p < .001$.
[c] $p < .0001$.

even at the standard threshold of $p < .001$. A ROI analysis of STS/STG at $p < .01$ indicated a single significant activation for Happy > Neutral in right aSTG (peak: 57, 2, −2).

3.3. Study 2

In order to examine separately the effects of emotion and AV integration on selective activation in the STS/STG region, a second study was conducted, using only fearful and neutral stimuli, with blocks separating both emotion and modality. Corroborating the findings from Study 1a, a conjunction analysis of AV versus V and AV versus A, separated by emotion ([AVF > VF]∩[AVF > AF]∩[AVN > VN]∩[AVN > AN]) demonstrated bilateral activation in the STS/STG, stronger in the right hemisphere (see Table 4, Fig. 4a–b); activations in the right hemisphere included four unique activations spanning anterior and posterior STS/STG and temporal pole, whereas the single significant activation in the left STS was in the anterior region. At a lower threshold ($p < .01$), one of the right posterior STS activations extended within one voxel of the peak found during the AV integration conjunction analysis in Study 1a, suggesting that the finding of greater activation in right pSTS during the bimodal condition was replicated in the Study 2 sample.

A conjunction analysis of Fearful compared to Neutral stimuli within each modality was conducted ([AVF > AVN]∩[VF > VN]∩[AF > AN]) in order to examine the effects of emotion while controlling for AV integration effects. Two significant activations to Fearful stimuli relative to Neutral stimuli were found in the left aSTG (see Table 4, Fig. 4c). These findings are consistent with the STG findings in Study 1b.

Direct comparison between fearful-only AV and unimodal stimuli allowed for investigation of the effect of AV integration without the possible confound of differences between perception of

