Homeostatic brain network mechanisms facilitate perceptual binding of audio-visual speech

Abhishek Mukherjee[†], Soibam Shyamchand Singh[†], Dipanjan Ray,
Partha Raghunathan, and Arpan Banerjee[*]

National Brain Research Centre, NH8, Manesar, Gurgaon 122052, India.

[†] *These authors contributed equally to this work (co-first authors).*

[*] *To whom correspondence should be addressed. E-mail: arpan@nbrc.ac.in.*

In daily lives, speech perception requires binding of spatiotemporally disjoint auditory and visual cues. On the other hand, functional segregation and integration are the two complementary mechanisms that capture brain information processing. Here, we demonstrate using fMRI recordings that subjective perceptual experience of multisensory speech stimuli is dependent on a homeostatic balance of segregation and integration mechanisms. Previous reports conceptualized posterior superior temporal sulcus as the key brain region for binding signals from multiple sensory streams. However, we report an enhancement of segregated information processing in distributed brain regions, defined as the perceptual binding network. The seed-based whole brain functional connectivity of each node in this network was anti-correlated with higher propensity for illusory perception. Interestingly, the perceptual binding network was anti-correlated with other intrinsic brain networks, such as dorsal attention and default mode networks during cross-modal perception. The pattern disappeared for people who rarely reported the illusory perception, further strengthening the hypothesis of homeostatic balance. The cognitive theories of Bayesian causal inference and predictive coding hypothesis could explain the balance of segregative and integrative mechanisms during cross-modal perception.

Keywords

multisensory, whole-brain functional connectivity, pSTS, fMRI, segregation, integration, homeostasis, perception, predictive coding, Bayesian

Introduction

The illusion of a puppet vocalizing human speech that in reality comes from the sounds of a ventriloquist has intrigued mankind for ages (Alais and Burr, 2004). While spatial incongruency is a well-studied aspect of ventriloquism, perceptual binding of temporally dissonant stimuli is subtle (Bertelson and Aschersleben, 2003), such as when viewing a poorly dubbed video our perceptual system immediately perceives the asynchrony in sound and lip movements even when we do not know the language. Some previous studies have also reported that temporal dissonance within a certain

window between multiple sensory components can be integrated to facilitate illusory perceptual experience (Bertelson and Aschersleben, 2003; Munhall et al., 1996; Shams et al., 2002; Stevenson et al., 2012), but do collective behavior of brain networks modulate perceptual binding is an unanswered question.

In this article, we hypothesised that integration of information among relevant brain areas is complementary to activations in each of them individually for the cross-modal perceptual binding necessary for the emergence of illusory perception to happen, and this relationship may be parametrically controlled by behavioral contexts. Previous studies have used illusory perceptual experience from the paradigms of sound-flash illusion (Watkins et al., 2007) and McGurk illusion (McGurk and Macdonald, 1976) to suggest posterior superior temporal sulcus (pSTS) as the key hub for cross-modal percept binding (Beauchamp et al., 2004; Nath and Beauchamp, 2011), whereas other studies have not found pSTS activations to be enhanced specific to temporal synchrony of auditory and visual modalities (Jones and Callan, 2003). From a large-scale network perspective, noisy stimulus in one modality has also been shown to increase functional brain connectivity between pSTS and areas processing the other modality; for example, a blurry video will increase strength of pSTS–auditory connectivity (Nath and Beauchamp, 2011). Such shifts in connectivity arising from stimulus associations can be linked to causal inference models of perception (Koerding et al., 2007; Magnotti et al., 2018) and lately biophysically realistic network mechanisms have been proposed (Cuppini et al., 2017; Kumar et al., 2020; Thakur et al., 2016). Nonetheless, a cohesive story of how subjective perceptual binding emerges from synergistic functional integration and segregation mechanisms in brain networks (Tomasi et al., 2014) is currently elusive.

Using functional magnetic resonance imaging (fMRI) recordings from human participants ($N = 52$; 25 males, 27 females; mean age = 24.5 years, SD = 3.12) we explored how functional integration and segregation (Friston et al., 1993) facilitate subjective cross-modal perception. The McGurk incongruent video—auditory /pa/ superimposed on the lip movement of /ka/ from a human vocalizer (McGurk and Macdonald, 1976)—was shown to volunteers. To parametrically control the subjective perception, different versions of this video were presented with temporal lags introduced between the auditory and visual sensory streams ranging from -300 (A leading V) to 450 ms (V leading A) in intervals of 150 ms (Fig. 1*a*), along with a matched number of congruent /ta/ (auditory /ta/–visual /ta/) presented at 0 ms lag. The choosing of an asymmetrical range of audio-visual (AV) lags ($-$L2, $\ldots$, L0, $\ldots$, L3) was guided by findings from a previous study (Munhall et al., 1996) reporting about a bias in perceptual experience toward visual-leading scenarios. Volunteers reported whether they heard /ta/, /pa/, or "others" in a forced-choice paradigm. Thus, a behavioral control of subjective illusory /ta/ perception, hallmark of cross-modal perceptual binding, was achieved with AV lags as the control parameter (Fig. 1*b* and Fig. S1). Subsequently, the fMRI data were investigated via two alternative routes. First, the activations maps (Friston et al., 1994) were extracted that quantify *segregated information processing* in each AV lag block, capturing stimulus-driven changes. Further, the AV lag blocks from each participant were sorted in terms of response percentages to construct maximum and minimum illusory perception blocks (Max /ta/ and Min /ta/). Once compiled, the activation maps in Max /ta/ and Min /ta/ were compared against Pure /ta/, thus capturing the brain activations sorted vis-à-vis perceptual response. Second, to capture the *functional integration mechanisms* (Friston et al., 1993) BOLD time series data were extracted, for both stimulus-based and perception-based routes, and pairwise Pearson correlation analysis between each active seed region and all other remaining brain parcels was performed (see *Materials and Methods*).

