www.nature.com/scientificreports

SCIENTIFIC REPORTS

Selective Audiovisual Semantic Integration Enabled by Feature-Selective Attention

Yuanqing Li[1,5], Jinyi Long[1,5], Biao Huang[2], Tianyou Yu[1,5], Wei Wu[1,5], Peijun Li[2], Fang Fang[3] & Pei Sun[4]

An audiovisual object may contain multiple semantic features, such as the gender and emotional features of the speaker. Feature-selective attention and audiovisual semantic integration are two brain functions involved in the recognition of audiovisual objects. Humans often selectively attend to one or several features while ignoring the other features of an audiovisual object. Meanwhile, the human brain integrates semantic information from the visual and auditory modalities. However, how these two brain functions correlate with each other remains to be elucidated. In this functional magnetic resonance imaging (fMRI) study, we explored the neural mechanism by which feature-selective attention modulates audiovisual semantic integration. During the fMRI experiment, the subjects were presented with visual-only, auditory-only, or audiovisual dynamical facial stimuli and performed several feature-selective attention tasks. Our results revealed that a distribution of areas, including heteromodal areas and brain areas encoding attended features, may be involved in audiovisual semantic integration. Through feature-selective attention, the human brain may selectively integrate audiovisual semantic information from attended features by enhancing functional connectivity and thus regulating information flows from heteromodal areas to brain areas encoding the attended features.

An audiovisual object in the real world may contain multiple semantic features, such as the gender and emotional features of a speaker's face and voice. During the recognition of an audiovisual object, the human brain integrates the semantic information from these features obtained by the visual and the auditory modalities, i.e., audiovisual semantic integration may occur in the brain. Audiovisual integration facilitates rapid, robust, and automatic object perception and recognition[1–3]. Comparisons of visual-only and auditory-only stimuli have revealed that congruent audiovisual stimuli lead to stronger neural responses than either type of stimulus alone in the posterio a specific network in which the parietal and perhaps lateral frontal cortices appear to be optimally situated to mediate the integration and attentional selection of motion information across modalities[12]. In audiovisual face perception, cross-modal attention influences cross-modal binding during speech reading[13,14]. [illegible], attention and audiovisual integration interact with each other in a sophisticated manner. However, feature-selective attention in audiovisual conditions and the relationship between feature-selective attention and high-level audiovisual semantic integration remain to be explored.

In a single (visual or auditory) modality, feature-selective attention may lead to selective processing of the attended features of an object in the brain[7–9,15–17]. Nobre *et al.*[8] demonstrated that ERPs are modulated by feature-selective attention and that irrelevant features are inhibited during the early stages of perceptual analysis in humans. In monkeys, Mirabella *et al.*[17] observed that neurons in visual area V4 exhibit selectivity to elemental object features. Based on these studies that have employed unimodal stimuli, here we explore the hypothesis that a similar feature-selective attention mechanism in an audiovisual condition is involved in audiovisual semantic integration.

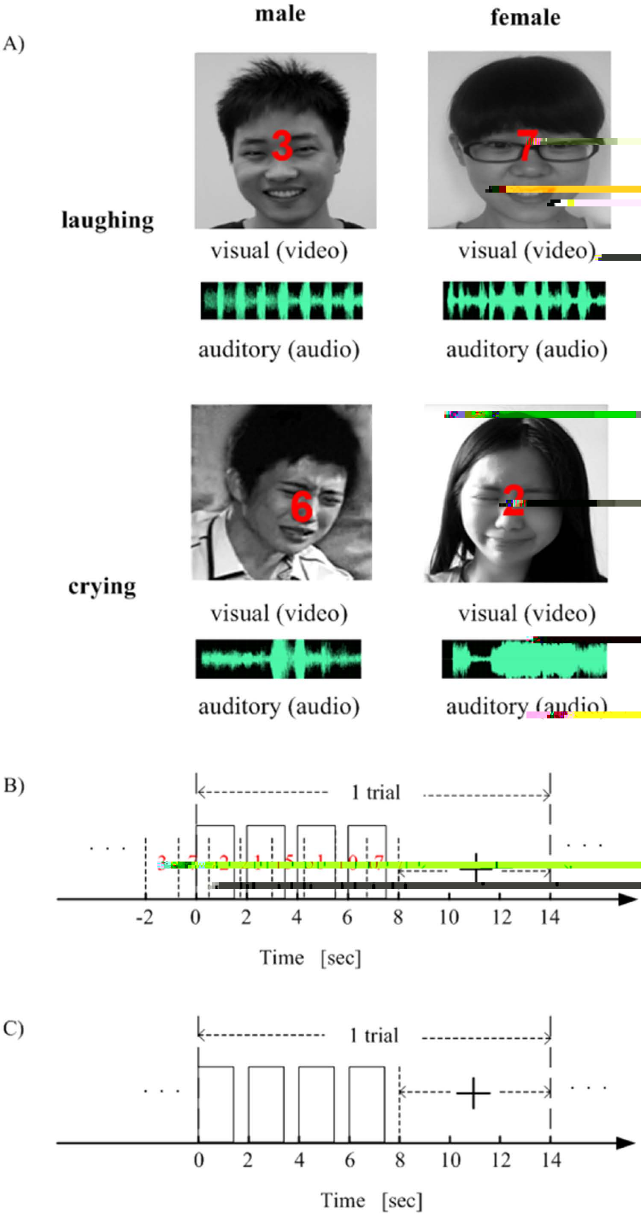

Figure 1. Experimental stimuli and time courses. (**A**) Four examples of audiovisual stimuli; the red numbers indicate runs with the number task only. (**B**) Time course of a trial for the runs with the number task, in which the stimuli included randomly presented numbers and video/audio/movie clips. (**C**) Time course of a trial for the runs with the gender, emotion, or bi-feature task. For both (**B,C**), the presentation of a stimulus (video/audio/movie clip) lasted 1,400 ms and was repeated four times during the first eight seconds in a trial. A visual cue (+) appeared at the 8th second and persisted for six seconds.