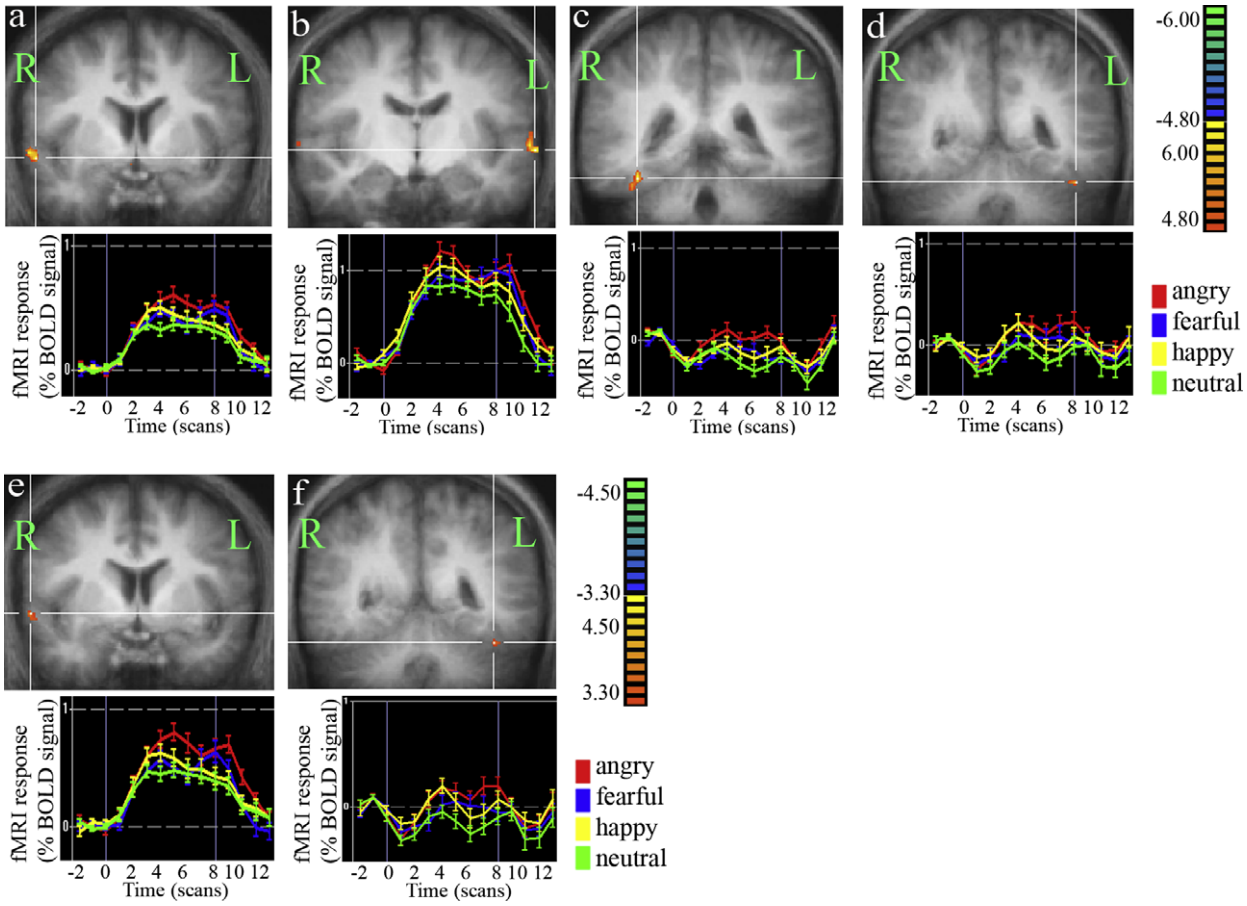

Fig. 3. Increased activation in bilateral STG (a: 54, 2, −5; b: −63, −10, 1) and bilateral FG (c: 36, −37, −14, d: −42, −46, −17) for the balanced average of Angry, Fearful, and Happy > Neutral; $t(9) = 4.8$, $p<.001$. Increased activation in right STG (e: 58, 2, −2) and left FG (f: −39, −46, −17) for (Angry > Neutral)∩(Fearful > Neutral)∩(Happy > Neutral); $t(9) = 3.3$, $p < .01$. Cross-hairs on brain images represent the voxel of peak activation. Time course shown below each coronal image represents activation in the peak voxel. On the *x*-axis, 0 indicates block onset, and 8 indicates block offset.

Table 4
Study 2: Emotion and AV integration effects

Brain region	Cluster size (mm^3)	Talairach			t value
		x	y	z	
AV Integration: (AVF > AF)∩(AVF > VF)∩(AVN > AN)∩(AVN > VN)[d]					
Right pSTS	184	59	−43	10	3.76 [b]
Right pSTS[e]	84	45	−43	13	3.94 [b]
Right temporal pole	60	51	17	−9	4.48 [b]
Left aSTS	30	−63	−13	−5	3.80 [b]
Right pSTS	22	45	−34	10	3.58 [b]
Emotion: (AVF > AVN)∩(AF > AN)∩(VF > VN)[d]					
Left anterior STG	115	−57	−13	1	3.43 [a]
Left anterior STG	42	−57	−4	−2	3.10 [a]
Fearful-only AV Integration (AVF > AF)∩(AVF > VF)[f]					
Right pSTS	1661	42	−46	10	5.70 [c]
Left aSTS	1288	−63	−25	-5	6.42 [c]
Right superior frontal gyrus	716	12	44	40	5.53 [c]
Posterior thalamus	680	12	−25	-2	5.08 [c]
Right pSTS	203	42	−31	7	4.79 [c]
Right temporal pole	80	51	17	−8	4.48 [c]
AV-only Emotion: AVF > AVN[d]					
Right anterior STG	2767	60	−7	4	6.06 [c]
Left anterior lateral fissure	1707	−63	−13	4	5.29 [c]
Right FG	542	42	−37	−11	5.06 [c]
Right posterior STG	175	60	−37	19	4.47 [c]

[a] $p < .01$.
[b] $p < .001$.
[c] $p < .0001$.
[d] A priori ROI analysis on group averaged maps.
[e] Activation spreads within one voxel of the peak identified in Study 1 (AV > A)∩(AV > V) at $p < .01$.
[f] Whole-brain analysis on group averaged maps.

emotionally charged and neutral stimuli. A conjunction analysis, (Fearful AV > Fearful V)∩(Fearful AV > Fearful A), indicated that the AV stimuli led to greater activation in right-hemisphere pSTS, right temporal pole, and left aSTS relative to fearful unimodal stimuli (see Table 4). Additional activations were seen in other regions, including right superior frontal gyrus, and a region in the posterior thalamus.

In order to evaluate the effects of emotion while holding AV integration constant, fearful AV and neutral AV were contrasted. This comparison also investigated whether the fusiform gyrus findings in Study 1b were replicated, since FG would not be expected to demonstrate activation in the emotion conjunction, since audio-only conditions were included. Areas of activation in right anterior and posterior STG, right FG, and left anterior lateral fissure demonstrated greater activation to fearful than neutral stimuli (see Table 4 and Fig. 4d).

