Cerebral Cortex, October 2018;28: 3623–3637

doi: 10.1093/cercor/bhx235
Advance Access Publication Date: 25 September 2017
Original Article

ORIGINAL ARTICLE

The Effects of Audiovisual Inputs on Solving the Cocktail Party Problem in the Human Brain: An fMRI Study

Yuanqing Li[1,2], Fangyi Wang[1,2], Yongbin Chen[1,2], Andrzej Cichocki[3,4] and Terrence Sejnowski[5]

[1]Center for Brain Computer Interfaces and Brain Information Processing, South China University of Technology, Guangzhou 510640, China, [2]Guangzhou Key Laboratory of Brain Computer Interaction and Applications, Guangzhou 510640, China, [3]Riken Brain Science Institute, Wako shi 3510198, Japan, [4]Skolkovo Institute of Science and Technology (SKOTECH), Moscow 143026, Russia and [5]Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

Address correspondence to Yuanqing Li. Email: auyqli@scut.edu.cn

Abstract

At cocktail parties, our brains often simultaneously receive visual and auditory information. Although the cocktail party problem has been widely investigated under auditory-only settings, the effects of audiovisual inputs have not. This study explored the effects of audiovisual inputs in a simulated cocktail party. In our fMRI experiment, each congruent audiovisual stimulus was a synthesis of 2 facial movie clips, each of which could be classified into 1 of 2 emotion categories (crying and laughing). Visual-only (faces) and auditory-only stimuli (voices) were created by extracting the visual and auditory contents from the synthesized audiovisual stimuli. Subjects were instructed to selectively attend to 1 of the 2 objects contained in each stimulus and to judge its emotion category in the visual-only, auditory-only, and audiovisual conditions. The neural representations of the emotion features were assessed by calculating decoding accuracy and brain pattern-related reproducibility index based on the fMRI data. We compared the audiovisual condition with the visual-only and auditory-only conditions and found that audiovisual inputs enhanced the neural representations of emotion features of the attended objects instead of the unattended objects. This enhancement might partially explain the benefits of audiovisual inputs for the brain to solve the cocktail party problem.

Key words: audiovisual inputs, cocktail party problem, decoding, neural representation, reproducibility.

Introduction

Despite the high levels of noise at cocktail parties, including conversations and laughter, we are easily able to recognize, understand and focus on the voice of the person with whom we are speaking. This task is known as the cocktail party problem (Cherry 1953), and it has received significant attention in many multidisciplinary areas such as neuroscience, psychology, acoustics and computer science. The cocktail party problem also appears in multiple engineering areas, such as communications, medical signal processing, speech signal processing, and brain signal analysis, which is still challenging and may be solved with a class of blind source separation algorithms, including independent component analysis (ICA) and sparse representation (Bell and Sejnowski 1995; Lewicki and Sejnowski 2000; Brown et al. 2001; Li et al. 2004). However, many engineering methods for blind source separation address

a problem that is quite different from that solved by the brain in cocktail party settings (McDermott 2009). For instance, when solving the cocktail party problem, the brain often selectively attends to one speech and ignores others. This selective attention mechanism has not been reflected in blind source separation algorithms. Furthermore, how to solve the cocktail party problem in the brain is far from clear. To date, most studies of the cocktail party problem in human brains have considered primarily auditory-only settings, and 2 associated challenges have been extensively discussed (McDermott 2009). The first is sound segregation by the brain, namely, the ability of the auditory system to derive the properties of individual sounds from a mixture of sounds entering the ears. The second challenge is selective attention to the sound source of interest. A number of studies have explored the modulatory effects of selection attention on sound segregation in the brain. For instance, it has been suggested that attention induces a top-down selectivity on neural activity to form a representation only of the attended sound stream (Ahveninen et al. 2011; Zion-Golumbic and Schroeder 2012). Applying a decoding method to multi-electrode surface recordings from the cortex of subjects, Mesgarani and Chang demonstrated that signals in the auditory cortex can be used to reconstruct the spectrotemporal patterns of attended speech tokens better than those of ignored speech tokens (Mesgarani and Chang 2012). Several studies have revealed attentional modulation of the high gamma response (70–150 Hz, ECoG) to speech (Zion-Golumbic and Schroeder 2012). Ding and Simon (2012) showed similar attention-modulated neuronal selectivity in the low-frequency phase (1–8 Hz, MEG) locking to speech. Zion-Golumbic et al. (2013b) found that brain activity (ECoG) dynamically tracks speech streams using both low-frequency phase and high-frequency amplitude fluctuations and that attention "modulates" the representation by enhancing the cortical tracking of attended speech streams. Other studies have found many perceptual or cognitive cues that facilitate sound segregation and selective attention to the target speech/voice. For instance, if a mixture of auditory signals contains energy at multiple frequencies that start or stop at the same time, those frequencies are likely to belong to the same sound (McDermott 2009). Voice differences between the target speaker and other speakers can also provide cues for segregating concurrent independent sound sources (MicheyI and Oxenham 2010; Du et al. 2011). Moreover, spatial unmasking can result from top-down processes by facilitating selective spatial attention to the target (Freyman et al. 2001; Rakerd et al. 2006; Huang et al. 2009).

The human brain often—even daily—solves the cocktail party problem in audiovisual environments. However, in contrast to auditory-only settings, fewer studies have discussed the cocktail party problem in audiovisual settings. Nevertheless, several studies have reported that understanding speech, particularly under multispeaker conditions, can be significantly facilitated by viewing the speaker's face (Schwartz et al. 2004; Senkowski et al. 2008; Bishop and Miller 2009). First, visual information regarding the spatial locations of speakers can act as spatial cues to guide attention to the relevant parts of the auditory input, enabling listeners to suppress the portions of the auditory input that do not belong to the target signal (Haykin and Chen 2005; Kidd et al. 2005a). Senkowski et al. (2008) examined visuo-spatial attention in multiple speaker scenarios and observed that attention to visual inputs from flanking speaker interfered with speech recognition performance. Furthermore, lip reading can improve the ability to understand speech in a noisy environment, as has been shown in comparisons of audiovisual perception and audio-only perception (Grant 2001; Schwartz et al. 2004). Zion-Golumbic et al. (2013a) investigated audiovisual effects on envelope-tracking responses under 2 conditions: when listening to a single speaker and when attending to 1 speaker while ignoring a concurrent irrelevant speaker. These authors demonstrated that visual inputs enhanced selective speech envelope tracking in the auditory cortex in a "cocktail party" environment.

The benefits of visual modality discussed above may be associated with audiovisual integration in the brain. Many studies have shown that audiovisual integration in the brain may facilitate rapid, robust and automatic object perception and recognition (Calvert and Thesen 2004; Campanella and Belin 2007; Schweinberger et al. 2011). The underlying neural mechanisms of audiovisual integration have also been extensively explored. For instance, it was demonstrated in previous functional magnetic resonance imaging (fMRI) studies that congruent audiovisual stimuli lead to much stronger BOLD responses in the posterior superior temporal sulcus/middle temporal gyrus (pSTS/MTG) compared with visual-only and auditory-only stimuli (Bushara et al. 2003; Calvert and Thesen 2004; Macaluso et al. 2004; Macaluso and Driver 2005). An fMRI study studied how the brain applies visual information to improve comprehension in naturalistic conditions and found that audiovisual integration likely improves comprehension by enhancing communication among the left temporal–occipital boundary, the left medial-temporal lobe, and the left superior temporal sulcus (STS) (Bishop and Miller 2009). Although our brains often receive visual and auditory information simultaneously when faced with a cocktail party problem, the effects of audiovisual integration and its neural mechanisms have not been extensively explored.

In this study, we explored the effects of audiovisual inputs in solving the cocktail party problem from the viewpoint of neural representations. We hypothesized that audiovisual inputs might enhance the neural representation of the attended object in a cocktail party environment, which may partially explain the underlying neural mechanism regarding the behavioral benefits of audiovisual inputs. To test this hypothesis, we conducted an fMRI experiment in which subjects were presented visual-only, auditory-only or congruent audiovisual dynamic facial stimuli. Each congruent audiovisual stimulus (a synthesized movie clip) consisted of a combination of 2 movie clips, each of which was a dynamical audiovisual face associated with a positive or negative emotion (crying or laughing). The video and audio portions of the synthesized movie clips were extracted and used as visual-only and auditory-only stimuli, respectively. Specifically, a visual-only stimulus included 2 dynamic faces (displayed on the left and right sides of the screen), and the corresponding auditory-only stimulus was the mixture of the 2 voices produced by the 2 faces. The subjects were tasked with selectively attending to 1 object (a face, voice, or face-voice pair) and judging its emotion category. It is well known that multivariate pattern analysis (MVPA) approaches can be used to separate and localize spatially distributed patterns that are generally too weak to be detected by univariate methods, such as general linear model (GLM) analysis (Friston et al. 1994; Polyn et al. 2005; Goebel and van Atteveldt 2009; Pereira et al. 2009; Zeng et al. 2012). With the help of MVPA methods, we may explore how percepts, memories, thought, and knowledge are represented in patterns of brain activity, and decode physical/semantic features of stimuli such as emotion, gender and familiarity of faces, and the conceptual categories associated with sentences (Li et al. 2015, 2016; Ghio et al. 2016). In this work, we applied an MVPA method to the collected fMRI data to assess the neural representations of the emotion features of the stimuli in the visual-only, auditory-only, and audiovisual conditions.

Specifically, we obtained a brain pattern for each stimulus that reflected the neural representation of the emotion feature of this stimulus. Based on these brain patterns, we decoded the emotion categories of the targets/non-targets (attended objects/unattended objects) from the fMRI data and calculated the reproducibility of brain patterns that corresponded to the crying and laughing categories of the targets/non-targets. Higher decoding accuracy and reproducibility index implied more encoded information and enhanced neural representations. Our experimental results indicated that audiovisual inputs induced enhanced neural representations of the emotion features of the attended targets in the simulated cocktail party environment.

Experimental Procedure and Methods

Subjects

Thirteen healthy native Chinese males (22–49 years of age with normal or corrected-to-normal vision and normal hearing) participated in this study. All participants provided written informed consent prior to the experiment. The experimental protocol was approved by the Ethics Committee of South China Normal University, Guangzhou, China.