For each of the three runs with the number task, in addition to the corresponding audiovisual, visual-only, or auditory-only facial stimuli from the movie clips, numbers in red appeared sequentially at the center of the screen (see Fig. 1A). The subjects' task was to attend to the numbers instead of the other stimuli (see Table 1). We designed a difficult number task for the subjects in which they were asked to find and count the repeated numbers to ensure that they fully ignored the features of the visual-only, auditory-only, or audiovisual facial stimuli. Therefore, the subjects performed this task with low accuracy, as shown in Fig. S3. At the beginning of each block, there were four seconds before the first trial, and a short instruction in Chinese (see Table 1) was displayed on the screen in the first two seconds (the last two seconds were used to display numbers, as indicated below). At the beginning of each trial, a visual-only, auditory-only or audiovisual facial stimulus was presented to the subjects for 1,400 ms, followed by a 600-ms blank period. This two-second cycle with the same stimulus was repeated four times, followed by a six-second blank period. Therefore, one trial lasted 14 seconds. In addition to the above stimuli, eight numbers in red appeared one by one at the center of the screen, each a random integer from 0 to 9. Each number lasted 900 ms, and the interval between two subsequent numbers was 350 ms. Two numbers appeared 2 seconds before the beginning of this trial. The subjects were asked to find and count the repeated numbers. After the stimulation, a fixation cross appeared on the screen. The subjects then responded by pressing the right-hand key according to the instruction for this block (see Table 1). The fixation cross changed color at the 12th second, indicating that the next trial would begin shortly (see Fig. 1B). In total, a run lasted 1,350 seconds.

The procedure for the three runs with the gender/emotion task was similar to that for the runs with the number task, except that no numbers appeared on the screen and the subjects performed a gender/emotion judgment task (See Table 1). Specifically, the subjects were asked to focus their attention on either the gender or the emotion of the
he-0.001 nn-he cf27in ıoai6(l ci i l)-3.8(e i c)-3(l-)-5.1(io)-3.1(a)72(l)6.9( ).8(ei d1.8(n ei)6n f)8.8(oe)0.2(-(e)-5(c-3.

voxels, time series detrending, and normalization of the time series in each block to zero mean and unit variance. All preprocessing steps were performed using SPM8[23] and custom functions in MATLAB 7.4 (MathWorks, Natick, Massachusetts, USA).

Univariate GLM analysis. This experiment included four experimental tasks (number, gender, emotion, and bi-feature). For each experimental task, three runs corresponding to the visual-only, the auditory-only, and the audiovisual stimulus conditions were performed. To confirm that audiovisual sensory integration occurred for each experimental task and determine the heteromodal areas associated with audiovisual integration, we performed voxel-wise group analysis of the fMRI data based on a mixed-effect two-level GLM in SPM8. In particular, using the data from the three number runs, we performed GLM analysis to explore the audiovisual integration at the sensory level when the subjects fully ignored the visual-only, auditory-only, or audiovisual facial stimuli while only attending to the numbers. The GLM analysis included the following data processing. The fMRI data for each subject were subjected to a first-level GLM, and the estimated beta coefficients across all subjects were then combined and analyzed using a second-level GLM. The following statistical criterion was used to determine brain areas for audiovisual sensory integration: [AV $>$ max (A,V) ($p < 0.05$, FWE-corrected)] $\cap$ [V $>$ 0 or A $>$ 0 ($p < 0.05$, uncorrected)][1,4,6,24–27], where $\cap$ denotes the intersection of two sets. For each subject, each task, and each stimulus condition, we also computed the percent signal changes of the pSTS/MTG clusters via region-of-interest (ROI)-based analysis (implemented by the MATLAB toolbox MarsBaR-0.43[28]). Specifically, we identified the clusters consisting of significantly activated voxels in the bilateral pSTS/MTG via group GLM analysis as above. First, a GLM model was estimated from the mean BOLD signal of the clusters, and the percent signal change in the clusters was then computed as the ratio between the maximum of the estimated event response and the baseline.

MVPA procedure for the calculation of the reproducibility ratio and decoding accuracy. For each subject, there were a total of 12 runs with four experimental tasks and three stimulus conditions. For each run, we calculated a reproducibility ratio corresponding to the gender feature and one corresponding to the emotion feature by applying an MVPA method to the fMRI data. The reproducibility ratio is an index that measures the similarity of the neural activity patterns within a class (e.g., the male class in the gender dimension) and the difference in neural activity patterns between two classes (e.g., male vs. female in the gender dimension). The higher the reproducibility ratio, the stronger the similarity of brain patterns within each class and the larger the difference between the two classes of brain patterns associated with the two gender or two emotion categories. Using the fMRI data, we also decoded the gender and emotion categories of the stimuli perceived by the subjects. The neural representations of gender and emotion features were analyzed by comparing the reproducibility ratios or decoding accuracy rates for different stimulus conditions (visual-only, auditory-only, and audiovisual) and experimental tasks (number, gender, emotion, and bi-feature). In particular, the subjects only attended to the number during the three number runs, but the MVPA was based on the gender and emotion features of the visual-only, auditory-only, or audiovisual facial stimuli. In this manner, we analyzed the neural representations of gender and emotion features when none was attended. Below, we explain the MVPA procedure for gender categories (the MVPA procedure for emotion categories was similar).