4. Discussion

Perception of bimodal emotional stimuli increased activation in the bilateral superior temporal sulcus/superior temporal gyrus (STS/STG) relative to unimodal emotional conditions. The STS/STG has been implicated in numerous tasks, including integration of audio-visual (AV) sensory and language tasks that are devoid of emotional content (e.g., motion perception, McGurk effect, respectively) and components of social engagement (e.g., eye gaze, biological motion). AV emotional stimuli may involve the STS/STG for multiple aspects of social perception, including the integration of auditory and visual perceptual cues and emotion processing, a critical component of social engagement. However, results from the present studies suggest distinct mechanisms for AV integration and perception of emotion; although regions of the STS/STG are recruited for both types of tasks, two distinct aspects of audio-visual emotion processing appear to rely on STS/STG.

Fig. 5 shows the ROI peak findings along the STS/STG for both studies, illustrating the dissociable effects of AV integration and emotionality, as well as convergence of findings between Study 1 and Study 2. Based on the summary of activation in the STS/STG region, it is clear that the effects of emotion are distinct from the effects of AV integration. The effects of emotion for Study 1 and 2 (shaded symbols) are consistently demonstrated in aSTG bilaterally. A role for the aSTG in emotion perception makes sense, since rostral regions of the STG have projections to multiple nuclei in the amygdala (Stefanacci & Amaral, 2002). Although projections are found along most of the STG/STS (except for the most caudal point), the greatest density are found more anteriorly.

The effects of AV integration (unshaded symbols) are more widely distributed, but appear distinct from the emotion peaks nonetheless. AV integration leads to greater activation in bilateral pSTS relative to unimodal conditions in Study 1a. In the replication of this integration effect in Study 2, in which fearful and neutral alone are compared to their unimodal counterparts, activation occurs in right temporal pole, right pSTS, and left aSTS. When only fearful AV is compared to each fearful unimodal condition, the peaks are nearly identical. These findings are consistent with other investigations of audio-visual emotion perception (Kreifelts et al., 2007; Pourtois et al., 2005). Although the temporal pole finding was not predicted a priori from our literature review, it makes sense in that the temporal pole includes the most anterior portion of STG. Furthermore, the temporal pole has been hypothesized to play a role integration of sensory information from multiple modalities (Olson, Ploaker, & Ezzyat, 2007).

Other a priori regions-of-interest (ROIs) for integration of bimodal cues based on the literature were not found to be significantly more active to AV stimuli compared to unimodal stimuli in this study; these areas included the fusiform gyrus, cingulate gyrus, insula, amygdala, and superior colliculi. The amygdala may show rapid habituation (Buchel, Morris, Dolan, & Friston, 1998). In fact, when statistical analysis was limited to only the first run for each participant, bilateral amygdala activation was found for the AV condition relative to both unimodal conditions ($p < .01$, uncor-

