Crossmodal Integration Enhances Neural Representation of Task-Relevant Features in Audiovisual Face Perception

Yuanqing Li1, Jinyi Long1, Biao Huang2, Tianyou Yu1, Wei Wu1, Yongjian Liu3, Changhong Liang2 and Pei Sun4

1Center for Brain Computer Interfaces and Brain Information Processing, South China University of Technology, Guangzhou 510640, China, 2Department of Radiology, Guangdong General Hospital, Guangzhou 510080, China, 3Department of MR, Foshan Hospital of Traditional Chinese Medicine, Foshan 528000, China and 4Department of Psychology, Tsinghua University, Beijing 100084, China

Address all correspondence to Yuanqing Li, Center for Brain Computer Interfaces and Brain Information Processing, South China University of Technology, Guangzhou, 510640, China. Email: auyqli@scut.edu.cn; Pei Sun, Department of Psychology, Tsinghua University, Beijing, 100084, China. Email: peisun@tsinghua.edu.cn

Previous studies have shown that audiovisual integration improves identification performance and enhances neural activity in heteromodal brain areas, for example, the posterior superior temporal sulcus/middle temporal gyrus (pSTS/MTG). Furthermore, it has also been demonstrated that attention plays an important role in crossmodal integration. In this study, we considered crossmodal integration in audiovisual facial perception and explored its effect on the neural representation of features. The audiovisual stimuli in the experiment consisted of facial movie clips that could be classified into 2 gender categories (male vs. female) or 2 emotion categories (crying vs. laughing). The visual/auditory-only stimuli were created from these movie clips by removing the auditory/visual contents. The subjects needed to make a judgment about the gender/emotion category for each movie clip in the audiovisual, visual-only, or auditory-only stimulus condition as functional magnetic resonance imaging (fMRI) signals were recorded. The neural representation of the gender/emotion feature was assessed using the decoding accuracy and the brain pattern-related reproducibility indices, obtained by a multivariate pattern analysis method from the fMRI data. In comparison to the visual-only and auditory-only stimulus conditions, we found that audiovisual integration enhanced the neural representation of task-relevant features and that feature-selective attention might play a role of modulation in the audiovisual integration.

Keywords: audiovisual face perception, brain pattern, crossmodal integration, decoding, feature-selective attention, reproducibility

Introduction

During social communication, we acquire different types of information such as speech content, age, gender, and emotion, from a speaker through different modalities that can be visual or auditory. All of these types of multisensory information are believed to be integrated in our brain. Numerous behavioral studies have shown that audiovisual integration facilitates person perception and recognition. For instance, bimodal congruent face–voice stimuli lead to faster and more accurate categorization of emotion expressions (Collignon et al. 2008), voice recognition (Schweinberger et al. 2007), and identity information processing (Campanella and Belin 2007; Schweinberger et al. 2007). Compared to unimodal faces or voices, congruent face–voice pairs not only lead to beneficial behaviors, but also produce different brain activity patterns. For instance, congruent audiovisual emotional stimuli can modulate the activity in the bilateral superior temporal gyrus (STG), fusiform gyrus (FG), left amygdala, and right thalamus (Kreifelts et al. 2007; Jeong et al. 2011) and enhance the connectivity between audiovisual integration areas and associative auditory and visual cortices (Kreifelts et al. 2007).

It is still under debate how semantic information from different modalities is integrated in the brain. As shown by human neuroimaging studies, enhanced brain activities are observed in posterior superior temporal sulcus (STS)/middle temporal gyrus (pSTS/MTG), a heteromodal area, in the congruent audiovisual stimulus conditions (e.g., Calvert et al. 2000; Bushara et al. 2003; Beauchamp, Argall, et al. 2004; Kreifelts et al. 2007; Stevenson et al. 2010). However, this area appears to be relatively insensitive to the meaning of congruent multimodal objects (Beauchamp, Argall, et al. 2004; Beauchamp, Lee, et al. 2004; Taylor et al. 2006). Furthermore, it was shown that while pSTS/MTG acts as a presemantic, heteromodal region for crossmodal perceptual features, perirhinal cortex integrates these features into higher level conceptual representations (Taylor et al. 2006). The conceptual representation of a stimulus feature may be described as a pattern in the human brain (Cox and Savoy 2003; Kriegeskorte et al. 2006; Formisano et al. 2008; Mitchell et al. 2008). From the viewpoint of effective neural representation, the brain activity patterns corresponding to different semantic categories of stimuli should be differentiable, whereas the brain activity patterns corresponding to the same semantic category of stimuli should be reproducible. Until now, it remains unclear whether or not crossmodal integration improves the discriminability of brain activity patterns corresponding to different semantic categories of stimuli and the reproducibility of the same semantic category of brain activity patterns.

Recent findings indicate that attention can modulate integration across various stages (Koelewijn et al. 2010; Talsma et al. 2010). For instance, a close relationship between crossmodal attention and crossmodal binding during speech reading has been demonstrated (Saito et al. 2005). Although a large number of studies have addressed the question of how attentional shifts in on one modality can affect orienting in other modalities, the role that attention plays during multisensory integration itself seems to be largely overlooked (Talsma et al. 2010). During audiovisual facial perception, we may attend to only one feature (e.g., emotion) of the speaker’s face and voice while ignoring other features, and this phenomenon is related to so-called feature-selective attention, a special form of feature-based attention (Nobre et al. 2006; Mirabella et al. 2007; Chelazzi et al. 2010). To our knowledge, no study has
addressed the issue of how feature-selective attention affects the neural representations of different features during cross-modal integration.