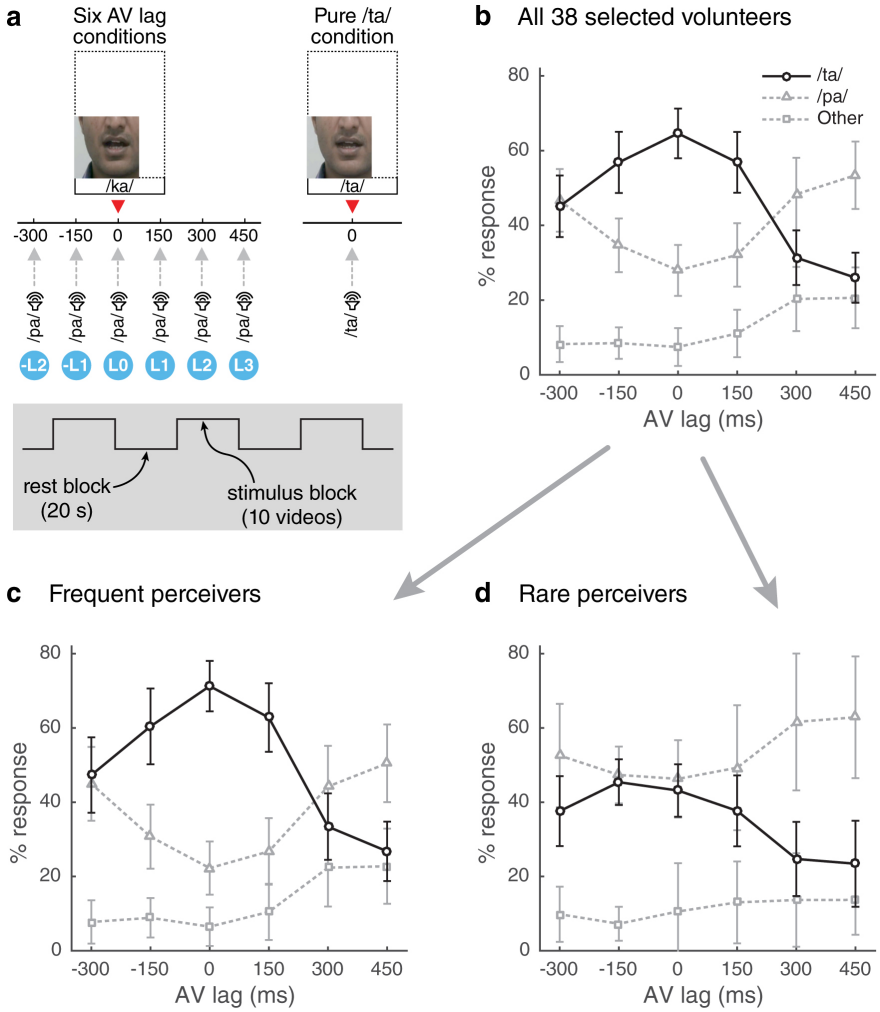


Figure 1. (*a*) Stimulus design of the study. (*b*) Behavioral responses averaged across all 38 selected volunteers. Behavioral responses averaged across 29 frequent perceivers (*c*), who self-reported illusory /ta/ perception for at least 60% of all responses in any AV lag condition, and averaged across the remaining 9 volunteers (*d*) who were categorized as rare perceivers. Error bars represent the standard error of the mean (95% CI) across the respective population.

Results

Parametric Control of Illusory Perception. More than chance level responses to illusory /ta/ (cross-modal perception) were observed in all participants at multiple AV lags for the McGurk incongruent stimulus, demonstrating that indeed the degree of AV temporal synchrony affects subjective perception (Fig. 1 and Fig. S1). In order to study the neural basis of subjective variability, we chose to categorize participants who reported perceiving /ta/ for at least 60% of the total responses at any AV lag as *frequent perceivers* ($N = 42$) and the participants who did not meet this criterion as *rare perceivers* ($N = 10$). Some participants had to be removed from further analysis (see *Materials and Methods* for detailed reasons).

Analysis of variance (ANOVA) performed on frequent perceivers ($N = 29$) revealed that response percentage of illusory /ta/ was significantly varying at 95% confidence levels (used in all

statistical significance tests in the article) across AV lags ($F(5, 166) = 14.4896, p < 0.0001$). Frequent perceivers also showed significant variation across AV lags for unisensory /pa/ responses ($F(5, 166) = 5.5067, p = 0.0001$) as well as for "other" responses ($F(5, 166) = 3.4140, p = 0.0058$). On the other hand, rare perceivers ($N = 9$) showed significant variation only for cross-modal /ta/ perception ($F(5, 47) = 3.5425, p = 0.0084$) but not for /pa/ ($F(5, 47) = 0.8678, p = 0.5098$) or "other" responses ($F(5, 47) = 0.220, p = 0.9521$). Welch's t-test was further performed to investigate the differences in perceptual responses between frequent and rare perceivers for each AV lag condition. Significant differences in illusory /ta/ perception between frequent and rare perceivers occurred at $-$L1 ($p = 0.0217$), L0 ($p < 0.0001$), and L1 ($p = 0.0016$). Unisensory /pa/ perception also showed significant differences at $-$L1 ($p = 0.0110$), L0 ($p = 0.0027$), and L1 ($p = 0.0468$) (Fig. S1). There were no significant differences in the reporting of "other" response across all conditions.

Brain Mapping of Perceptual Binding. Analysis of fMRI images using statistical parametric mapping (SPM) software (Friston et al., 1994) was performed with the task (AV lag blocks) and baseline (inter-trial rest blocks) conditions. Group-level inference from random effect analysis revealed that frequent perceivers ($N = 29$) displayed maximal volume of significantly active brain areas during L0 and L1 (Fig. 2). The extent of task-activated brain areas diminished bilaterally with increasing AV separation (Fig. 2 and Table S1). When sorted by perceptual category, Max /ta/ condition showed the maximal brain activation compared to both Min /ta/ and Pure /ta/ conditions. The task-activated clusters in frequent perceivers could be localized to 32 brain regions of interest (ROIs), with reference to the Brainnetome parcellation scheme (Fan et al., 2016) (Table S4 and Fig. S2; see also *Materials and Methods* and Table S3). The percent signal difference (PSD) between the task and baseline conditions was computed for the 32 ROIs (Fig. S3). Relatively high PSDs were observed in L0 and L1 (except for bilateral MVOcC and thalamus), and for Max /ta/ perception (except for left MFG, right OrG, bilateral ACC, MVOcC, and thalamus). When compared with frequent perceivers, rare perceivers ($N = 9$) showed weaker activation and peak volumes were more restricted to fewer anatomical landmarks (Fig. S4). Rare perceivers however displayed relatively higher activation in lateral superior frontal region (right hemispheric dominant), and bilateral orbitofrontal and inferior parietal cortical regions (Table S2; see *Supplementary Note 1* for details). Moreover, rare perceivers also displayed relatively high PSDs at L1 and L2, mostly peaking at L2 (Fig. S5; see also *Supplementary Note 2*).

