Behavioral/Cognitive

Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face–Voice Emotional Integration

Rebecca Watson,[1,2] **Marianne Latinus,**[2,3] **Takao Noguchi,**[4] **Oliver Garrod,**[2] **Frances Crabbe,**[2] **and Pascal Belin**[2,3,5]

[1]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht 6229 EV, The Netherlands, [2]Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, G12 8QB, United Kingdom, [3]Neuroscience Institute of Timone, Coeducational Research Unit 7289, National Center of Scientific Research–Aix-Marseille University, F-13284 Marseille, France, [4]Department of Psychology, University of Warwick, Coventry, CV4 7A, United Kingdom, and [5]International Laboratories for Brain, Music, and Sound, University of Montreal and McGill University, Montreal, Quebec, Canada, H2V 4P3

The integration of emotional information from the face and voice of other persons is known to be mediated by a number of "multisensory" cerebral regions, such as the right posterior superior temporal sulcus (pSTS). However, whether multimodal integration in these regions is attributable to interleaved populations of unisensory neurons responding to face or voice or rather by multimodal neurons receiving input from the two modalities is not fully clear. Here, we examine this question using functional magnetic resonance adaptation and dynamic audiovisual stimuli in which emotional information was manipulated parametrically and independently in the face and voice via morphing between angry and happy expressions. Healthy human adult subjects were scanned while performing a happy/angry emotion categorization task on a series of such stimuli included in a fast event-related, continuous carryover design. Subjects integrated both face and voice information when categorizing emotion—although there was a greater weighting of face information—and showed behavioral adaptation effects both within and across modality. Adaptation also occurred at the neural level: in addition to modality-specific adaptation in visual and auditory cortices, we observed for the first time a crossmodal adaptation effect. Specifically, fMRI signal in the right pSTS was reduced in response to a stimulus in which facial emotion was similar to the vocal emotion of the preceding stimulus. These results suggest that the integration of emotional information from face and voice in the pSTS involves a detectable proportion of bimodal neurons that combine inputs from visual and auditory cortices.

Key words: **emotion perception; functional magnetic resonance adaptation; multisensory integration**

Introduction

Stimulation in natural settings usually recruits a number of different sensory channels in parallel. Particularly important with regards to social interaction is the recognition of emotional cues from both the face and voice of others. A number of regions of the human brain have been implicated in the integration of these affective cues, including "convergence zones," such as the posterior superior temporal sulcus (pSTS; Pourtois et al., 2005; Kreifelts et al., 2007; Hagan et al., 2009, 2013; Robins et al., 2009), putative "unisensory" regions (i.e., the primary visual and auditory cortices; de Gelder et al., 1999; Pourtois et al., 2000, 2002), and limbic structures, such as the amygdala (Dolan et al., 2001; Klasen et al., 2011). These regions tend to show greater activity as measured with neuroimaging techniques such as fMRI in response to bimodal face–voice emotional stimulation than to either modality alone, leading to them being classified as "multisensory."

However, it remains unclear whether the enhanced responses of these presumed affective multisensory regions reflect populations of truly bimodal neurons (that receive affective input from both the visual and auditory modalities) or is simply evoked by groups of interdigitized, unimodal visual and auditory neurons. Because of its limited spatial resolution, in which activity from hundreds and thousands of (potentially, heterogeneous) neurons within a voxel is averaged out, the traditional fMRI method cannot distinguish between these two possibilities. Various fMRI studies defined multisensory regions using specific statistical criteria, but a number of these allow in theory for a purely additive audio and visual response, which could be evoked simply by overlapping face- and voice-sensitive neurons.

Researchers attempted to circumvent the problem of limited spatial resolution in fMRI by using functional magnetic resonance adaptation (fMR-A; Grill-Spector et al., 1999, 2006). The logic is as follows: if repetition of a given feature in the stimulation results in a reduction of fMRI signal in a given voxel, then that voxel is assumed to contain neurons that are specifically involved in processing, or representing, the repeated feature. It was suggested that fMR-A might also be helpful in distinguishing between voxels in the multisensory cortex consisting of only uni-

Received Oct. 18, 2013; revised Feb. 21, 2014; accepted March 12, 2014.

Author contributions: R.W., M.L., and P.B. designed research; R.W., T.N., O.G., and F.C. performed research; R.W., M.L., and P.B., analyzed data; R.W. and P.B. wrote the paper.

This work was funded by Biotechnology and Biological Sciences Research Council Grant BBJ003654/1.

The authors declare no competing financial interests.

Correspondence should be addressed to Rebecca Watson, Department of Cognitive Neuroscience, Maastricht University, Oxfordlaan 55, Maastricht 6229 EV, The Netherlands. E-mail: rebecca.watson@maastrichtuniversity.nl.

DOI:10.1523/JNEUROSCI.4478-13.2014

 0270-6474/14/346813-09$15.00/0

sensory neuronal subpopulations and voxels composed of a mixture of unisensory and multisensory populations (Goebel and van Atteveldt, 2009; Tal and Amedi, 2009). Varying adaptation and recovery responses could shed light on the subvoxel organization of proposed affective multisensory regions: multisensory neurons should adapt to crossmodal repetitions of emotion information (e.g., an angry voice followed by an angry face), whereas unisensory neurons should not.

Here we used fMR-A to investigate crossmodal adaptation to affective information in faces and voices. We used dynamic audiovisual stimuli in which affective information was independently and parametrically manipulated in each modality and used these stimuli in an efficiency-optimized, "continuous carryover" design (Aguirre, 2007) as a means to test whether crossmodal adaptation effect could be observed, i.e., a significant influence of emotional information in one modality on neural response to emotional information in the other modality. Specifically, we hypothesized that, if a cerebral region contained a sufficiently large proportion of multisensory auditory–visual neurons involved in processing emotional information, as opposed to interspersed populations of unisensory neurons, that region should show detectable evidence of crossmodal adaptation.

Materials and Methods

Participants

Eighteen participants (10 males, eight females; mean ± SD age, 25 ± 3.7 years) were scanned in the fMRI experiment. All had self-reported normal or corrected-to-normal vision and hearing. The study was approved by the ethical committee of the University of Glasgow and conformed to the Declaration of Helsinki. All volunteers provided informed written consent before and received payment at the rate of £6/h for participation.

Stimuli

Stimuli consisted of 25 novel face–voice stimuli used previously by our laboratory. Stimulus construction and pretesting has been described previously in full by Watson et al. (2013) and therefore will be described in brief here.

Raw audiovisual content, recorded using a Did3 capture system (Dimensional Imaging Ltd.), consisted of two actors (one male, one female) expressing anger and happiness in both the face and voice. The sound "ah" was chosen because it contains no linguistic information. Actors were initially instructed to express each emotion with low, medium, and high intensity, with standardized timing when possible. Two final clips per actor (one anger, one happiness) were selected (on the basis of high-intensity production, similar duration) for use.

Audiovisual clips were then split into their audio and visual components, for within-modality morphing. A landmarked face mesh was applied to each frame of the sequence, which was then used as a basis for morphing in MATLAB (Mathworks). Face morphing consisted of morphing between angry dynamic facial expressions (one male and one female) and happy dynamic facial expressions (one male and one female) within gender. The resulting output was two different face continua (one per gender), each consisting of five within-modality stimuli, morphed between 90% anger and 90% happiness in 20% steps.