In view of these 2 questions, we explored the effect of cross-modal integration on the neural representation of task-relevant features in audiovisual facial perception. According to previous studies, we hypothesized that audiovisual integration only enhances the neural representation of task-relevant features. In our experiment, we used facial movie clips that could be classified orthogonally into 2 gender categories (male vs. female) or 2 emotion categories (crying vs. laughing) as the congruent audiovisual stimuli; as unimodal visual/auditory stimuli, we used the same movie clips after removing the appropriate auditory or visual contents. The subjects were instructed to make a judgment on the gender/emotion category for each movie clip in the congruent audiovisual, visual-only, or auditory-only stimulus condition while functional magnetic resonance imaging (fMRI) signals were recorded. Through applying a multi-voxel pattern analysis (MVPA) method to the fMRI data, we performed category decoding in the gender or emotion dimension. During the gender or emotion category decoding, we estimated the neural activity patterns elicited by the presented stimuli, which reflected the neural representation of the gender or emotion feature contained in these stimuli, and their category information. We assessed these neural activity patterns using 3 indices, that is, decoding accuracy, within-class reproducibility (for the same category of brain activity patterns) and between-class reproducibility (for different categories of brain activity patterns), and test our hypothesis accordingly.

Materials and Methods

Subjects
Nine healthy native male Chinese (aged 23–45 years) participated in the study. All participants had normal or corrected-to-normal vision and gave their informed written consent prior to the experiment. The experimental protocol was approved by the Ethics Committee of Guangdong General Hospital, China.

Experimental Stimuli and Design
Eighty movie clips of human faces including video and audio recordings were selected from internet sources. Semantically, these 80 movie clips could be partitioned orthogonally into 2 groups based on either the gender (40 male vs. 40 female Chinese faces) or the emotion (40 crying vs. 40 laughing faces). The estimated ages of the persons in the movie clips ranged from 20 to 70 years old as evaluated by another independent group of 3 subjects. After appropriate image processing (Windows movie maker), each edited movie clip was in gray scale, lasting 1400 ms and subtending 10.7° × 8.7°. The luminance levels of the videos were matched by adjusting the total power value (the sum of the squares of the pixel gray values; see examples in Fig. 1A) of each video. Similarly, the audio power levels were matched by adjusting the total power value of each audio clip. These edited movie clips consisting of both video and audio recordings were used as the audiovisual stimuli in our experiment. The unimodal visual/auditory stimuli were from the same movie clips as above but with either the audio or video portion removed. As the audiovisual stimuli in this study were always congruent, hereafter, we did not explicitly state the word “congruent” in the audiovisual/multimodal stimulus condition.

The visual stimulus was projected onto a screen using an LCD projector (SA-9900 fMRI Stimulation System, Shenzhen Sinorad Medical Electronics, Inc.). Subjects viewed the stimulus through a mirror mounted on the head coil, and the auditory stimulus was delivered through a pneumatic headset (SA-9900 fMRI Stimulation System, Shenzhen Sinorad Medical Electronics, Inc.) with a special design to minimize the interference of scanner noise. Before the scanning, the sound level of the headset was adjusted such that the subject could hear the auditory stimulus clearly and comfortably.

We utilized a 2 × 3 factorial design, with the task (gender judgment or emotion judgment) as the first factor and the stimulus condition (audiovisual, visual-only, or auditory-only) as the second factor. Each subject performed 6 experimental runs corresponding to the 6 pairs of tasks and stimulus conditions respectively, with the order pseudorandomized. Each run included 10 blocks and each block contained 8 trials. The 6 runs took place in 3 different days (2 per day) for each subject to avoid fatigue. During the experiment, the subjects were asked to focus their attention on either the gender or the emotion of the presented stimuli (audiovisual, visual-only, or auditory-only stimuli) and make a corresponding judgment (male vs. female for gender judgment task, or crying vs. laughing for emotion judgment task) to each stimulus. When the subject performed a gender/emotion judgment task, the gender/emotion feature was defined as task-relevant while the emotion/gender feature was defined as task-irrelevant. As an

![Figure 1](https://example.com/fig1.png)

Figure 1. (A) Four examples of audiovisual stimuli. (B) Time course of a trial. The presentation of a stimulus lasted 1400 ms and was repeated 4 times during the first 8 s in a trial. A visual cue (+) appeared at the eighth second and persisted for 6 s. The stimuli were presented in 2 runs, 1 for the gender judgment task and 1 for the emotion judgment task.
example, in the following we describe the experimental procedure of the run corresponding to the audiovisual stimulus condition with emotion judgment task (the other runs were performed with similar procedures). At the beginning of the run, 5 volumes (lasting 10 s) were acquired without stimulation. The 80 audiovisual stimuli were randomly assigned to the 80 trials, with the gender and emotion categories of the stimuli balanced within each block. There was a 20-s blank period (gray screen and no auditory stimulation) between adjacent blocks. At the beginning of each block, a short instruction ("cry 1 and laugh 2" or "cry 2 and laugh 1") was displayed for 4 s on the screen. The instruction "cry 1 and laugh 2" required that the subject should press key 1 and key 2 for crying and laughing emotions, respectively, whereas the instruction "cry 2 and laugh 1" implied that the subject needed to press key 2 and key 1 for crying and laughing emotions, respectively. The 2 keys were pseudorandomly assigned to the 2 emotion categories in each block. Similarly, for gender judgment task in the other runs, the instructions displayed were either "male 1 and female 2" or "male 2 and female 1," which instructed the subject to press the respective keys to indicate the gender categories of the stimuli. At the beginning of each trial, a stimulus was presented to the subject for 1400 ms, followed by a 600-ms blank period. This 2-s cycle with the same stimulus was repeated 4 times for effectively eliciting a brain activity pattern and was followed by a 6-s blank period. After the stimulation, a fixation cross appeared on the screen, and the subject was asked to make an emotion judgment by pressing one of the 2 keys. The fixation cross changed color at the 12th second, indicating that the next trial would begin shortly (Fig. 1B). In total, a run lasted 1350 s.

**Behavioral Experiment Outside Scanner**

An additional group of 12 subjects (9 men and 3 women, aged 23–36 years) participated in a separate behavioral experiment that was conducted to further examine whether there were benefits of multisensory integration at the behavioral level. Different groups of subjects were recruited in the behavioral and fMRI experiments in order to avoid the confounding introduced by subjects’ familiarity of the stimulus materials. The design of the behavioral experiment was identical to that of the fMRI experiment except that the stimulus was presented only once within each trial such that the response time (RT) accurately reflected subject’s speed of judgment. After the experiment, the RT and the percentage of correct judgment were calculated for each pair of stimulus condition and task.