Experimental Stimuli

We browsed the Internet and selected 80 short movie clips of Chinese faces that included video and audio recordings. The 80 movie clips contained crying or laughing faces but no words were spoken, and the clips were semantically partitioned into 2 groups based on gender (40 male vs. 40 female Chinese faces) or emotion (40 crying vs. 40 laughing faces). We further processed these stimulus clips using Windows Movie Maker. Each edited movie clip was in gray scale with a duration of 1400 ms and subtending 10.7° × 8.7°. The luminance levels of the videos were matched by adjusting the total power value of each video (the sum of the squares of the pixel gray values). Similarly, the audio power levels were matched by adjusting the total power value of each audio clip. Next, we constructed 80 new movie clips by combining 2 movie clips into 1, and this group of 80 constructed clips had the following characteristics: 1) each synthesized movie clip was a combination of a male and a female clip, and the emotion categories were randomly paired; 2) in 40 synthesized movie clips, the female faces were located on the left, and the male faces were on the right, while in the other 40 synthesized movie clips, the female faces were located on the right, and the male faces were on the left; 3) the audio of each synthesized movie clip corresponded to the mixture of the audio signals of the male and female clips; and 4) each of the original 80 clips were used twice in constructing the 80 synthesized clips.

These synthesized movie clips, containing both videos and audios, were used as the audiovisual stimuli. Furthermore, the visual-only stimuli corresponded to the video clips extracted from the 80 synthesized movie clips, and the auditory-only stimuli corresponded to the audio clips extracted from the 80 synthesized movie clips (see examples in Fig. 1A). Visual stimulation was presented binocularly using OLED video goggles (NordicNeuroLab, Bergen, Norway; SVGA, 800 × 600 pixels; refresh rate: 85 Hz; FOV: 30 horizontal, 23 vertical; stimulus luminance: 70–110 cd/m^2). Sound stimuli were delivered through MRI-compatible AudioSystem headphones (NordicNeuroLab, Bergen, Norway). The participants operated an MRI-compatible response box using their dominant right hand to respond.

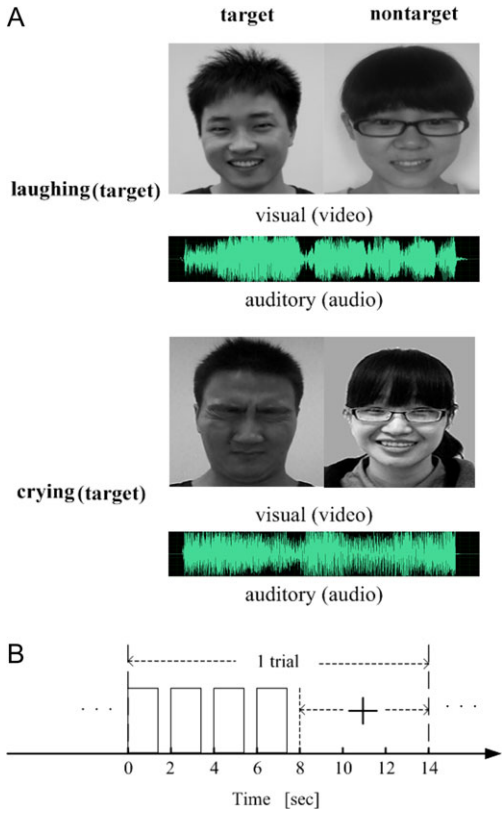

Figure 1. Experimental stimuli and time courses. (*A*) Two examples of audiovisual stimuli. (*B*) Time course of a trial for the audiovisual, visual-only, or auditory-only conditions, where the presentation of a stimulus (video/audio/movie clip) lasted 1400 ms and was repeated 4 times during the first 8 s of the trial. A visual cue ("+") appeared at the 8th second and lasted for 6 s.

Experimental Procedure

We utilized a 1 × 3 factorial design with the task (emotion judgment) as the first factor and the stimulus condition (audiovisual, visual-only, or auditory-only) as the second factor. Each subject performed 3 experimental runs corresponding to the 3 stimulus conditions presented in a pseudo-randomized order. Each run included 8 blocks, and each block contained 10 trials. The 3 runs took place over 1 day for each subject. During the experiment, the subjects were asked to attend to an object contained in each presented stimulus (audiovisual, visual-only, or auditory-only stimulus) and to make a corresponding judgment regarding its emotion (crying vs. laughing). In this study, the attended/unattended objects contained in the visual-only, auditory-only, or audiovisual stimuli were also called targets/non-targets. As an example, 1 run of the experimental procedure corresponding to the audiovisual stimulus condition (the other runs were performed with similar procedures) is described as follows. 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 emotion categories of the target of the stimuli balanced within each block. A 20-s blank period (consisting of a gray screen with no auditory stimulation) was included between adjacent blocks. At the beginning of each block, a short instruction ("attend to the emotion feature of the male, cry 1 and laugh 2", "attend to the emotion feature of the male, cry 2 and laugh 1", "attend to the emotion feature of the female, cry 1 and laugh 2", or "attend to the emotion feature of the female, cry 2 and laugh 1") was displayed for 4 s on the screen.

The instruction "cry 1 and laugh 2" indicated that the subject should press key 1 or key 2 for crying or laughing emotions, respectively, whereas the instruction "cry 2 and laugh 1" indicated that the subject should press key 2 or key 1 for crying or laughing emotions, respectively. The 2 keys were pseudo-randomly assigned to the 2 emotion categories in each block. As shown in Figure 1B, 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 judge the emotion of the attended object by pressing 1 of the 2 keys. The fixation cross changed color at the 12th second, indicating that the next trial would begin shortly. In total, a run lasted 1310 s. fMRI data were collected during the experiment.

fMRI Data Collection

fMRI data were collected using a 3 T Siemens Trio scanner with a 12-channel phase array head coil at South China Normal University. A 3D anatomical T1-weighted scan (FOV, 256 mm; matrix, 256 × 256; 176 slices; and slice thickness: 1 mm) was acquired before the functional scan for each subject. During the functional experiment, gradient-echo echo-planar (EPI) T2*-weighted images (32 slices acquired in an ascending interleaved order; TR = 2000 ms, TE = 30 ms; FOV: 224 mm, matrix: 64 × 64, slice thickness: 3.5 mm) were acquired, covering the entire brain.

Data Processing

Pre-processing

The fMRI data were pre-processed using SPM8 (Friston et al. 1994) and custom functions in MATLAB 7.4 (MathWorks, Natick, MA, USA). Specifically, for each run, the first 5 volumes collected before magnetization equilibrium was reached were discarded prior to the analysis. The following pre-processing steps were then performed on the fMRI data collected during each run: head-motion correction, slice-timing correction, co-registration between the functional and structural scans, normalization to the MNI standard brain, data masking to exclude non-brain voxels, time-series detrending, and normalization of the time series in each block to a zero mean and unit variance.

MVPA Procedure

Using the collected fMRI data, we performed an MVPA analysis, which was similar to that described in our previous study (Li et al. 2015), to assess the neural representations of the emotion features of the experimental stimuli. The 3 runs completed by each subject corresponded to the visual-only, auditory-only, audiovisual conditions. For each run, we first calculated 2 reproducibility indices of the brain patterns corresponding to the crying and laughing categories of the attended objects (targets) by applying the MVPA method to the fMRI data. A higher reproducibility indicates a stronger similarity within each class of brain patterns associated with the crying or laughing category. Based on the brain patterns extracted from the fMRI data, we also decoded the emotion categories (crying vs. laughing) of the targets perceived by the subject. A similar analysis was also performed for the non-targets in each run. Below, we explain the MVPA procedure associated with the targets for each run.

The calculation of the reproducibility indices and the decoding accuracy of the targets in each run were performed based on an 8-fold cross-validation, as illustrated in Figure 2. Specifically, the data from the 80 trials were evenly partitioned into 8 non-overlapping datasets. For the kth-fold of the cross-validation (k = 1,…,8), the kth dataset (10 trials) was used for the test, and the remaining 7 datasets (70 trials) were used for voxel selection and classifier training. Following the 8-fold cross-validation, the average reproducibility indices and the decoding accuracy rates were calculated across all folds. The data processing for the kth-fold included the following steps: 1) Voxel selection. A spherical searchlight algorithm was applied to the training dataset for voxel selection (Kriegeskorte et al. 2006). Specifically, this algorithm was sequentially centered at each voxel with a 3-mm radius searchlight that highlighted 19 voxels. Within each searchlight that corresponded to a voxel, we computed a Fisher ratio based on a Fisher linear discriminant analysis to indicate the level of discrimination between the 2 categories of targets (crying vs. laughing) in the local neighborhood of that voxel. A Fisher ratio map was thus obtained for the entire brain. We selected K informative voxels with the highest Fisher ratios. Since the performed multivariate analysis operates with voxels in a neighborhood, the searchlight approach is also referred to as local pattern effects mapping. For the searchlight approach, a strong assumption that discriminative information is located within small brain regions is made, and this assumption is appropriate in light of the known functional organization of the brain (Kriegeskorte et al. 2006). 2) Neural activity pattern estimation. Using the selected voxels, a K-dimensional pattern vector was constructed for each trial in the training and test datasets. Specifically, because of the delayed hemodynamic response, we calculated each element of the pattern vector as the mean BOLD response of a selected voxel over seconds 6–14 of the trial (the last 4 volumes of each trial). 3) Reproducibility index calculation. We used $\cos\theta$ as a reproducibility index to assess similarities among neural activity patterns elicited by the stimuli, where θ is the angle between 2 pattern vectors, and larger $\cos\ \theta$ values indicate higher similarities. Specifically, we extracted 10 pattern vectors corresponding to the 10 trials of the test dataset, which included 5 vectors in each class (crying or laughing stimuli). For each pair of pattern vectors within the same class, we calculated a reproducibility index. The mean of the reproducibility indices from each class was defined as a reproducibility index for the kth-fold. Two reproducibility indices were thus obtained for the crying and laughing targets. 4) Decoding/Prediction. To predict emotion categories for the kth-fold, a linear support vector machine (SVM) classifier was trained based on the pattern vectors of the labeled training data (+1 and −1 for the crying and laughing targets, respectively). The emotion category of each trial of the test data was then predicted by applying the SVM to the corresponding pattern vector. After the 8-fold cross-validation, we obtained the decoding accuracy of each run corresponding to the targets. Furthermore, by varying the above number K of selected voxels, we could obtain the curves of reproducibility and decoding accuracy. The calculation of the reproducibility indices and the decoding accuracy for the non-targets in each run were also performed through a similar 8-fold cross-validation, as described above (also see Fig. 2).