Audio output from the audiovisual recordings was processed in Adobe Audition 2.0 (Adobe Systems) and then morphed using the MATLAB-based morphing algorithm STRAIGHT (Kawahara, 2006). Vocal morphing ran in parallel with the facial morphing, in that the equivalent individual voices were also morphed between 90% anger and 90% happiness in 20% steps, resulting in two vocal continua, one per actor, each consisting of five vocal stimuli.

Within actor, the five dynamic face and five voice stimuli were all equal length. To ensure that all stimuli were of equal length, we edited video and audio clips between actors. Editing was conducted in Adobe Premiere 2.0 (Adobe Systems) and consisted of inserting or deleting video frames to match predefined time points (e.g., mouth opening, mouth closing) across clip. We made efforts to ensure that editing occurred between frames with as little difference in movement as possible to retain the naturalness of the video clip. The editing produced 10 adjusted video clips, each 780 ms long. The audio samples were then also adjusted in accordance with the temporal landmarks identified in the video clips to create 10 vocalizations (five for each actor) of equal length. Within actor, the five visual and five auditory clips were then paired together in all possible combinations. This resulted in a total of 25 audiovisual stimuli for each actor, parametrically varying in congruence between face and voice affective information. Stimuli are illustrated in Figure 1.

Design and procedure

Continuous carryover experiment

In the main experiment, stimuli were presented using the Psychtoolbox in MATLAB via electrostatic headphones (NordicNeuroLab) at a sound pressure level of 80 dB as measured using a Lutron Sl-4010 sound level meter. Before they were scanned, subjects were presented with sound samples to verify that the sound pressure level was comfortable and loud enough considering the scanner noise. Audiovisual movies were presented in two scanning runs (over 2 different days) while the blood oxygenation level-dependent (BOLD) signal was measured in the fMRI scanner. We used a continuous carryover experimental design (Aguirre, 2007). This design allows for measurement of the direct effects (i.e., that of face and voice emotion morph) and the repetition suppression effect, which can be observed in pairs of voices or faces (like in the typical fMRI adaptation experiments).

The stimulus order followed two interleaved $N = 25$ type 1–index 1 sequences (one for each of the speaker continua; interstimulus interval, 2 s), which shuffles stimuli within the continuum so that each stimulus is preceded by itself and every other within-continuum stimulus in a balanced manner. The sequence was interrupted seven times with 20 s silent periods, which acted as a baseline, and at the end of a silent period, the last five stimuli of the sequence preceding the silence were repeated before the sequence continued. These stimuli were removed in our later analysis. Participants were instructed to perform a two-alternative forced-choice emotion classification task (responses, angry or happy) using two buttons of an MR-compatible response pad (NordicNeuroLab). They were also instructed to pay attention to both the face and voice but were told they could use the information presented in whatever way they wanted to make their decision on emotion. Reaction times (relative to stimulus onset) were collected using MATLAB with a response window limited to 2 s.

Localization of multisensory regions

In addition to the main experiment, we also used an independent multisensory localizer to identify regions involved in multisensory processing. We further performed an isolated region of interest (ROI) analysis within these areas to assess whether there were significant crossmodal adaptation effects. Therefore, we were consequently able to infer the neuronal properties (i.e., multisensory vs interdigitized unisensory neurons) of these independently established multisensory regions. During an 11 min scanning run, participants were presented with a variety of dynamic audiovisual and unimodal stimuli in 18 different 16 s blocks (for additional details of stimuli, refer to Watson et al., 2014). Thus, each block contained eight different stimuli. These blocks were broadly categorized as follows: (1) faces paired with their corresponding vocal sounds (AV-P); (2) objects (visual and audio) (AV-O); (3) voices alone (A-P); (4) objects (audio only) (A-O); (5) faces alone (V-P); and (6) objects (visual only) (V-O).

Thus, categories 1 and 2 were audiovisual, 3 and 4 were audio only, and 5 and 6 were visual only. There were three different stimulus blocks within each type, each containing different visual/auditory/audiovisual stimuli. A 16 s null event block comprising silence and a gray screen was also created. Each of the 18 blocks was repeated twice, and the blocks were presented pseudorandomly: each block was always preceded and followed by a block from a different category (e.g., a block from the "faces-alone" category could never be preceded/followed by any other block from the faces-alone category). The null event block was repeated six times and interspersed randomly within the presentations of the stimulus blocks.

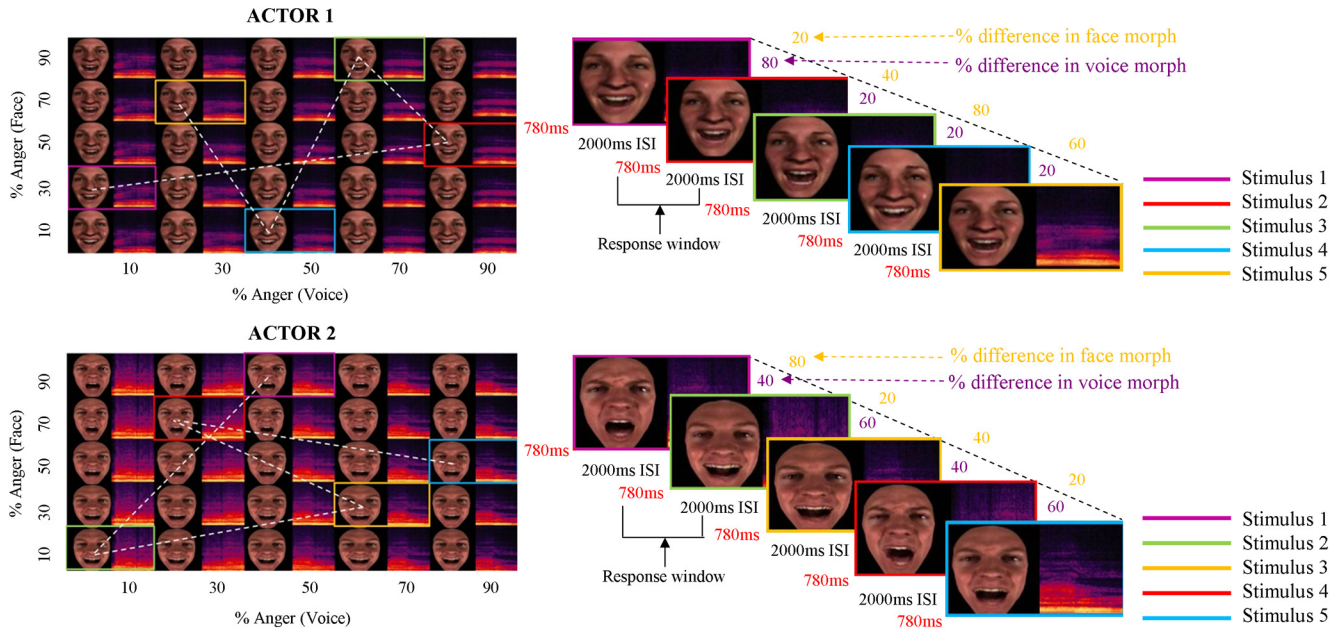

Figure 1. Experimental stimuli: two sets (1 per actor) of dynamic and time-synchronized audiovisual stimuli were presented within a continuous carryover design (Aguirre, 2007) in interleaved type 1–index 1 sequences over two experimental sessions (sequential presentation indicated by dotted lines in the left panels). Every face morph (extending from 90% angry to 90% happy) was paired with every voice morph (extending from 90% angry to 90% happy) within actor, so to create two sets of 25 face–voice stimuli, parametrically varying in congruence (examples in colored rectangles). In a carryover design, every stimulus precedes and follows every other stimulus, such that each stimulus serves as an adaptor for the following stimulus. The right panels show parts of an example type 1–index 1 sequence. In each experimental run, sequences were blocked by actor. Participants performed a two-alternative forced-choice task (angry or happy) on emotion. The right panels indicate examples of within-block sequences of stimuli. ISI, Interstimulus interval.

