BRILL

Multisensory Research 31 (2018) 481–500

MULTISENSORY RESEARCH

brill.com/msr

Segregation and Integration of Cortical Information Processing Underlying Cross-Modal Perception

G. Vinodh Kumar [1,*], **Neeraj Kumar** [1], **Dipanjan Roy** [2] **and Arpan Banerjee** [1]

[1] Cognitive Brain Lab, National Brain Research Centre, NH 8, Manesar, Gurgaon 122051, India

[2] Centre of Behavioural and Cognitive Sciences, University of Allahabad, Allahabad 211002, India

Received 6 December 2016; accepted 17 April 2017

Abstract

Visual cues from the speaker's face influence the perception of speech. An example of this influence is demonstrated by the McGurk-effect where illusory (cross-modal) sounds are perceived following presentation of incongruent audio–visual (AV) stimuli. Previous studies report the engagement of specific cortical modules that are spatially distributed during cross-modal perception. However, the limits of the underlying representational space and the cortical network mechanisms remain unclear. In this combined psychophysical and electroencephalography (EEG) study, the participants reported their perception while listening to a set of synchronous and asynchronous incongruent AV stimuli. We identified the neural representation of subjective cross-modal perception at different organizational levels — at specific locations in sensor space and at the level of the large-scale brain network estimated from between-sensor interactions. We identified an enhanced positivity in the event-related potential peak around 300 ms following stimulus onset associated with cross-modal perception. At the spectral level, cross-modal perception involved an overall decrease in power at the frontal and temporal regions at multiple frequency bands and at all AV lags, along with an increased power at the occipital scalp region for synchronous AV stimuli. At the level of large-scale neuronal networks, enhanced functional connectivity at the gamma band involving frontal regions serves as a marker of AV integration. Thus, we report in one single study that segregation of information processing at individual brain locations and integration of information over candidate brain networks underlie multisensory speech perception.

Keywords

EEG, network, multisensory, perception, ERP, spectral, coherence, timing

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

DOI:10.1163/22134808-00002574

1. Introduction

Combination of information from different senses enhances our perceptual and response ability. For example, although speech perception is based on the processing of the auditory signals, speech intelligibility can be influenced when it is accompanied by the visual articulatory gestures of the speaker. This can either result in enhancement of the auditory perception (Helfer, 1997; Sumby and Pollack, 1954) or modulate it when accompanied with *semantically-incongruent* lip movements (McGurk and MacDonald, 1976). Numerous research papers have explored the cortical correlates of multisensory perception, and demonstrated the involvement of specific modules and distributed cortical networks. However, it remains unclear at what scales these networks are engaged and what the most pertinent substrate is for representing the mechanism of multisensory perception.

The conventional view of sensory processing is that convergence and integration of information across different modalities occurs in specific cortical modules post extensive processing within sensory-specific subcortical and cortical regions. However, evidence from recent studies shows that multisensory integration extends beyond modularity and suggests that multisensory convergence is considerably widespread in the brain (Bizley and King, 2012; Calvert and Thesen, 2004; McIntosh, 2004). Furthermore, even the primary sensory areas have been claimed as a part of the emerging network of multisensory regions (Allman *et al*., 2009; Bizley and King, 2012). From the perspective of localization of function via integration hypothesis (Bressler and Kelso, 2001; Horwitz, 2005; Luria, 1995; McIntosh, 2004), it is fundamentally important to understand the network-level mechanisms at various spatiotemporal scales over which multisensory information processing is represented.

Behavioral and neuroimaging studies in the domain of speech perception have extensively used McGurk effect to gain insights on mechanism of audio–visual (AV) integration and multisensory perception (Green *et al*., 1991; Hasson *et al*., 2007; Jones and Callan, 2003; Kaiser, 2004; Keil *et al*., 2012; Kumar *et al*., 2016; Saint-Amour *et al*., 2007; Sekiyama *et al*., 2003; Stevenson *et al*., 2010; Van Wassenhove *et al*., 2005; Wallace *et al*., 1993). During the McGurk effect, an auditory speech sound */ba/* superimposed onto the visual lip movement of */ga/* gives rise to an illusory (cross-modal) percept of */da/* (McGurk and MacDonald, 1976). A substantial amount of evidence employing the McGurk effect demonstrates activation of specific cortical modules like the pSTS (posterior Superior Temporal Sulcus) (Jones and Callan, 2003; Nath and Beauchamp, 2011, 2012; Sekiyama *et al*., 2003), frontal and parietal areas (Callan *et al*., 2003; Skipper *et al*., 2007) being responsible for the cross-modal perception. On the other hand, studies employing connectivity measures on functional imaging and electrophysiological data primarily reveal interactions

among cortical regions of interest (Keil *et al*., 2012) or characterize the properties of the global network (Kumar *et al*., 2016) endorsing the mechanism of functional integration. However, to our knowledge no study has reported that both mechanisms are operational on a putative data set along with their variability across trials. Therefore, investigating the interplay between the modular components of an extended cortical network of multisensory regions concomitantly with dynamic changes within the components would help us develop a comprehensive account of underlying mechanisms involved in multisensory perception. In the present study, we used an incongruent McGurk pair (audio */pa/* superimposed on a video of the face articulating */ka/*) to induce the cross-modal percept */ta/*. Further, we introduced a temporal asynchrony in the onset of audio and visual events of the McGurk stimuli to diminish the rate of cross-modal responses */ta/*, in comparison to the unimodal response of */pa/*, thus creating two perceptual categories which can be further studied from the perspective of integration and segregation of information processing in the brain at different spatial scales. We observed the representation of dynamical information processing at each spatial scale, the individual sensor level in EEG data (using time series and spectro-temporal representations of sensor-level power) and large-scale brain networks (using the imaginary coherence to extract the between-sensor interactions), indicating that multi-scale representation of the AV integration is pertinent for a comprehensive understanding of multisensory speech processing.