Synergistic Network Organization. To investigate the network interactions from the 32 ROIs, the seed-based whole-brain functional connectivity (SWFC) was computed from each ROI. For frequent perceivers, the averaged SWFC decreased maximally at L0 followed by L1 (Fig. 3*a* shows example pSTS seeds; Fig. S6). Among perception-sorted categories, Max /ta/ exhibited lowest SWFC followed by Min /ta/ and Pure /ta/ (Fig. 3*b*). The spatial organization of SWFC analysis with pSTS parcels as seeds is also shown in Fig. S7. For the rare perceivers, the decrease in average SWFC can be to some extent associated with decreasing AV lag, but not with increasing illusory /ta/ perception (Fig. 3*c*,*d*).

The relationship between SWFC and the propensity of perceiving illusory /ta/ was further evaluated using SWFC–behavior correlation (SBC; see *Materials and Methods*). For frequent perceivers all the 32 parcels displayed negative SBC, which means the SWFC as a function of AV lags is anti-correlated with the propensity of illusory /ta/ perception (Table S5). Zooming into the nature of interactions among the anti-correlated brain areas (with behavior) by choosing a threshold correlation < -0.5, that accounted for top 40%, we obtained 13 seed ROIs (9 in the left hemisphere: SFG, IFG, OrG, PrG, STG, MTG, pSTS, IPL, LOcC; 4 in the right hemisphere: PrG, pSTS, INS, ACC) (Fig. 3*e*,*f*; Fig. S6). Henceforth, we will refer to these 13 ROIs as the perceptual binding network (PBN).

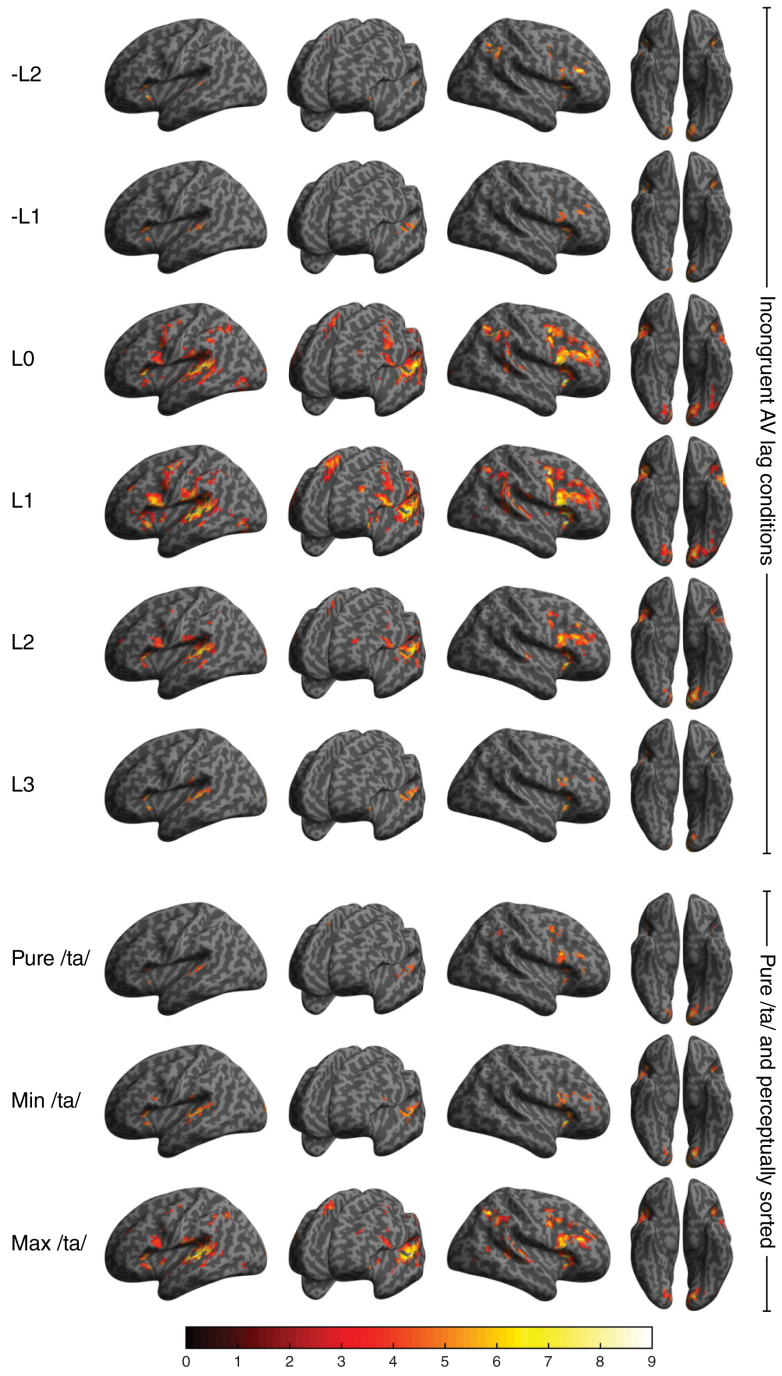
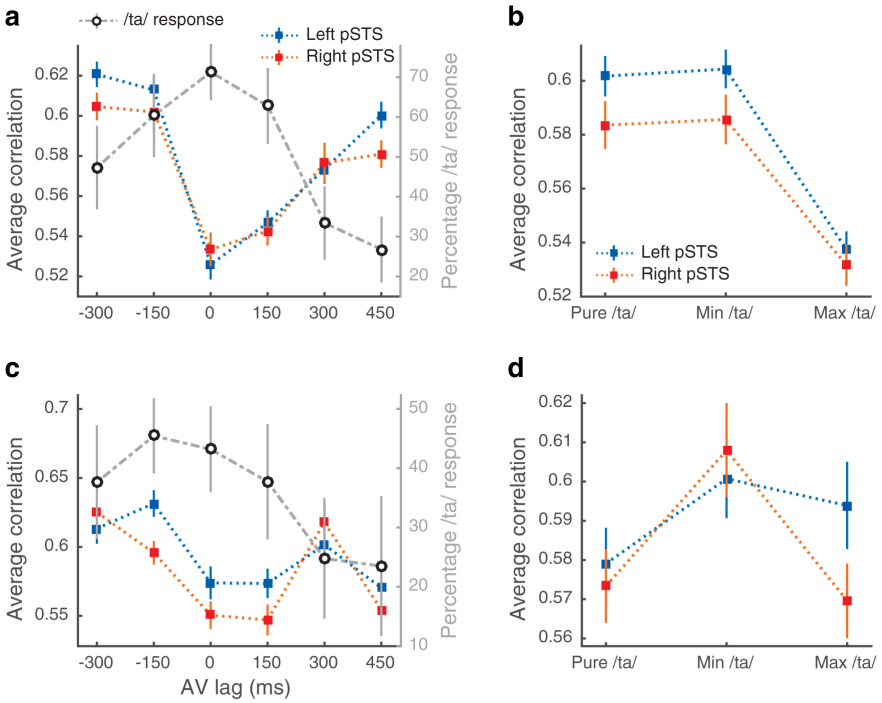


Figure 2. Group-level task-enhanced (task > rest) active brain areas for frequent perceivers (peak $p \leqslant 0.01$, FDR corrected; cluster extent threshold of 10 voxels). Color bar represents the T-score.