**fMRI data Acquisition and Preprocessing**

fMRI Experiments were performed using a GE Signal Excite HD 3-Tesla MR scanner in Guangdong General Hospital, China. Prior to the functional experiment, a 3D anatomical T1-weighted scan (FOV, 280 mm; matrix, 256 × 256; 128 slices; slice thickness: 1.8 mm) was acquired for each subject per day. During the functional experiment, gradient-echo echo-planar (EPD) T2*-weighted images (25 slices with ascending noninterleaved order; TR = 2000 ms, TE = 35 ms, flip angle = 70 deg; FOV: 280 mm, matrix: 64 × 64, slice thickness: 5.5 mm, no gap) were acquired and covered the whole brain. As described in the preceding text, each subject performed 6 runs. A total of 675 volumes and the corresponding behavioral data were acquired in each run.

In each run, the first 5 volumes collected before magnetization equilibrium were received from analysis. For each subject, preprocessing consisted of head motion correction, slice timing correction, co-registration between the functional scans and the structural scan, normalization to an MNI standard brain, data masking to exclude nonbrain voxels, time series detrending, and normalization of the time series in each block to zero mean and unit variance. All preprocessing steps were conducted using SPM5 (Friston et al. 1994) and custom functions in MATLAB 7.4 (Mathworks, Natick).

**MVPA Procedure**

MVPA approaches can effectively pool the information available across many fMRI voxels, allowing a feature of the presented stimuli to be decoded from coarse-scale population responses. Typically, an increase in the strength of feature-selective fMRI responses is reflected by improved decoding performance (Jehee et al. 2011). Furthermore, by focusing on distributed activity patterns, MVPA approaches enable us to separate and localize spatially distributed patterns that are generally too weak to be detected by univariate methods, such as the general linear model (GLM) (Friston et al. 1994; Polyn et al. 2005; Goebel and van Atteveldt, 2009; Pereira et al. 2009; Zeng et al. 2012). In previous fMRI studies, MVPA approaches have been successfully used to decode stimulus features from fMRI signals (e.g., Haxby et al. 2001; Cox and Savoy 2003; Kamitani and Tong, 2005; Formisano et al. 2008; Kay et al. 2008; Mitchell et al. 2008; Miyawaki et al. 2008; Li, Mayhew, et al. 2009; Li, Namburi, et al. 2009).

In the experiment for each subject, there were 2 runs corresponding to the gender and emotion judgment tasks for each of the audiovisual, visual-only, and auditory-only stimulus conditions. For each run, fMRI data were used to decode the gender/emotion categories of the stimuli perceived by the subject. In each decoding calculation, an MVPA method was applied to the fMRI data of a run through a 10-fold cross-validation. As an example, below we describe the data processing steps for one decoding calculation in full detail. For the 10-fold cross-validation, the 80-trial data were equally partitioned into 10 nonoverlapping datasets, each corresponding to 1 of the 10 blocks. For the kth fold of the cross-validation (k = 1, . . . , 10), the kth dataset (8 trials) was used for prediction and performance evaluation, and the other 9 datasets (72 trials) were used for training, as described below.

1. "Initial voxel selection with a spherical searchlight algorithm."

   The initial voxel selection was based on the training dataset. A spherical searchlight algorithm that was sequentially centered at each voxel with a 3-mm radius searchlight highlighting 19 voxels was used (Kriegeskorte et al. 2006). Within each searchlight, a Fisher ratio was computed through Fisher linear discriminant analysis (FLDA) as a multivariate contrast statistic that could pool the discriminative information of all the contained voxels. The Fisher ratio was then recorded in a statistical map for the voxel at the center of the searchlight. As a result, this method yielded a spatially continuous map indicating the level of discrimination between the 2 gender or emotion categories (male vs. female or crying vs. laughing) in the local neighborhood of each voxel.

2. "Neural activity pattern estimation." Based on the resulting Fisher ratio map, K informative voxels with the highest Fisher ratios were selected. A K-dimensional pattern vector was then constructed for each trial, of which each element represented the mean BOLD response of a selected voxel from 6 to 14 s of each trial (the last 4 volumes to take into account the delay in the hemodynamic response). This pattern vector, which depended on 2 factors, that is, spatial locations of the selected voxels and their average signal amplitudes across multiple time points, reflected the neural representation of the gender or emotion feature contained in the presented stimulus.

3. "Prediction." We trained a linear SVM classifier using the pattern vectors of the labeled training data (72 trials). For each trial of the test dataset (the 8 trials not used in the training stage), a pattern vector was extracted as described above. The gender/emotion category was predicted by applying the trained SVM classifier to the pattern vector. After a 10-fold cross-validation, the average decoding/prediction accuracy rate was calculated across all folds.

4. "Calculation of within-class and between-class reproducibility indices." In (Schurger et al. 2010), the angle between 2 pattern vectors was used as a reproducibility index to measure their similarity. In the present study, we used cos θ as a reproducibility index to further assess the neural activity patterns elicited by the presented stimuli, where θ was the angle between 2 pattern vectors. The larger the cos θ, the higher the similarity. Each pattern vector, denoted by a column vector Pi, belonged to either of the 2 classes denoted by C1 and C2 (male vs. female in the gender dimension or crying vs. laughing in the emotion dimension). We calculated the average within-class and between-class reproducibility indices RN and RB for the kth fold of cross-validation as below (we extracted 8 pattern vectors corresponding to the 8 trials from the test
The permutation procedure was similar to that in (Kriegeskorte et al. 2000; Calvert and Thesen, 2004; Beauchamp, 2005), where the [Pi] denotes the intersection of 2 sets. \[ C \cap \]_dc. [AV > max(A,V) (P < 0.05, uncorrected)] found regions with the response for multimodal stimulus condition exceeding the maximum of the responses for auditory-only and visual-only stimulus conditions, and was implemented by performing the conjunction null test of AV > V and AV > A in SPM5.