Imaging parameters

Functional images covering the whole brain (32 slices; field of view, 210 × 210 mm; voxel size, 3 × 3 × 3 mm) were acquired on a 3 T Tim Trio Scanner (Siemens) with a 12 channel head coil, using an echo planar imaging (EPI) sequence (interleaved; TR, 2 s; TE, 30 ms; flip angle, 80°) in both the carryover and localizer experiments. In total, we acquired 1560 EPI volumes for the carryover experiment, split into two scanning sessions consisting of 780 EPI volumes, and 336 EPI volumes were acquired for the multimodal localizer. For both the carryover experiment and experimental localizer, the first 4 s of the functional run consisted of "dummy" gradient and radio frequency pulses to allow for steady-state magnetization during which no stimuli were presented and no fMRI data were collected. MRI was performed at the Centre for Cognitive Neuroimaging (Glasgow, UK).

At the end of each fMRI session, high-resolution T1-weighted structural images were collected in 192 axial slices and isotropic voxels (1 × 1 × 1 mm; field of view, 256 × 256 mm^2; TR, 1900 ms; TE, 2.92 ms; time to inversion, 900 ms; flip angle, 9°).

Imaging analysis

SPM8 software (Wellcome Department of Imaging Neuroscience, London, UK) was used to preprocess and analyze the imaging data. First, the anatomical scan was anterior commissure–posterior commissure centered, and this correction applied to all EPI volumes.

Functional data were motion corrected using a spatial transformation that realigned all functional volumes to the first volume of the run and subsequently realigned the volumes to the mean volume. The anatomical scan was coregistered to the mean volume and segmented. The anatomical and functional images were then normalized to the Montreal Neurological Institute (MNI) template using the parameters issued from the segmentation keeping the voxel resolution of the original scans (1 × 1 × 1 and 3 × 3 × 3 mm, respectively). Functional images were then smoothed with a Gaussian function (8 mm FWHM).

EPI time series were analyzed using the general linear model as implemented in SPM8. For each subject (first-level analysis), localizer and experimental data were modeled separately.

Localizer data

EPI time series were analyzed using the general linear model as implemented in SPM8. Functional data were analyzed in one two-level random-effects design. The first-level, fixed-effects individual participant analysis involved a design matrix containing a separate regressor for each block category ($n = 6$). These regressors contained boxcar functions representing the onset and offset of stimulation blocks convolved with a canonical hemodynamic response function. To account for residual motion artifacts, the realignment parameters were also added as nuisance covariates to the design matrix. Using the modified general linear model, parameter estimates for each condition at each voxel were calculated and then used to create contrast images for each category relative to baseline: AV-P > baseline, AV-O > baseline, A-P > baseline, A-O > baseline, V-P > baseline, and V-O > baseline. These six contrast images, from each participant, were taken forward into the second-level two factors (modality and category) ANOVA. The order of conditions was as follows: audiovisual (face + voice); audiovisual (object + sound); audio only (vocal); audio only (nonvocal); visual only (object); and visual only (face).

We then tested for general audiovisual regions with the conjunction analysis AV > V 32 ∩ AV > A (conjunction null hypothesis; Nichols et al., 2005), including both people and object information (i.e., AV-P + AV-O > V-P + V-O ∩ AV-P + AV-O > A-P + A-O). This localized regions that showed a higher response to audiovisual stimuli compared with both visual and audio-only stimuli.

Localizer results are reported at a threshold of $p < 0.05$ (cluster size corrected).

Main functional run: adaptation (continuous carryover)

Functional data were analyzed using four two-level random-effects designs: two that examined unimodal carryover effects, and two that examined crossmodal carryover effects.

Unimodal adaptation. For both face and voice unimodal carryover effects, brain activity time-locked to stimulus onset and duration was modeled in separate design matrices against one parametric modulator, which accounted for the absolute percentage difference between the (1)

face (i.e., unimodal face adaptation) or (b) voice morph levels (i.e., unimodal voice adaptation) of consecutive bimodal stimuli.

Crossmodal adaptation. In two separate design matrices (one for each possible direction of the crossmodal effect, i.e., face-to-voice or voice-to-face), brain activity was modeled against three parametric modulators: (1) the first accounted for the absolute difference between the face morph levels of consecutive stimuli (unimodal face adaptation); (2) the second accounted for the absolute difference between the voice morph levels of consecutive stimuli (unimodal voice adaptation); and (3) the third accounted for the crossmodal carryover effect that was either the absolute percentage difference between the (a) voice morph of a stimulus and the face morph of the following stimulus (i.e., voice-to-face crossmodal adaptation) or (b) the absolute percentage difference between the face morph of a stimulus and the voice morph of the following stimulus (i.e., face-to-voice crossmodal adaptation). In these design matrices, the latter crossmodal regressor was orthogonalized with respect to the first two unimodal regressors in the SPM routine. In this way, we were able to regress out the variance associated with unimodal effects before examining crossmodal effects. This was to ensure that we did not misinterpret effects apparently related to crossmodal adaptation but in fact attributable to unimodal adaptation.

In all four design matrices (unimodal face adaptation, unimodal voice adaptation, face-to-voice adaptation, and voice-to-face adaptation), a linear expansion allowed us to investigate regions in which the signal varied in account with the percentage difference in morph between stimuli, with a hypothesized linear modulation of signal as the degree of morph level difference increased parametrically. Contrasts for the effects at the first level for each design matrix were entered into four separate second-level, group random-effects analysis, in which we conducted a one-sample *t* test over first-level contrast images from all participants.

Whole-brain analyses are reported within an audiovisual versus baseline mask (mask threshold, $p < 0.001$, voxel uncorrected) at a threshold of $p < 0.05$ (FWE voxel corrected).

ROI analysis

In parallel with the whole-brain analysis, we performed an ROI-based analysis to specifically examine regions highlighted as involved in audiovisual integration in the separate multimodal localizer. Tests of unimodal and crossmodal effects within this ROI were conducted within MarsBar (ROI toolbox for SPM).