In the case of rare perceivers, several seed regions displayed positive SBC (Table S5 and Fig. S6; see also *Supplementary Note 3*).

Functional integration within the PBN was evaluated using linear partial correlation coefficients computed from the BOLD time series of constituent ROIs. pSTS(L) was found to maintain positive correlation with other temporal regions MTG(L), STG(L), and pSTS(R) (Fig. S8). In almost all the conditions, PrG(L) was positive correlated with SFG(L), IFG(L), IPL(L), and PrG(R) that further



e

Left hemisphere		Right hemisphere	
Seed	r_{SBC}*	Seed	r_{SBC}*
SFG	-0.52	PrG	-0.58
IFG	-0.60	pSTS	-0.51
OrG	-0.56	INS	-0.58
PrG	-0.67	ACC	-0.59
STG	-0.53		
MTG	-0.53		
pSTS	-0.56		
IPL	-0.54		
LOcC	-0.64		

*$p \ll 0.0001$ (Bonferroni-corrected).

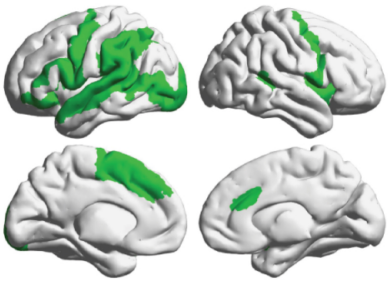


Figure 3. Average seed-based whole-brain functional connectivity (SWFC) of example pSTS seeds for frequent perceivers: AV lag conditions (*a*), and Pure /ta/ and perceptually sorted categories (*b*). Panels (*c*) and (*d*) represent the corresponding pSTS seed average SWFC across rare perceivers. Error bars represent the standard error of the mean (95% CI). (*e*) Frequent perceivers' seed ROIs showing large anti-correlation (< -0.5; top 40% selected) between average SWFC and percentage illusory /ta/ response across AV lags; the 13 nodes represents the perceptual binding network (PBN) (Table S5; see also *Materials and Methods* for details). (*f*) Surface plot showing the seed ROIs of the PBN detailed in panel (*e*).

positively correlated with INS(R). The average positive correlation between PrG(L) and SFG(L) is minimum at L0, although a trend analysis across all AV lags did not reveal any significant differences ($F(5, 61) = 1.17, p = 0.3336$). ACC(R) also maintained a positive correlation with both INS(R) and SFG(L). We further observed positive correlation between IFG(L) and OrG(L), and between MTG(L) and LOcC(L) in all the conditions. Interestingly, negative correlation was observed between

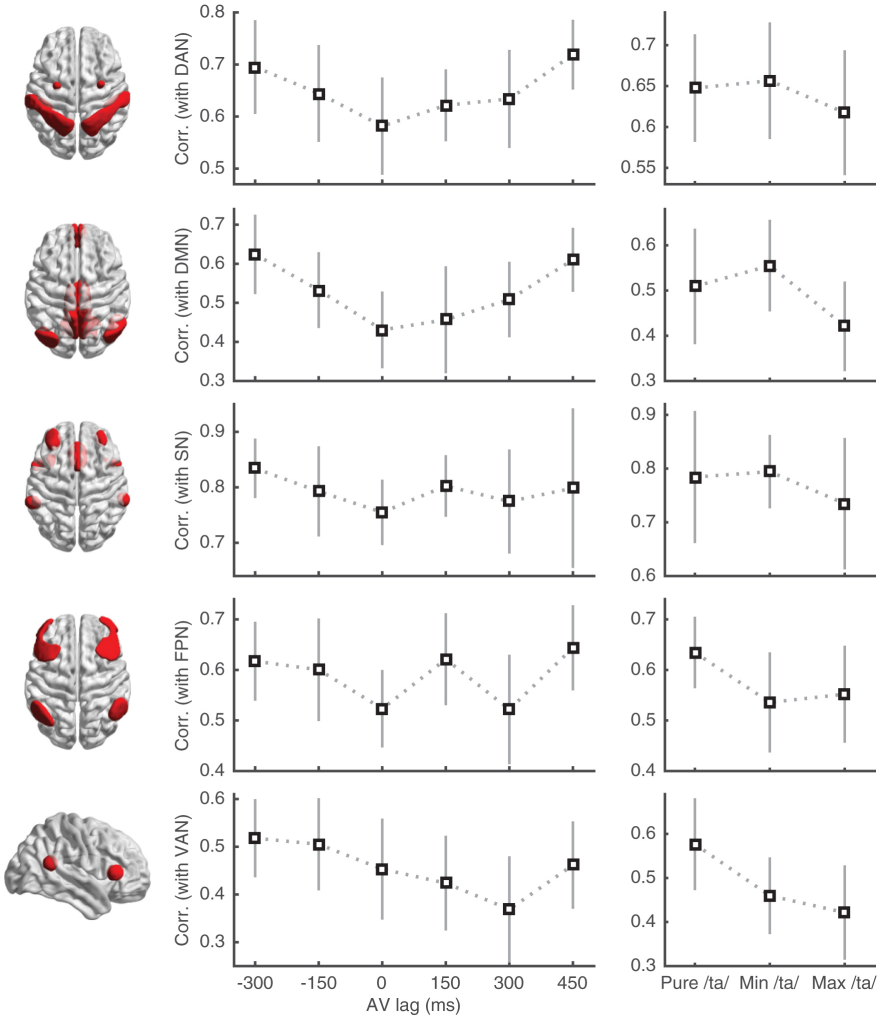


Figure 4. Between-network connectivity computed from projections on first principal component for the perceptual binding network (PBN) and the resting-state brain networks—dorsal attention network (DAN), default mode network (DMN), saliency network (SN), fronto-parietal network (FPN), and ventral attention network (VAN).