For each run, 10-fold cross-validation was performed for the calculation of the reproducibility ratio and decoding accuracy corresponding to the two gender categories (refer to Fig. S1 in Supplemental Information). Specifically, the data from 80 trials were equally partitioned into 10 non-overlapping datasets. For the *k*th fold of the cross-validation ($k = 1, \ldots, 10$), the *k*th dataset (eight trials) was used for the test, and the other nine datasets (72 trials) were used for voxel selection and classifier training. After the 10-fold cross-validation, the average reproducibility ratio and decoding accuracy rate were calculated across all folds. The data processing procedure for the *k*th fold included the following:

1) Voxel selection based on the training data. A spherical searchlight algorithm that was sequentially centered at each voxel with a 3-mm radius searchlight highlighting 19 voxels was applied to the training dataset for voxel selection[29]. Within each searchlight corresponding to a voxel, we computed a Fisher ratio through Fisher linear discriminant analysis, and this ratio indicated the level of discrimination between the two gender categories in the local neighborhood of this voxel. A Fisher ratio map was thus obtained for the whole brain. *K* informative voxels with the highest Fisher ratios were then selected (e.g., $K = 1500$ in this study).

2) Pattern extraction. Using the *K* selected voxels, we constructed a *K*-dimensional pattern vector for each trial of the training data in which each element represented the mean BOLD response of a selected voxel from the 6th to the 14th second of this trial (the last four volumes, to account for the delay in the hemodynamic response; each trial lasted 14 he e $\theta_{i,j}$ i he angle be een o pa e n ec o, P_i

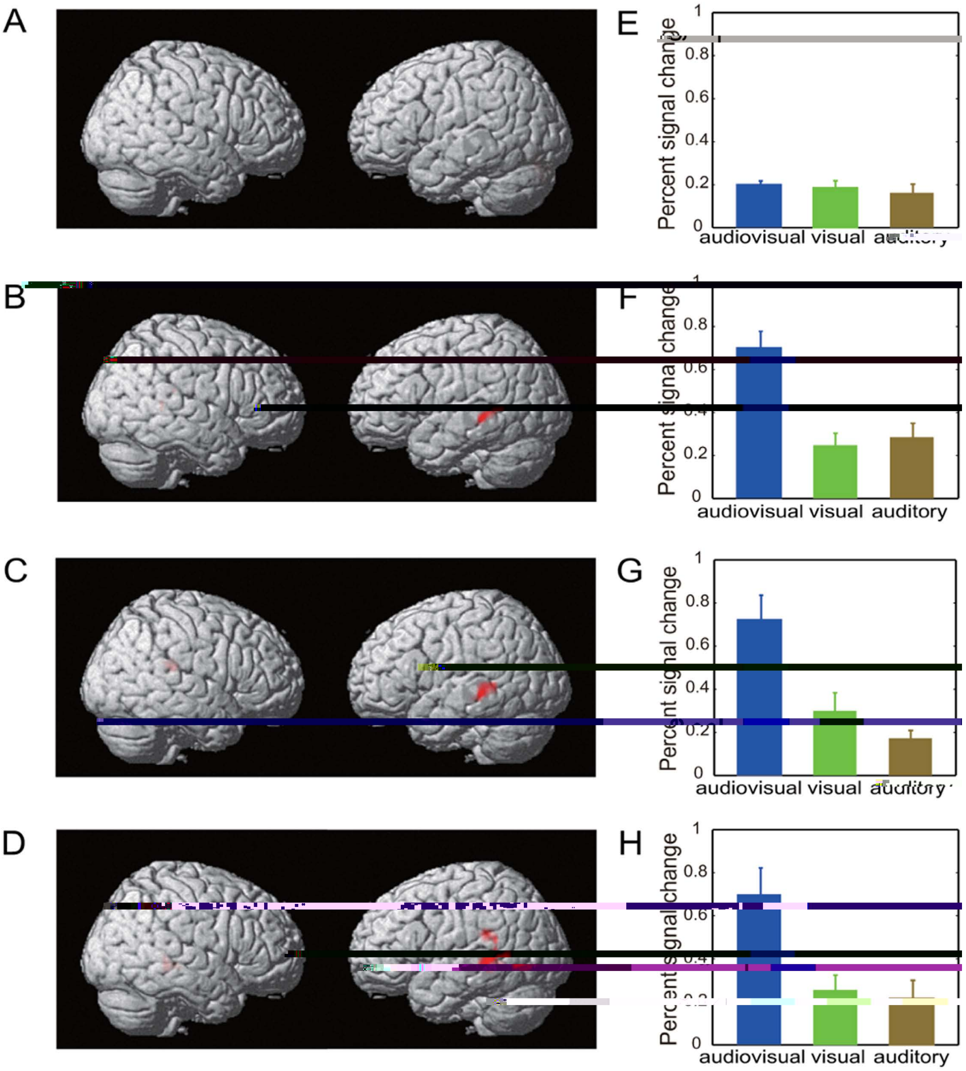

Figure 2. Brain areas for audiovisual sensory integration that met the criterion [AV>max (A,V) ($p < 0.05$, FWE-corrected)]∩[V>0 or A>0 ($p < 0.05$, uncorrected)]. (**A**) No brain areas exhibited audiovisual sensory integration for the number task. (**B**) Brain areas exhibiting audiovisual sensory integration for the gender task, including the left pSTS/MTG (Talairach coordinates of the cluster center: (−57, −34, −5); cluster size: 76). (**C**) Brain areas exhibiting audiovisual sensory integration for the emotion task, including the left pSTS/MTG (cluster center: (−60, −40, 1); cluster size: 98) and the right pSTS/MTG (cluster center: (45, −34, 19); cluster size: 13). (**D**) Brain areas exhibiting audiovisual sensory integration for the bi-feature task, including the left pSTS/MTG (cluster center: (−54, −

differentiated for different experimental tasks or different semantic features. [illegible], audiovisual sensory integration rather than audiovisual semantic integration occurred in the identified heteromodal areas of the pSTS/MTG, consistent with previous results[10].