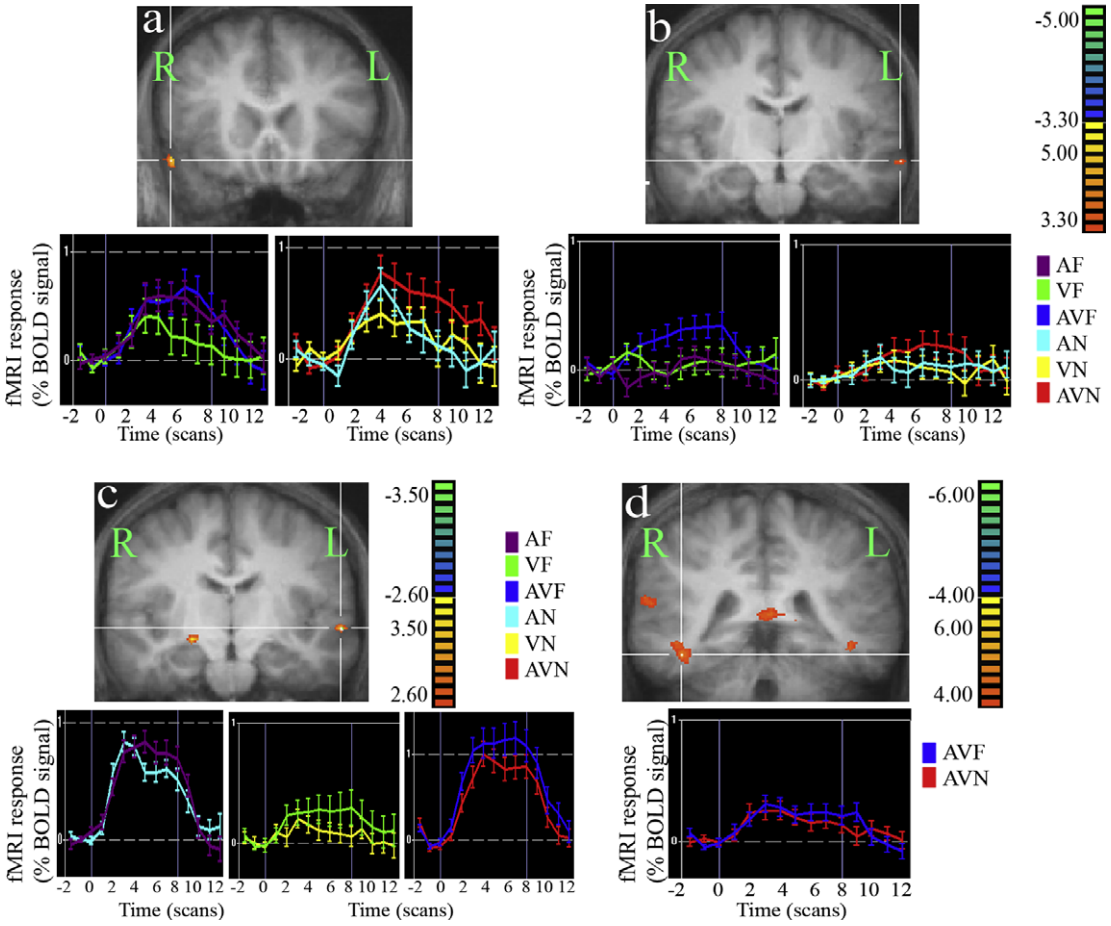

Fig. 4. Study 2 results: Increased activation in bilateral STS (a: 51, 17, −9; b: −63, −13, −5) for AV stimuli, (AVF > AF)∩(AVF > VF)∩(AVN > AN)∩(AVN > VN), $t(4859) = 3.3$, $p < .001$. Increased activation to fearful stimuli versus neutral stimuli in left STG (c: −57, −13, 1), (AVF > AVN)∩(AF > AN)∩(VF > VN), $t(4859) = 2.6$, $p < .01$. When examining AV conditions alone, AVF > AVN in right FG (d: 42, −37, −11), $t(4859) = 3.9$, $p < .0001$. Cross-hairs on brain images represent the voxel of peak activation. Time courses shown below each coronal image present activation in the peak voxel; multiple time course graphs indicate the comparisons of interest. On the *x*-axis, 0 indicates block onset, and 8 indicates block offset.