Localization of Informative Brain Areas Associated with the Neural Representations

During the above MVPA procedure, we selected/determined a voxel set for each fold of the cross-validation. Based on this voxel set, we obtained a brain pattern vector for each trial. Using these pattern vectors, we then performed the classification and reproducibility calculation. A brain pattern vector

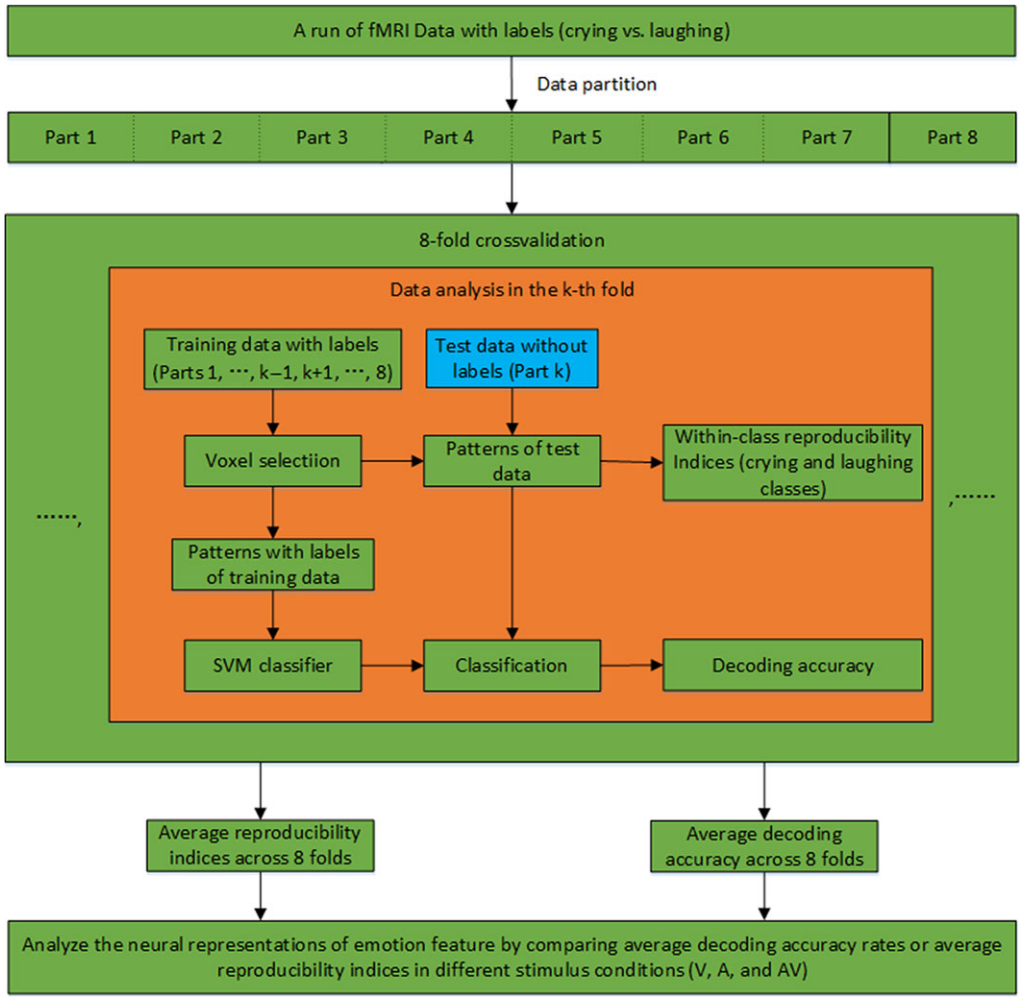

Figure 2. MVPA procedure for calculating the reproducibility indices and decoding accuracy in an experimental run.

could be seen as a neural representation of the emotion feature of its corresponding trial (Haxby et al. 2014). The voxels involved in the pattern vectors were those contributing to the neural representations. However, the voxel sets obtained in different folds generally were not fully overlapped. One reason was that some voxels were incorrectly selected because of noise. Below, we present an algorithm to localize the informative voxels as accurately as possible.

Using the data from each run, we localized a voxel set, which provided information regarding the 2 emotion categories of the targets (crying vs. laughing). As an example, we described the localization of informative voxels based on the data in the audiovisual condition as described below. For each subject, we performed an 8-fold cross-validation to decode the emotion categories of the targets, as described previously. We trained an SVM classifier in each fold and obtained an SVM weight map for the entire brain (the unselected voxels were assigned a weight of zero). The absolute values of the SVM weights were normalized to [0,1] by dividing the absolute value of each weight with the maximum and then used to construct a whole-brain weight map for each fold, which reflected the importance of the voxels for decoding the emotion categories. By averaging the weight maps across all folds and all subjects, an actual group weight map was obtained for differentiating the emotion categories. We then performed 1000 permutations to obtain 1000 group weight maps for the emotion categories. Each group weight map was constructed as described above with the exception that, for each subject, the labels of all trials were randomly assigned. To control the family wise error rate, the maximum voxel weight was obtained for each of the 1000 group weight maps calculated in the permutations and a null distribution was constructed using the 1000 maximum voxel weights (Nichols and Hayasaka 2003). The actual group weight map was then converted to a p map based on the null distribution. Specifically, the P value of a voxel was estimated as the rank of the actual map's value at this voxel in the null distribution divided by 1000. The resulting p map was thresholded with $P < 0.05$, then we obtained a set of voxels with significant P values, which were informative for the emotion categories of the targets in the audiovisual condition.

Functional Connectivity Analysis

Using the above localization algorithm, we obtained an informative voxel set for each stimulus condition, which reflected the informative brain areas contributing to the neural representations of emotion features. In the audiovisual stimulus condition, the visual information and auditory information might be integrated in the brain. The key heteromodal brain areas associated with audiovisual integration were the pSTS/MTG (Kreifelts et al. 2007; Jeong et al. 2011; Müller et al. 2012). Note that the heteromodal

areas associated with audiovisual integration cannot be identified by our localization algorithm that was designed principally to determine the informative voxels regarding the differentiation of the 2 emotion categories of the targets (crying vs. laughing). Therefore, we identified the heteromodal areas by performing GLM analysis and comparing the audiovisual condition with the visual-only and auditory-only conditions, as shown in Supporting Materials. In the following, we analyzed the functional connectivities between the informative brain areas contributing to the neural representations and the heteromodal brain areas, pSTS/MTG, associated with audiovisual integration using the fMRI data from the visual-only, auditory-only, and audiovisual conditions. The functional connectivities might reflect the information flow between the 2 classes of brain areas and partially explain how audiovisual integration affected the neural representations of emotion features in the audiovisual condition.

As an example, the functional connectivity analysis using the data from the audiovisual condition is described below. First, we obtained the informative voxel set for the emotion category of the targets in the audiovisual condition, as indicated above. Based on a GLM analysis, we compared the audiovisual condition with the visual-only and auditory-only conditions and identified the heteromodal areas of the left pSTS/MTG (cluster center: (−57, −45, 12), cluster size: 85) and the right pSTS/MTG (cluster center: (60, −42, 15), cluster size: 227) (see Supporting Materials). Next, we performed a multivariate Granger causality (GC) analysis to assess the functional connectivities between the brain areas (25 clusters shown in Table 1) related to differentiation of the emotion categories of the targets and the heteromodal areas (2 clusters mentioned above) (Hopfinger et al. 2000; Hamilton et al. 2011). Specifically, for each pair of clusters (say Clusters A and B), one from the 25 informative clusters and the other from the 2 heteromodal areas, we obtained 2 average time series by averaging the time series of all the voxels within each of the 2 clusters. Using the 2 average time series, we calculated 2 GC values, one represented the strength of the connection from Cluster A to Cluster B, whereas the other represented the strength of the connection from Cluster B to Cluster A (Seth 2010). Totally, we obtained 100 GC values for each subject. Furthermore, we performed a non-parameter statistical test on these GC values with a significance level of $P = 0.05$ (FDR corrected) (Seth 2010). More details of GC analysis is provided in Supporting Materials. A significant GC value implied a directional connection, that is, from Cluster A to Cluster B. On the contrary, a non-significant GC value indicated that the strength of its corresponding connection was too weak. In this case, we consider that this connection did not exist. If the 2 GC values corresponding to a pair of clusters were significant, we assumed that there was a bi-directional connection between the 2 clusters. In line with the foregoing, we also calculated the functional connectivities between the brain areas related to differentiation of the emotion categories in each unimodal condition (the visual-only or auditory-only condition) and the heteromodal areas.

Results

In this section, the behavioral results in the fMRI experiment were first presented. Next, we presented the fMRI data analysis results, including the reproducibility of brain patterns, decoding accuracies, and the localized informative brain areas, which were associated with the neural representations of stimulus features. Furthermore, the functional connectivity results were shown to explain the effects of the neural representations, induced by audiovisual inputs.