Results

Behavioral results

Direct effects

Each participant's mean categorization values for each audiovisual emotion morph stimulus (collapsed across actor) was submitted to a two-factor (face morph and voice morph) repeated-measures ANOVA, with five levels per factor (percentage of "anger" information in the morph). This was to assess the overall contributions of face and voice emotion morph on categorical response. The repeated-measures ANOVA highlighted a main effect of voice morph ($F_{(1.14,19.4)} = 15.3$, $p < 0.002$, $\eta^2p = 0.473$) and face morph ($F_{(2.02,34.3)} = 348$, $p < 0.0001$, $\eta^2p = 0.953$) and also a significant voice × face interaction ($F_{(5.78,98.1)} = 6.78$, $p < 0.0001$, $\eta^2p = 0.285$). Thus, it appears that face morph had a larger driving effect overall on categorization ratings, but its influence differed depending on what particular voice with which a face was paired.

In a series of planned comparisons, we further examined at which points there were significant differences in categorization ratings between stimuli. We proposed that maximum incongruence between face and voice (i.e., 80% difference) would cause significant shifts in categorization compared with "endpoint" congruent stimuli (i.e., 10% angry face–10% angry voice; 90% angry face–90% angry voice). To test these hypotheses, we performed the following paired-sample *t* tests: (1) 10% angry face–10% angry voice versus 10% angry face–90% angry voice; (2) 10% angry face–90% angry voice versus 90% angry face–90% angry voice; (3) 90% angry face–90% angry voice versus 90% angry face–10% angry voice; and (4) 90% angry face–10% angry voice versus 10% angry face–10% angry voice. After a Bonferroni's correction for multiple comparisons (level of significance, $p < 0.0125$), all comparisons were significant (comparison 1: $t_{(17)} = -24.0$, $p < 0.0001$; comparison 2: $t_{(17)} = -3.42$, $p < 0.004$; comparison 3: $t_{(17)} = 27.6$, $p < 0.0001$; and comparison 4: $t_{(17)} = 2.87$, $p < 0.0125$, respectively).

Second, each participant's mean reaction time values for each stimulus (collapsed across actors) were submitted to another two-factor (face morph and voice morph) repeated-measures ANOVA, with five levels per factor (percentage of anger information in the morph). As with categorical data, this was to assess the overall contribution of face and voice emotion morph—or the "direct effects" of face and voice morph—on reaction times. The ANOVA of reaction time data highlighted a main effect of voice morph ($F_{(2.91,49.6)} = 11.8$, $p < 0.0001$, $\eta^2p = 0.409$) and face morph ($F_{(2.34,39.7)} = 70.6$, $p < 0.0001$, $\eta^2p = 0.806$) and also a significant interaction between the two modalities ($F_{(2.90,39.4)} = 7.40$, $p < 0.0001$, $\eta^2p = 0.303$). Similar to the previous analysis, face morph drove the speed of categorization more than voice morph, albeit with different modulating effects at particular points in the 3D categorization space.

As in our categorization analysis, we proposed that maximum incongruence between face and voice (i.e., 80% difference) would take significantly longer to categorize compared with endpoint congruent stimuli. However, we also expected that some stimuli that were congruent, but with a lower clarity value (i.e., 50% angry face–50% angry voice), would take longer to categorize than endpoint congruent stimuli. To test these hypotheses, we performed the following paired-sample *t* tests: (1) 10% angry face–10% angry voice versus 10% angry face–90% angry voice; (2) 10% angry face–90% angry voice versus 90% angry face–90% angry voice; (3) 90% angry face–90% angry voice versus 90% angry face–10% angry voice; (4) 90% angry face–10% angry voice versus 10% angry face–10% angry voice; (5) 50% angry face–50% angry voice versus 10% angry face–10% angry voice; and (6) 50% angry face–50% angry voice versus 90% angry face–90% angry voice. After a Bonferroni's correction for multiple comparisons (level of significance, $p < 0.008$), comparisons 1, 4, 5, and 6 were significant (comparison 1: $t_{(17)} = -4.72$, $p < 0.0001$; comparison 4: $t_{(17)} = 3.25$, $p < 0.006$; comparison 5: $t_{(17)} = 10.67$, $p < 0.0001$; and comparison 6: $t_{(17)} = 6.29$, $p < 0.0001$, respectively), but comparisons 2 and 3 were not (comparison 2: $t_{(17)} = 1.30$, $p = 0.210$; and comparison 3: $t_{(17)} = -5.80$, $p = 0.569$, respectively).

All direct behavioral results are illustrated in Figure 2.

Adaptation effects

Here the interest was to investigate whether and how difference in emotion morph affected speed of emotion categorization. We conducted a hierarchical regression analysis for each subject, in which there were four regressors (face-to-face emotion morph difference, voice-to-voice emotion morph difference, face-to-voice emotion morph difference, and voice-to-face emotion morph difference), two of which were covariates in our model (face-to-face emotion morph difference and voice-to-voice emotion morph difference, i.e., the unimodal effects), and the dependent variable was reaction time. The first five stimulus values from each experimental block (apart from block one) were removed. This analysis provided two models: one that included only unimodal adaptation regressors, and a second that included all four adaptation regressors. In this way, and in parallel with the fMRI analysis, we ensured that any variance associated with the