2. Materials and Methods

2.1. Participants

Nineteen healthy volunteers (10 males and 9 females, in the range of 22–29 years of age; mean age 25, SD = 2) participated in the study. All participants gave written informed consent, and they had no neurological or audiological problems. They all had normal or corrected-to-normal vision and were right-handed. The study was carried out following the ethical guidelines and prior approval of the Institutional Review Board of the National Brain Research Centre, India. The data from four volunteers were not included in the study because they reported to hear only the auditory stimuli and did not perceive the McGurk effect when audio–visual stimuli were incongruent.

2.2. Stimuli and Trials

2.2.1. Stimuli

Each participant responded to 360 trials which consisted of videos of a native Hindi-speaking male articulating the syllables */ka/* and */ta/* (see Fig. 1). One-fourth (90 trials) of the trials consisted of congruent video (visual */ta/*

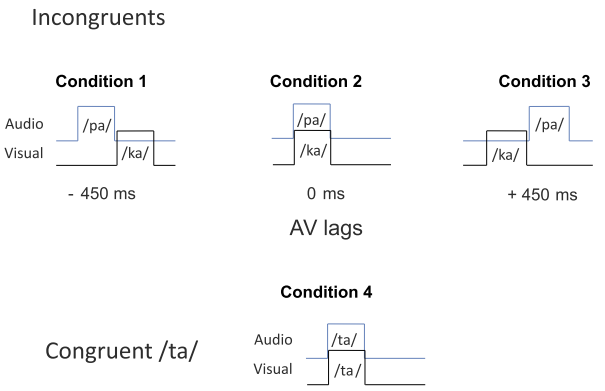

Figure 1. Stimuli: Each condition represents a video of speaker articulating a speech sound. AV lags show the temporally incongruent placement of the audio */pa/* with respect to the articulation (lip movement) of */ka/*. The congruent */ta/* represents a video with audio */ta/* dubbed onto a video of a person articulating */ta/*.

auditory */ta/*). The remaining three-fourths of the trials comprised incongruent videos (visual */ka/* auditory */pa/*) presented with AV lags: −450 ms (audio leads the articulation), 0 ms (synchronous) and +450 ms (articulation leads the audio), each encompassing one-fourth of the overall trials. The audio syllable was extracted from a video of the speaker articulating */pa/* using the software Audacity (www.audacityteam.org). Subsequently, the extracted audio syllable was superimposed onto the muted video of the speaker articulating the syllable */ka/* using the software Videopad Editor (www.nchsoftware.com). The stimuli were rendered into a 800 × 600 pixels movie with a digitization rate of 29.97 frames per second. Stereo soundtracks were digitized at 48 kHz with 32 bit resolution. Presentation software (Neurobehavioral System Inc.) was used to present the stimuli using a 17″ LED monitor. Sound was delivered using sound tubes at an overall intensity of ∼60 dB.

2.2.2. *Experimental Design*

The experiment was divided into three blocks. Each block consisted of 120 trials comprising four kinds of videos (30 trials of each): a congruent video and the three incongruent McGurk pair videos with AV lags. Inter-stimulus intervals were pseudo-randomly varied between 1200 ms and 2800 ms to minimize expectancy effects. The subjects were instructed to report what they heard while watching the speaker using a set of three keys: */pa/*, */ta/* or 'anything else'.

The subjects also performed a behavioral task post EEG scan. The task consisted of 60 trials, comprising 30 trials of the auditory syllables */ta/* and */pa/* each. The subjects were instructed to report what they heard using the choices */ta/* and */pa/*.

2.3. Data Acquisition and Analysis

2.3.1. EEG

Continuous EEG scans were acquired using a Neuroscan system (Synamps2, Compumedics, Inc.) with 64 Ag/AgCl scalp electrodes sintered on an elastic cap in a 10–20 montage. Recordings were made against a centroid (Cz) reference and digitized at a sampling rate of 1000 Hz. Channel impedances were kept at values < 5 kΩ.

2.3.2. Preprocessing of EEG Signals

The EEG data acquired was initially re-referenced to linked mastoids and filtered using a bandpass of 0.2–45 Hz. Subsequently, the continuous EEG was divided into epochs (-400 ms to 900 ms surrounding the onset of the first stimulus, i.e., the sound or articulation) and sorted based on the responses, */ta/*, */pa/* and 'other', respectively. Epochs were baseline-corrected by removing the temporal mean of the EEG signal on an epoch-by-epoch basis. Subsequently, we performed artifact rejection to eliminate the response contamination from ocular and muscle-related activities. However, depending on the analysis, we used two different thresholds. For statistical analysis of the event-related potentials, to minimize false positives arising from high amplitude in the low-frequency waveforms, epochs with a maximum signal amplitude above 50 μV or a minimum below -50 μV were removed from all electrodes. For spectral and network analysis, we used a signal amplitude threshold of ± 100 μV for artifact rejection as amplitude differences in waveforms will have no relevance in the spectral domain.