Table 1 Informative brain areas for discriminating emotion categories in the audiovisual condition ($P < 0.05$, FDR corrected)

Brain region	Side	MNI coordinates			Weight	mm^3
		X	Y	Z		
Precentral gyrus	L	−51	3	33	0.0206	594
Superior frontal gyrus (dorsolateral)	R	18	36	42	0.044	540
Middle frontal gyrus	L	−36	45	21	0.1574	405
Middle frontal gyrus	R	39	51	15	0.0226	486
Supplementary motor area	L	−3	0	78	0.1288	675
Supplementary motor area	R	6	−15	60	0.0364	729
Superior frontal gyrus (medial)	R	6	72	9	0.1634	756
Insula	R	39	9	6	0.0265	432
Anterior cingulate and paracingulate gyri	L	−3	21	21	0.011	432
Calcarine fissure and surrounding cortex	L	−9	−78	9	0.022	459
Cuneus	L	−6	−96	27	0.1631	1053
Lingual gyrus	R	15	−72	−12	0.245	540
Middle occipital gyrus	L	−21	−102	3	0.2181	459
Middle occipital gyrus	R	27	−72	30	0.0199	567
Inferior occipital gyrus	R	27	−99	−6	0.2696	729
Fusiform gyrus	L	−24	−3	−42	0.1683	513
Fusiform gyrus	R	45	−24	−30	0.1813	432
Postcentral gyrus	R	63	−12	21	0.1769	1215
Superior parietal gyrus	L	−24	−45	60	0.1604	729
Supramarginal gyrus	R	45	−39	33	0.2119	1026
Precuneus	R	9	−51	48	0.0262	783
Superior temporal gyrus	L	−51	−39	12	0.022	594
Superior temporal gyrus	R	54	−45	15	0.131	864
Middle temporal gyrus	L	−54	−3	−15	0.0306	567
Inferior temporal gyrus	R	45	−12	−39	0.1791	1161

Behavioral Results

The average behavioral accuracy rates with standard deviations obtained in our fMRI experiment were 94.90 ± 3.00%, 70.77 ± 9.28% and 96.15 ± 3.16% for the visual-only condition, the auditory-only condition, and the audiovisual condition, respectively. A 1-way repeated-measures ANOVA was performed to assess the response accuracy rates and indicated a significant main effect of the stimulus condition ($P < 0.0001$, $F(2,24) = 95.108$). Furthermore, post hoc Bonferroni-corrected paired t-tests indicated that the accuracy rate was significantly lower for the auditory-only condition compared with the audiovisual condition and the visual-only condition (all $P < 0.05$). There was no significant difference between the visual-only condition and the audiovisual condition ($P > 0.05$). The average response times with standard deviations were 0.9065 ± 0.1827, 1.6145 ± 0.2771, and 0.8548 ± 0.1754 s for the visual-only condition, the auditory-only condition, and the audiovisual condition, respectively. The response time for each trial was calculated based on the onset of the fourth stimulus repetition. A 1-way repeated-measures ANOVA was performed to assess the response time and indicated a significant main effect of the stimulus condition ($P < 0.0001$, $F_{2,24} = 263$). Furthermore, post hoc Bonferroni-corrected paired t-tests indicated that the response time of the audiovisual condition was significantly lower than the response times of the visual-only condition and the auditory-only condition (all $P < 0.05$). Furthermore, the response time of the visual-only condition was also significantly lower than that of the auditory-only condition ($P < 0.05$). Regarding response accuracy and time, these results showed the behavioral benefits of audiovisual inputs over the auditory-only inputs for human brain to solve the cocktail party problem.

Reproducibility Results

Using the MVPA method, we calculated 2 reproducibility curves with respect to the number of selected voxels (from 25 to 3000) that corresponded to the crying and laughing targets categories for each of the 3 runs (see Experimental Procedure and Methods). These curves are shown in Figure 3A and C. Additionally, Figure 3B and D shows the reproducibility indices obtained with the 1600 selected voxels for the crying and laughing targets, respectively. Figure 3 indicates that the reproducibility indices were significantly higher for the audiovisual condition than for the visual-only and auditory-only conditions. In the following analysis, we used the 1600 selected voxels as an example to present the statistical results. A 1-way

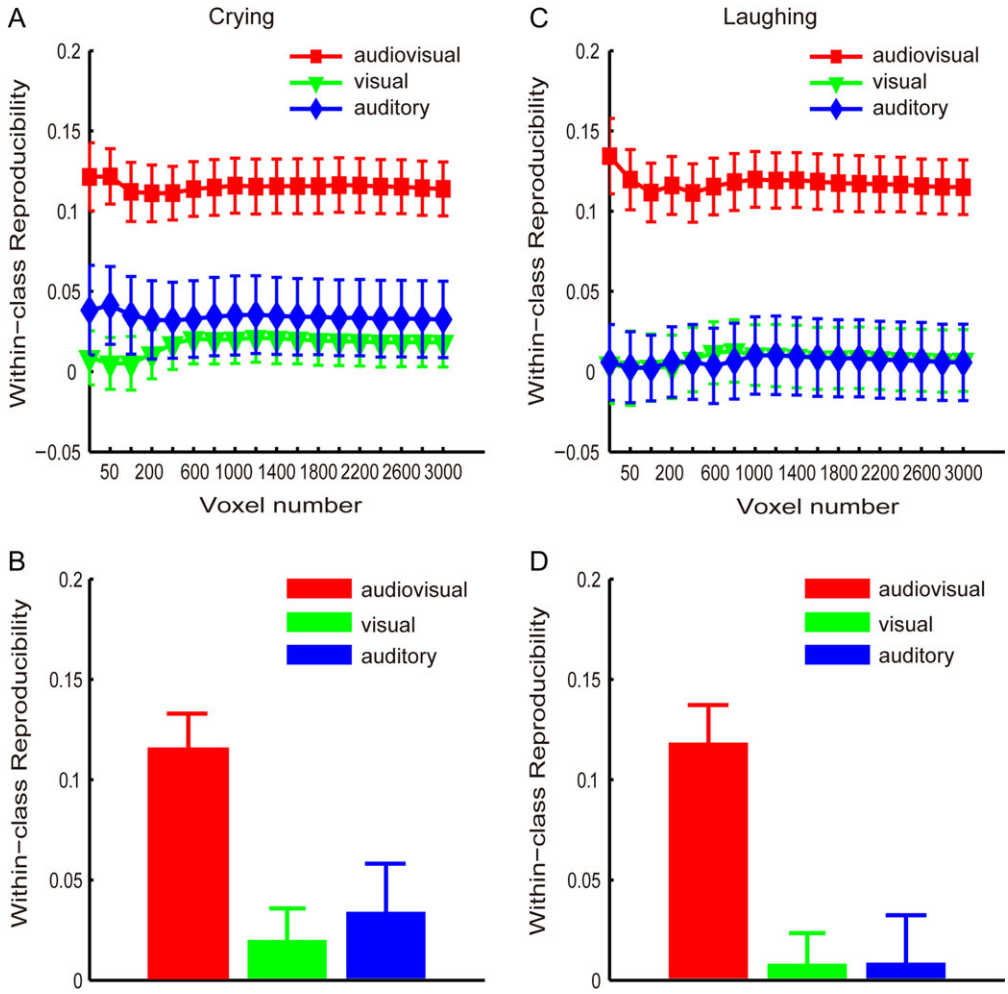

Figure 3. Reproducibility indices (means and standard errors across all subjects) of the brain patterns corresponding to the targets in the audiovisual, visual-only, and auditory-only stimulus conditions. (*A*) and (*B*): "crying" targets; (*C*) and (*D*): "laughing" targets. (*A*) and (*C*): curves of the reproducibility indices with respect to the numbers of voxels; (*B*) and (*D*): reproducibility indices obtained with 1600 voxels.

repeated-measures ANOVA was performed to assess the reproducibility indices corresponding to the crying targets and indicated a significant main effect of the stimulus condition ($P < 0.01$, $F_{2,24} = 8.05$; Fig. 3B). Furthermore, post hoc Bonferroni-corrected paired t-tests indicated that the reproducibility index of the audiovisual condition was significantly higher than the reproducibility indices of the visual-only condition ($t(12) = 4.75$, $P < 0.01$) and auditory-only condition ($t(12) = 2.89$, $P < 0.05$). There was no significant difference between the visual-only and auditory-only conditions ($t(12) = 0.50$, $P \approx 1$). Another 1-way repeated-measures ANOVA was performed to assess the reproducibility indices corresponding to the laughing targets and also indicated a significant main effect of the stimulus condition ($F_{2,24} = 12.41$, $P < 10^{-3}$; Fig. 3D). According to the post hoc Bonferroni-corrected paired t-tests, the reproducibility index was significantly higher for the audiovisual condition than for the visual-only ($t(12) = 4.58$, $P < 0.01$) and auditory-only conditions ($t(12) = 3.67$, $P < 0.05$), and there was no significant difference between the visual-only and auditory-only conditions ($t(12) = 0.045$, $P \approx 1$).

We also calculated 2 reproducibility curves with respect to the number of selected voxels (from 25 to 3000) that corresponded to the crying and laughing categories of the non-targets for each of the runs (1, 2, 3), which are shown in Figure 4A and C, respectively. Figure 4B and D shows the reproducibility indices obtained with 1600 selected voxels for the crying and laughing non-targets, respectively. Figure 4 indicates that the reproducibility indices were not significantly higher for the audiovisual condition compared with the visual-only and auditory-only conditions. A 1-way repeated-measures ANOVA was performed to assess the reproducibility indices obtained with 1600 selected voxels corresponding to the crying/laughing non-targets and indicated that the main effect of the stimulus condition was non-significant ($F_{2,24} = 0.65$, $P = 0.53$, Fig. 4B; $F_{2,24} = 0.25$, $P = 0.78$, Fig. 4D).