To complement the statistical analysis, we also computed the percent signal changes of the pSTS/MTG clusters for each subject, each task, and each stimulus condition by conducting region of interest (ROI)-based analysis (performed by the MATLAB toolbox MarsBar-0.43 (Brett et al. 2002)). The clusters consisting of significantly activated voxels in the bilateral pSTS/MTG were determined by the group GLM analysis described above. For each subject, each task, and each stimulus condition, a GLM model was first estimated from the mean BOLD signal of the cluster. The percent signal change in the clusters was then computed as the ratio of the maximum of the estimated event response to the baseline.

**Results**

**Behavioral Results**

Figure 2 shows the behavioral results, that is, the RTs and the percentages of correct judgment, from the fMRI experiment (left in each subplot) and the separate behavioral experiment (right in each subplot). For the gender judgment task (Fig. 2A, C), 1-way repeated measure ANOVA revealed that there was a significant main effect of stimulus condition (the audiovisual, visual-only, and auditory-only conditions) on both RT (inside scanner: \( P < 0.001, F_{2, 8} = 15.245 \); outside scanner: \( P < 0.05, F_{2, 11} = 5.681 \)) and the percentage of correct judgment (inside scanner: \( P < 0.001, F_{2, 8} = 15.245 \); outside scanner: \( P < 0.05, F_{2, 11} = 18.278 \)). Post hoc Bonferroni-corrected paired \( t \)-tests showed that the RT was significantly lower (inside scanner: \( P < 0.05, t_{(8)} = 4.1079 \) for audiovisual stimulus vs. visual-only stimulus; \( P < 0.01, t_{(8)} = 4.2699 \) for audiovisual stimulus vs. auditory-only stimulus; outside scanner: \( P < 0.05, t_{(11)} = 2.9107 \) for audiovisual stimulus vs. visual-only stimulus; \( P < 0.05, t_{(11)} = 3.2585 \) for audiovisual stimulus vs. auditory-only stimulus) and the percentage of correct judgment for the gender judgment task was significantly higher for the audiovisual stimulus condition than for the visual-only or auditory-only stimulus condition (inside scanner: \( P < 0.05, t_{(8)} = 3.2576 \) for audiovisual stimulus vs. visual-only stimulus; \( P < 0.01, t_{(8)} = 4.9247 \) for auditory-only stimulus vs. auditory-only stimulus; outside scanner: \( P < 0.05, t_{(11)} = 2.9272 \) for audiovisual stimulus vs. visual-only stimulus; \( P < 0.01, t_{(11)} = 5.7208 \) for audiovisual stimulus vs. auditory-only stimulus).

For the emotion judgment task (Fig. 2B, D), 1-way repeated measure ANOVA revealed that there was a significant main effect of stimulus condition (the audiovisual, visual-only, and auditory-only conditions) on both RT (inside scanner: \( P < 0.001, F_{2, 8} = 12.217 \); outside scanner: \( P < 0.001, F_{2, 11} = 10.809 \)) and the percentage of correct judgment (inside scanner: \( P < 0.01, F_{2, 8} = 7.058 \); outside scanner: \( P < 0.001, F_{2, 11} = 11.785 \)). Post hoc Bonferroni-corrected paired \( t \)-tests showed that the RT was significantly lower for the audiovisual stimulus condition than for the visual-only or auditory-only stimulus condition (inside scanner: \( P < 0.05, t_{(8)} = 3.9898 \) for audiovisual stimulus vs. visual-only stimulus; \( P < 0.05, t_{(8)} = 3.6807 \) for audiovisual stimulus vs. auditory-only stimulus; outside scanner: \( P < 0.01, t_{(11)} = 4.1206 \) for audiovisual...
stimulus vs. visual-only stimulus; \( P < 0.01 \) corrected, \( t_{(11)} = 4.8283 \) for audiovisual stimulus vs. auditory-only stimulus). However, only for the separate behavioral experiment, the percentage of correct judgment for the emotion judgment task was significantly higher for the audiovisual stimulus condition than for the visual-only or auditory-only stimulus condition (outside scanner; \( P < 0.05 \) corrected, \( t_{(11)} = 3.3364 \) for audiovisual stimulus vs. visual-only stimulus; \( P < 0.01 \) corrected, \( t_{(11)} = 4.9489 \) for audiovisual stimulus vs. auditory-only stimulus). There was no significant difference between the percentage of correct judgment for the audiovisual stimulus condition and that for the visual-only stimulus condition for the fMRI experiment; this might be due to the relatively small number of participating subjects.

Additionally, from Figure 2 it can be seen that the behavioral performance of the subjects was worse for the fMRI experiment than for the separate behavioral one. We may explain this according to the noisy environment inside the scanner and the higher level of fatigue for the subjects resulted by the longer duration of the fMRI experiment.

Decoding Accuracy and Reproducibility

For each of the 9 subjects and each of the audiovisual, visual-only, and auditory-only stimulus conditions, we conducted 2 experimental runs, 1 for the gender judgment task and 1 for the emotion judgment task. For each run, we separately decoded the gender categories (“male” and “female”) and the emotion categories (“crying” and “laughing”) of the stimuli from the collected fMRI signals using the MVPA method. Each decoding calculation was carried out through a 10-fold cross-validation procedure for each subject, and an average accuracy rate across all folds was obtained with \( K \) selected voxels (see Materials and Methods for more details).