PrG(L) and MTG(L) for all intermediate lags (L0, $\pm$L1, and L2) and Max /ta/ conditions. PrG(L) also maintained a negative correlation with OrG(L) for L2 and L3, however no significant variation was observed across all AV lags or perceptual responses ($p = 0.0914, 0.2102$).

Finally, functional integration among the PBN and the resting-state networks (RSNs) was estimated by performing a between-network analysis. Pre-processed BOLD time series for the PBN were projected onto their first principal component (Friston et al., 1993) and the same procedure was followed for each RSN: dorsal attention network (DAN), default mode network (DMN), saliency network (SN), fronto-parietal network (FPN), and ventral attention network (VAN), to obtain a lower dimensional representation of network dynamics. Correlation coefficients were computed between temporal projections of the PBN and each RSN for all AV lags and perception-sorted categories. Interestingly, all the RSNs showed high correlation with the PBN, but the correlations got weaker close to AV stimuli synchrony or with increasing illusory perception for DAN, DMN, and SN, and resembled the SBC patterns in frequent perceivers (Fig. 4).

Discussion

Identifying integration and segregation (Tomasi et al., 2014; Tononi et al., 1994) of large-scale brain networks is increasingly becoming an important tool to understand information processing mechanisms underlying various cognitive tasks (Cohen and D'Esposito, 2016; Mohr et al., 2016). A more riveting hypothesis is that normative human cognition can be defined as a homeostatic balance between integrative and segregative mechanisms with imbalances leading to neuropsychiatric disorders (Lord et al., 2017). The present study demonstrates how such a balance facilitates perceptual binding of incongruent audio-visual/multisensory stimuli. This is a significant departure from the earlier understanding of multisensory integration from the perspective of inverse effectiveness which has been reinforced by several studies using different neuroimaging modalities (Meredith and Stein, 1983; Molholm et al., 2002; Stevenson and James, 2009).

Previous studies focused on regions like STS for multisensory processing and integration (Nath and Beauchamp, 2011; Stevenson and James, 2009). However, several unanswered questions related to multisensory integration and more subtly cross-modal binding (how stimulation of one sensory stream affects others) requires a theoretical understanding of distributed network mechanisms (Yau et al., 2015). Our paradigm allowed us to distinguish multisensory processing (involved in all AV lags) from the cross-modal perceptual binding (predominantly in L0 and L1), and the importance of the relationship of pSTS with other distributed regions—together defined as the perceptual binding network (PBN). Following previous studies (Beauchamp, 2005; Beauchamp et al., 2004), pSTS activation is enhanced during cross-modal illusory /ta/ perception, compared to similar activation patterns during Pure /ta/ and Min /ta/ perceptual categories. Interestingly, we found left thalamic activations increase for Max /ta/ in frequent perceivers contrary to right thalamus for rare perceivers. Further, for Min /ta/, right thalamus is more dominant in frequent perceivers compared to left thalamus in rare perceivers. Such observations points to a stronger role of cortico-thalamic circuitry in perceptual binding. Recent studies have demonstrated that decrease in levels of alpha coherence in electroencephalographic (EEG) activity is indicative of propensity to McGurk-illusion perception (Kumar et al., 2020), wherein the cortico-thalamic basis of alpha rhythms is well-known (da Silva et al., 1973). fMRI time-scale may not allow the temporal resolution to identify the dynamic mechanisms of the cortico-thalamic circuitry, but future electrophysiological studies can surely address this question.

The second important finding of this study is the decrease in correlation between each activated ROI and rest of the brain areas peaking when maximal cross-modal perceptual binding is reported. Thus, a network responsible for the perceptual binding (PBN) could be defined as a subset of all activated areas from SPM analysis, constituting of the parcels that are most anti-correlated with the perceptual behavior (Fig. 3*e* and Table S5). Subsequently, PBN was also found to be correlated with DAN, DMN, and to some extent SN; but the correlation pattern varied across AV lags reaching a minimum at the peak of illusory perception (Fig. 4). Also, it is important to note that the observations of enhanced functional connectivity (FC) among the multisensory and unisensory regions observed by earlier studies (Nath and Beauchamp, 2011) can be explained from the spatial topographies of functionally connected regions with pSTS (Fig. S7) at L0 and L1. Thus, an increasing activation of the nodes within PBN occurs simultaneously with their systemic disengagement from rest of the brain defining the integrative mechanisms that are predominantly compensatory to the task-driven haemodynamic oxygen metabolism. To the best of our knowledge, this is the first demonstration of temporal binding windows (Stevenson et al., 2012) at the level of functional integration and segregation in the brain.

The two major findings we presented can be interpreted from Bayesian causal inference (Koerding et al., 2007; Magnotti et al., 2018; Rohe et al., 2019) and predictive coding hypothesis (Friston

and Kiebel, 2009; Mumford, 1991). First, segregation in distributed brain areas within PBN validates the presence of mechanisms that allows integration of disparate sensory objects by imposing a common causal structure following earlier studies (Rohe et al., 2019). The imposition of the causal structure becomes harder within the temporal binding window and the requirement of a homeostatic balance (Lord et al., 2017) necessitates higher segregation and decreased local integration for each node of the PBN. Second, following predictive coding hypothesis, the decreased FC between the context-driven PBN and the context-free RSNs carry signatures of feedback functional projections from hierarchically higher order systems onto the PBN which also peak within the temporal binding window. Indeed, an immediate goal of subsequent studies should be to employ neurostimulation techniques such as TMS and tDCS (Yau et al., 2015) with a goal to perturb the homeostatic balance of integration–segregation mechanisms between PBN and context independent RSNs and identify their deeper causal architecture.

MATERIALS AND METHODS

Participants and Stimuli

Fifty-two healthy right-handed human volunteers (25 female, 27 male; age range = 24.5 years, SD = 3.12) participated in the functional magnetic resonance imaging (fMRI) study. All participants provided informed consent in a format approved by the Institutional Human Ethics Committee (IHEC), National Brain Research Centre.