Decoding Results

For each experimental run, we separately decoded the emotion categories (i.e., "crying" and "laughing") of the targets from the collected fMRI data using the MVPA method (see Experimental Procedure and Methods). We systematically varied the number of selected voxels from 25 to 3000 to decode the emotion categories, and the results are shown in Figure 5A. The decoding results obtained from 1600 selected voxels are shown in

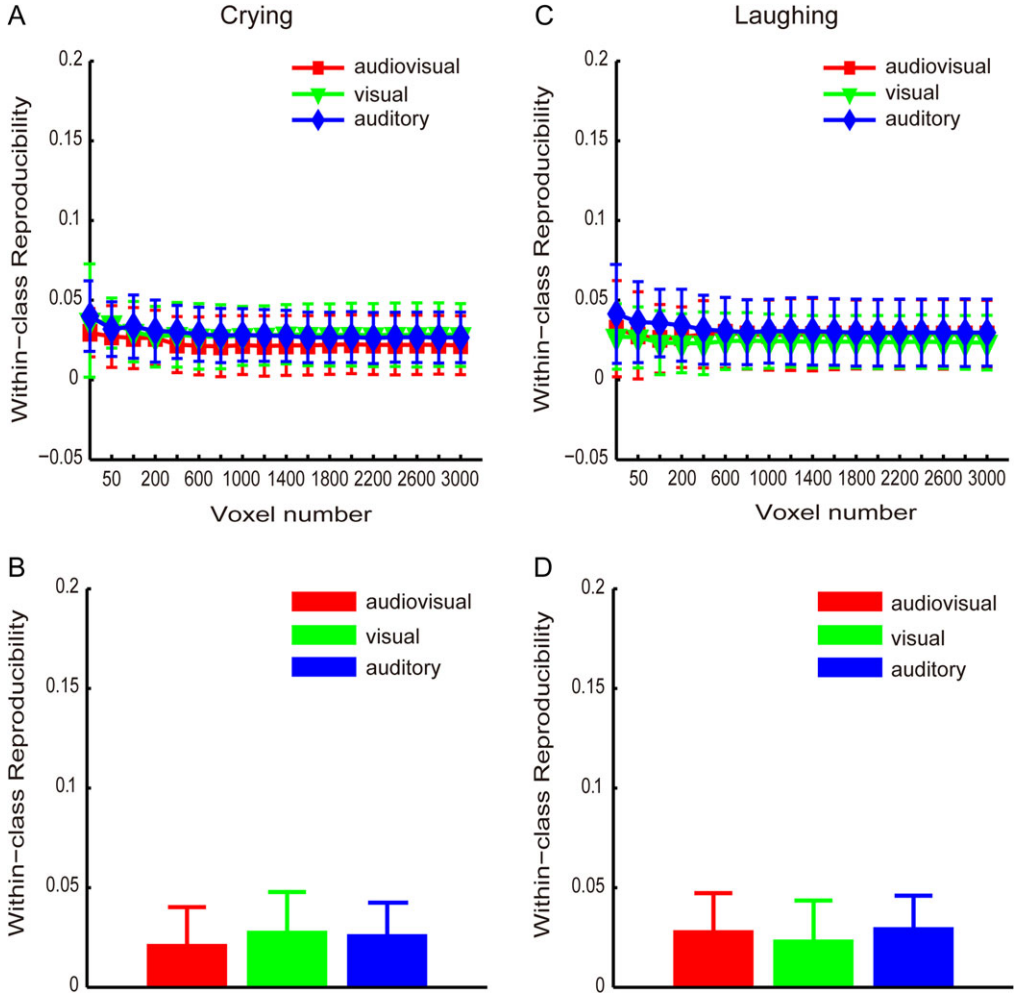

Figure 4. Reproducibility indices (means and standard errors across all subjects) of the brain patterns corresponding to the non-targets in the audiovisual, visual-only, and auditory-only stimulus conditions. (A) and (B): "crying" non-targets; (C) and (D): "laughing" non-targets. (A) and (C): curves of the reproducibility indices with respect to the numbers of voxels; (B) and (D): reproducibility indices obtained with 1600 voxels.

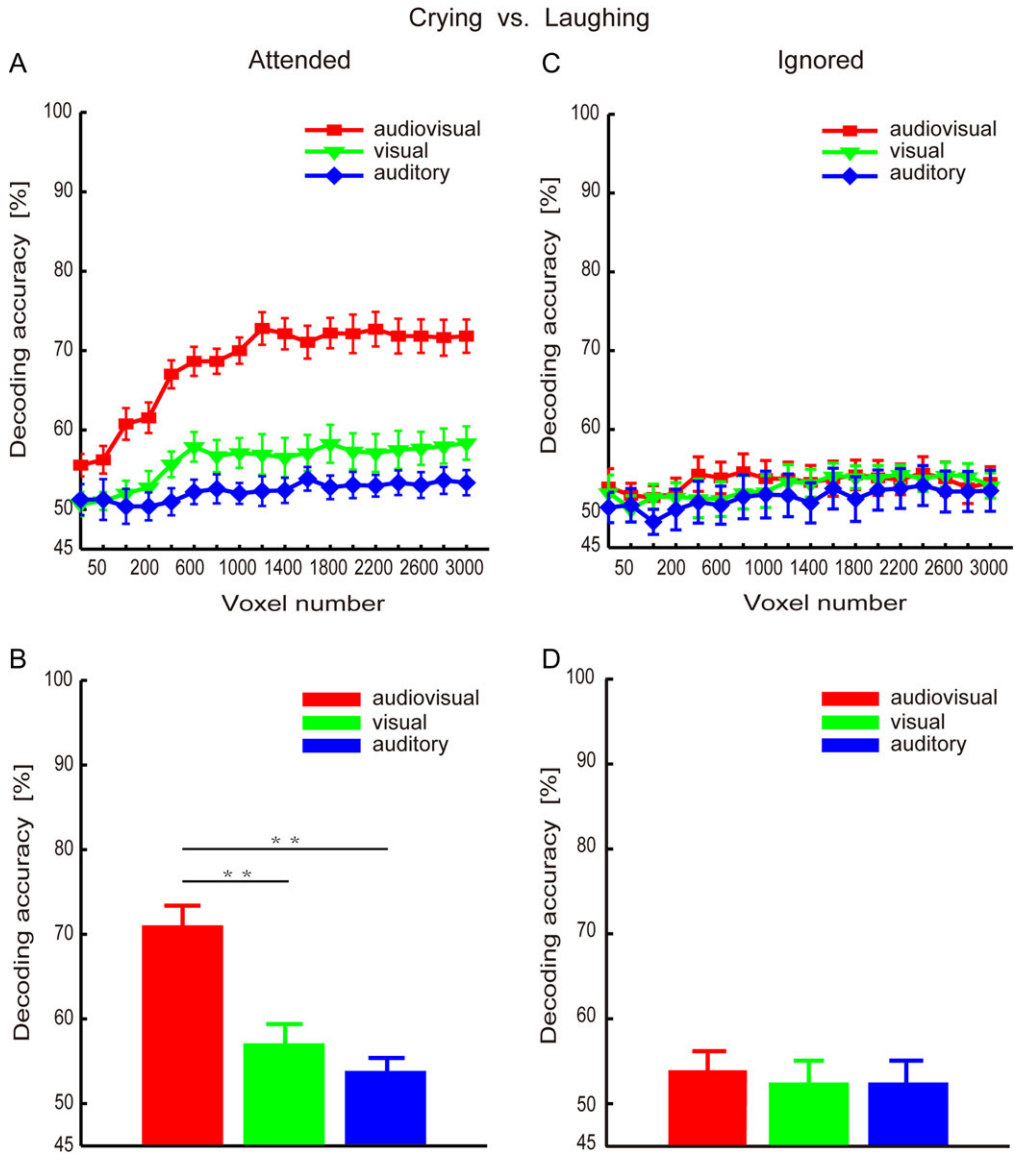

Figure 5. Decoding accuracy rates (means and standard errors across all subjects) for the audiovisual, visual-only, and auditory-only stimulus conditions. (*A*) and (*B*): Decoding accuracy rates for the targets; (*C*) and (*D*): Decoding accuracy rates for the non-targets; (*A*) and (*C*): Decoding accuracy curves with respect to the number of voxels. (*B*) and (*D*): Decoding accuracy rates based on 1600 voxels. Note: (i) The decoding accuracy rates were significantly higher for the audiovisual condition than for the visual-only and auditory-only conditions; (ii) There was no significant difference between the visual-only and auditory-only conditions.

Figure 5*B*. We observed that the decoding accuracies were higher for the audiovisual condition than for the visual-only condition and auditory-only conditions. We used 1600 selected voxels as an example to present the statistical results shown below. A 1-way repeated-measures ANOVA indicated a significant main effect of the stimulus condition ($P < 10^{-7}$, $F_{2,24} = 38.05$). Post hoc Bonferroni-corrected paired t-tests for the stimulus condition indicated that the decoding accuracy was significantly higher for the audiovisual condition than for the visual-only condition ($t(12) = 6.51$, $P < 10^{-4}$) and auditory-only condition ($t(12) = 7.70$, $P < 10^{-4}$). There was no significant difference between the visual-only condition and the auditory-only condition ($t(12) = 1.72$, $P = 0.33$).

For each experimental run, we also decoded the emotion categories (i.e., "crying" and "laughing") of the non-targets from the collected fMRI data. The decoding accuracy curves with the number of selected voxels (from 25 to 3000) are shown in Figure 5*C*, and the decoding results obtained from 1600 selected voxels are shown in Figure 5*D*. A 1-way repeated-measures ANOVA was performed to assess the decoding accuracy rates obtained with 1600 selected voxels and revealed that the decoding accuracies were not significantly higher for the audiovisual condition than for the visual-only and auditory-only conditions ($F_{2,24} = 0.23$, $P = 0.80$).

Informative Brain Areas

Using the data collected in the audiovisual condition, we obtained voxels that were informative for discriminating emotion categories (see Experimental Procedure and Methods). The distribution of these informative voxels is shown in Table 1 (25 clusters) and Figure 6. We also obtained the informative voxels for discriminating emotion categories in the visual-only

condition (27 clusters) and the auditory-only condition (23 clusters), which are not presented here.

Functional Connectivity

Using the collected data, a GLM analysis was performed to identify heteromodal areas in the left (cluster center: (−57, −45, 12)) and the right pSTS/MTG (cluster center: (60, −42, 15)) (see Supporting Materials). For each experimental run, we calculated the functional connectivities between the above heteromodal areas and the obtained brain areas that were informative for discriminating emotion categories (Table 1 shows the informative brain areas for the audiovisual run; see Experimental Procedure and Methods). These results are shown in Figure 7 and demonstrate that there were more functional connections between the bilateral heteromodal areas and the brain areas encoding the emotion feature for the audiovisual condition than for the visual-only and auditory-only conditions. Thus, compared with the visual-only and auditory-only stimuli, the audiovisual stimuli induced enhanced functional connectivity and thus regulated information flow between the bilateral heteromodal areas and the brain areas that encoded the emotion feature of the attended objects.