2.3.3. Event-Related Potential (ERP) Analysis

The preprocessed EEG data were further sorted according to the responses using customized MATLAB codes. After pooling across all subjects, the ERPs for each condition contained a minimum of 128 trials, were averaged and plotted across all electrodes. As we specifically focused on the difference in the ERP pattern between the */ta/* and */pa/* responses, the sorted epochs for each stimulus condition were compared statistically. Ms-by-ms paired t-tests were performed between the */ta/* and */pa/* responses across all electrodes to evaluate the spatio-temporal properties of AV integration. For each scalp electrode, the first time point where the t-test yielded a p-value < 0.05 and continued to do so for at least 20 consecutive data points (20 ms) was considered significantly different. The method serves as an alternative to Bonferroni correction for multiple comparisons, which would increase the possibility of false negatives (Murray *et al.*, 2002).

2.3.4. Spectral Analysis

A time–frequency spectrogram of EEG signals at each electrode was computed on a single-trial basis and sorted based on the responses, */ta/*, */pa/* and

'other', respectively. We computed the spectral power at different frequencies over time using customized MATLAB (www.mathworks.com) codes and the Chronux toolbox (www.chronux.org). The time bandwidth product and the number of tapers were set at 3 and 5, respectively, and a fixed time window of 0.3 s was applied while using the Chronux function mtspecgramc.m to compute the time–frequency spectrogram of the sorted time series in EEG data.

The time–frequency spectrogram computed for the perceptual categories */ta/* and */pa/* were compared channel by channel employing cluster-based permutation tests (Maris *et al*., 2007). During the cluster-based permutation tests, 1000 iterations of trial randomization were carried out to generate the permutation distribution at a frequency band at a time point. Subsequently, a two-tailed test with a threshold of 0.025 was used to evaluate the positive (increased spectral power) and negative (decreased spectral power) clusters at the respective sensors.

2.3.5. Network Analysis

To comprehend the cortico-cortical interactions underlying AV integration, we assessed the imaginary component of pairwise sensor-level coherence introduced by Nolte and colleagues (Nolte *et al*., 2004). This functional connectivity estimate captures the 'true' brain interactions that occur with a certain time lag, neglecting the spurious interactions arising from common references, volume conduction and crosstalk. Imaginary coherence refers to the complex part of the coherency C_{ij} that quantifies the phase relationship between two time series $\hat{x}_i(t)$ and $\hat{x}_j(t)$ at a specific frequency f. Coherency $C_{ij}(f)$ is the normalized cross-spectrum between two signal pairs, which in the current study are the EEG signals from different sensor pairs i and j.

$$C_{ij}(f) = \frac{S_{ij}(f)}{\sqrt{S_{ii}S_{jj}}}, \tag{1}$$

where S_{ij} is the cross-spectrum obtained by performing the complex conjugate of the Fourier transforms of $\hat{x}_i(t)$ and $\hat{x}_j(t)$.

Imaginary coherence was evaluated in the time window of 0.9 s post the onset of the first stimulus (audio or visual) for each perceptual category (*/ta/* and */pa/*) at all the AV lags. We employed the Chronux function crossSpecMatc.m to obtain the normalized cross-spectral matrix for all sensor combinations. Subsequently, we extracted the imaginary part of the cross-spectral values that constitute the imaginary coherence. The values of the imaginary coherence in the three frequency bands (alpha, beta and gamma) were further averaged using the Circular statistics function circ_mean.

Imaginary coherence computed for */ta/* and */pa/* responses were further compared between each channel pair for significant difference at different frequency bands (alpha, beta and gamma) explicitly by means of the cluster-based permutation test (Maris *et al*., 2007).

For each channel pair, the imaginary coherence difference between /*ta*/ and /*pa*/ was evaluated using the Fisher's Z transformation

$$Z(f) = \frac{\tanh^{-1}(C_1(f)) - \tanh^{-1}(C_2(f)) - (\frac{1}{2m_1-2} - \frac{1}{2m_2-2})}{\sqrt{\frac{1}{2m_1-2} + \frac{1}{2m_2-2}}}, \qquad (2)$$

where $2m_1, 2m_2 =$ degrees of freedom; $Z(f) \approx N(0, 1)$ is a unit normal distribution; and C_1 and C_2 are the imaginary coherence values at frequency band f.

The coherence Z-statistic matrix obtained from the above computation formed the observed Z-statistics. Consequently, 1000 iterations of trial randomization were carried out to generate the permutation distribution at a frequency band for each channel pair. Subsequently, a two-tailed test with a threshold of 0.001 was used to evaluate the channel pairs that showed significantly different interactions between the two perceptual categories. The same statistical tests were carried out to test the differences at different AV lags.

3. Results

3.1. Behavior

We converted the behavioral responses corresponding to McGurk stimuli with the AV lags to percentage measures for each perceptual category (/*pa*/, /*ta*/ or 'other') from all subjects using customized Matlab codes. To qualify a participant as an illusory (cross-modal) perceiver, we set a minimum threshold of 60% of /*ta*/ response in any AV lag, −450, 0 and +450 ms. Fifteen participants qualified and four participants failed to perceive above the set threshold. Data from only 15 perceivers were used for further group-level analysis (Note 1). We observed that a maximum percentage of illusory (/*ta*/) responses occurred at 0 ms AV lag (Fig. 2). The percentage of /*pa*/ responses was also at minimum at 0 ms AV lag. We ran a repeated-measures two-way ANOVA on the percentage responses with AV lags and the perceptual categories (/*ta*/ and /*pa*/) as the variables. We observed that there was no influence of AV lags in the percentage of responses of /*ta*/ and /*pa*/ [$F(2, 89) = 0.84,\ p = 0.44$]. However, we found a significant difference in the percentage responses between the two perceptual categories [$F(1, 89) = 19.46,\ p < 0.0001$]. Also, the interaction between perceptual categorization and AV lags was significant [$F(2, 89) = 23.83,\ p < 0.0001$]. Furthermore, we performed a post-hoc test using the Scheffe method on the perceptual categories. We observed a significant difference in the percentage responses between the two perceptual categories at the 95% confidence level. We also performed a paired Student's t-test on the percentage of responses (/*ta*/ and /*pa*/) at each AV lag. Insignificant differences of 10.20% and 11.40% were observed between /*ta*/ and /*pa*/