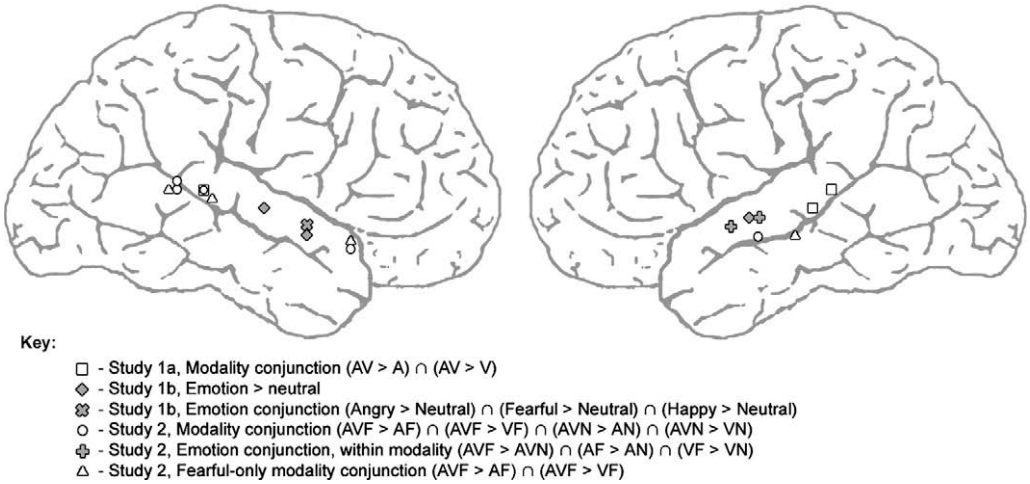

Fig. 5. Schematic brain showing all activation peaks in the STS/STG region. Peak activations for emotion and AV integration are distinct across Study 1 and Study 2; the effects of emotion, shown in gray symbols, are evident in bilateral aSTG, whereas the effects of AV integration, shown in white symbols, are distributed in bilateral pSTS and right aSTS/STG.

rected, meeting threshold for an a priori ROI); activation was stronger in the right hemisphere. Using only the first run greatly reduces power, and it may be that a larger sample is necessary to further investigate the role of the amygdala in bimodal integration. Other subcortical ROIs may also require increased power to detect effects.

The findings of greater activity in the pSTS to AV stimuli relative to the unimodal conditions suggest that there is enhancement, or

additivity, when visual and auditory modalities are combined into a single percept. However, the current study did not find evidence of super-additivity in the pSTS, which is defined as the bimodal condition (AV) producing greater activation than the sum of its parts (unimodal A and V). Super-additivity was first measured and defined on a cellular level, using direct single-cell recording. As was discussed recently by Ethofer and colleagues (Ethofer, Pourtois, & Wildgruber, 2006), it is challenging to use human functional neuroimaging techniques to evaluate integration in the sense of super- and sub-additivity (Laurienti, Perrault, Stanford, Wallace, & Stein, 2005). Because the BOLD response for each voxel averages across thousands of neurons, it is not possible to discriminate a situation in which a proportion of the cells are active to visual stimuli and a proportion of unique neurons are active to auditory stimuli, from a voxel containing neurons that are truly super-additive. In both cases, the voxel would show increased activation to bimodal stimuli. It is unlikely that a majority of cells in any given voxel will demonstrate super-additivity, given that only a minority of cells demonstrate super-additivity in single-cell recording studies. Furthermore, as others have pointed out (van Atteveldt et al., 2004), there may be a ceiling effect of the BOLD response which limits the ability to assess super-additivity at the voxel level.

Other findings identified in the emotion analyses (Study 1b) included bilateral fusiform gyrus (FG) when comparing the average of all emotion conditions to the neutral condition. This finding is consistent with other studies that demonstrate that more intense emotional faces lead to greater activity in FG relative to weaker emotional faces and neutral faces (Glaescher, Tuescher, Weiller, & Buechel, 2004), likely because of modulatory effects of attention, not because of a direct role for the fusiform in computations about facial expressions (Schultz, 2005; Tranel, Damasio, & Damasio, 1998). Examination of individual emotions relative to neutral suggested that each emotion may lead to a unique pattern of activation; the STG findings in Study 1b appear to be driven by the angry condition. Fearful stimuli demonstrated increased activation in the left superior frontal sulcus, and happy did not demonstrate any significant differences on a whole-brain level analysis. In Study 2, the comparison of the fearful condition with neutral did not reproduce findings in the left superior frontal sulcus; rather, activation was identified in the left aSTG, which is more consistent with the averaged findings from all emotions relative to neutral in Study 1b, and in the right superior frontal gyrus. Furthermore, AV fearful stimuli relative to each unimodal fearful condition led to greater activation of the posterior thalamus (12, −25, 2), similar to the finding from a paradigm examining auditory influence on visual motion perception (Bushara et al., 2003); this replication suggests that nuclei in the posterior thalamus may be involved in any audio-visual integration, not specific to emotional cues. Given the small sample size of Study 2 and the limited emotions included in the design, this effect should be replicated before conclusive interpretation. Finally, in the comparison of fearful AV to each fearful unimodal stimuli in Study 2, activation was found in the right superior frontal gyrus and posterior thalamus, in addition to bilateral aSTS/STG and right pSTS. Activations in other a priori ROIs (e.g., insula, cingulate cortex) were not identified in these studies.