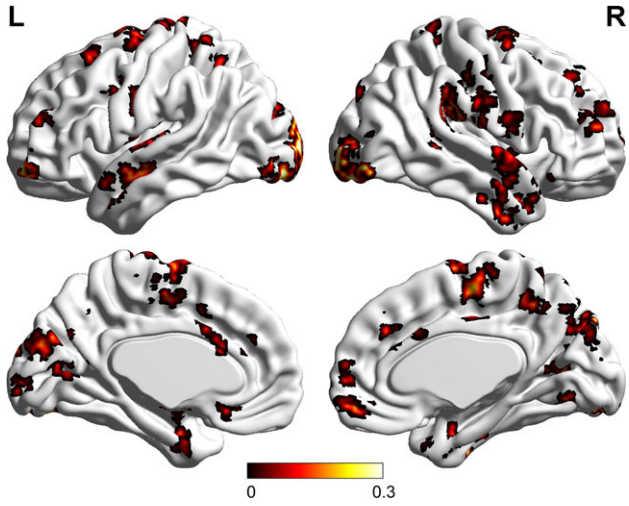

Figure 6. The distribution of the localized informative brain areas for discriminating emotion categories in the audiovisual condition. L: left; R: right; Colors: significant weights after group average, which reflected the importance of the voxels for decoding the emotion categories (see Experimental Procedure and Methods: Data processing).

Discussion

When the brain is faced with a cocktail party problem, visual inputs (e.g., face, lips, or the visuo-spatial information of the speaker) may play important roles, such as providing useful spatial cues to guide attention to the behaviorally relevant speaker (Haykin and Chen 2005; Kidd et al. 2005a; Senkowski et al. 2008), enhancing the ability of the auditory cortex to track the temporal speech envelope of the attended speaker (Zion-Golumbic et al. 2013a) and improving the intelligibility or resolving the perceptual ambiguity of speech in a noisy environment (Grant 2001; Schwartz et al. 2004). Multiple studies have also indicated that visual inputs affect neural responses to speech, both in early sensory cortices and in higher order speech-related areas (Besle et al. 2004; Davis et al. 2008; McGettigan et al. 2012).

In this study, we conducted an fMRI experiment to investigate the cocktail party problem in the audiovisual condition. Synthesized movie clips consisting of both video and audio recordings were used as the audiovisual stimuli in the fMRI experiment, and the video and audio portions were extracted from the synthesized movie clips for use as visual-only and auditory-only stimuli, respectively. Furthermore, each synthesized movie clip consisted of a combination of 2 movie clips (see Experimental Procedure and Methods). The subjects were required to attend to an object contained within each visual-only, auditory-only or audiovisual stimulus and to judge its emotion category (crying or laughing). First, the behavioral results reflected the advantages of audiovisual stimuli over the visual-only and auditory-only stimuli. To explore the neural representation effects of audiovisual inputs on the cocktail party problem, we assessed the emotion information of the stimuli encoded by the brain. Specifically, we calculated the reproducibility indices of brain patterns for the stimuli in each stimulus condition and decoded the emotion categories by applying an MVPA to the fMRI data. We found that the neural representations were enhanced for the attended audiovisual objects compared with the attended visual-only and auditory-only objects, which might partially explain the behavioral benefits of audiovisual inputs. Furthermore, we localized the informative brain areas that contributed to the enhanced neural representations and the heteromodal areas (pSTS/MTG) involved in audiovisual integration. We found that the functional connectivities between the informative areas and the heteromodal areas were enhanced. This might

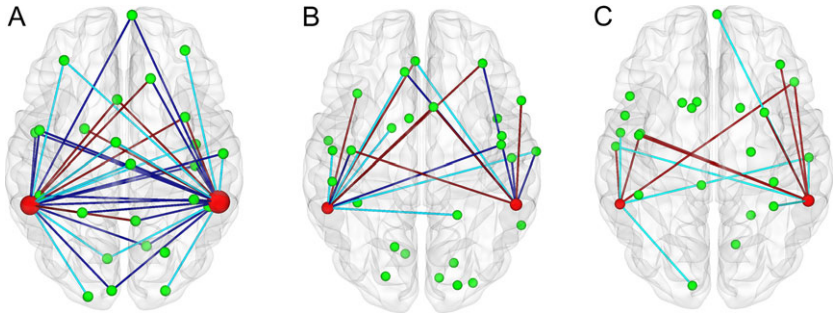

Figure 7. The functional connectivities between the heteromodal areas (left and right pSTS/MTG) and the brain areas encoding the emotion features of targets in the audiovisual (*A*), visual-only (*B*), and auditory-only (*C*) conditions. Green spheres: brain areas encoding the emotion features. Magenta spheres: heteromodal areas (the sizes represented the numbers of connections). Green lines: connections from the heteromodal areas to the informative brain areas (14, 6, and 7 connections for *A*, *B*, and *C*, respectively). Brown lines: connections from the informative brain areas to the heteromodal areas (8, 7, and 8 connections for *A*, *B*, and *C*, respectively). Blue lines: bi-directional connections (20, 6, and 0 connections for *A*, *B*, and *C*, respectively).

explain how neural representations of attended objects were enhanced by audiovisual inputs.

Behavioral Results

Our fMRI experiment simulated an audiovisual cocktail party environment, in which the behavioral effects were observed by comparing the audiovisual condition with the auditory-only one. Specifically, the behavioral accuracy rate was significantly higher, whereas the response time was significantly lower for the audiovisual condition than for the auditory-only condition. The poor behavioral performance in terms of the accuracy and response time for the auditory-only condition was likely due to scanner noise and the relatively difficult experimental task in the auditory-only condition, that is, it is more difficult to recognize emotion based exclusively on mixed voices than when facial images are presented. When the human brain faces with a cocktail party problem, visual inputs (e.g., face, lips, or the visuospatial information of the speaker) may play important roles, such as providing useful spatial cues to guide attention to the behaviorally relevant speaker (Haykin and Chen 2005; Kidd et al. 2005a; Senkowski et al. 2008) and improving the intelligibility or resolving the perceptual ambiguity of speech in a noisy environment (Grant 2001; Schwartz et al. 2004). The behavioral results in this study also indicated that the visual information not only facilitated selective attention but also played an important role in resolving perceptual ambiguity in the cocktail party environment. In our experiment, there was no significant difference of response accuracy between the visual-only and the audiovisual condition. This was because it was an easy task for the subjects to recognize the emotion of the visual-only stimuli (videos of faces) without adding noises into them, as in a real cocktail party environment. However, the behavioral benefits of audiovisual inputs over auditory-only inputs not only originated from the visual modality but also from the auditory modality. The main reasons included the following 3-folds. First, the response time was significantly lower for the audiovisual condition than for the visual-only condition. Second, there occurred audiovisual integration in the audiovisual condition, as shown by our GLM analysis (see also Supporting Materials). Considering single face-voice pairs, several studies have reported that audiovisual integration led to faster and more accurate categorization of emotion expressions (Collignon et al. 2008) and identity information processing (Campanella and Belin 2007; Schweinberger et al. 2007). These results may partially support our behavioral observations although each audiovisual stimulus in our experiment contained 2 face-voice pairs (the subjects selectively attended 1). Third, compared with the visual-only and auditory-only conditions, the neural representations of the emotion features of the attended objects were enhanced, as shown in our MVPA analysis.

Neural Representations Enhanced by Audiovisual Stimuli in the Cocktail Party Environment

By focusing on distributed activity patterns, MVPA approaches enable us to separate, localize, and analyze spatially distributed brain patterns that are generally weak (Haxby et al. 2011, 2014). Haxby et al. (2011) introduced a common framework organized around the concept of high-dimensional representational spaces, which may integrate the common MVPA methods, including multivariate pattern (MVP) classification, representational similarity analysis (RSA), and stimulus-model-based encoding and decoding. In a neural representational space, each neural response is expressed as a vector with each pattern feature being a measure of local activity, for example, a voxel or a single neuron (Haxby et al. 2014). Based on the neural activity patterns, vectors in a neural representational space, MVP classification demonstrates reliable distinctions among brain states, whereas RSA analyzes the geometry of representations in terms of the similarities among brain states (Kriegeskorte et al. 2008). Furthermore, stimulus-model-based encoding and decoding algorithms can predict the patterns of neural response to novel stimuli and decode stimulus features of stimuli based on the novel neural response patterns (Kay et al. 2008; Mitchell et al. 2008). In this study, we assessed the neural representations of emotion features of the stimuli using an MVPA method. Specifically, we calculated the reproducibility indices of brain patterns for the stimuli in each stimulus condition and decoded the emotion categories from the fMRI data. Our MVPA results revealed that the brain patterns corresponding to the attended objects yielded a significantly improved reproducibility index for the audiovisual condition (face-voice pairs) than for the visual-only and auditory-only conditions (Fig. 3), whereas reproducibility was not improved for the brain patterns associated with the ignored objects (Fig. 4). Furthermore, our results indicated that, when considering the attended objects, the decoding accuracy rate calculated from the fMRI data corresponding to the audiovisual condition was significantly higher than that obtained in the visual- and auditory-only conditions (Fig. 5*A*, *B*). An increased decoding accuracy in the audiovisual condition was not observed for the ignored objects (Fig. 5*C*, *D*). Together, these results demonstrated the neural modulation effects of audiovisual integration on the cocktail party problem. Specifically, audiovisual integration improved the reproducibility of the brain patterns corresponding to the attended objects and strengthened the differentiation between the 2 classes of brain patterns corresponding to the 2 emotion categories of the attended objects, which thus implied an enhancement of neural representation.