MVPA results of the reproducibility ratios and decoding accuracy rates. Using an MVPA method, for each of the 12 runs of the experiment with four attentional tasks and three stimulus conditions, we calculated the reproducibility ratios corresponding to the gender categories (males vs. females) and the emotion categories (crying vs. laughing) of the stimuli respectively. Furthermore, each calculation of reproducibility ratio was based on 1500 selected voxels (see Materials and Methods); the results of reproducibility ratios are shown in Fig. 3. We also systematically varied the number of selected voxels from 25 to 1500 to calculate the reproducibility ratios and obtained similar results (see Fig. S4).

For the reproducibility ratios of the gender/emotion categories, two-way repeated measures ANOVAs revealed significant main effects of stimulus condition (gender categories: $p < 10^{-17}$, $F(2, 8) = 88.73$; emotion categories: $p < 10^{-16}$, $F(2, 8) = 51.37$) and experimental task (gender categories: $p < 10^{-17}$, $F(3, 8) = 81.13$; emotion categories:

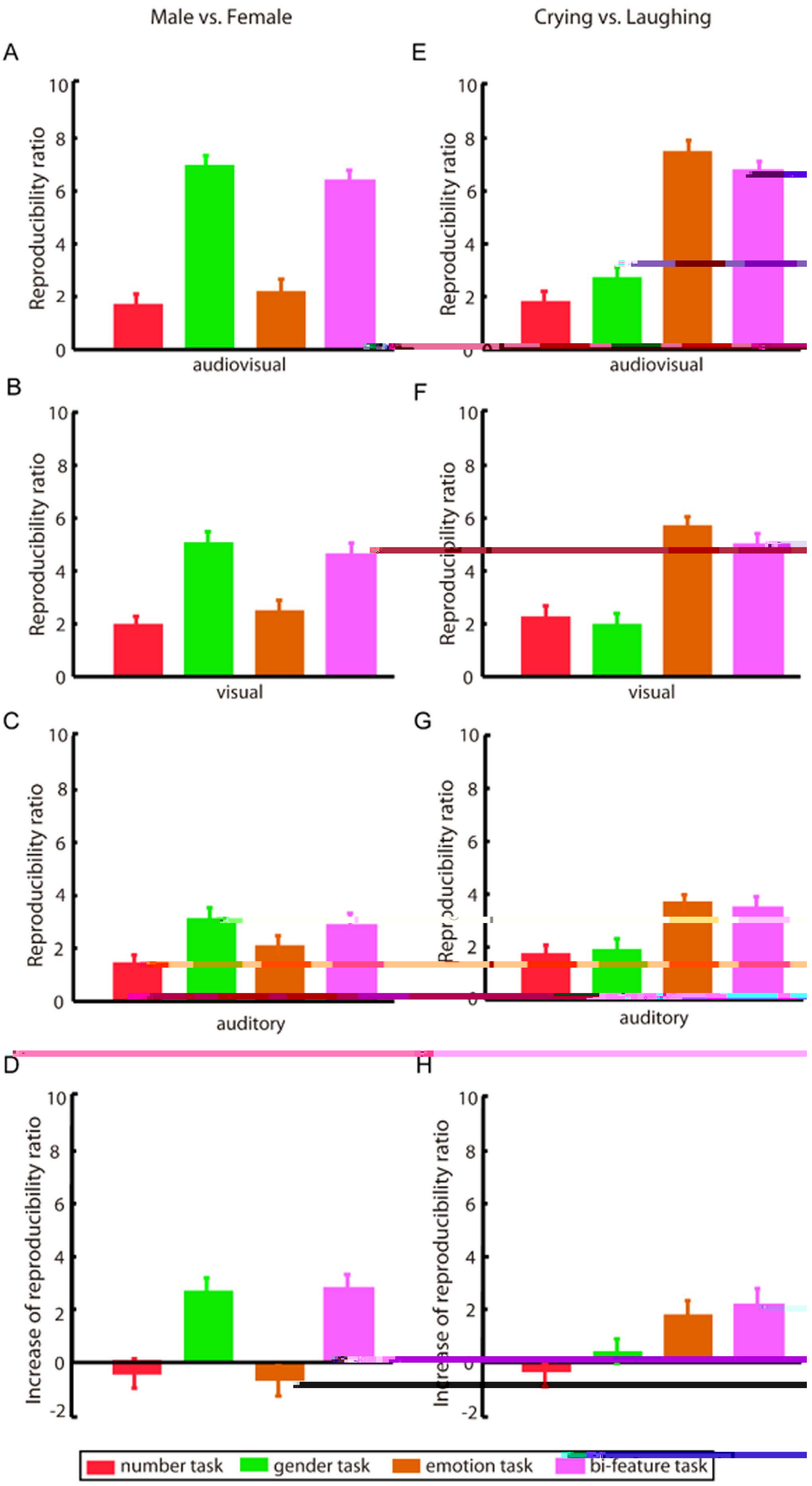

Figure 3. Reproducibility ratios (means and standard errors across all subjects) and the corresponding comparison results. Left/Right: gender/emotion categories; the first 3 rows: audiovisual, visual-only, and auditory-only stimuli conditions, respectively; the 4th row: the reproducibility ratio in the audiovisual condition minus the maximum of the reproducibility ratios in the visual-only and auditory-only conditions.