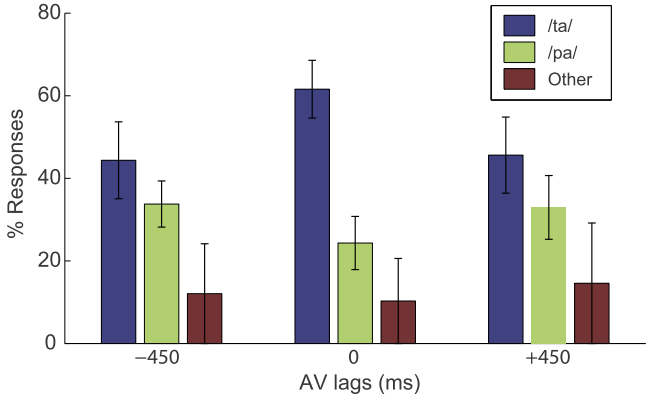

Figure 2. Behavior: Percentage of perceptual categorization for */pa/*, */ta/* and 'other' percepts as a function of AV lags, normalized and grouped over all 15 perceivers. The error bars represent 95% confidence interval.

responses at −450 ms AV lag [$t(14) = 0.63$, $p = 0.27$] and +450 ms AV lag [$t(14) = 0.45$, $p = 0.67$], respectively. However, at 0 ms AV lag we observed that the percentage of */ta/* responses was significantly higher by 36.58% than the percentage of */pa/* responses [$t(14) = 10.20$, $p < 0.0001$]. The hit rate of */ta/* responses during congruent */ta/* was observed to be 0.97. Also, the hit rate of */ta/* and */pa/* during auditory-alone conditions was observed to be 0.96 and 0.98, respectively.

3.2. Event-Related Activity

The difference wave obtained by subtracting the event-related responses of */pa/* from the responses of */ta/* (*/ta/* − */pa/*) for the AV lags −450 ms, 0 ms and +450 ms at all scalp electrodes are shown in Fig. 3A. In the difference wave, we observed a positive peak between ∼300–380 ms in frontal-polar, frontal and centro-parietal sensors at −450 ms AV lag and in frontal-polar, central, temporal, centro-parietal and parieto-occipital sensors at 0 ms AV lag, respectively. However, we did not observe any such peaks in the difference wave at +450 ms AV lag.

To compute the sensors eliciting significantly different amplitude during */ta/* responses than */pa/* responses, we performed millisecond-by-millisecond *t*-tests between the two conditions. To ignore transient responses, the criteria for significance were chosen such that at the onset latency the first point in the time series was where the *p*-value was less than 0.05 and remained so for at least 20 ms consecutively. The cluster plots in Fig. 3B exhibit such temporal windows. At −450 ms AV lag we observed a difference in the frontal and central sensors at ∼370 ms followed by which we observed a difference in the temporal, centro-parietal, parieto-occipital and occipital sensors ranging from

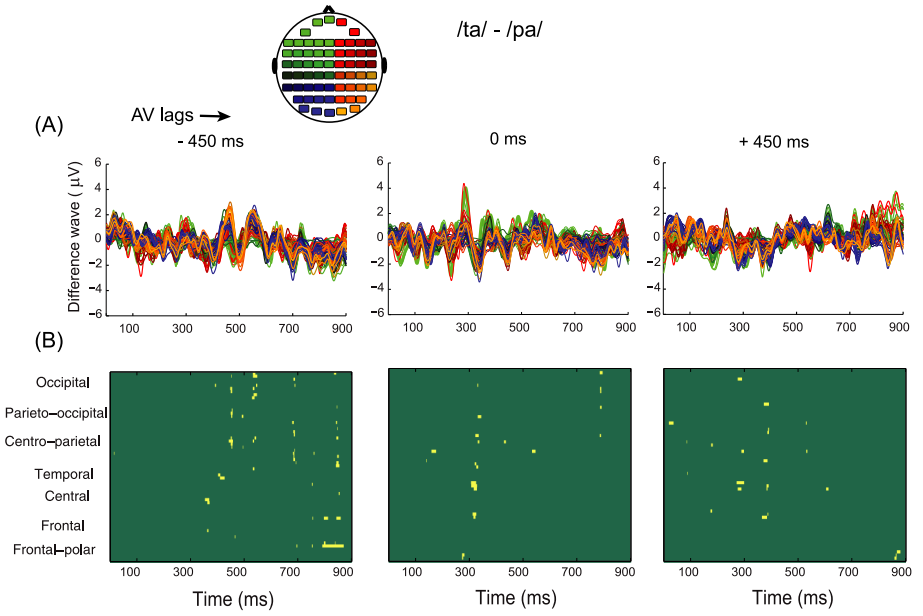

Figure 3. Event-Related Potential: (A) The difference wave between ERPs sorted out from */ta/* and */pa/* response trials at −450 ms, 0 ms and +450 ms AV lag. The topoplot at the top left displays the color code used for plotting ERPs assigned to respective scalp channel locations. For example, the green and red positive peaks around 300 ms represent the peak of activity in the left frontal and right frontal sensors. (B) Statistical cluster plots of the difference between the perceptual categories (*/ta/* and */pa/*) for each stimulus. The clusters indicate the time points where the p-values were < 0.05 for more than 20 ms. General sensor positions are arranged from frontal to posterior regions (bottom to top).