Seven video stimuli each of 2 second duration were prepared. Six of them were incongruent audio-visual (AV) objects where audio recordings of a male speaker vocalizing /pa/ is dubbed on a video showing the lip movement while articulating the vocalization of /ka/ (/pa/–/ka/). The remaining one video was of a congruent AV object (/ta/–/ta/). In the six incongruent videos, the gap between the onset of auditory and visual streams were varied from −300 to 450 ms in steps of 150 ms (Fig. 1*a*). Negative sign indicates that the auditory stimulus onset precedes that of visual lip movement (A leading), and the positive sign on the other hand implies the auditory stimulus succeeds the lip articulation (V leading). An asymmetric range of AV lags were chosen, because Munhall and colleagues (Munhall et al., 1996) reported the dominance of illusory perception toward positive lags where the onset of lip articulation precedes auditory onset. The congruent /ta/–/ta/ stimulus had synchronous onset of AV stimuli. In the stimulus video, the articulation of /ka/ or /ta/ always started from a neutral lip-closed position. Videos were created using VideoPad video-editing software (NCH Software, CO, USA) at a frame rate of 25 frames/second and a resolution of 1280 × 720 pixels. The auditory /pa/ and /ta/ syllables were of 0.549 and 0.531 s duration and were edited in Audacity software (Sourceforge.net) to minimize the background noise. The audio sampling rate was 44 kHz and had a bit rate of 128 kbps. The resulted seven stimuli thus have five asynchronous and two synchronous stimuli—McGurk stimuli (/pa/–/ka/) with −300, −150, 0, 150, 300, 450 ms AV lags, and Pure /ta/ (/ta/–/ta/) stimulus with 0 ms AV lag. Henceforth, we shall also refer the AV lag conditions in shorter notation as −L2, −L1, L0, L1, L2, L3 for −300, −150, 0, 150, 300, 450 ms lag, respectively.

Task Design

The task design is illustrated in Fig. 1*a*. Inside the scanner, the stimuli were presented in a block design, and each activation block of length 20 s comprised 10 video stimuli of one kind. In total there were 28 activation blocks in the whole experiment: 4 activation blocks for each of the 7 stimulus categories. The 28 activation blocks were alternated with another 28 inter-trial resting blocks, each also of 20 s duration. We randomized the presentation of the 7 stimulus categories with each category comprising 4 alternated activation blocks of one kind. The stimuli were presented through an INVIVO MR-compatible CRT screen attached to the head-coil and an MRI-compatible headphones (Philips, The Netherlands). Presentation software (Neurobehavioral Systems, CA) was used to display the stimuli. During the presentation, participants were instructed to attend to the AV stimuli watching at the speaker all the times, and report their perceptual experience of whether they heard "/pa/", "/ta/" or "any other" through a three-button fibre-optic response device by Curdes (Current Designs, PA, USA), which was used to record the responses of participants.

fMRI Data Acquisition

Structural T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) scan was performed in the beginning of each participant's scanning session using a 3 Tesla Philips Achieva MRI scanner. For each participants, seven functional T2* scans (TR = 2 s, TE = 35 ms, flip angle = 90°) corresponding to the seven different stimuli (−L2 to L3 and Pure /ta/) were acquired. An 8-channel head coil was used to acquire the T2*-weighted images. The voxel size of the acquired images was $3.59 \times 3.59 \times 5$ mm^3 and the voxel size of the reconstructed images was $1.80 \times 1.80 \times 5$ mm^3. The slice thickness was 5 mm and inter-slice distance was 1 mm.

Behavioral Data Analysis

The button press responses from the 52 participants were converted into normalized percentage measures of each perceptual response category, /pa/, /ta/, or "other," separately for each AV lag and Pure /ta/ condition (Fig. S1*a*). A maximum of 40 responses were expected from each task. Tasks with less than 80% responses were rejected. Subsequently, participants who reported perceiving illusory /ta/ for at least 60% of the total responses at any AV lag were grouped as frequent perceivers, and those with less than 60% /ta/ response were grouped as rare perceivers. For the current study, 42 volunteers were categorized as frequent perceivers based on this threshold, and the remaining 10 as the rare perceivers. The volunteers were further screened based on their /ta/ response percentage during congruent Pure /ta/ (/ta/–/ta/) stimulus condition. For both frequent and rare perceivers, the minimum threshold of /ta/ response was set to 40%. Six frequent perceivers and one rare perceiver were rejected from this criterion. Two more frequent perceivers were rejected because one had reported >95% /ta/ response in five AV lag conditions and another >95% "other" response in two AV lag conditions. Thus, only 34 frequent perceivers and 9 rare perceivers were selected based on their behavioral responses. Furthermore, five more frequent perceivers were rejected because of excessive movement during MRI scanning. Finally, 29 frequent and 9 rare perceivers were selected for further analyses. Statistical analyses of group responses for each perceptual experience /pa/, /ta/, and "other" were performed using analysis of variance (ANOVA), with lags as independent variable.

fMRI Data Analysis

Pre-processing. For each participant, all the acquired functional brain volumes were realigned and co-registered with the structural scan using a six-parameter rigid-body transformation (three translations and three rotations). Those participants with maximum movement greater than 2 units (i.e., 2 mm in translational axes, or 2° in rotational axes) in any of the six degrees of freedom were rejected from further analysis. The structural image of each participant's brain was then segmented into grey and white matter maps, and using which the functional brain volumes were spatially normalized to Montréal Neurological Institute (MNI) standard brain. The normalized functional images were then spatially smoothed minimally with a 4 mm full-width at half-maximum (FWHM) isotropic Gaussian kernel (kernel size equivalent to two times the voxel dimension). All the preprocessing was done using Statistical Parametric Mapping Software 8 (SPM8) in MATLAB (MathWorks, MA). The first brain volume of each participant was removed to allow the magnetization to reach to a steady state.

Stimulus-based activation. After pre-processing, we performed the first-level intra-subject fixed-effects analysis using t-test on the activation blocks versus the inter-trial blocks, with the experimental design (event onset and their time course) as essential regressors and the realignment parameters for

head movements as nuisance regressors. To accommodate the inter-participant variability within each AV lag and perceptual condition, the contrast images (task block > inter-trial rest block) from the first-level analysis were entered into a second-level random-effects group analysis, using one-sample t-tests. We reported the significantly activated clusters during each AV lag and perceptual condition for both frequent (Fig. 2 and Table S1) and rare perceivers (Fig. S4 and Table S2), with the labeling of the brain areas done according to the parcels in Brainnetome atlas (http://atlas.brainnetome.org/; Table S3). Fig. S4*b–e* displayed a qualitative comparison of task-active clusters between frequent and rare perceivers.

For defining regions of interest (ROIs), task-active brain parcels from frequent perceivers during L0 and L1 conditions were first chosen (each parcel's volume and dimension defined using Brainnetome atlas). These two lags were chosen because there were maximal significant activation of brain volumes as well as maximal average illusory /ta/ response during these lags (Table S1; Fig. 1*c* and Fig. 2). In the case of frequent perceivers, the active brain parcels selected with reference to L0 and L1 conditions also covered all the parcels active during other stimulus as well as perceptual conditions. The active parcels were then grouped into 32 ROIs based on division through major anatomical landmarks (Table S4). To further investigate the change in within-ROI activation volume for different task conditions, we created 29 leave-one-out samples from the frequent perceivers' group. Next, second-level random-effect group analyses were performed on these 29 subsamples, and the average within-ROI activation volume was reported in Fig. S2.