These are the first neuroimaging studies to utilize dynamic, integrated audio-visual emotional stimuli in which the semantic content of the auditory track can be considered to be consistent with each affective condition in the study, as opposed to stimuli containing affectively neutral semantic information artificially paired with affective prosody. Although the difference between emotionally ambiguous and emotionally neutral content may be considered to be minor, emotionally ambiguous stimuli allow the participant to focus more completely on the affective content, rather than expend any effort processing whether the semantic content makes sense in the affective tone; this is critical for avoiding confounding activations during an fMRI study. Study 2's small sample precludes use of random effects analyses, which limits the generalization of results. However, results from Study 2 support the finding that the effects of emotion and AV integration in the STS/STG region are clearly dissociable.

The current studies represent an important step toward improving our understanding of the neural computations that process emotion in dynamic and multimodal social interactions. Future directions include the investigation of congruent relative to incongruent dynamic emotional stimuli, building on the work of Dolan and colleagues (2001) and Ethofer and colleagues (2006). This paradigm will measure unique activations when AV cues are bound in a single congruent stimulus relative to activations due to the simultaneous presentation of incongruent AV cues, during which integration is not expected to take place. For example, if an angry face is paired with a happy voice, it is predicted that these cues will not be bound into a seamless multimodal percept, and neural activations will differentiate between congruent and incongruent conditions. It will also be of interest to compare incongruent stimuli that appear natural (i.e., the incongruence goes undetected), which may be a model for complex emotional presentations in which the speaker's facial expression and tone of voice do not match, as in a person displaying sarcasm, relative to the more unnatural incongruence which is artificially generated.

Furthermore, future behavioral and neuroimaging studies will examine the neural mechanisms underlying the facial bias when presented with incongruent emotional cues in facial expression and affective prosody. The bias toward perception of the facial expression was found in a behavioral study of emotion perception using dynamic audio-visual stimuli (Santorelli & Robins, 2006); future studies will explore the role of automatic facial mimicry in the facial expression bias, as well as the brain regions that play a role in the processing of congruent and incongruent emotional cues. In addition, differences in emotion perception between males and females should be examined in future studies, as in studies with other emotional stimuli (e.g., Hofer, Siedentopf, & Ischebeck, 2007).

Finally, these paradigms can be applied to research investigating clinical samples known to have deficits in emotion perception, including autism, mood and anxiety disorders, and schizophrenia. Deficits in emotion perception have been documented in many clinical populations (for reviews of emotion perception in schizophrenia and autism, please see Abdi & Sharma, 2004; Edwards, Jackson, & Pattison, 2002; Kohler & Brennan, 2004; Travis & Sigman, 1998). However, it has been difficult to develop tools to accurately assess emotion perception in real-world social situations, given that real-world social situations are inherently multimodal. Individuals must integrate dynamic and rapidly-changing cues such as tone of voice, facial expression, body postures. Clinical tools, such as the Diagnostic Assessment of Nonverbal Accuracy, Second Edition (Nowicki, 2004), present static photographs and isolated spoken text with prosodic cues, which fail to replicate the complexity of social interaction, and thereby limit generalizability to emotion perception in social contexts. Use of dynamic multimodal stimuli, which mimic real-world emotional cues more than previous emotion designs, will facilitate understanding of the neural mechanisms underlying the deficits in emotion perception in these groups. The stimuli developed for the present studies may also be used to develop more ecologically valid tools for measuring emotion perception in individuals with impairment in emotion perception. Increased ecological validity, both in research paradigms and clinical assessment tools, will improve our understanding of the disruptions to the emotion perception networks, which in turn may lead to improved intervention.

Acknowledgments

This work was supported in part by the following funding sources: National Alliance for Autism Research, Marie Bristol-Power Postdoctoral Fellowship; Yale University School of Medicine, James Hudson Brown–Alexander B. Coxe Postdoctoral Fellowship in the Medical Sciences, NIMH T32 MH18268, the STC Program of the National Science Foundation under Agreement No. IBN-9876754. We thank Harder and Co. for assistance with creation of the stimuli used in these studies, Rhea Paul for her input on the study, and all of our colleagues at the Yale University School of Medicine Magnetic Resonance Research Center, including the MR technologists. We thank Ted Long for assistance with data collection and Sylvia Glasscoe for computer programming. We also thank Erin McClure and Tara McKee for their thoughtful comments on manuscript drafts.

References

Abdi, Z., & Sharma, T. (2004). Social cognition and its neural correlates in schizophrenia and autism. *Cns Spectrums, 9*(5), 335–343.

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience, 13*(2), 232–240.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*(5), 809–823.