450 ms to 900 ms. Also, around 900 ms we observed a difference in the frontal sensors. At 0 ms AV lag, we observed a difference prominently around 300 ms post stimulus onset in frontal-polar, frontal, central, temporal, centro-parietal and the parieto-ocipital sensors. Similarly, at +450 ms AV lag, we observed a difference predominantly between ∼300 and 400 ms across the entire brain.

3.3. Power of Oscillatory Activity

The relative difference in the time–frequency spectrogram between */ta/* and */pa/* responses (*/ta/* − */pa/*) at each sensor obtained after the cluster-based permutation test is shown in Fig. 4. Figures 4A, B and C plot the differences in the spectral power between */ta/* and */pa/* responses at −450 ms, 0 ms and +450 AV lag, respectively. At −450 ms AV lag, we observed negative clusters predominantly in the theta and alpha bands denoting a decrease in the spectral power in the left frontal, left temporal and bilateral occipital sensors. However, at 0 ms AV lag we observed positive clusters, denoting an increase in the spectral power in the theta and alpha frequency bands predominantly

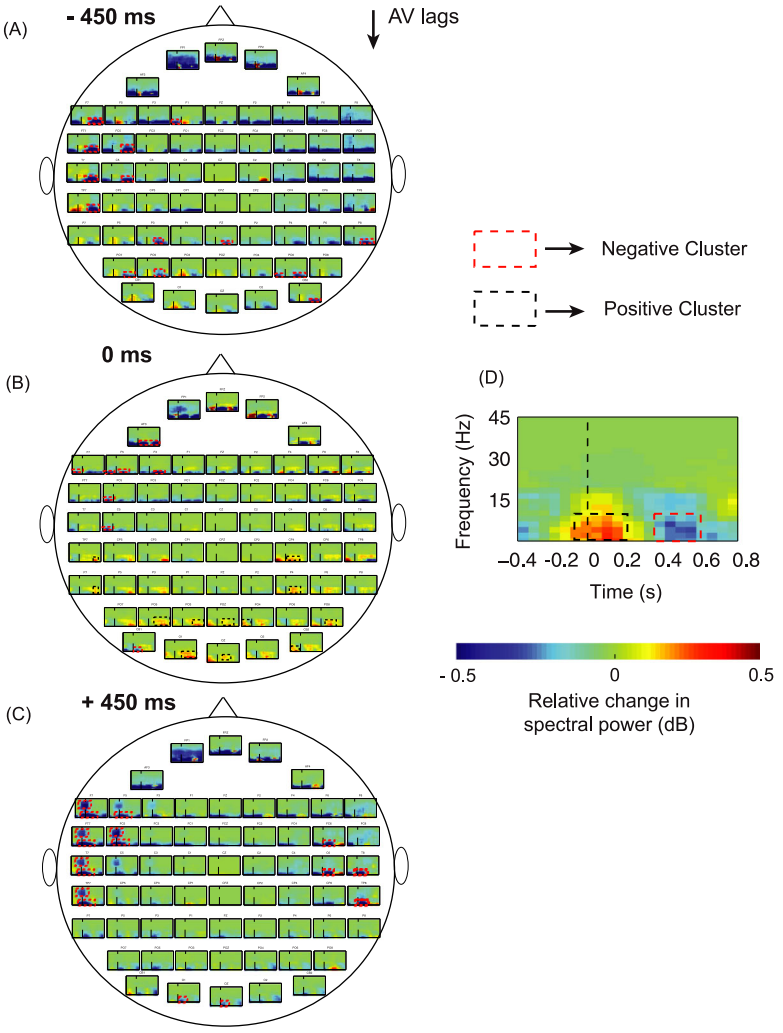

Figure 4. Power spectral analysis: Time–frequency spectrogram difference at each sensor time locked to the onset of the first stimulus (400 ms pre-stimulus and 900 ms post stimulus): (A) −450 ms AV lag, (B) 0 ms AV lag and (C) +450 ms AV lag. The red and black dotted boxes represent the areas in the respective sensors that exhibit a significant difference between the perceptual categories (/*ta*/ and /*pa*/). Panel (D) represents an enlargement of the spectrogram at each sensor showing islands of increased and decreased power.

in the occipital sensors and left temporo-parietal and right centro-parietal sensors in addition to suppression of alpha and theta power in left frontal areas. At +450 ms AV lag, we observed a bilateral decrease in the spectral power in the frontal and temporal sensors. On the left frontal and temporal sensors, negative clusters were observed in the theta, alpha, beta and the gamma bands.

However, in the right temporal sensors negative clusters were observed in the theta bands.