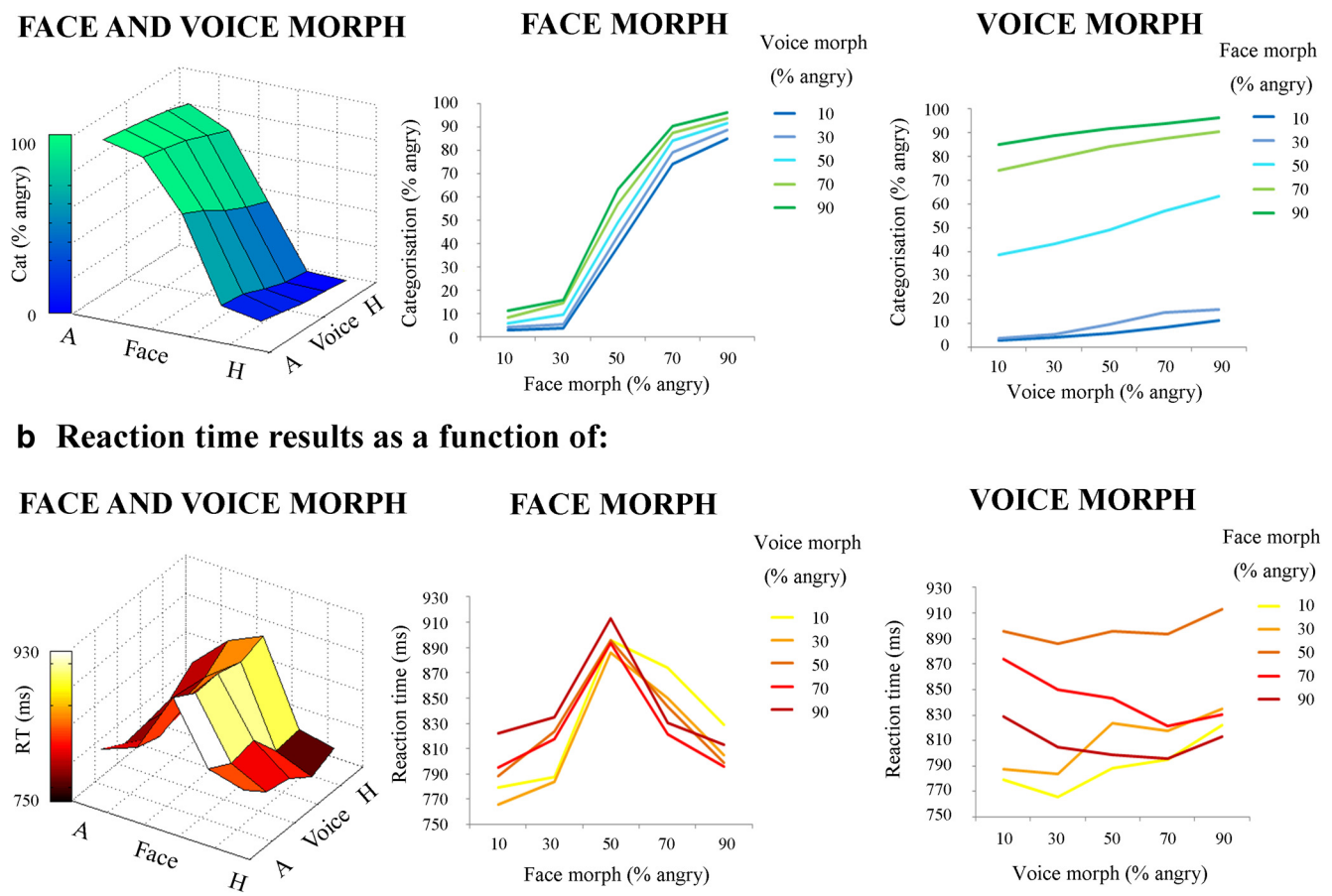

Figure 2. Behavioral results: direct effects of face and voice emotion morph. ***a***, Categorization results. Categorization (percentage angry responses) as a function of face morph (middle), voice morph (right), and both (left). ***b***, Reaction time results. Reaction time (milliseconds) as a function of face morph (middle), voice morph (right), and both (left). Both face and voice morph were morphed between 10% happy and 90% happy in 20% steps. Both categorization and reaction time results are averaged across actor. Note the greater influence of facial versus vocal emotional cues on behavioral responses. A, Angry; H, Happy.

two crossmodal adaptation regressors was independent of any unimodal effects. To analyze the significance at the group level, we entered individual β values for each regressor into separate one-sample *t* tests, in which they were compared with a hypothetical mean of zero. We first observed that, in the first model, there were significant unimodal adaptation effects (face: $t_{(17)} = -5.019$, $p < 0.0001$; voice: $t_{(17)} = 8.510$, $p < 0.0001$). Second, we found that there was a significant crossmodal adaptation effect but only in one direction: voice-to-face emotion morph difference ($t_{(17)} = 13.283$, $p < 0.0001$) significantly modulated reaction time, but face-to-voice emotion morph difference did not ($t_{(17)} = 1.353$, $p = 0.194$).

fMRI results

Multimodal localizer

A conjunction analysis of the auditory and visual conditions using the "max" criterion (AV > A ∩ AV > V; highlighting multimodal regions in which response to bimodal stimuli is greater than to each modality alone) identified a single cerebral region located in the posterior superior temporal gyrus (pSTG)/STS of the right hemisphere ($p < 0.05$, FWE cluster size corrected; Figs. 3, 4*c*; Table 1*a*). This cluster defined an ROI for tests of crossmodal adaptation in the main functional run (see below, ROI analysis).

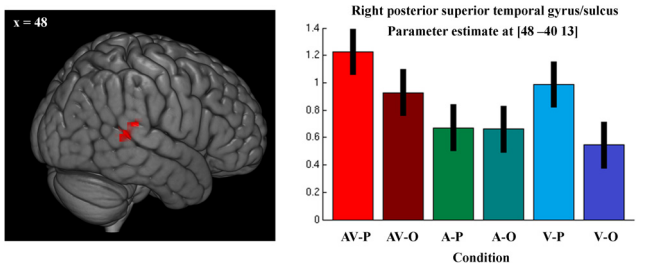

Figure 3. Imaging results: multimodal localizer. Left, A cluster in the right STG/STS responding more to audiovisual, compared with either visual or auditory, information alone localized using a conjunction analysis (AV > A ∩ AV > V; conjunction null hypothesis; Nichols et al., 2005). Results were thresholded at $p < 0.05$ (cluster corrected). Right, Condition effects at the peak voxel of the cluster.

Affective adaptation (continuous carryover)

Unimodal adaptation. We observed significant ($p < 0.05$, FWE voxel corrected) unimodal face adaptation in the left putamen and right fusiform gyrus (FG; Table 1*bi*) and significant voice adaptation effects in the bilateral STG/STS and right inferior frontal gyrus (Fig. 4*a*,*b*; Table 1*bii*). Generally, the response heightened as the degree of difference in happiness–anger morph between consecutive stimuli became larger, either in the auditory or visual modality.