(a) Gender decoding accuracy

We systematically varied \( K \), the number of selected voxels, from 25 to 1600 for decoding the gender categories and the results are shown in Figure 3A. For each run, the decoding accuracy rates of the gender categories with 1600 selected voxels are shown in Figure 3C,E. Here, as an example, we used 1600 selected voxels to present the detailed decoding results. From Figure 3A, we could obtain similar results for \( K \in [200, 1600] \). Furthermore, we compared the decoding accuracy rates of the task-relevant features with those of the task-irrelevant features in each of the audiovisual, visual-only, and auditory-only stimulus conditions. Specifically, we calculated the differences in the decoding accuracies of gender categories between the gender and the emotion judgment tasks under the audiovisual, visual-only, and auditory-only stimulus conditions, as shown in Figure 3G.
We next performed a statistical analysis on the decoding results obtained with $K = 1600$. For gender decoding accuracy rates (Fig. 3C,E), a 2-way repeated-measures ANOVA revealed that there were significant main effects of stimulus condition (the audiovisual, visual-only, and auditory-only conditions) and task (the gender and emotion judgment tasks) ($P < 10^{-7}$, $F_{2, n} = 35.85$ for the stimulus conditions; $P < 10^{-5}$, $F_{1, s} = 29.31$ for the judgment tasks), and a significant interaction effect.
between the 2 factors ($P<0.001$, $F_{2, s}=7.59$). Furthermore, Post hoc Bonferroni-corrected paired t-tests showed that the decoding accuracy rate for the task-relevant gender feature was significantly higher for the audiovisual stimulus condition than for the visual- or auditory-only stimulus condition ($P<0.005$ corrected, $t_{(8)}=3.99$ for audiovisual stimulus vs. visual-only stimulus; $P<0.005$ corrected, $t_{(8)}=3.96$ for audiovisual stimulus vs. auditory-only stimulus), and that there was no significant difference between the visual-only and the auditory-only stimulus condition (Fig. 3C). There were also no significant differences between each pair of the decoding accuracy rates of the task-irrelevant gender feature in the 3 stimulus conditions (Fig. 3E). Additionally, the decoding accuracy rates of the gender feature in the 3 stimulus conditions were significantly higher for the gender judgment task than for the emotion judgment task ($P<10^{-4}$ corrected, $t_{(8)}=7.04$ for audiovisual stimulus; $P<0.01$ corrected, $t_{(8)}=4.34$ for visual-only stimuli; $P<0.01$ corrected, $t_{(8)}=3.38$ for auditory-only stimuli). Bonferroni corrected for multiple comparison. Figure 3C,E). Furthermore, we compared the extent of the increases in the decoding accuracy rates of gender categories between the gender judgment and the emotion judgment tasks for the audiovisual, visual-only, and auditory-only stimulus conditions (Fig. 3G), and a 1-way repeated-measures ANOVA showed that there was a main effect of stimulus condition ($P<10^{-4}$, $F_{2, s}=12.88$). Post hoc Bonferroni-corrected paired t-tests showed that the increase in the gender decoding accuracy rate was significantly higher for the audiovisual stimulus condition than for the visual-only and auditory-only stimulus conditions ($P<0.01$ corrected, $t_{(8)}=3.66$ for audiovisual stimulus vs. visual-only stimulus; $P<0.005$ corrected, $t_{(8)}=4.0$ for audiovisual stimulus vs. auditory-only stimulus), and there was no significant difference between the visual-only and auditory-only conditions was observed.

(b) Emotion decoding accuracy

The results for decoding the emotion categories with $K$ varying from 25 to 1600 are shown in Figure 3B. For each run, the decoding accuracy rates of the emotion categories with 1600 selected voxels are shown in Figure 3D,F. We calculated the differences in the decoding accuracies of emotion categories between the gender and the emotion judgment tasks under the audiovisual, visual-only, and auditory-only stimulus conditions, which are shown in Figure 3H.

A 2-way repeated-measures ANOVA showed that there were significant main effects of stimulus condition and task ($P<10^{-6}$, $F_{2, s}=16.58$ for stimulus conditions; $P<10^{-5}$, $F_{1, s}=64.18$ for judgment tasks) and a significant interaction effect between stimulus condition and task ($P<10^{-6}$, $F_{2, s}=15.74$). Pairwise multiple comparisons showed that the emotion decoding accuracy rate was significantly higher for the audiovisual stimulus condition than for the visual-only and auditory-only stimulus conditions ($P<0.001$ corrected, $t_{(8)}=4.84$ for audiovisual stimulus vs. visual-only stimulus; $P<0.005$ corrected, $t_{(8)}=4.45$ for audiovisual stimulus vs. auditory-only stimulus), and there was no significant difference between the visual-only and auditory-only conditions (Fig. 3D). There were also no significant differences between any pair of the 3 decoding accuracy rates of the task-irrelevant emotion feature (Fig. 3F). Furthermore, the decoding accuracy rates of the emotion feature in the 3 stimulus conditions were significantly higher for the emotion judgment task than for the gender judgment task ($P<10^{-6}$ corrected, $t_{(8)}=15.82$ for audiovisual stimulus; $P<10^{-4}$ corrected, $t_{(8)}=8.55$ for visual stimulus; $P<0.01$ corrected, $t_{(8)}=4.34$ for auditory stimulus; Fig. 3D,F). We compared the extent of the increases in the decoding accuracy rates of emotion categories between the emotion judgment task and the gender judgment in the audiovisual, visual-only, and auditory-only stimulus conditions (Fig. 3H), and a 1-way repeated-measures ANOVA showed that there was a main effect of stimulus condition ($P<10^{-4}$, $F_{2, s}=14.35$). Additionally, Post hoc Bonferroni-corrected paired t-tests showed that the increase in emotion decoding accuracy rate was significantly higher for the audiovisual stimulus condition than for the visual-only and auditory-only stimulus conditions ($P<0.005$ corrected, $t_{(8)}=3.95$ for audiovisual stimulus vs. visual-only stimulus; $P<0.01$ corrected, $t_{(8)}=3.48$ for audiovisual stimulus vs. auditory-only stimulus), and no significant difference between the visual-only and auditory-only stimulus conditions was observed.