3.4. Functional Connectivity

To assess the functional connectivity underlying AV integration, we non-parametrically compared the imaginary coherence between (*/ta/*) and unisensory (*/pa/*) responses from all the pairwise sensor combinations. We observed significant changes in connectivity ($p < 0.001$) at −450 ms AV lag (Fig. 5A) in the alpha band between left parietal-occipital, parietal-temporal and right occipital sensors; in the beta band between left frontal-temporal, frontal-parietal and right frontal-temporal sensors and in the gamma band between bilateral frontal, left frontal-temporal and frontal-parietal sensors. At 0 ms AV lag (Fig. 5B), significant differences in the connectivity were observed in the alpha band bilaterally between frontal-parietal sensors, unilateral right frontal-temporal, frontal-occipital temporal and temporal-occipital sensors; in the beta band between left frontal-temporal and right frontal-parietal sensors; in the gamma band between bilateral frontal-parietal and frontal-temporal sensors, right frontal-occipital, temporal-parietal and parietal-occipetal sensors. At +450 ms AV lag (Fig. 5C), significant differences in interaction were observed in the beta band between left temporal-parietal, temporal-occipital and among occipital sensors; in the gamma band among left frontal sensors and among right occipital sensors.

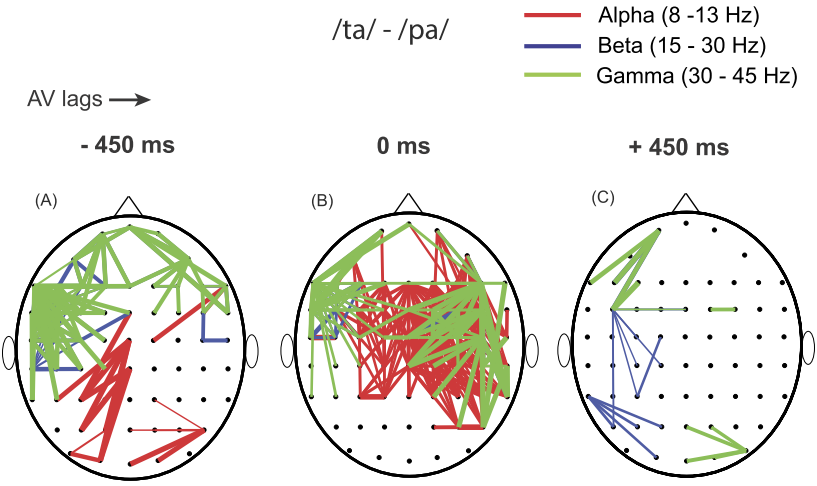

Figure 5. Functional connectivity changes: Imaginary coherence difference between */ta/* and */pa/* response trials. Dots indicate the channel location and the lines indicate channel pairs with statistically significant ($p < 0.001$, see details in text) imaginary coherence changes at different frequency bands as indicated by the color codes in the top right at (A) −450 ms, (B) 0 ms, and (C) +450 ms AV lags.

4. Discussion

In the present study, we used EEG to investigate the spatiotemporal structure of cortical activity underlying multisensory speech perception. We exploited the trial-by-trial variability in the perception of McGurk stimuli to identify the neural representation of multisensory speech perception at different scales. We compared the neural correlates of unisensory and cross-modal perception using identical stimuli at the ERP, spectral and large-scale functional network level. Thus, we could capture the trial-by-trial variability of a participant as well as the segregation-based information-processing mechanisms at the individual sensor level (from ERP, spectral methods) and integration-based information-processing mechanisms (using imaginary coherence) in one single study. The main findings of the study are: (1) A positive peak in the latency range of 300–400 ms serves as a temporal marker of AV integration; (2) decreased post stimulus theta (4–7 Hz), alpha (8–12 Hz) and beta (13–30 Hz) band activity across frontal, temporal sensors and enhanced theta and alpha band activity across occipital sensors act as a spectral signature for cross-modal perception; (3) enhanced functional connectedness at the gamma band with the frontal sensors is pivotal for cross-modal perception.

Previous studies have shown that by presenting certain *semantically-incongruent* AV stimuli, one can induce an illusory (cross-modal) perceptual experience in the participants (Keil *et al*., 2012; MacDonald and McGurk, 1978; McGurk and MacDonald, 1976; Nath and Beauchamp 2011; Van Wassenhove *et al*., 2007 and several others). In the current study, we constructed incongruent AV stimuli by superimposing auditory */pa/* onto video of the speaker articulating */ka/* to induce an illusory percept of */ta/*. Furthermore, studies have also demonstrated that the illusory experience can be modulated by the introduction of AV lags (Munhall at al., 1996; Van Wassenhove *et al*., 2007). Therefore we introduced an AV lag of ±450 ms to our incongruent AV stimuli to generate three conditions overall: −450 ms (audio preceding video), 0 ms (synchronous onsets of audio and video) and +450 ms (video preceding audio) AV lag. We observed that the stability of the illusory percept varied with the introduction of the AV lags. Synchronous AV stimuli resulted in a response of illusory perception that was stable and at a significantly higher frequency of occurrence than the unisensory percept */pa/*, whereas AV lags of −450 ms and +450 ms resulted in lowering of the illusory percept and a higher occurrence of the unisensory percept */pa/*. Additionally, we observed a hit rate of */ta/* responses above 90% for congruent */ta/* stimuli and above 95% during our post-hoc 'auditory alone' behavioral experiment. Our behavioral response results corroborate existing studies of the McGurk effect (Munhall at al., 1996; Van Wassenhove *et al*., 2007) that demonstrate the effect of AV

lags on perception. Furthermore, variability in the perception of identical incongruent stimuli served as an efficient handle to compare and understand the processing of multisensory speech stimuli (Thakur *et al*. 2016).