$p < 10^{-17}$, $F(3, 8) = 68.26$) (Fig. 3A–C,E–G). There was also a significant interaction effect between the two factors of stimulus conditions and experimental tasks (gender categories: $p < 10^{-17}$, $F(6, 8) = 30.07$; emotion categories: $p < 10^{-8}$, $F(6, 8) = 10.05$). Post hoc Bonferroni-corrected paired t-tests on the stimulus conditions revealed the following: (i) for each task-relevant feature (gender categories with the gender or the bi-feature task, left panel of Fig. 3; emotion categories with the emotion or the bi-feature task, right panel of Fig. 3), the reproducibility ratios were significantly higher for the audiovisual stimulus condition than for the visual-only or auditory-only stimulus condition (all $p < 0.001$ corrected); and (ii) for each task-irrelevant feature (gender categories with the number or the emotion task, left panel of Fig. 3; emotion categories with the number or the gender task, right panel of Fig. 3), there were no significant differences between the audiovisual and the visual-only or auditory-only stimulus condition (all $p > 0.05$). Furthermore, post hoc Bonferroni-corrected paired t-tests on the experimental tasks revealed that (i) in each of the audiovisual, visual-only and auditory-only stimulus conditions, the reproducibility ratios for gender/emotion categories were significantly higher for each relevant task (gender categories: the gender or the bi-feature task, left panel of Fig. 3; emotion categories: the emotion or the bi-feature task, right panel of Fig. 3) than for each irrelevant task (gender categories: the number or the emotion task, left panel of Fig. 3; emotion categories: the number or the gender task, right panel of Fig. 3) (all $p < 0.05$, corrected) and that (ii) in each of the audiovisual, visual-only and auditory-only stimulus conditions, there were no significant differences in the reproducibility ratios for gender/emotion categories between two relevant tasks or between two irrelevant tasks (all $p > 0.05$).

For each run of the experiment, we further calculated the decoding accuracies of the gender categories (males vs. females) and the emotion categories (crying vs. laughing) (see Materials and Methods), which are presented in Fig. S5. The decoding results also reveal the enhancement effect produced by the audiovisual stimuli only for task-relevant features (see Fig. S5).

When the brain is receiving both auditory and visual signals, more reproducible representations may be produced even if no audiovisual integration occurs. We thus conducted a control experiment that included an incongruent audiovisual run for the gender task and one for the emotion task. The experimental procedure for each run was similar to that of the congruent audiovisual run with gender/emotion task of the main experiment except that the audiovisual stimuli were incongruent in the gender or emotion dimensions. The experimental results demonstrated that compared with the visual-only and auditory-only stimulus conditions, the incongruent audiovisual stimuli did not enhance the neural representation of the attended features (see the control experiment in the Supplemental Information for details).

MVPA results for informative voxels, cross-reproducibility ratios, and functional connectivity. By applying an MVPA method to the data collected in the audiovisual condition with bi-feature task, we obtained the informative voxels for gender/emotion category discrimination (see Materials and Methods). The distributions of these informative voxels are presented in Tables 2 and 3 for gender categories and emotion categories, respectively.

Based on the sets of the informative 3(d f(d)-23.7o)124

Brain region	Tal coordinates			max weight	Numbers of voxels in the clusters
	x	y	z		
Right Precuneus	12	−50	52	0.087	23
Left Middle Frontal Gyrus	−38	36	30	0.067	26
Right Middle Frontal Gyrus	40	27	43	0.084	32
Right Middle Temporal Gyrus	60	−21	−10		

38 36 22 21

($p < 10^{-9}$, F(2, 8) = 36.97 for gender categories; $p < 10^{-11}$, F(2, 8) = 46.13 for emotion categories). Furthermore, post hoc Bonferroni-corrected paired t-tests demonstrated that the cross-reproducibility ratios were significantly higher for the relevant tasks than for the irrelevant tasks (gender categories: $p < 0.001$ corrected, t(8) = 16.23 for gender task vs. number task; $p < 0.001$ corrected, t(8) = 15.49 for gender task vs. emotion task; emotion categories: $p < 0.001$ corrected, t(8) = 16.05 for emotion task vs. number task; $p < 0.001$ corrected, t(8) = 14.36 for emotion task vs. gender task) and that there was no significant difference between the number task and the irrelevant emotion/gender task (all $p > 0.05$) (Fig. 4). Based on the sets of informative voxels for the gender/emotion categories, we also performed gender category and emotion category decoding for each of the audiovisual stimuli within number, gender and emotion tasks; the corresponding cross-decoding accuracy rates are presented in Fig. S6. From Tables 2 and 3 and Figs 3 and S6, we can conclude the following: (i) the informative voxels in Table 2/Table 3 are involved in the processing of the gender/emotion features in the audiovisual conditions; (ii) the corresponding voxels in Table 2/Table 3 are informative only when the gender/emotion feature is attended.

For the purpose of functional connectivity calculation, we selected four voxel clusters, each with a size 62 from the heteromodal areas: left STS/MTG (cluster center: (−52 −22 8)), right STS/MTG (cluster center: (54 −18 9)), left perirhinal cortex (cluster center: (−26, −20, −22)), and right perirhinal cortex (cluster center: (26, −18, −22)), as described in the related references[10,32]. For each of the audiovisual stimuli within number, gender and emotion tasks, we calculated the functional connectivity with two directions between the heteromodal areas and the informative brain areas in Table 2 (for gender categories) or Table 3 (for emotion categories) via Granger causality analysis a