Percent signal difference (PSD). For each event in the block design, the ROI percent signal change was evaluated by estimating maximal height of the HRF fitted event time course (beta × event regressor), relative to the actual mean signal of the ROI. The signal was estimated for an event duration of 2 seconds. The percent signal change was extracted using a MarsBaR batch script (http://marsbar.sourceforge.net). For each of the 32 ROIs, the PSD in each AV lag and perceptual condition was then evaluated by subtracting the percent signal change of inter-trial rest blocks from that of the activation blocks. The changes in average PSD for frequent and rare perceivers across conditions are reported in Figs. S3 and S5, respectively (see also Table S4 for maximal average PSD conditions in frequent perceivers).

Functional Connectivity Analysis

Seed-based functional connectivity. For the seed-based connectivity analysis, each participant's grey matter volume was parcellated using the Brainnetome atlas (http://atlas.brainnetome.org/). Among the 210 cortical regions of the atlas, the two subregions of pSTS in each hemisphere were merged into a single ROI. In the subcortical region, only the whole thalamic nuclei in each hemisphere were considered. Taken together, the final set of 210 ROIs covered the entire cortical region and the thalamus (Table S3). For each participant, the mean BOLD time series from each 210 ROI as well as from 32 seed ROIs (Table S4) were extracted using REX toolbox (Whitfield-Gabrieli, 2009) (https://web.mit.edu/swg/software.htm). From each extracted time series, stimulus blocks corresponding to each AV lag and "Pure /ta/" condition were selected based on the experimental design, and then concatenated. The concatenated time series were then preprocessed through linear detrending (linear or linear breakpoint detrending). Functional connectivities between each seed region and the remaining of 210 ROIs (i.e., exclusive of seed parcels) were then evaluated using the Pearson correlation measure between corresponding time series (Lee Rodgers and Nicewander, 1988; Pearson, 1895). The averaged connectivities from each seed ROI to the whole brain, for each AV lag and

perceptual conditions, were evaluated employing Fisher-z-transformed FDR-corrected ($p \leqslant 0.05$) correlation values. We will term this measure as seed-based whole-brain functional connectivity (SWFC).

The relationship between the trend that SWFC changes across AV lag conditions (Fig. 3*a*,*c*; red and blue lines for the example pSTS ROIs) and that of the changes in the degree of illusory /ta/ perception (Fig. 3*a*,*c*, gray line) was further evaluated by estimating the Pearson correlation coefficient between the two trends. Hereafter, we will refer it as SWFC–behavior correlation (SBC). Each reported seed SBC is the average significant correlation ($p \leqslant 0.05$, Bonferroni corrected) value from 50 bootstrap samples created by random sampling 70% participants within each category (frequent and rare perceivers). All ROIs in frequent perceivers displayed negative significant SBC, while many ROIs in rare perceivers showed positive SBC (Table S5). In reference to frequent perceivers, we further selected top 40% of the ROIs showing high negative SBC, resulting to a sub-network 13 nodes (Fig. 3*e*) which we will refer to as perceptual binding network (PBN) (a comparative SWFC trend shown in Fig. S6).

Interaction within the PBN. Using the task block time series extracted from each 13 node of the PBN (Fig. 3*e*), linear partial correlation for each pair of nodes was evaluated. By employing Fisher-z transformation, the significant correlation values ($p \leqslant 0.05$, uncorrected) across 29 frequent perceivers were then averaged. Only those average correlation values across a minimum of 8 participants (i.e., number of significant correlation $\geqslant 8$) and $\geqslant$0.2 were reported in Fig. S8.

Interaction between the PBN and resting-state networks (RSNs). For the frequent perceivers we also investigated the functional interaction between the PBN and human default mode network (DMN), saliency network (SN), fronto-parietal network (FPN), dorsal attention network (DAN), and ventral attention network (VAN). The ROI masses for DMN, SN, FPN, and DAN were taken from CONN toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012) which elucidated these RSNs through ICA analyses of 497 participants from HCP dataset (www.nitrc.org/projects/conn). For the VAN, two spherical ROI masses each of 10 mm diameter were created using MarsBaR toolbox: one for the right temporo-parietal junction (Fox et al., 2006) (centered around MNI coordinates: 54, −50, 19) and another for the right ventrolateral prefrontal cortex (Sylvester et al., 2013) (centered around MNI coordinates: 51, 27, 8). Task-block BOLD time series for the respective ROIs were extracted using REX toolbox. Next, through principal component decomposition the projected time series from each network constituents to their first principal component was evaluated. Finally, network interactions were estimated by using Pearson correlation measure between corresponding first principal component time series. We reported the evolution of the network interaction across AV lag and perceptual conditions in Fig. 4.

References

D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.*, 14:257–262, 2004.

M. S. Beauchamp. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr. Opin. Neurobiol.*, 15:145–153, 2005.

M. S. Beauchamp, B. D. Argall, J. Bodurka, J. H. Duyn, and A. Martin. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.*, 7:1190–1192, Nov. 2004. ISSN 1097-6256. doi: 10.1038/nn1333.

P. Bertelson and G. Aschersleben. Temporal ventriloquism: crossmodal interaction on the time dimension. 1. Evidence from auditory-visual temporal order judgment. *Int. J. Psychophysiol.*, 50: 147–155, Oct. 2003. ISSN 0167-8760. doi: 10.1016/s0167-8760(03)00130-2.

J. R. Cohen and M. D'Esposito. The segregation and integration of distinct brain networks and their relationship to cognition. *J. Neurosci.*, 36:12083–12094, Nov. 2016. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2965-15.2016.

C. Cuppini, L. Shams, E. Magosso, and M. Ursino. A biologically inspired neurocomputational model for audiovisual integration and causal inference. *Eur. J. Neurosci.*, 46:2481–2498, Nov. 2017. ISSN 1460-9568. doi: 10.1111/ejn.13725.

F. H. L. da Silva, T. H. M. T. van Lierop, C. F. Schrijer, and W. S. van Leeuwen. Organization of thalamic and cortical alpha rhythms: Spectra and coherences. *Elec. Clin. Neuro.*, 3(6):627 – 639, 1973.