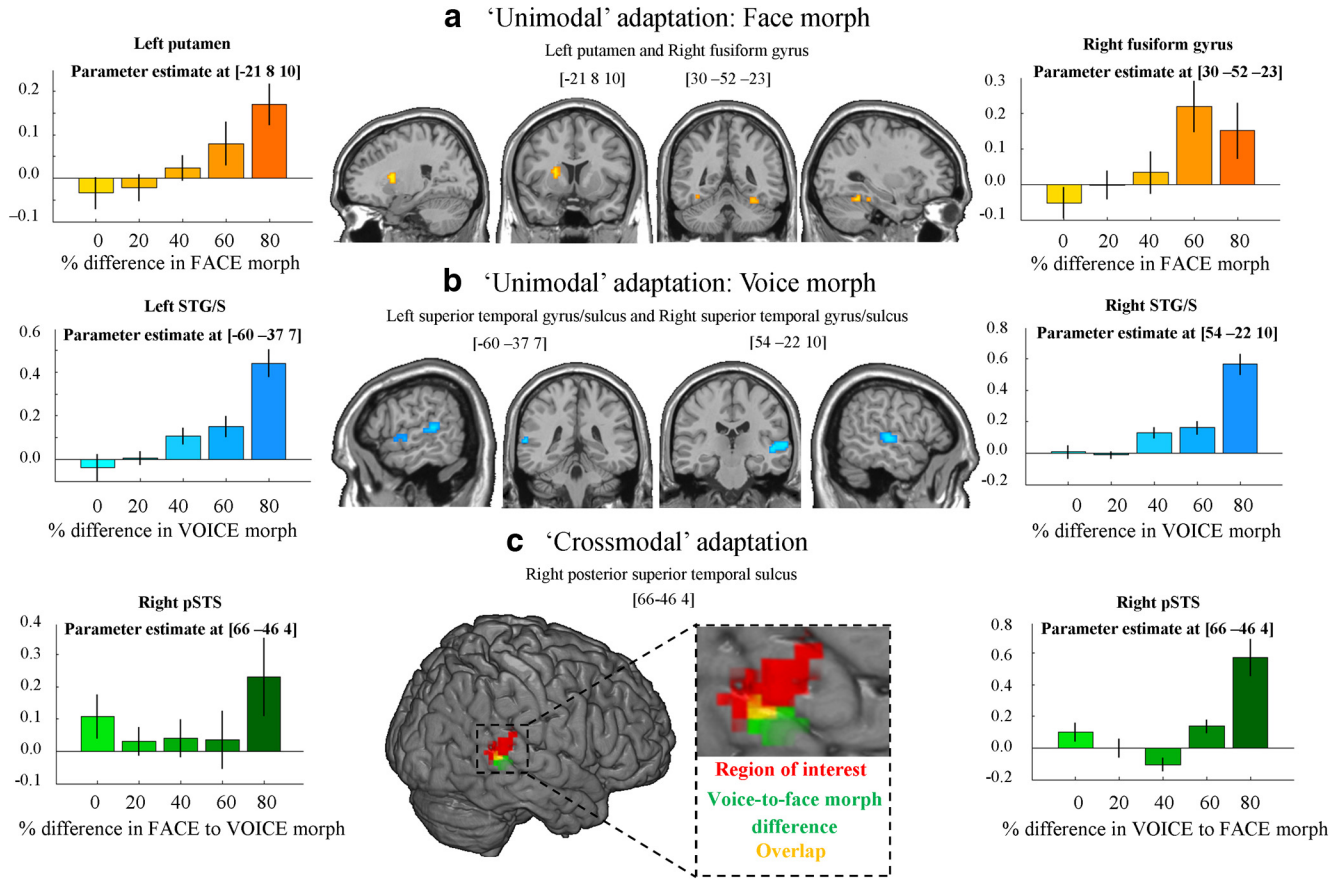

Figure 4. Imaging results. ***a***, Unimodal face adaptation. Activation in left putamen and right FG in response to varying percentage difference in face morph between consecutive stimuli. Left and right, Parameter estimate at the peak activated voxel of left putamen and right FG, respectively, as a result of varying percentage difference in face morph. Results were thresholded at $p < 0.05$ (FWE voxel corrected) and a minimum cluster size of more than five contiguous voxels. ***b***, Unimodal voice adaptation. Activation in bilateral STG/STS in response to varying percentage difference in voice morph between consecutive stimuli. Left and right, Parameter estimate at the peak activated voxel in left and right STG/STS, respectively, as a result of varying percentage difference in voice morph. Results were thresholded at $p < 0.05$ (FWE voxel corrected) and a minimum cluster size of more than five contiguous voxels. ***c***, Crossmodal adaptation. Red, Results from the independent functional multimodal localizer. An ROI analysis showed that voice-to-face emotion morph difference evoked a significant response in this region ($p < 0.025$, $t = 2.12$). Green, Activation in right pSTS as a result of varying percentage difference between voice and the following face morph of consecutive stimuli. Results were thresholded at $p < 0.001$ (voxel uncorrected) with a minimum cluster size of more than five contiguous voxels. Yellow, Overlap between the localizer and voice-to-face morph difference activation. Left and right, Parameter estimate at the peak activated voxel of right pSTS as a result of varying percentage difference in face-to-voice morph, and voice-to-face morph, respectively. It should be noted that face-to-voice morph difference did not evoke a significant response in this region.

Crossmodal adaptation. No crossmodal effects were significant at $p < 0.05$ (FWE voxel corrected) after having partialed out unimodal effects. However, at a more liberal threshold of $p < 0.001$ (voxel uncorrected), a crossmodal carryover effect was observed in the posterior part of the right pSTS (Fig. 4*c*; Table 1*ci*). Interestingly, this effect was asymmetrical as for the behavioral effect: activity was observed for voice-to-face emotion morph difference but not for face-to-voice emotion morph difference. That is, BOLD signal in response to an AV stimulus was greater in this region when there was a large difference between the facial emotion of the current stimulus and the vocal emotion of the previous one but not when the vocal emotion of the current stimulus differed from that of the previous face.

ROI analysis

A separate ROI analysis was conducted to test for crossmodal adaptation effects specifically at locations independently identified using the "multimodal" localizer (see above). Within this ROI in the pSTS/STG, there was a significant effect of unimodal voice adaptation ($p < 0.005$, $t = 2.92$), and unimodal face adaptation was just above the level of significance ($p = 0.055$, $t = 1.68$). Furthermore, we observed a significant crossmodal adaptation effect in addition to unimodal effects. Again, this effect was asymmetrical: it was observed for voice-to-face emotion morph difference ($p < 0.025$, $t = 2.12$) but not face-to-voice emotion morph difference ($p = 0.24$, $t = 0.72$).

Discussion

The aim of this study was to examine the cerebral correlates of the integration of affective information from the face and voice. Dynamic face-to-voice stimuli, parametrically varying in emotion, were used in conjunction with a continuous carryover design to enable measurements of adaptation effects (Aguirre, 2007). Overall, we demonstrate crossmodal adaptation in the right pSTS, suggesting the presence of multisensory, integrative neurons in this area.

Behavioral results indicated that emotion categorization and speed of categorization were modulated in line with parametric shifts in affective content of the face and voice. Significantly, both modalities affected emotion perception—an integration effect—but face morph exerted a far larger influence on behavioral responses, both categorical and reaction times. This is in line with other studies, in which emotion categorization has been found consistently to be more accurate and quicker for faces (Hess et al., 1988; Collignon et al., 2008; Bänziger et al., 2009).

Table 1. Imaging results

Brain regions	Coordinates (mm)			*k*	*t* statistic
	x	*y*	*z*		
a, Multimodal localizer					
STG/STS	48	−40	13	153	5.23
b, Unimodal adaptation					
i, Adaptation to face emotion					
Putamen	−21	8	10	14	7.46
FG	30	−52	−23	8	6.40
ii, Adaptation to voice emotion					
STG/STS	54	−22	10	51	7.98
STG/STS	−60	−37	7	24	7.37
Inferior frontal gyrus	48	23	22	11	6.32
c, Crossmodal adaptation					
i, Adaptation to voice-to-face emotion					
STS	66	−46	4	9	4.20
ii, Adaptation to face-to-voice emotion					
No significant voxels					

a, Results from multisensory functional localizer experiment. Contrasts were cluster thresholded at $p < 0.05$ (FWE corrected). MNI coordinates and *t* scores are from the peak voxel of a cluster. *b*, Unimodal adaptation results. *bi*, Adaptation to face emotion. *bii*, Adaptation to voice emotion. Contrasts were thresholded to display voxels reaching a significance level of $p < 0.05$ (FWE corrected) and an additional minimum cluster size of more than five contiguous voxels. Contrasts were also masked by an AV versus baseline contrast thresholded at $p < 0.001$ (voxel uncorrected). MNI coordinates and *t* scores are from the peak voxel of a cluster. *c*, Crossmodal adaptation results. *ci*, Voice-to-face adaptation. *cii*, Face-to-voice adaptation. Contrasts were thresholded to display voxels reaching a significance level of $p < 0.001$ (uncorrected) and an additional minimum cluster size of more than five contiguous voxels. Contrasts were masked by an AV versus baseline contrast thresholded at $p < 0.001$ (voxel uncorrected). MNI coordinates and *t* scores are from the voxel of a cluster.