Audiovisual inputs can modulate the neural representations associated with the stimuli or tasks, which has been shown in several studies. In particular, regarding motion perception, several studies based on intracranial recordings, fMRI and EEG in humans demonstrated that audiovisual information could enhance sensory representations in occipital regions (Poirier et al. 2005; Alink and Singer 2008; Sadaghiani et al. 2009; Kayser et al. 2017). Bonath et al. performed an fMRI study and showed that the ventriloquist illusion and its neural/spatial representation in the auditory cortex could be modulated by audiovisual synchrony (Bonath et al. 2014). Using human EEG recordings, Crosse et al. examined how visual speech enhanced the cortical representation of auditory speech and found that the influence of visual inputs on the neural tracking of the audio speech signals was significantly greater in noisy than in quiet listening conditions, consistent with the principle of inverse effectiveness (Crosse et al. 2016). A neurophysiological model was proposed to explain that a visual stimulus could modulate the activities of auditory cortical neurons via direct feedforward or lateral visual inputs into auditory cortex, providing a mechanism through which auditory and visual stimuli could be bound together, enhancing their representations (Bizley et al. 2016). In our experiment, we also observed that the neural representations were enhanced for the attended audiovisual objects compared with the attended visual-only and auditory-only objects, which was partially in line with the aforementioned results. However, our approach differed from the studies regarding the effects of audiovisual inputs on neural representations in 3-folds: First, we performed the fMRI experiment to simulate an audiovisual cocktail party environment and explored the effects of audiovisual

inputs on solving the cocktail party problem. Second, we used an MVPA method to directly extract brain patterns that carried the semantic information of emotion categories of the stimuli from the fMRI data, and assessed the neural representations of emotion features by calculating the reproducibility indices of these brain patterns and decoding the emotion categories of the stimuli based on these brain patterns. Third, combining the behavioral observations, we concluded that the phenomenon of neural representation enhancement might be used by the human brain to solve the cocktail party problem in audiovisual environments and at least partially explain the behavioral benefits of audiovisual inputs.

Informative Brain Areas Involved in the Enhanced Neural Representations

Using the fMRI data collected in the audiovisual stimulus condition, we localized the voxels that were informative for the emotion category decoding, as shown in Table 1. The identified informative brain areas included the inferior occipital gyrus, the right/left fusiform gyrus, the right/left STG, the left MTG, the right precuneus, and the right/left medial frontal gyrus. These regions have been reported as contributing to face perception. In particular, it has been suggested that the inferior occipital gyrus contributes to the early stage of face information processing, whereby information is further transferred to the STS and the fusiform gyrus (Sergent et al. 1992; Puce et al. 1998; Haxby et al. 2000; Golby et al. 2001; Freeman et al. 2010). The identified informative brain areas also included the amygdala, the insula and the precentral gyrus, which were proven to play key roles in emotion processing in previous studies (LaBar et al. 1998; Iidaka et al. 2001; Vuilleumier et al. 2001; Pessoa et al. 2002; Phillips et al. 2004; Stein et al. 2007; Fusar-Poli et al. 2009). In trials showing emotional expressions, a univariate analysis indicated brain activation primarily in bilateral amygdala, fusiform gyrus, MTG/STS, and inferior occipital gyrus (Jansma et al. 2014). Therefore, our results are partially supported by these existing evidences related to face information processing.

Functional Connectivities Between Heteromodal Areas Involved in Audiovisual Integration and the Informative Areas Associated with the Enhanced Neural Representations

The neural mechanisms of audiovisual integration have been extensively investigated using neuroimaging techniques. For instance, several brain regions, including the pSTS/MTG, have been identified to be associated with audiovisual integration (Calvert et al. 2000; Frassinetti et al. 2002; Bushara et al. 2003). Furthermore, it has been shown that the neural activity of the pSTS/MTG can be enhanced when congruent audiovisual stimuli are presented than when visual-only and auditory-only stimuli are presented. In contrast, the enhanced neural activity of the pSTS/MTG may indicate audiovisual integration (Frassinetti et al. 2002; Bushara et al. 2003; Calvert and Thesen 2004; Macaluso and Driver 2005). In this study, we performed GLM analysis based on the fMRI data collected in the visual-only, auditory-only, and audiovisual conditions (see Supporting Materials). We observed increased neural activity of the pSTS/MTG for the audiovisual condition than for the visual-only and auditory-only conditions (see Figure S1, Supporting Materials), suggesting the occurrence of audiovisual integration. The heteromodal areas pSTS/MTG were also identified.

Audiovisual integration emerges at multiple processing stages within the cortical hierarchy including associative auditory and visual areas and heteromodal areas such as pSTS/MTG. There exist feedforward and lateral connections in early visual and auditory areas, as well as feedback connections from higher brain areas such as the multisensory STS and IPS, which play an important role in mediating predictive coding or object recognition in audiovisual conditions (Bizley et al. 2016). In particular, Lewis and Noppeney reported that audiovisual synchrony improved motion discrimination, which was associated with enhanced connectivity between early visual and auditory areas (Lewis and Noppeney 2010). Werner and Noppeney demonstrated that the automatic auditory response amplification was mediated via both direct and indirect (via STS) connectivity to visual cortices (Werner and Noppeney 2010). In an fMRI study, Noesselt et al. showed some connectivity between visual and auditory cortices and multisensory STS, and suggested a significantly increased influence from multisensory STS on primary visual and auditory areas A1 and V1 specifically during audiovisual temporal correspondence (Noesselt et al. 2007). It was also showed that effective connectivity between audiovisual integration areas and associative auditory and visual cortices was enhanced during audiovisual stimulation of emotional voices and faces (Kreifelts et al. 2007). Furthermore, by performing several fMRI and behavioral experiments, Nath and Beauchamp found increased functional connectivity between the STS and auditory cortex when the auditory modality was more reliable (less noisy) and increased functional connectivity between the STS and visual cortex when the visual modality was more reliable (Nath and Beauchamp 2011). Taken together, these results offered further insight into the neural process accomplishing multimodal integration. In the current study, through the Granger causal connectivity analysis, we found that the functional connectivities between the heteromodal areas pSTS/MTG and the identified informative brain areas involved in the enhanced neural representations were augmented in the audiovisual condition compared with the visual-only and auditory-only conditions (Fig. 7). This finding first extended the above results mainly involving the connectivities between the visual and auditory cortices and the heteromodal areas. Furthermore, our result indicated that the enhancement of neural representation was associated with audiovisual integration and might explain how neural representations of attended objects were enhanced by audiovisual inputs. Specifically, our brains might enhance the neural representations of the features of the attended audiovisual objects by modulating functional connectivity and information flows between the heteromodal areas and the informative brain areas.

Roles of Selective Attention for the Enhancement of the Neural Representations of Emotion Features

Multiple behavioral studies with auditory-only cocktail party paradigms showed that selective attention facilitated sound segregation in our brains (Elhilali et al. 2009; Ahveninen et al. 2011; Zion-Golumbic and Schroeder 2012). Furthermore, neuroimaging studies have suggested that selective attention enhances the cortical tracking of attended speech streams by modulating both low-frequency phase and high-frequency amplitude fluctuations, which are beneficial for achieving sound segregation (Ding and Simon 2012; Mesgarani and Chang 2012; Zion-Golumbic and Schroeder 2012; Zion-Golumbic et al. 2013b). In fact, the modulatory effects of selective attention on the neural representations of speech do not merely reflect the acoustical properties of the stimuli but are instead closely related to the perceived aspects of

speech (Mesgarani and Chang 2012)—although the latter has not been extensively investigated. In the audiovisual cocktail party condition, selective attention is involved in both the visual and auditory modalities, where spatial unmasking can be a result of top-down processes that aid in selectively facilitating spatial attention to the target (Freyman et al. 2001; Kidd et al. 2005b; Rakerd et al. 2006; Huang et al. 2009). In this study, we observed the modulation effects of selective attention on the neural representations of the emotion features of the stimuli. Specifically, comparing the audiovisual condition with the visual-only and auditory-only conditions, we found that the neural representations of the emotion features of the attended objects instead of the ignored objects were enhanced by audiovisual inputs. Therefore, this enhancement of neural representations induced by audiovisual inputs was selective, which reflected the roles of selective attention.

Limitations and Future Study

Our experimental paradigm simulated an audiovisual cocktail party environment. However, we did not add auditory noises, including reverberation from a cocktail party environment with much more than 2 speakers, into our experimental stimuli. This was mainly due to the high fMRI scanner noise always existing in our experiment. Furthermore, "crying" and "laughing" voices were used as stimuli instead of natural speech stimuli, and the emotion categories of the crying and laughing stimuli other than the speech contents were decoded in this study. In the future, fMRI experiments based on natural speech stimuli and noises from a cocktail party environment should be conducted to further explore the effects of audiovisual integration on solving the cocktail problem. To achieve this objective, we will also develop new MVPA algorithms for effectively decoding the contents of speech from the fMRI data, which are collected with an experimental paradigm closer to a real cocktail party environment. Additionally, we will increase the number of subjects to enhance the experimental results.

In summary, our experimental results indicated that the neural representations of the semantic features (emotion features) of the attended objects (face-voice pairs) instead of the unattended objects were enhanced in the audiovisual cocktail party conditions compared with the visual-only and auditory-only conditions. This enhancement effect might have arisen from the enhanced functional connectivities between the informative brain areas, which were involved in the neural representations, and the heteromodal areas (pSTS/MTG), suggesting that it was associated with audiovisual integration. Furthermore, the selectivity of neural representation enhancement regarding the attended objects was due to the roles of the object-selective attention. These findings may partially explain the neural mechanism of the behavioral benefits that result from audiovisual inputs in cocktail party environments.

Supplementary Material

Supplementary data are available at *Cerebral Cortex* online.

Funding

This work was supported by the National Key Research and Development Program of China (Grant no. 2017YFB1002505), the National Natural Science Foundation of China (Grant nos. 61633010 and 91420302), and Guangdong Natural Science Foundation (Grant no. 2014A030312005). A. Cichocki's work was partially supported by the Ministry of Education and Science of the Russian Federation (Grant no. 14.756.31.0001).

References

Ahveninen J, Hamalainen M, Jaaskelainen IP, Ahlfors SP, Huang S, Lin FH, Raij T, Sams M, Vasios CE, Belliveau JW. 2011. Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proc Natl Acad Sci USA. 108:4182–4187.

Alink A, Singer WL. 2008. Capture of auditory motion by vision is represented by an activation shift from auditory to visual motion cortex. J Neurosci. 28:2690–2697.

Bell AJ, Sejnowski TJ. 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7:1129–1159.

Besle J, Fort A, Delpuech C, Giard MH. 2004. Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci. 20:2225–2234.

Bishop CW, Miller LM. 2009. A multisensory cortical network for understanding speech in noise. J Cognitive Neurosci. 21: 1790–1804.

Bizley JK, Maddox RK, Lee AKC. 2016. Defining auditory-visual objects: behavioral tests and physiological mechanisms. Trends Neurosci. 39:74.

Bonath B, Noesselt T, Krauel K, Tyll S, Tempelmann C, Hillyard SA. 2014. Audio-visual synchrony modulates the ventriloquist illusion and its neural/spatial representation in the auditory cortex. Neuroimage. 98:425–434.