(c) Reproducibility

The average within-class and between-class reproducibility indices for the gender and the emotion judgment tasks are shown in Figure 4. For the average within-class reproducibility indices, a 2-way repeated-measures ANOVA revealed that there were significant main effects of stimulus condition (the audiovisual, visual-only, and auditory-only conditions) and task (the gender and emotion judgment tasks) ($P<10^{-5}$, $F_{2, s}=15.17$ for the stimulus conditions; $P<0.005$, $F_{1, s}=6.74$ for the judgment tasks). Post hoc Bonferroni-corrected pairwise t-tests indicated that the average within-class reproducibility indices for both the gender and the emotion judgment tasks (Fig. 4A,B) were significantly higher for the audiovisual stimulus condition than for the visual- or auditory-only stimulus condition (audiovisual stimuli vs. visual-only stimuli: $P<0.01$ corrected, $t_{(8)}=3.67$ for the gender judgment task and $P<0.01$ corrected, $t_{(8)}=3.702$ for the emotion judgment task; audiovisual stimuli vs. auditory-only stimuli: $P<0.02$ corrected, $t_{(8)}=3.16$ for the gender judgment task and $P<0.02$ corrected, $t_{(8)}=3.08$ for the emotion judgment task). Furthermore, for the average between-class reproducibility indices, a 2-way repeated-measures ANOVA revealed that there were significant main effects of stimulus condition (the audiovisual, visual-only, and auditory-only conditions) and task (the gender and emotion judgment tasks) ($P<10^{-6}$, $F_{2, s}=28.52$ for the stimulus conditions; $P<10^{-6}$, $F_{1, s}=34.06$ for the judgment tasks). Post hoc Bonferroni-corrected pairwise t-tests indicated that the average between-class reproducibility indices for both the gender and the emotion judgment tasks (Fig. 4C,D) were significantly lower for the audiovisual stimulus condition than for the visual- or auditory-only stimulus condition (audiovisual stimuli vs. visual-only stimuli: $P<0.01$ corrected, $t_{(8)}=3.96$ for the gender judgment task and $P<0.01$ corrected, $t_{(8)}=3.84$ for the emotion judgment task; audiovisual stimuli vs. auditory-only stimuli: $P<0.01$ corrected, $t_{(8)}=3.48$ for the gender judgment task and $P<0.01$ corrected, $t_{(8)}=3.44$ for the emotion judgment task).

**Distribution of Informative Voxels for Gender/Emotion Category Discrimination**

For the gender/emotion judgment task in the audiovisual stimulus condition, we searched for informative voxels...
involved in decoding the 2-gender/emotion categories using a permutation test at the group level (see Methods). We found that these informative voxels were distributed in many brain areas, as shown in Table 1 (for gender category decoding) and Table 2 (for emotion category decoding). In the audiovisual stimulus condition, many common brain areas, including the right/left STG, the right MTG, the right/left parahippocampal gyrus, the right precuneus, and the right/left medial frontal gyrus (MFG), participated in the gender and the emotion category decodings. Conversely, several brain areas (e.g., the left precentral gyrus) were only involved in gender category decoding, whereas several other brain areas (e.g., the amygdala) were exclusively involved in emotional category decoding (Tables 1 and 2).

### Table 1

<table>
<thead>
<tr>
<th>Brain region</th>
<th>Tal coordinates</th>
<th>Maximum weight</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left precentral gyrus</td>
<td>−37 19 12</td>
<td>0.0704</td>
<td>36</td>
</tr>
<tr>
<td>Left fusiform gyrus</td>
<td>−22 −85 −12</td>
<td>0.0711</td>
<td>8</td>
</tr>
<tr>
<td>Left parahippocampal gyrus</td>
<td>−28 −24 −11</td>
<td>0.1386</td>
<td>27</td>
</tr>
<tr>
<td>Right parahippocampal gyrus</td>
<td>24 −16 −10</td>
<td>0.0722</td>
<td>22</td>
</tr>
<tr>
<td>Right precuneus</td>
<td>22 −78 33</td>
<td>0.0707</td>
<td>35</td>
</tr>
<tr>
<td>Left lingual gyrus</td>
<td>−14 −90 −15</td>
<td>0.0702</td>
<td>21</td>
</tr>
<tr>
<td>Right insula</td>
<td>40 −5 −1</td>
<td>0.0734</td>
<td>31</td>
</tr>
<tr>
<td>Right putamen</td>
<td>18 10 −10</td>
<td>0.0821</td>
<td>17</td>
</tr>
<tr>
<td>Left anterior cingulate</td>
<td>−12 30 22</td>
<td>0.0828</td>
<td>22</td>
</tr>
<tr>
<td>Left posterior cingulate</td>
<td>−1 −43 17</td>
<td>0.0769</td>
<td>14</td>
</tr>
<tr>
<td>Right superior parietal lobule</td>
<td>31 −68 43</td>
<td>0.0717</td>
<td>24</td>
</tr>
<tr>
<td>Right inferior parietal lobule</td>
<td>48 −60 41</td>
<td>0.0734</td>
<td>23</td>
</tr>
<tr>
<td>Right middle frontal gyrus</td>
<td>30 50 −14</td>
<td>0.0888</td>
<td>39</td>
</tr>
<tr>
<td>Left medial frontal gyrus</td>
<td>−31 36 −14</td>
<td>0.0761</td>
<td>24</td>
</tr>
<tr>
<td>Left superior temporal gyrus</td>
<td>−42 −6 −17</td>
<td>0.0814</td>
<td>22</td>
</tr>
<tr>
<td>Right superior temporal gyrus</td>
<td>46 −20 4</td>
<td>0.0748</td>
<td>34</td>
</tr>
<tr>
<td>Left middle temporal gyrus</td>
<td>−57 −67 6</td>
<td>0.0979</td>
<td>26</td>
</tr>
<tr>
<td>Right middle temporal gyrus</td>
<td>57 −28 −5</td>
<td>0.0813</td>
<td>19</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Brain region</th>
<th>Tal coordinates</th>
<th>Maximum weight</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right cuneus</td>
<td>12 −85 14</td>
<td>0.0751</td>
<td>32</td>
</tr>
<tr>
<td>Right precuneus</td>
<td>6 −56 62</td>
<td>0.0956</td>
<td>24</td>
</tr>
<tr>
<td>Left medial frontal gyrus</td>
<td>−4 −27 43</td>
<td>0.0762</td>
<td>29</td>
</tr>
<tr>
<td>Right medial frontal gyrus</td>
<td>2 64 14</td>
<td>0.0784</td>
<td>22</td>
</tr>
<tr>
<td>Right middle temporal gyrus</td>
<td>60 −21 −10</td>
<td>0.0979</td>
<td>34</td>
</tr>
<tr>
<td>Left superior temporal gyrus</td>
<td>−36 3 −20</td>
<td>0.0741</td>
<td>38</td>
</tr>
<tr>
<td>Right superior temporal gyrus</td>
<td>38 5 −18</td>
<td>0.0734</td>
<td>22</td>
</tr>
<tr>
<td>Right fusiform gyrus</td>
<td>28 −37 −17</td>
<td>0.0825</td>
<td>28</td>
</tr>
<tr>
<td>Left parahippocampal gyrus</td>
<td>−20 −2 −11</td>
<td>0.0849</td>
<td>43</td>
</tr>
<tr>
<td>Right parahippocampal gyrus</td>
<td>16 −30 −11</td>
<td>0.0861</td>
<td>24</td>
</tr>
<tr>
<td>Left amygdala</td>
<td>−26 −10 −7</td>
<td>0.0731</td>
<td>20</td>
</tr>
<tr>
<td>Right amygdala</td>
<td>27 2 −19</td>
<td>0.0760</td>
<td>28</td>
</tr>
<tr>
<td>Left lentiform nucleus</td>
<td>−28 −13 −5</td>
<td>0.0751</td>
<td>30</td>
</tr>
<tr>
<td>Left insula</td>
<td>40 −14 7</td>
<td>0.0813</td>
<td>22</td>
</tr>
<tr>
<td>Left putamen</td>
<td>−25 −10 −7</td>
<td>0.0757</td>
<td>19</td>
</tr>
<tr>
<td>Left anterior cingulate</td>
<td>−6 42 12</td>
<td>0.0803</td>
<td>22</td>
</tr>
<tr>
<td>Right inferior parietal lobule</td>
<td>37 −45 50</td>
<td>0.0786</td>
<td>24</td>
</tr>
<tr>
<td>Left superior parietal lobule</td>
<td>−10 −59 64</td>
<td>0.0828</td>
<td>30</td>
</tr>
</tbody>
</table>
Multimodal Audiovisual Integration