Boucher, J., Lewis, V., & Collis, G. (2000). Voice processing abilities in children with autism, children with specific language impairments, and young typically developing children. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 41*(7), 847–857.

Buchanan, T. W., Lutz, K., Mizazade, S., Specht, K., Shah, N. J., Zilles, K., et al. (2000). Recognition of emotional prosody of verbal components of spoken language: An fMRI study. *Cognitive Brain Research, 9*, 227–238.

Buchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998). Brain systems mediating aversive conditioning: An event-related fMRI study. *Neuron, 20*(5), 947–957.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience, 6*(2), 190–195.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15*(1), 57–70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649–657.

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in Cognitive Sciences, 2*(7), 247–253.

Calvert, G. A., & Thesen, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris, 98*(1–3), 191–205.

Celani, G., Battacchi, M. W., & Arcidiacono, L. (1999). The understanding of the emotional meaning of facial expressions in people with autism. *Journal of Autism and Developmental Disorders, 29*(1), 57–66.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments.. *Behavioral Research Methods, Instruments, and Computers, 25*(2), 257–271.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*(10), 460–467.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion, 14*(3), 289–311.

Dolan, R. J., Morris, J. S., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences of the United States of America, 98*(17), 10006–10010.

Edwards, J., Jackson, H., & Pattison, P. (2002). Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review. *Clinical Psychology Review, 22*(6), 789–832.

Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., et al. (2006). Impact of voice on emotional judgment of faces: An event-related fMRI study. *Human Brain Mapping, 27*(9), 707–714.

Ethofer, T., Pourtois, G., & Wildgruber, D. (2006). Investigating audiovisual integration of emotional signals in the human brain. *Progress in Brain Research, 156*, 345–361.

Frost, R., Repp, B. H., & Katz, L. (1989). Can speech perception be influenced by simultaneous presentation of print? *Journal of Memory and Language, 27*(6), 741–755.

Fujiki, M., Spackman, M., Brinton, B., & Illig, T. (2008). Ability of children with language impairment to understand emotion conveyed by prosody in a narrative passage. *International Journal of Language and Communication Disorders, 43*(3), 330–345.

Gepner, B., Deruelle, C., & Grynfeltt, S. (2001). Motion and emotion: A novel approach to the study of face processing by young autistic children. *Journal of Autism and Developmental Disorders, 31*(1), 37–45.

Glaescher, J., Tuescher, O., Weiller, C., & Buechel, C. (2004). Elevated responses to constant facial emotions in different faces in the human amygdala: An fMRI study of facial identity and expression. *BMC Neuroscience, 5*(November 17).

Goebel, R. (2000). *Brain Voyager 2000*. Maastricht, The Netherlands: Brain Innovation.

Goebel, R. (2004). *Brain Voyager QX*. Maastricht, The Netherlands: Brain Innovation.

Gur, R. C., Schroeder, L., Turner, T., McGrath, C., Chan, R. M., Turetsky, B. I., et al. (2002). Brain activation during facial emotion processing. *Neuroimage, 16*(3), 651–662.

Hariri, A. R., Bookheimer, S. Y., & Mazziotta, J. C. (2000). Modulating emotional responses: Effects of a neocortical network on the limbic system. *Neuroreport, 11*(1), 43–48.

Harwood, N. K., Hall, L. J., & Shinkfield, A. J. (1999). Recognition of facial emotional expressions from moving and static displays by individuals with mental retardation. *American Journal on Mental Retardation, 104*(3), 270–278.

Haviland, J. M., Walker-Andrews, A. S., Huffman, L. R., Toci, L., & Alton, K. (1996). Intermodal perception of emotional expressions by children with autism. *Journal of Developmental and Physical Disabilities, 8*(1), 77–88.

Hobson, R. P., Ouston, J., & Lee, A. (1988). Emotion recognition in autism—Coordinating faces and voices. *Psychological Medicine, 18*(4), 911–923.

Hofer, A., Siedentopf, C., & Ischebeck, A. e. a. (2007). Sex differences in brain activation patterns during processing of positively and negatively valenced emotional words. *Psychological Medicine, 37*(1), 109–119.

Iidaka, T., Omori, M., Murata, T., Kosaka, H., Yonekura, Y., Okada, T., et al. (2001). Neural interaction of the amygdala with the prefrontal and temporal cortices in the processing of facial expressions as revealed by fMRI. *Journal of Cognitive Neuroscience, 13*(8), 1035–1047.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMR1 study of the McGurk effect. *Neuroreport, 14*(8), 1129–1133.