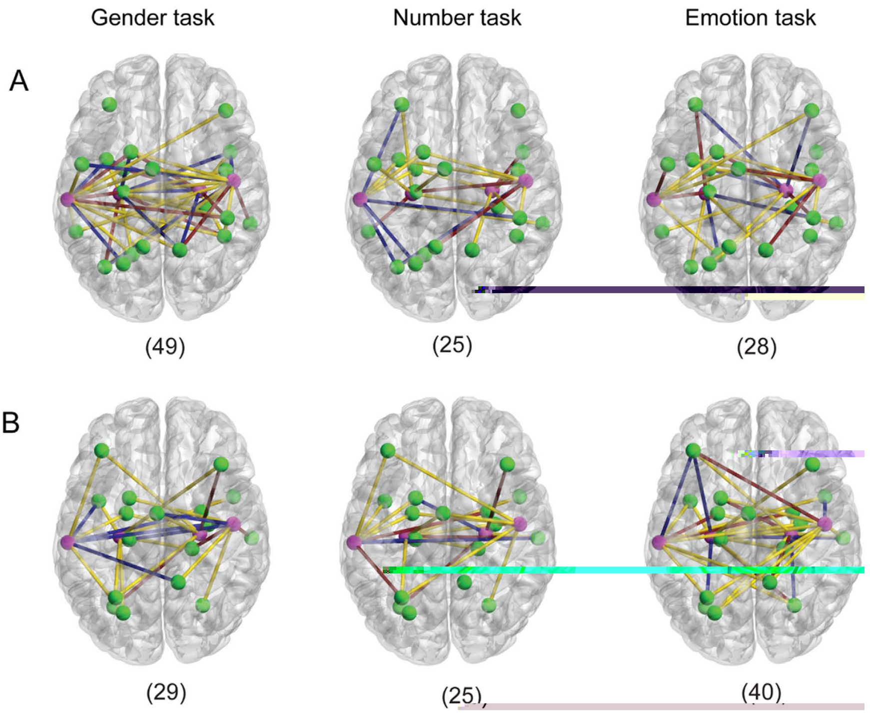

Figure 5. The functional connectivity between the heteromodal areas and the brain areas encoding the gender feature (A) or the emotion feature (B). Green spheres: brain areas from Table 2 in (**A**) or Table 3 in (**B**). Magenta spheres: heteromodal areas. Yellow lines: connections from the heteromodal areas to the informative brain areas. Blue lines: connections from the informative brain areas to the heteromodal areas. Purple lines: connections with bi-direction. Numbers in brackets: total numbers of functional connections.

the group level (see Materials and Methods). As shown in Fig. 5, there were more functional connections from the heteromodal areas to the brain areas encoding the gender/emotion feature (Table 2/Table 3) for the relevant task (gender/emotion task) than for the irrelevant tasks (number and emotion/gender tasks). We thus observed that in the audiovisual condition, feature-selective attention enhanced the functional connectivity and thus regulated the information flows from the heteromodal areas to the brain areas encoding the attended feature. Furthermore, this enhancement of the functional connectivity may imply that both the heteromodal areas and the brain areas encoding the attended feature are involved in audiovisual semantic integration.

Discussion.

In the present study, we explored the neural modulation of audiovisual semantic integration by feature-selective attention. During the fMRI experiments, the subjects were instructed to neglect all features, attend to a single feature (gender or emotion), or simultaneously attend to two features (both gender and emotion) of a series of facial movie clips in the visual-only, audio-only and audiovisual stimulus conditions. To assess the semantic information of a feature encoded in the brain, we calculated a reproducibility ratio for each feature, experimental task and stimulus condition by applying an MVPA method to the fMRI data, and we further analysed the functional connectivity between the brain areas encoding the semantic feature and the heteromodal areas. Our results suggested that in the audiovisual condition, feature-selective attention may function as a prerequisite for the audiovisual semantic integration of a feature and that the human brain might selectively integrate the semantic information of the attended features by enhancing the functional connectivity and thus intensifying the information flows from the heteromodal areas to the brain areas encoding the features. Furthermore, the reproducibility ratio may serve as an index for evaluating the audiovisual semantic integration of a feature.

Feature-selective attention: enhancing the neural representations of the attended features in the audiovisual condition.

Considering the audiovisual condition with number, gender, emotion, and bi-feature tasks, we observed that the reproducibility ratios and decoding accuracy rates were higher for the attended features than for unattended features (Figs 3 and 4, S4–S6). This result indicates that feature-selective attention enhanced the neural representations of the attended features and thus increased both the similarity of the neural activity patterns within a class (e.g., male or female class) and the difference between the two classes of the neural activity patterns (e.g., male vs. female). To focus on relevant information and ignore what is irrelevant, the human brain is equipped with a selection mechanism accomplished by the cognitive function of attention[34]. Specifically, in the visual-only or audio-only conditions, the brain selectively processes one or several features via feature-selective attention[7,9,15–17]. Our results revealed that in the audiovisual condition, the feature-selective attention mechanism still permits selective processing of the attended features. In contrast to the visual-only or audio-only conditions, feature-selective attention in the audiovisual condition selectively enhanced the functional connectivity from the heteromodal areas and the brain areas encoding the attended feature (Fig. 5). This enhancement modulated the corresponding information flows and played an important role in achieving the enhancement of neural representations of the attended feature in the audiovisual condition.

Feature-selective attention: a prerequisite for the audiovisual integration of a semantic feature. First, our data analysis results for the experimental runs with the number task supported the conclusion that feature-selective attention is a prerequisite for the audiovisual integration of a semantic feature. As shown in Fig. 2 (A,E), when none of the features of the audiovisual stimuli were attended, audiovisual sensory integration was not observed, not to mention higher level audiovisual semantic integration. Second, using the data for the audiovisual runs with the bi-feature task, we separately localized the brain areas associated with the gender and emotion category differentiation (Tables 2 and 3, respectively). Previous studies have demonstrated that some of the selected brain areas, specifically the STS and the fusiform gyrus, are involved in facial information processing[35–38]. For each of the audiovisual runs with the number, gender and emotion tasks, we calculated cross-reproducibility ratios and cross-decoding accuracy rates for the gender and emotion features using the selected voxels in Tables 2 and 3. We thus demonstrated that these voxels encoded the semantic information of a feature (gender or emotion) only when the feature was attended (Figs 4 and S6). A distributed network including the dorsal medial superior temporal and ventral intraparietal areas is involved in the multisensory integration of visual and vestibular information[39]. Accordingly, we infer that the audiovisual semantic integration corresponding to a feature might be accomplished by a distributed network including the heteromodal areas and the brain areas encoding the feature (Fig. 5). When a feature of an audiovisual object is not attended, our results indicate that the corresponding informative brain areas are not involved in the processing of this feature (Figs 4 and S6), potentially inhibiting the audiovisual semantic integration for this unattended feature.