4.1. Segregation of Information Processing Underlying Illusory Perception

4.1.1. Timing of Neural Information Processing

Converging evidence suggests that conscious perception is marked by a higher P300 component (Pitts *et al*., 2014; Railo *et al*., 2011; Rutiku *et al*., 2015). Our results demonstrate a robust positive peak in the temporal window of 300–400 ms as seen in the ERP difference plot (Fig. 3A) at −450 ms and 0 ms AV lags. The results are further validated by cluster plots of ERPs obtained from millisecond-by-millisecond paired *t*-tests (Fig. 3B) between /*ta*/ and /*pa*/ at all the AV lags. Although no robust peak around 300 ms was observed during +450 ms AV lag, cluster plots demonstrate a difference across central, temporal, centro-parietal and occipital sensors around 300 ms post stimulus onset. Importantly, significant differences in the ERP start only post 300 ms stimulus onset at −450 ms and 0 ms. In addition, interestingly the difference persists longer at ±450 AV lag than at 0 ms AV lag, where the difference was observed in most sensors closely around the 300 ms window. We attribute the persistence of difference beyond 300 ms at ±450 AV lag to the neurophysiological processes involved in binding the information across the two modalities. Considering the asynchronous AV stimuli, one can hypothesize that the neurophysiological process is the working memory that holds the first incoming stimulus (audio or visual) before integrating with the upcoming stimulus. Behavioral studies by Van Wassenhove and colleagues (Van Wassenhove *et al*., 2007) demonstrate 200 ms of asynchrony as the temporal window of AV integration. However, electrophysiological studies understanding preparatory processes show the elicitation of ERP components upto 600–800 ms in response to a cue followed by a target stimulus (Simson *et al*., 1977). In light of this finding we can endorse our speculation of the persistent difference post 300 ms at ±450 AV lag arising from the underlying binding processes. The smaller difference window observed at 0 ms AV lag indicates an integration mechanism that is distinct from the processing when the AV stimuli are time-lagged. These mechanisms can be understood further by inspecting the signals at different scales. Furthermore, we also observe a difference before 300 ms, primarily at the central and parietal electrodes at +450 ms AV lag. These might arise from the anticipatory processes trying to predict the auditory representation following articulatory cues. Our findings here primarily point to the P300 component as the temporal marker of cross-modal perception.

4.1.2. Spectro-Temporal Structure of Brain Rhythms at Each Sensor

Oscillatory cortical activity modulates and drives perception (VanRullen, 2016). Non-parametric statistical comparison of the time–frequency spectrogram between the perceptual categories (*/ta/ − /pa/*) (Fig. 4) highlights the durations and frequencies at each sensor that have significantly different signal power change. The patterns of spectral difference allow us to speculate on the mechanism of AV integration which we discuss in the following paragraph. At −450 AV lag, we observed a suppression in the theta and alpha bands primarily in the frontal, left-temporal and occipital sensors. Similarly, at +450 ms AV lag, we observed a bilateral suppression of spectral power in the frontal and temporal sensors in the theta, alpha, beta and gamma bands.

Theta band activity has been implicated in the encoding of new information and retrieval of episodic memories (Klimesch, 1999; Nyhus and Curran, 2010). Furthermore, suppression of alpha band power has been implicated in attention and language comprehension processes by enabling controlled access to knowledge (Bastiaansen and Hagoort, 2006; Hanslmayr *et al*., 2011; Herrmann and Knight, 2001; Klimesch, 2012; Payne *et al*., 2013; Sigala *et al*., 2014). From an information processing perspective, event-related desynchronization in a local area indicates the onset of preparatory processes (Herrmann and Knight, 2001). Also, differences across the sensors might reflect the activity in the underlying sensory-specific and working memory areas endorsing the fuzzy logical model of perception, in which each input is first independently evaluated with prototypes stored in memory followed by its integration and perception (Massaro, 1989). Our claim arises in the first place from the nature of the stimuli (−450 ms and +450 ms AV lag) as in both cases either the audio */pa/* precedes the articulation or vice-versa. Furthermore, suppression of beta band power has been implicated in top-down control of attention (Engel and Fries, 2010). Additionally, gamma band oscillations have been associated with visual perception, attention and the processing of auditory and spatial information (Kaiser and Lutzenberger, 2005; Kaiser *et al*., 2006). Therefore, the suppression in the beta and gamma bands observed in the left temporal sensors at +450 ms AV lag might be associated with the attention network guiding the perceptual processing. Interestingly, at 0 ms AV lag, we observed a difference in the spectral power predominantly in the occipital sensors followed by the frontal and temporal sensors. We observed enhanced theta and alpha band activity in the occipital sensors; however, we observed suppression in those bands in the frontal and left temporal sensors. Here, a plausible hypothesis behind the emergence of cross-modal perception is the engagement of associative memory networks aided by the synchronous presentation of visual stimuli that integrate the well-learnt audio–visual cues (Albright, 2012).

4.1.3. Integration of Information Underlying Multisensory Perception

To gain insight into the integration of information that occurs in the functional network that disambiguates the two perceptual states, we evaluated the variation in the coherence of ongoing oscillatory activity. In an earlier study (Kumar *et al*., 2016), we showed evidence of a global network being operational during multisensory perception. However, the local sub-networks giving rise to such large-scale interactions were unknown as the real part of coherency is susceptible to volume conduction effects. In the current article, we focus our analysis on the complex part of the coherency, i.e., the imaginary coherence, because this measure is sensitive only to synchronization of two processes that occur with a time lag and are minimally affected by volume conduction (Nolte *et al*., 2004). Also, it reduces the false positive estimates of interactions existent in functional connectivity measures such as absolute coherence and phase synchrony (Guggisberg *et al*., 2008).