Brown GD, Yamada S, Sejnowski TJ. 2001. Independent component analysis at the neural cocktail party. Trends Neurosci. 24:54–63.

Bushara KO, Hanakawa T, Immisch I, Toma K, Kansaku K, Hallett M. 2003. Neural correlates of cross-modal binding. Nat Neurosci. 6:190–195.

Calvert GA, Campbell R, Brammer MJ. 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol. 10:649–657.

Calvert GA, Thesen T. 2004. Multisensory integration: methodological approaches and emerging principles in the human brain. J Physiol Paris. 98:191–205.

Campanella S, Belin P. 2007. Integrating face and voice in person perception. Trends Cogn Sci. 11:535–543.

Cherry EC. 1953. Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am. 25: 975–979.

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

Crosse MJ, Di LG, Lalor EC. 2016. Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J Neurosci. 36:9888–9895.

Davis C, Kislyuk D, Kim J, Sams M. 2008. The effect of viewing speech on auditory speech processing is different in the left and right hemispheres. Brain Res. 1242:151–161.

Ding N, Simon JZ. 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol. 107:78–89.

Du Y, He Y, Ross B, Bardouille T, Wu XH, Li L, Alain C. 2011. Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. Cereb Cortex. 21:698–707.

Elhilali M, Xiang J, Shamma SA, Simon JZ. 2009. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol. 7:e1000129.

Frassinetti F, Bolognini N, Ladavas E. 2002. Enhancement of visual perception by crossmodal visuo-auditory interaction. Exp Brain Res. 147:332–343.

Freeman JB, Rule NO, Adams RB, Ambady N. 2010. The neural basis of categorical face perception: graded representations of face gender in fusiform and orbitofrontal cortices. Cereb Cortex. 20:1314–1322.

Freyman RL, Balakrishnan U, Helfer KS. 2001. Spatial release from informational masking in speech recognition. J Acoust Soc Am. 109:2112–2122.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ. 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp. 2:189–210.

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

Ghio M, Vaghi MM, Perani D, Tettamanti M. 2016. Decoding the neural representation of fine-grained conceptual categories. Neuroimage. 132:93–103.

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

Golby AJ, Poldrack RA, Brewer JB, Spencer D, Desmond JE, Aron AP, Gabrieli JDE. 2001. Material-specific lateralization in the medial temporal lobe and prefrontal cortex during memory encoding. Brain. 124:1841–1854.

Grant KW. 2001. The effect of speechreading on masked detection thresholds for filtered speech. J Acoust Soc Am. 109: 2272–2275.

Hamilton JP, Chen G, Thomason ME, Schwartz ME, Gotlib IH. 2011. Investigating neural primacy in major depressive disorder: multivariate Granger causality analysis of resting-state fMRI time-series data. Mol Psychiatry. 16:763–772.

Haxby JV, Connolly AC, Guntupalli JS. 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu Rev Neurosci. 37:435–456.

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ. 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron. 72:404–416.

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

Haykin S, Chen Z. 2005. The cocktail party problem. Neural Comput. 17:1875–1902.

Hopfinger JB, Buonocore MH, Mangun GR. 2000. The neural mechanisms of top-down attentional control. Nat Neurosci. 3:284–291.

Huang Y, Huang Q, Chen X, Wu XH, Li L. 2009. Transient auditory storage of acoustic details is associated with release of speech from informational masking in reverberant conditions. J Exp Psychol Human. 35:1618–1628.

Iidaka T, Omori M, Murata T, Kosaka H, Yonekura Y, Okada T, Sadato N. 2001. Neural interaction of the amygdala with the prefrontal and temporal cortices in the processing of facial expressions as revealed by fMRI. J Cognitive Neurosci. 13: 1035–1047.

Jansma H, Roebroeck A, Münte TF. 2014. A network analysis of audiovisual affective speech perception. Neuroscience. 256: 230–241.

Jeong JW, Diwadkar VA, Chugani CD, Sinsoongsud P, Muzik O, Behen ME, Chugani HT, Chugani DC. 2011. Congruence of happy and sad emotion in music and faces modifies cortical audiovisual activation. Neuroimage. 54:2973–2982.

Kay KN, Naselaris T, Prenger RJ, Gallant JL. 2008. Identifying natural images from human brain activity. Nature. 452: 352–355.

Kayser SJ, Philiastides MG, Kayser C. 2017. Sounds facilitate visual motion discrimination via the enhancement of late occipital visual representations. Neuroimage. 148:31–41.

Kidd G, Arbogast TL, Mason CR, Gallun FJ. 2005a. The advantage of knowing where to listen. J Acoust Soc Am. 118:3804–3815.

Kidd G, Mason CR, Brughera A, Hartmann WM. 2005b. The role of reverberation in release from masking due to spatial separation of sources for speech identification. Acust Acta Acust. 91:526–536.

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

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. Proc Natl Acad Sci USA. 103: 3863–3868.

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2:4.

LaBar KS, Gatenby JC, Gore JC, LeDoux JE, Phelps EA. 1998. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron. 20: 937–945.

Lewicki MS, Sejnowski TJ. 2000. Learning overcomplete representations. Neural Comput. 12:337–365.

Lewis R, Noppeney U. 2010. Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. J Neurosci. 30:12329–12339.

Li Y, Cichocki A, Amari S. 2004. Analysis of sparse representation and blind source separation. Neural Comput. 16: 1193–1234.

Li Y, Long J, Huang B, Yu T, Wu W, Liu Y, Liang C, Sun P. 2015. Crossmodal integration enhances neural representation of task-relevant features in audiovisual face perception. Cereb Cortex. 25:384–395.

Li Y, Wang F, Huang B, Yang W, Yu T, Talsma D. 2016. The modulatory effect of semantic familiarity on the audiovisual integration of face-name pairs. Hum Brain Mapp. 37: 4333–4348.

Müller VI, Cieslik EC, Turetsky BI, Eickhoff SB. 2012. Crossmodal interactions in audiovisual emotion processing. Neuroimage. 60:553–561.

Macaluso E, Driver J. 2005. Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci. 28:264–271.

Macaluso E, George N, Dolan R, Spence C, Driver J. 2004. Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage. 21:725–732.

McDermott JH. 2009. The cocktail party problem. Curr Biol. 19: R1024–R1027.

McGettigan C, Faulkner A, Altarelli I, Obleser J, Baverstock H, Scott SK. 2012. Speech comprehension aided by multiple modalities: Behavioural and neural interactions. Neuropsychologia. 50:762–776.

Mesgarani N, Chang EF. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 485:233–236.

MicheyI C, Oxenham AJ. 2010. Objective and subjective psychophysical measures of auditory stream integration and segregation. J Assoc Res Otolaryngol. 11:709–724.

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meanings of nouns. Science. 320: 1191–1195.

Nath AR, Beauchamp MS. 2011. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. J Neurosci. 31:1704–1714.

Nichols T, Hayasaka S. 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat Methods Med Res. 12:419–446.

Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze HJ, Driver J. 2007. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J Neurosci. 27: 11431–11441.

Pereira F, Mitchell T, Botvinick M. 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage. 45:S199–S209.

Pessoa L, McKenna M, Gutierrez E, Ungerleider LG. 2002. Neural processing of emotional faces requires attention. Proc Natl Acad Sci USA. 99:11458–11463.

Phillips ML, Williams LM, Heining M, Herba CM, Russell T, Andrew C, Brammer MJ, Williams SCR, Morgan M, Young AW, et al. 2004. Differential neural responses to overt and covert presentations of facial expressions of fear and disgust. Neuroimage. 21:1484–1496.

Poirier C, Collignon O, Devolder AG, Renier L, Vanlierde A, Tranduy D, Scheiber C. 2005. Specific activation of the V5 brain area by auditory motion processing: an fMRI study. Cognitive Brain Res. 25:650–658.

Polyn SM, Natu VS, Cohen JD, Norman KA. 2005. Category-specific cortical activity precedes retrieval during memory search. Science. 310:1963–1966.

Puce A, Allison T, Bentin S, Gore JC, McCarthy G. 1998. Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci. 18:2188–2199.

Rakerd B, Aaronson NL, Hartmann WM. 2006. Release from speech-on-speech masking by adding a delayed masker at a different location. J Acoust Soc Am. 119:1597–1605.

Sadaghiani S, Maier JX, Noppeney U. 2009. Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. J Neurosci. 29:6490–6499.

Schwartz J-L, Berthommier F, Savariaux C. 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition. 93:B69–B78.

Schweinberger SR, Kloth N, Robertson DMC. 2011. Hearing facial identities: Brain correlates of face-voice integration in person identification. Cortex. 47:1026–1037.

Schweinberger SR, Robertson D, Kaufmann JM. 2007. Hearing facial identities. Q. J Exp Psychol. 60:1446–1456.

Senkowski D, Saint-Amour D, Gruber T, Foxe JJ. 2008. Look who's talking: the deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. Neuroimage. 43:379–387.

Sergent J, Ohta S, MacDonald B. 1992. Functional neuroanatomy of face and object processing. A positron emission tomography study. Brain. 115:15–36.

Seth AK. 2010. A MATLAB toolbox for Granger causal connectivity analysis. J Neurosci Meth. 186:262–273.

Stein MB, Simmons AN, Feinstein JS, Paulus MP. 2007. Increased amygdala and insula activation during emotion processing in anxiety-prone subjects. Am J Psychiat. 164: 318–327.

Vuilleumier P, Armony JL, Driver J, Dolan RJ. 2001. Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. Neuron. 30:829–841.

Werner S, Noppeney U. 2010. Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. J Neurosci. 30: 2662–2675.

Zeng LL, Shen H, Liu L, Wang LB, Li BJ, Fang P, Zhou ZT, Li YM, Hu DW. 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain. 135:1498–1507.

Zion-Golumbic E, Cogan GB, Schroeder CE, Poeppel D. 2013a. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J Neurosci. 33: 1417–1426.

Zion-Golumbic E, Schroeder CE. 2012. Attention modulates 'speech-tracking' at a cocktail party. Trends Cogn Sci. 16: 363–364.

Zion-Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, et al. 2013b. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron. 77:980–991.