Feature-selective audiovisual semantic integration. In this study, from the perspectives of neural information encoding and functional connectivity, we demonstrated the modulation effects of feature-selective attention on audiovisual semantic integration. Specifically, when one or two features of the audiovisual objects were attended, the enhancement of the neural response level in the heteromodal areas of the pSTS/MTG indicated the occurrence of audiovisual sensory integration (Fig. 2B–D,F–H), providing the basis for the audiovisual semantic integration corresponding to the attended features. MVPA analysis demonstrated that for only the attended features, the semantic information encoded in the brain was improved by the audiovisual stimuli compared with the visual-only and the auditory-only stimuli (Figs 3, S4, and S5). We previously considered the case in which a single feature of the stimuli was attended[22], as in the experiments with the gender and emotion tasks in this study. Compared with the visual-only and auditory-only stimulus conditions, we observed that the congruent audiovisual stimuli enhanced the neural representation of the attended features. However, how this enhancement is implemented in the brain remains unclear. In this study, we extended this conclusion for the case in which none of the features was attended or more than one feature of the stimulus was attended. Furthermore, the Granger causal connectivity analysis indicated that not only the heteromodal areas but also the brain areas encoding the attended feature may be involved in the audiovisual semantic integration. In the audiovisual condition, feature-selective attention enhanced/reduced the functional connectivity from the heteromodal areas and the brain areas encoding the attended/unattended feature (Fig. 5) and therefore modulated the information flows among these areas. This modulation may be responsible for the enhancement of the semantic information of the attended features by the audiovisual stimuli. Through this modulation of feature-selective attention, the human brain may selectively integrate the semantic information for the attended features of the audiovisual facial stimuli. By contrast, for the unattended features, the corresponding audiovisual semantic integration was inhibited.

Reproducibility ratio: an index for the audiovisual semantic integration of a feature. To form high-level conceptual representations of the semantic features of an audiovisual object, the brain performs audiovisual semantic integration, which may be based on audiovisual integration at the sensory level[10]. Numerous neuroimaging and electrophysiological studies have demonstrated that congruent audiovisual stimuli can enhance neural activities, e.g., in the bilateral superior temporal gyrus (STG)[18–21]. Conversely, in the audiovisual condition, the enhancement of brain activities in heteromodal areas such as the pSTS/MTG may serve as an indicator of audiovisual sensory integration[4,24–26]. Regarding audiovisual semantic integration, numerous studies have discussed the influences of semantic factors on audiovisual integration (see reference[40] and the references therein). However, no studies have addressed the differentiation of the effects of audiovisual semantic integration for different semantic features. The difficulty may lie in the assessments of the integrated and unintegrated information contained in the brain signals. In this study, we observed that the audiovisual semantic integration effects associated with different feature-selective attention tasks could not be differentiated based on the levels of neural activities in the pSTS/MTG (see Results and Fig. 2). This result is consistent with the function of the pSTS/MTG as a presemantic, heteromodal region for crossmodal perceptual features[10]. MVPA approaches open the possibility of separating and localizing spatially distributed patterns, which generally are too weak to be detected by univariate methods, such as GLM[23,41–43]. Using an MVPA method, we calculated a reproducibility ratio corresponding to a feature to assess the semantic information encoded in the brain; the corresponding semantic information was enhanced only for the attended features when the audiovisual stimulus condition was compared with the visual-only and auditory-only stimulus conditions (Figs 3, S4 and S5). We thus observed the differential effects of audiovisual semantic integration for the attended and unattended features. Furthermore, the reproducibility ratio might be used as an index for evaluating the audiovisual semantic integration of a feature.

Finally, we describe several limitations of this study to illustrate future directions. First, we employed a relatively complex experimental design, which led to the collection of large amounts of data. For each subject, the collection of the functional and structural MRI data lasted about six hours, not including preparation time. Because of the difficulty in data collection, we used a relatively small number of subjects. But statistically significant experimental results were still obtained. Second, only visual-only, audio-only and audiovisual facial stimuli were considered in this study. In the future, we must amplify our experimental design, increase the number of subjects, and further consider non-facial stimuli to extend our conclusions.