The pSTS/MTG is an important brain area associated with the audiovisual integration. Through group GLM analysis (see Materials and Methods section), we determined that several brain areas, including the bilateral pSTS/MTG and the right middle frontal gyrus, satisfied the following criterion for judging multimodal audiovisual integration: \[A > 0 \text{ or } V > 0 \quad (P < 0.05, \text{ FWE-corrected}) \cap [A V > \max (A, V) \quad (P < 0.05, \text{ uncorrected})].\] The distribution of these brain areas are shown in Figure 5A for the gender judgment task and in Figure 5B for the emotion judgment task. The subject-average percent signal changes in the audiovisual, visual-only, and auditory-only stimulus conditions for the bilateral pSTS/MTG activation clusters are shown in Figure 6A for the gender judgment task and in Figure 6B for the emotion judgment task. It follows from Figure 6A, B that for both judgment tasks, the average percent signal changes are significantly higher for the audiovisual stimulus condition than for the visual-only and auditory-only stimulus conditions (\(P < 10^{-4}\) corrected, \(t_{(8)} = 10.9521\) for gender judgment task; \(P < 10^{-5}\) corrected, \(t_{(8)} = 12.4310\) for emotion judgment task).

Discussion

In the present study, we explored the effects of crossmodal integration on the neural representations of features of audiovisual faces in the human brain. During the fMRI experiment, the subjects were instructed to judge the gender or emotion of a series of facial movie clips under the audiovisual, visual-only or auditory-only stimulus conditions. The neural representation of a feature was assessed by the category decoding accuracy, as well as the within-class reproducibility and the between-class reproducibility, which were obtained by an MVPA method. Compared with the visual-only and auditory-only stimulus conditions, the audiovisual stimulus condition lead to higher category decoding accuracy and better within-class reproducibility for both gender and emotion judgments. This indicates that crossmodal integration enhances the neural representation of audiovisual faces.
conditions, we showed that both the category decoding rates of task-relevant features and the within-class reproducibility indices were significantly higher, while the between-class reproducibility indices were significantly lower for the audiovisual stimulus condition (Figs 3 and 4). In comparison, similar results were not observed for the task-irrelevant features (Fig. 3). Thus we may conclude that crossmodal integration enhances the neural representations of task-relevant features.

Previous behavioral studies have demonstrated that multisensory integration may facilitate perception and recognition (Calvert and Thesen 2004). Meanwhile, the neural mechanisms of audiovisual integration have been explored using neuroimaging techniques and several brain regions, including the pSTS/MTG, were identified as heteromodal sensory areas (Calvert et al. 2000; Frassinetti et al. 2002; Bushara et al. 2003). Particularly, increased neural activity was observed in the pSTS/MTG when the audiovisual stimulus condition was compared with the visual-only and auditory-only stimulus conditions, and this increase in neural activity has often been referred to as a superadditive effect. In our experiment, we also observed such an increase of neural activity in the pSTS/MTG (Figs 5 and 6), which may be served as an indication of occurrence of cross-modal integration (e.g., Calvert et al. 2000; Frassinetti et al. 2002; Bushara et al. 2003; Calvert and Thesen 2004; Macaluso and Driver 2005). Our new observation was that the neural representation of the task-relevant emotion or gender feature was enhanced by the crossmodal integration in audiovisual face perception. By subtracting the decoding accuracy for the task-irrelevant feature as a baseline, we found that the increase in decoding accuracy rate for the audiovisual condition was significantly larger than the increases in decoding accuracy rates for the visual-only and auditory-only conditions (Fig. 3G, H). Furthermore, the increased within-class reproducibility indices for the task-relevant features in the audiovisual stimulus condition implied the higher similarity of the neural activity patterns within a class, while the decreased between-class reproducibility indices implied the larger difference between the 2 classes of the neural activity patterns (Fig. 4). As a result, the decoding performance was improved, indicating increased amount of category information carried by these neural activity patterns.