L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird, P. T. Fox, S. B. Eickhoff, C. Yu, and T. Jiang. The human Brainnetome Atlas: a new brain atlas based on connectional architecture. *Cereb. Cortex*, 26:3508–3526, Aug. 2016. ISSN 1460-2199. doi: 10.1093/cercor/bhw157.

M. D. Fox, M. Corbetta, A. Z. Snyder, J. L. Vincent, and M. E. Raichle. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proc. Natl. Acad. Sci. U.S.A.*, 103(26): 10046–10051, 2006.

K. Friston and S. Kiebel. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 364:1211–1221, May 2009. ISSN 1471-2970. doi: 10.1098/rstb.2008.0300.

K. J. Friston, C. D. Frith, P. F. Liddle, and R. S. Frackowiak. Functional connectivity: the principal-component analysis of large (PET) data sets. *J. Cereb. Blood Flow Metab.*, 13:5–14, Jan. 1993. ISSN 0271-678X. doi: 10.1038/jcbfm.1993.4.

K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.*, 2(4):189–210, 1994. doi: 10.1002/hbm.460020402.

J. A. Jones and D. E. Callan. Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport*, 14:1129–1133, June 2003. ISSN 0959-4965. doi: 10.1097/00001756-200306110-00006.

K. P. Koerding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams. Causal inference in multisensory perception. *PLoS One*, 2:e943, Sept. 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000943.

G. V. Kumar, S. Dutta, S. Talwar, D. Roy, and A. Banerjee. Biophysical mechanisms governing large-scale brain network dynamics underlying individual-specific variability of perception. *Eur. J. Neurosci.*, Apr. 2020. ISSN 1460-9568. doi: 10.1111/ejn.14747.

J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *Am. Stat.*, 42(1):59–66, 1988.

L.-D. Lord, A. B. Stevner, G. Deco, and M. L. Kringelbach. Understanding principles of integration and segregation using whole-brain computational connectomics: implications for neuropsychiatric disorders. *Philos. Trans. A Math. Phys. Eng. Sci.*, 375:20160283, June 2017. ISSN 1364-503X. doi: 10.1098/rsta.2016.0283.

J. F. Magnotti, K. B. Smith, M. Salinas, J. Mays, L. L. Zhu, and M. S. Beauchamp. A causal inference explanation for enhancement of multisensory integration by co-articulation. *Sci. Rep.*, 8:18032, Dec. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-36772-8.

H. McGurk and J. Macdonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

M. A. Meredith and B. E. Stein. Interactions among converging sensory inputs in the superior colliculus. *Science*, 221(4608):389–391, Jul 1983.

H. Mohr, U. Wolfensteller, R. F. Betzel, B. Misic, O. Sporns, J. Richiardi, and H. Ruge. Integration and segregation of large-scale brain networks during short-term task automatization. *Nat. Commun.*, 7:13217, Nov. 2016. ISSN 2041-1723. doi: 10.1038/ncomms13217.

S. Molholm, W. Ritter, M. M. Murray, D. C. Javitt, C. E. Schroeder, and J. J. Foxe. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Res. Cogn. Brain Res.*, 14:115–128, June 2002. ISSN 0926-6410. doi: 10.1016/s0926-6410(02)00066-6.

D. Mumford. On the computational architecture of the neocortex. I. The role of the thalamo-cortical loop. *Biol. Cybern.*, 65:135–145, 1991. ISSN 0340-1200. doi: 10.1007/BF00202389.

K. G. Munhall, P. Gribble, L. Sacco, and M. Ward. Temporal constraints on the McGurk effect. *Atten. Percept. Psychophys.*, 58:351–362, 1996.

A. R. Nath and M. S. Beauchamp. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J. Neurosci.*, 31(5):1704–1714, 2011.

K. Pearson. Vii. note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.*, 58(347-352):240–242, 1895.

T. Rohe, A.-C. Ehlis, and U. Noppeney. The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nat. Commun.*, 10:1907, Apr. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09664-2.

L. Shams, Y. Kamitani, and S. Shimojo. Visual illusion induced by sound. *Brain. Res. Cogn. Brain. Res.*, 14:147–152, 2002.

R. A. Stevenson and T. W. James. Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage*, 44:1210–1223, Feb. 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2008.09.034.

R. A. Stevenson, R. K. Zemtsov, and M. T. Wallace. Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *J. Exp. Psychol. Hum. Percept. Perform.*, 38:1517–1529, Dec. 2012. ISSN 1939-1277. doi: 10.1037/a0027339.

C. M. Sylvester, D. M. Barch, M. Corbetta, J. D. Power, B. L. Schlaggar, and J. L. Luby. Resting state functional connectivity of the ventral attention network in children with a history of depression or anxiety. *J. Am. Acad. Child Adolesc. Psychiatry*, 52(12):1326–1336, 2013.

B. Thakur, A. Mukherjee, A. Sen, and A. Banerjee. A dynamical framework to relate perceptual variability with multisensory information processing. *Sci. Rep.*, 6:31280, Aug. 2016. ISSN 2045-2322. doi: 10.1038/srep31280.

D. Tomasi, R. Wang, G.-J. Wang, and N. D. Volkow. Functional connectivity and brain activation: a synergistic approach. *Cereb. Cortex*, 24:2619–2629, Oct. 2014. ISSN 1460-2199. doi: 10.1093/cercor/bht119.

G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. U.S.A.*, 91:5033–5037, May 1994. ISSN 0027-8424. doi: 10.1073/pnas.91.11.5033.

S. Watkins, L. Shams, O. Josephs, and G. Rees. Activity in human V1 follows multisensory perception. *Neuroimage*, 37:572–578, Aug. 2007. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.05.027.

S. Whitfield-Gabrieli. Region of interest extraction (rex) toolbox. *Boston, MA*, 497, 2009.

S. Whitfield-Gabrieli and A. Nieto-Castanon. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.*, 2(3):125–141, 2012.

J. M. Yau, G. C. DeAngelis, and D. E. Angelaki. Dissecting neural circuits for multisensory integration and crossmodal processing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370:20140203, Sept. 2015. ISSN 1471-2970. doi: 10.1098/rstb.2014.0203.

Acknowledgements This study was supported by NBRC Core funds and Ramalingaswami Fellowship (BT/RLF/Re-entry/31/2011) and Innovative Young Biotechnologist Award (IYBA, BT/07/IYBA/2013), Department of Biotechnology, Ministry of Science and Technology, Government of India to AB. SSS was supported by a National Post-Doctoral Fellowship from Science and Engineering Research Board, Government of India (SERB-NPDF; No. PDF/2016/003188).

Competing interests The authors declare no competing interest.