We also observed adaptation effects at the behavioral level. Significantly, the crossmodal adaptation effect occurred only in one direction: the emotion morph difference between a voice and the following face significantly modulated reaction times, whereas that of a face and the following voice did not. This priming effect of vocal information on facial information is consistent with previous research highlighting crossmodal adaptive effects in the domain of identity processing (Ellis et al., 1997; Hills et al., 2010). Additionally, it should be noted that a recent study in fact demonstrated adaptation effects between affective face adaptors and test voices (Skuk and Schweinberger, 2013).

Cerebrally, we first observed that both face-to-face and voice-to-voice emotion morph difference modulated cerebral activity, namely in the putamen and FG, and bilateral STG/STS and inferior frontal gyrus, respectively. These findings are consistent with previous research on face and voice emotion perception. For example, a recent meta-analysis (Fusar-Poli et al., 2009) linked processing of emotional faces to increased activation in the putamen, in particular, that of happy faces. Furthermore, the FG has been associated consistently with the perception of human faces (Puce et al., 1995; Kanwisher et al., 1999; Haxby et al., 2000) and has been shown to be more active during expressive (e.g., fearful) face processing than neutral faces (Morris et al., 1998; Vuilleumier et al., 2004; Sabatinelli et al., 2011). Regarding affective voice processing, studies showed that the middle temporal gyrus and STS activate more when people listen to angry as opposed to neutral speech (Grandjean et al., 2005; Sander et al., 2005) or when people attend to affective prosody compared with the semantic content of the spoken words (Mitchell et al., 2003). Furthermore, Ethofer et al. (2009) demonstrated recently successful decoding of vocal emotions from fMRI responses in bilateral voice-sensitive areas.

Multisensory neurons in the right pSTS

Central to our main hypothesis, we observed crossmodal adaptation effects during face-to-voice emotion integration. Within a wide-reaching search of any regions responding to audiovisual information (compared with baseline), we observed a crossmodal adaptation effect in the right pSTS, a region that has been well documented as a multimodal region, in both humans (Beauchamp et al., 2004; Ethofer et al., 2006; Kreifelts et al., 2007, 2010; Watson et al., 2013) and nonhuman primates (Benevento et al., 1977; Bruce et al., 1981). This effect was small, only significant at a relatively lenient threshold, but importantly was independent of any variance elicited by either of the unimodal carryover effects: our design allowed us to regress out both unimodal face and voice adaptation effects, ensuring that the variance associated with crossmodal adaptation was modeled separately from variance explained by unimodal adaptation effects.

Additionally, this finding was confirmed in a complementary ROI analysis. Using an independent functional localizer, we isolated a cluster in the right pSTG/STS that responded more to audiovisual information than to either the visual or auditory modality alone, using a conjunction analysis (Goebel and van Atteveldt, 2009; Kreifelts et al., 2010; Love et al., 2011). We then tested for crossmodal adaptation within this cluster only, which yielded a significant effect.

Thus, our results suggest the existence of a sufficiently large proportion of multisensory neurons in the right pSTS to be detected using fMRI. This finding converges with that of a previous study that observed a "patchy organization" in this region consisting of interspersed unisensory and multisensory neurons (Beauchamp et al., 2004). We build on that observation by showing that some such multisensory neurons may integrate information in the context of affective processing. Furthermore, more recently, Kreifelts et al. (2009) observed that audiovisual integration of affective signals peaked in the anterior pSTS, at an overlap of face- and voice-sensitive regions. They proposed that this implies a possible interaction of the underlying voice- and face-sensitive neuronal populations during the formation of the audiovisual percept. We argue that such an audiovisual percept could partly reflect the contribution of populations of multisensory neurons. However, note we do not suggest that right pSTS, a complex, heterogeneous zone, is exclusively composed of bimodal neurons, nor do we suggest that all of face–voice integration effects in right pSTS are mediated by these bimodal neurons.

An asymmetrical crossmodal adaptation effect

Interestingly, the observed crossmodal adaptation effect was asymmetrical: activity in both whole-brain and ROI analyses was driven by the difference between a voice and the following face but not the difference between a face and the following voice. Therefore, it appears that voice exerted a stronger adaptive effect on face than face did on voice. This is in line with our behavioral data, in which only the difference between a voice and the following face significantly modulated reaction times.

With regards to the asymmetry of the observed effect, one might presume that, if a neuron was multisensory and therefore "coding" for both stimulus dimensions, both voice-to-face emotion morph difference and face-to-voice emotion morph difference would exert similar effects on its response. Why this was not the case could be attributable to various possibilities. It should be noted that, as mentioned previously, faces had a stronger effect on emotion judgment than voices. Therefore, the smaller effect of voices may have meant that modulations by preceding faces were even less pronounced and therefore did not reach significance at the behavioral and neural levels. However, in this experiment, we explicitly chose not to manipulate our stimuli so to "equate" the level of difficulty of emotion categorization; rather, the stimuli were left to reflect a natural situation in which affective information conveyed by the face and voice is rarely of equal informative-

ness. Additional investigation regarding this could involve including manipulations, such as adding noise to stimuli to equate categorization difficulty, to investigate whether this provokes parallel adaptation effects between face and voice, and voice and face.

Furthermore, alongside unequal direct effects, there may also have been underlying asymmetries in the unimodal adaptive effect of each modality, in turn affecting the strength of crossmodal adaptation in either direction. A recent study (Ethofer et al., 2013) investigated adaptation to faces, voices, and combined face–voice affective stimuli: these authors found that, although modality-specific cortices, such as the face-sensitive and voice-sensitive cortex in the STS, showed a stronger response habituation for their respective preferred stimulus class, the multisensory pSTS and orbitofrontal cortex (OFC) showed an adaptive response that was equal for both faces and voices. In the pSTS response habituation was stronger for audiovisual stimuli than for face-only and voice-only stimuli, whereas in the OFC it was equal across all three modalities. It would be of interest to see whether, in these same regions in which adaptation to faces and voices was approximately equal, there would additionally be bidirectional crossmodal adaptation effects. However, at this point at least, our results seem to converge with that of this study in that the pSTS seems to be a main locus of audiovisual integration effects.

However, equally, we at the same time argue that there is perhaps no reason to assume that the effect should be perfectly symmetrical: indeed, rather than an "all-or-nothing" phenomenon, such multimodal neurons may receive different proportions of synapses from visual and auditory neurons, subsequently influencing the strength of the crossmodal adaptation effect in either direction. Furthermore, those visual and auditory inputs could be characterized by differential modulating effects or weighting on the neural response.