Keightley, M. L., Winocur, G., Graham, S. J., Mayberg, H. S., Hevenor, S. J., & Grady, C. L. (2003). An fMRI study investigating cognitive modulation of brain regions associated with emotional processing of visual stimuli. *Neuropsychologia, 41*(5), 585–596.

Kesler, M. L., Andersen, A. H., Smith, C. D., Avison, M. J., Davis, C. E., Kryscio, R. J., et al. (2001). Neural substrates of facial emotion processing using fMRI. *Cognitive Brain Research, 11*(2), 213–226.

Kohler, C., & Brennan, A. (2004). Recognition of facial emotions in schizophrenia. *Current Opinion in Psychiatry, 17*(2), 81–86.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *Neuroimage, 37*(4), 1445–1456.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., & Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research, 166*(3–4), 289–297.

Loveland, K. A., Tunalikotoski, B., Chen, R., Brelsford, K. A., Ortegon, J., & Pearson, D. A. (1995). Intermodal perception of affect in persons with autism or down-syndrome. *Development and Psychopathology, 7*(3), 409–418.

McGurk, H., & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mitchell, R. L. C., Elliott, R., Barry, M., Cruttenden, A., & Woodruff, P. W. R. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia, 41*(10), 1410–1421.

Nowicki, S. J. (2004). *A test and manual for the Diagnostic Analysis of Nonverbal Accuracy 2*. Emory University.

Olson, I. R., Gatenby, J. C., & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research, 14*, 129–138.

Olson, I. R., Ploaker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: A review of findings on social and emotional processing. *Brain, 130*, 1718–1731.

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1990). Are there emotion perception deficits in young autistic children? *Journal of Child Psychology and Psychiatry, 31*(3), 343–361.

Peppe, S., McCann, J., Gibbon, F., O'Hare, A., & Rutherford, M. (2007). Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech Language and Hearing Research, 50*(4), 1015–1028.

Pourtois, G., de Gelder, B., Bol, A., & Crommelinck, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex, 41*(1), 49–59.

Robins, D.L., & Schultz, R.T. (2004). Cross-modal integration of emotional information. Paper presented at the International Neuropsychological Society, February, 2004, Baltimore, MD.

Saito, D. N., Yoshimura, K., Kochiyama, T., Okada, T., Honda, M., & Sadato, N. (2005). Cross-modal binding and activated attentional networks during audio-visual speech integration: A functional MRI study. *Cerebral Cortex, 15*(11), 1750–1760.

Santorelli, N. T., & Robins, D. L. (2006). *Perception of emotional cues from facial expression and affective prosody*. Paper presented at the International Neuropsychological Society, Boston, MA.

Sato, W., Kochiyama, T., Yoshikawa, S., Naito, E., & Matsumura, M. (2004). Enhanced neural activity in response to dynamic facial expressions of emotion: An fMRI study. *Cognitive Brain Research, 20*(1), 81–91.

Schultz, R. T. (2005). Developmental deficits in social perception in autism: The role of the amygdala and fusiform face area. *International Journal of Developmental Neuroscience, 23*, 125–141.

Stefanacci, L., & Amaral, D. G. (2002). Some observations on cortical inputs to the macaque monkey amygdala: An anterograde tracing study. *Journal of Comparative Neurology, 451*(4), 301–323.

Sumby, W. H., & Pollack, I. (1955). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Tranel, D., Damasio, A. R., & Damasio, H. (1998). Intact recognition of facial expression, gender, and age in patients with impaired recognition of facial identity. *Neurology, 38*, 690–696.

Travis, L., & Sigman, M. (1998). Social deficits and interpersonal relationships in autism. *Mental Retardation and Developmental Disabilities Research Reviews, 4*(2), 65–72.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43*(2), 271–282.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., et al. (2004). Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. *Cerebral Cortex, 14*(12), 1384–1389.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., et al. (2005). Identification of emotional intonation evaluated by fMRI. *Neuroimage, 24*(4), 1233–1241.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13*(10), 1034–1043.

Yirmiya, N., Sigman, M., Kasari, C., & Mundy, P. C. (1992). Empathy and cognition in high-functioning children with autism. *Child Development, 63*(1), 150–160.