References

1. Calvert, G. A. & Thesen, T. Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* **98,** 191–205 (2004).
2. Campanella, S. & Belin, P. Integrating face and voice in person perception. *Trends Cogn. Sci.* **11,** 535–543 (2007).
3. Schweinberger, S. R., Robertson, D. & Kaufmann, J. M. Hearing facial identities. *Q. J. Exp. Psych.* **60,** 1446–1456 (2007).
4. Bushara, K. O. *et al.* Neural correlates of cross-modal binding. *Nat. Neurosci.* **6,** 190–195 (2003).
5. Macaluso, E., Frith, C. D. & Driver, J. Multisensory stimulation with or without saccades: fMRI evidence for crossmodal effects on sensory-specific cortices that reflect multisensory location-congruence rather than task-relevance. *NeuroImage* **26,** 414–425 (2005).
6. Macaluso, E., George, N., Dolan, R., Spence, C. & Driver, J. Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage* **21,** 725–732 (2004).
7. McClurkin, J. W. & Optican, L. M. Primate striate and prestriate cortical neurons during discrimination. I. Simultaneous temporal encoding of information about color and pattern. *J. Neurophysiol.* **75,** 481–495 (1996).
8. Nobre, A. C., Rao, A. & Chelazzi, L. Selective attention to specific features within objects: Behavioral and electrophysiological evidence. *J. Cognitive Neurosci.* **18,** 539–561 (2006).
9. Woodman, G. F. & Vogel, E. K. Selective storage and maintenance of an object's features in visual working memory. *Psychon. B. Rev.* **15,** 223–229 (2008).
10. Taylor, K. I., Moss, H. E., Stamatakis, E. A. & Tyler, L. K. Binding crossmodal object features in perirhinal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 8239–8244 (2006).
11. Talsma, D., Senkowski, D., Soto-Faraco, S. & Woldorff, M. G. The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* **14,** 400–410 (2010).
12. Lewis, J. W., Beauchamp, M. S. & DeYoe, E. A. A comparison of visual and auditory motion processing in human cerebral cortex. *Cereb. Cortex* **10,** 873–888 (2000).
13. Joassin, F. *et al.* Cross-modal interactions between human faces and voices involved in person recognition. *Cortex* **47,** 367–376 (2011).
14. Saito, D. N. *et al.* Cross-modal binding and activated attentional networks during audio-visual speech integration: a functional MRI study. *Cereb. Cortex* **15,** 1750–1760 (2005).
15. Ahveninen, J. *et al.* Task-modulated "what" and "where" pathways in human auditory cortex. *Proc. Natl. Acad. Sci. USA.* **103,** 14608–14613 (2006).
16. Maunsell, J. H. R. & Hochstein, S. Effects of behavioral state on the stimulus selectivity of neurons in area V4 of the macaque monkey. In: Channels in the visual nervous system: neurophysiology, psychophysics and models, (ed, Blum B), 447–470. London: Freund (1991).
17. Mirabella, G. *et al.* Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. *Neuron* **54,** 303–318 (2007).
18. Jeong, J. W. *et al.* Congruence of happy and sad emotion in music and faces modifies cortical audiovisual activation. *NeuroImage* **54,** 2973–2982 (2011).
19. Kreifelts, B., Ethofer, T., Grodd, W., Erb, M. & Wildgruber, D. Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage* **37,** 1445–1456 (2007).
20. Müller, V. I., Cieslik, E. C., Turetsky, B. I. & Eickhoff, S. B. Crossmodal interactions in audiovisual emotion processing. *Neuroimage* **60,** 553–561 (2011).
21. Müller, V. I. *et al.* Incongruence effects in crossmodal emotional integration. *Neuroimage* **54,** 2257–2266 (2011).
22. Li, Y. *et al.* Crossmodal Integration Enhances Neural Representation of Task-Relevant Features in Audiovisual Face Perception. *Cereb. Cortex* **25,** 384–395 (2015).
23. Friston, K. J. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2,** 189–210 (1994).
24. Calvert, G. A., Campbell, R. & Brammer, M. J. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* **10,** 649–657 (2000).
25. Frassinetti, F., Bolognini, N. & Làdavas, E. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp. Brain Res.* **147,** 332–343 (2002).
26. Macaluso, E. & Driver, J. Multisensory spatial interactions: a window onto functional integration in the human brain. *TRENDS Neurosci.* **28,** 264–271 (2005).
27. Beauchamp, M. S. Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics* **3,** 93–113 (2005).
28. Brett, M., Anton, J.-L., Valabregue, R. & Poline, J.-B. Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage* **16,** 1140–1141 (2002).
29. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. USA.* **103,** 3863–3868 (2006).
30. Nichols, T. & Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* **12,** 419–446 (2003).
31. Hamilton, J. P., Chen, G., Thomason, M. E., Schwartz, M. E. & Gotlib, I. H. Investigating neural primacy in Major Depressive Disorder: multivariate Granger causality analysis of resting-state fMRI time-series data. *Mol. Psychiatry* **16,** 763–772 (2011).
32. Hopfinger, J. B., Buonocore, M. H. & Mangun, G. R. The neural mechanisms of top-down attentional control. *Nat. Neurosci.* **3,** 284–291 (2000).
33. Seth, A. K. A MATLAB toolbox for Granger causal connectivity analysis. *J. Neurosci. Meth.* **186,** 262–273 (2010).
34. Talsma, D., Doty, T. J. & Woldorff, M. G. Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* **17,** 679–690 (2007).
35. Gobbini, M. I. & Haxby, J. V. Neural response to the visual familiarity of faces. *Brain Res. Bull.* **71,** 76–82 (2006).
36. Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. The distributed human neural system for face perception. *Trends Cogn. Sci.* **4,** 223–232 (2000).
37. Haxby, J. V. *et al.* Face encoding and recognition in the human brain. *Proc. Natl. Acad. Sci. USA* **93,** 922–927 (1996).
38. Leveroni, C. L. *et al.* Neural systems underlying the recognition of familiar and newly learned faces. *J. Neurosci.* **20,** 878–886 (2000).
39. Zhang, W. & Wu, S. Reciprocally Coupled Local Estimators Implement Bayesian Information Integration Distributively. in *Adv. Neural Infor. Processing Syst.* (eds C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger) **26,** 19–27 (2013).
40. Doehrmann, O. & Naumer, M. J. Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Res.* **1242,** 136–150 (2008).
41. Goebel, R. & van Atteveldt, N. Multisensory functional magnetic resonance imaging: a future perspective. *Exp. Brain Res.* **198,** 153–164 (2009).
42. Pereira, F., Mitchell, T. & Botvinick, M. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* **45,** 199–209 (2009).
43. Polyn, S. M., Natu, V. S., Cohen, J. D. & Norman, K. A. Category-specific cortical activity precedes retrieval during memory search. *Science* **310,** 1963–1966 (2005).

Acknowledgements

This work was supported by the National Key Basic Research Program of China (973 Program) under Grant 2015CB351703, the National High-tech R&D Program of China (863 Program) under Grant 2012AA011601, the National Natural Science Foundation of China under Grants 91420302, 81471654 and 61403147, and Guangdong Natural Science Foundation under Grant 2014A030312005.

Author Contributions

Y.L. designed research and wrote the paper; J.L. and W.W. analyzed the data; B.H., T.Y. and P.L. performed the research; F.F. and P.S. revised the paper; all authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Li, Y. *et al.* Selective Audiovisual Semantic Integration Enabled by Feature-Selective Attention. *Sci. Rep.* **6**, 18914; doi: 10.1038/srep18914 (2016).