Numerous studies have addressed the issue of how attentional shifts in one modality can affect orienting in other modalities (Spence and Driver 2004); however, the role that attention plays during multisensory integration itself has been largely overlooked (Talsma et al. 2010). We considered feature-selective attention occurring in audiovisual face perception. By subtracting the decoding accuracy for the task-irrelevant feature as a baseline, we found that the increase in decoding accuracy rate for the audiovisual condition was significantly larger than the increases in decoding accuracy rates for the visual-only and auditory-only conditions (Fig. 3G, H). The fact that attention can enhance neural representation/encoding has been demonstrated in several recent studies (Xu 2010; Jehee et al. 2011); however, our results showed that the degree of improvement in decoding accuracy in the audiovisual stimulus condition was significantly larger than that in the visual-only and auditory-only stimulus conditions (Fig. 3G,H), a finding that was related to crossmodal integration. This enhancement in the audiovisual stimulus condition may be explained under the framework for interactions between attention and crossmodal integration as proposed in (Talsma et al. 2010). On one hand, crossmodal integration has stimulus-driven influences on attention; specifically, the congruent multisensory stimuli tend to capture attention and processing resources, thereby enhancing the ability for attentional selection. On the other hand, top-down selective attention can modulate and facilitate crossmodal integration as demonstrated in this study.

Using the data collected during the audiovisual stimulus condition, we localized the voxels which were informative for the gender and emotion category decodings separately and found that they were distributed across different brain areas (Tables 1 and 2). As shown in Tables 1 and 2, several brain regions, including the right FG, the right/left STG, the right MTG, the right precuneus, and the right/left MFG, were involved in both the gender and the emotion category decodings. Our results are partially consistent with the existing evidences related to face information processing (Haxby et al. 1996; Leveroni et al. 2000; Gobbini and Haxby 2006). For example, it was proposed in (Haxby et al. 2000) that the inferior occipital gyrus contributes to the early stage of face information processing, whereby, information is further transferred to the STS and the FG, where different aspects of faces are processed separately. Invariant aspects of a face, such as identity, gender, and race are processed primarily in the FG region (e.g., Sergent et al. 1992; Golby et al. 2001; Freeman et al. 2010); whereas changeable aspects, such as emotional expression, gaze, and lip-speech, depend on the STS region (e.g., Tanel et al. 1988; Calvert et al. 1997; Puce et al. 1998; Hoffman and Haxby 2000).

We also found that several brain regions were involved only in either the gender or the emotion category decoding. For example, the amygdala only participated in the emotion category decoding and the precentral gyrus was only involved in the gender category decoding. We noticed that the amygdala and precentral gyrus may be engaged in both emotion and gender information processing as shown by previous studies. It was believed that amygdala played a key role in emotion processing (Vuilleumier et al. 2001; Pessoa et al. 2002; Stein et al. 2007). However, there is also evidence that the amygdala was activated in gender discrimination task (Morris et al. 1998; Killgore and Yurgelun-Todd 2004). Likewise, the precentral gyrus was found to be activated in the processing of emotional faces (e.g., disgusted, fearful, and angry faces) (LaBar et al. 1998; Iidaka et al. 2001; Phillips et al. 2004; Fusar-Poli et al. 2009), but also in the processing of gender information (Critchley et al. 2000) and neutral faces (Fusar-Poli et al. 2009). The clear division of the roles between these 2 regions found in our study might be attributed to the special experimental design and the way how the fMRI data were analyzed. In our experiment, when the gender/emotion feature was attended, the other was always required to be suppressed or neglected. For fMRI data analysis, MVPA was used in the present study to find distributed brain patterns that served to discriminate different categories of the stimuli, while GLM was used in the previous studies which attempted to localize brain regions significantly activated by these stimuli. For example, even though the precentral gyrus was activated in the 2 categories of stimuli (crying and laughing) in the emotion judgment task, the
magnitudes of the activations under the 2 categories stimuli might be so close that the region barely contributed any useful discriminative information.

In addition to the brain network evoked by the facial information processing, we also found other brain areas, for example, the parahippocampal gyrus and pSTS/MTG, that may play essential roles in audiovisual information integration. Taylor et al. showed that the pSTS/MTG functions as a presemantic, heteromodal sensory area, while the perirhinal cortex plays a critical role in binding the meaningful aspects of audiovisual objects in crossmodal integration (Taylor et al. 2006). There also exist other brain areas, such as the hippocampal formation, including the hippocampus, entorhinal, perirhinal, and parahippocampus that support high-level integration of semantic information (Lavenez and Amaral 2000). Finally, several attention-related brain areas were selected for both the gender and the emotion judgment tasks, for example, the left/right putamen, cingulate gyrus and left/right superior parietal lobule, which might play a role of modulation during the audiovisual integration (Hopfinger et al. 2000; Nobre et al. 2006).

In summary, the present study revealed that audiovisual integration enhanced the neural representation of task-relevant features in audiovisual dynamic face perception and that the feature-selective attention might modulate this crossmodal integration. It should be noted that our findings were limited to audiovisual dynamic face perception. Future experiments using other types of stimuli are needed to further demonstrate this enhancement effect of crossmodal integration on the neural representations of features, and to further clarify the relationship between feature-selective attention and crossmodal integration.

Funding

This work was supported by the National High-tech R&D Program of China (863 Program) under grant 2012AA011601, the National Natural Science Foundation of China under grants 91120305 and 81271560, and the High-Level Talent Project of Guangdong Province, China.

Notes

Conflict of Interest: None declared.

References