Finally, regarding the pattern of this asymmetrical crossmodal effect, we noted that, at the peak voxel at least, the effect appeared to be driven particularly by most extreme morph level difference (i.e., 80%), perhaps acting as a "tipping point" for the marked release from adaptation. In other words, in the case of crossmodal adaptation, it could be possible that there is a precise percentage difference in emotion between the modalities at which release from adaptation is triggered rather than a graded linear parametric effect, as appeared more clearly in the unimodal face and voice adaptation analyses. That said, it should also be noted that, with inclusion of the 80% difference condition, the plot of effects had a strong linear component, and therefore we would still suggest that the physiology of the effect would be reflected by a linear expansion of the parametric modulator. However, an interesting future direction might be to investigate how inclusion of specific percentage differences in affect morph level would affect the grading of the adaptive response. This would be particularly relevant to crossmodal adaptation, in which inclusion or exclusion of particular audiovisual stimuli (and therefore morph differences) may evoke or extinguish the adaptive effect in either direction or change the pattern of the effect (e.g., linear to quadratic response). In this way, we may be able to tap into the more fine-grained mechanisms of affective face–voice integration.

References

Aguirre GK (2007) Continuous carry-over designs for fMRI. Neuroimage 35:1480–1494. CrossRef Medline

Bänziger T, Grandjean D, Scherer KR (2009) Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). Emotion 9:691–704. CrossRef Medline

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190–1192. CrossRef Medline

Benevento LA, Fallon J, Davis BJ, Rezak M (1977) Auditory–visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. Exp Neurol 57:849–872. CrossRef Medline

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369–384. Medline

Collignon O, Girard S, Gosselin F, Roy S, Saint-Amour D, Lassonde M, Lepore F (2008) Audio-visual integration of emotion expression. Brain Res 1242:126–135. CrossRef Medline

de Gelder B, Böcker KB, Tuomainen J, Hensen M, Vroomen J (1999) The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. Neurosci Lett 260:133–136. CrossRef Medline

Dolan RJ, Morris JS, de Gelder B (2001) Crossmodal binding of fear in voice and face. Proc Natl Acad Sci USA 98:10006–10010. CrossRef Medline

Ellis HD, Jones DM, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. Br J Psychol 88:143–156. CrossRef Medline

Ethofer T, Pourtois G, Wildgruber D (2006) Investigating audiovisual integration of emotional signals in the human brain. Prog Brain Res 156:345–361. CrossRef Medline

Ethofer T, Van De Ville D, Scherer K, Vuilleumier P (2009) Decoding of emotional information in voice-sensitive cortices. Curr Biol 19:1028–1033. CrossRef Medline

Ethofer T, Bretscher J, Wiethoff S, Bisch J, Schlipf S, Wildgruber D, Kreifelts B (2013) Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus. Neuroimage 76:45–56. CrossRef Medline

Fusar-Poli P, Placentino A, Carletti F, Landi P, Allen P, Surguladze S, Benedetti F, Abbamonte M, Gasparotti R, Barale F, Perez J, McGuire P, Politi P (2009) Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J Psychiatry Neurosci 34:418–432. Medline

Goebel R, van Atteveldt N (2009) Multisensory functional magnetic resonance imaging: a future perspective. Exp Brain Res 198:153–164. CrossRef Medline

Grandjean D, Sander D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P (2005) The voices of wrath: brain responses to angry prosody in meaningless speech. Nat Neurosci 8:145–146. CrossRef Medline

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R (1999) Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24:187–203. CrossRef Medline

Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci 10:14–23. CrossRef Medline

Hagan CC, Woods W, Johnson S, Calder AJ, Green GG, Young AW (2009) MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. Proc Natl Acad Sci USA 106:20010–20015. CrossRef Medline

Hagan CC, Woods W, Johnson S, Green GG, Young AW (2013) Involvement of right STS in audio-visual integration for affective speech demonstrated using MEG. PLoS One 8:e70648. CrossRef Medline

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4:223–233. CrossRef Medline

Hess U, Kappas A, Scherer K (1988) Multichannel communication of emotion: synthetic signal production. In: Facets of emotion: recent research (Scherer K, ed), pp 161–182. Hillsdale, NJ: Erlbaum.

Hills PJ, Elward RL, Lewis MB (2010) Cross-modal identity aftereffects and their relation to priming. J Exp Psychol Hum Percept Perform 36:876–891. CrossRef Medline

Kanwisher N, Stanley D, Harris A (1999) The fusiform face area is selective for faces not animals. NeuroReport 10:183–187. CrossRef Medline

Kawahara H (2006) STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds. Acoust Sci Tech 27:349–353. CrossRef

Klasen M, Kenworthy CA, Mathiak KA, Kircher TT, Mathiak K (2011) Su-

pramodal representation of emotions. J Neurosci 31:13635–13643. CrossRef Medline

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

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. CrossRef Medline

Kreifelts B, Ethofer T, Huberle E, Grodd W, Wildgruber D (2010) Association of trait emotional intelligence and individual fMRI activation patterns during the perception of social signals from voice and face. Hum Brain Mapp 31:979–991. CrossRef Medline

Love SA, Pollick FE, Latinus M (2011) Cerebral correlates and statistical criteria of cross-modal face and voice integration. Seeing Perceiving 24: 351–367. CrossRef Medline

Mitchell RL, Elliott R, Barry M, Cruttenden A, Woodruff PW (2003) The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. Neuropsychologia 41:1410–1421. CrossRef Medline

Morris JS, Friston KJ, Büchel C, Frith CD, Young AW, Calder AJ, Dolan RJ (1998) A neuromodulatory role for the human amygdala in processing emotional facial expressions. Brain 121:47–57. CrossRef Medline

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660. CrossRef Medline

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. CrossRef Medline

Pourtois G, Debatisse D, Despland PA, de Gelder B (2002) Facial expressions modulate the time course of long latency auditory brain potentials. Brain Res Cogn Brain Res 14:99–105. CrossRef Medline

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 41:49–59. CrossRef Medline

Puce A, Allison T, Gore JC, McCarthy G (1995) Face-sensitive regions in human extrastriate cortex studied by functional MRI. J Neurophysiol 74:1192–1199. Medline

Robins DL, Hunyadi E, Schultz RT (2009) Superior temporal activation in response to dynamic audio-visual emotional cues. Brain Cogn 69:269–278. CrossRef Medline

Sabatinelli D, Fortune EE, Li Q, Siddiqui A, Krafft C, Oliver WT, Beck S, Jeffries J (2011) Emotional perception: meta-analyses of face and natural scene processing. Neuroimage 54:2524–2533. CrossRef Medline

Sander D, Grandjean D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P (2005) Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. Neuroimage 28:848–858. CrossRef Medline

Skuk VG, Schweinberger SR (2013) Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. PLoS One 8:e81691. CrossRef Medline

Tal N, Amedi A (2009) Multisensory visual–tactile object related network in humans: insights gained using a novel crossmodal adaptation approach. Exp Brain Res 198:165–182. CrossRef Medline

Vuilleumier P, Richardson MP, Armony JL, Driver J, Dolan RJ (2004) Distant influences of amygdala lesion on visual cortical activation during emotional face processing. Nat Neurosci 7:1271–1278. CrossRef Medline

Watson R, Latinus M, Noguchi T, Garrod O, Crabbe F, Belin P (2013) Dissociating task difficulty from incongruence in face-voice emotion integration. Front Hum Neurosci 7:744. CrossRef Medline

Watson R, Latinus M, Charest I, Crabbe F, Belin P (2014) People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. Cortex 50:125–136. CrossRef Medline