Upon non-parametric comparison of the imaginary coherence of */ta/* and */pa/* responses, we observed an enhanced functional connectivity in the alpha band at −450 ms AV lag among the parietal-temporal-occipital sensors and at 0 ms AV lag among bilateral frontal-parietal-temporal and occipital sensors. However, at +450 ms AV lag we did not observe any significant difference in the functional connectivity at the alpha band between the */ta/* and */pa/* responses. Thalamo-cortical and cortico-cortical interactions are thought to be the generators of the human alpha rhythms, with the magnitude of the alpha coherence dependent on the frequency selectivity of the underlying network and the similarity of the inputs. Besides, alpha band synchronization has been associated with short term attentional processes (Kelly *et al*., 2003). Therefore, in the light of the aforementioned studies, the enhanced functional connectivity observed at 0 ms AV lag can be attributed to the attentional network. In addition, at 0 ms AV lag, the AV inputs being synchronous, the enhanced connectivity also reflects the processes involved in scrutinizing the congruency of the AV inputs. At −450 ms AV lag the difference in the connectivity alerts the short-term attentional network operating to integrate the auditory information to the upcoming visual information. However, as auditory processing is faster than the visual (Jain *et al*., 2015; Shelton and Kumar, 2010), at +450 ms AV lag, the temporal lag makes time available for the visual processing of the lip movement and therefore we do not observe an enhanced connectivity emerging from the short-term attentional processes.

Inter-areal coherence of oscillatory activity in the beta frequency range (15–30 Hz) has been has been implicated in top-down processing (Wang, 2010). Furthermore, promoted by the dense anatomical connectivity, the neurons self-organize themselves into large-scale neuronal assemblies called neurocognitive networks (NCN), in reaction to the cognitive demands (Bressler and Richter, 2014). In this context, the increased interaction we observed between

the temporal-parietal, bilateral temporal-parietal and temporal-occipital sensors at −450 ms (Fig. 5A), 0 ms (Fig. 5B) and +450 ms AV lag (Fig. 5C), respectively, provides a long-range inter-areal linkage of distributed cortical areas in NCNs. These also enable the processing of the retrieval of well learnt audio–visual associations as suggested by Albright and colleagues (Albright, 2012).

Enhanced functional connectivity, primarily between the frontal and parietal sensors in the gamma band, was observed at all AV lags. Ther fronto-parietal network has been shown to selectively bias the processing of lower-order sensory systems (Corbetta and Shulman, 2002). Besides, gamma band coherence has been shown to be implicated in voluntary eye movements, saccades and linguistic processing (Balazs *et al*., 2015; Pulvermüller *et al*., 1995). Stimulus selection by attention also induces local gamma band synchronization (Hipp *et al*., 2011). Furthermore, our gaze fixation results on the current data reported in Kumar *et al*. (2016) show enhanced gaze fixation on the mouth (Note 2) during /*ta*/ perception. Combining these data, we hypothesize that selective attention paid to the mouth is the result of a top-down interaction that governs the perceptual processing. Most interestingly, the enhanced functional connectivity (slightly more extensive in right hemisphere) between fronto-temporal, fronto-parietal and fronto-occipital sensors signifies an increase in crosstalk between visual association areas and multisensory and integrative centers of the brain when AV information is synchronous. On the other hand, during the presentation of asynchronous AV stimuli at ±450 ms, a more left-hemisphere-dominant network is operational, presumably due to the presence of pseudo-linguistic stimuli (/*pa*/-/*ka*/-/*ta*/). From the perspective of predictive coding (Sauseng *et al*., 2015; Talsma, 2015), one can infer that the prediction error and the internal representation of the brain can be updated within a small temporal window to process the incoming incongruent AV stimulus. Future studies can explore the boundaries of the temporal windows over which predictive coding is possible.

Overall, we present a multi-scale representation of multisensory speech processing. Although we observe markers at the individual sensor level, our results indicate that a comprehensive account of underlying neural processes emerges only when one analyzes the physiological signals at multiple scales. In the current study, due to the nature of the stimuli we were not able compare between temporal lags. However, future studies can explore such lags at the source level to build a complete picture of multisensory speech processing.

Acknowledgements

This research was funded by NBRC core, a Ramalingaswami fellowship grant (BT/RLF/Re-entry/31/2011) and an Innovative Young Bio-technologist Award

(IYBA) (BT/07/IYBA/2013) from the Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India to AB. DR is supported by a Ramalingaswami fellowship (BT/RLF/Re-entry/07/2014) from the Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India.

Notes

1. The data were used in a different set of analyses in Kumar *et al*. (2016).
2. A detailed analysis of gaze fixations was presented in Kumar *et al*. (2016).

References

Albright, T. D. (2012). On the perception of probable things: neural substrates of associative memory, imagery, and perception, *Neuron* **74**, 227–245.

Allman, B. L., Keniston, L. P. and Meredith, M. A. (2009). Adult deafness induces somatosensory conversion of ferret auditory cortex, *Proc. Natl Acad. Sci. USA* **106**, 5925–5930.

Balazs, S., Kermanshahi, K., Binder, H., Rattay, F. and Bodis-Wollner, I. (2015). Gamma-band modulation and coherence in the EEG by involuntary eye movements in patients in unresponsive wakefulness syndrome, *Clin. EEG Neurosci.* **47**, 196–206.

Bastiaansen, M. and Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension, *Progr. Brain Res.* **159**, 179–196.

Bizley, J. K. and King, A. J. (2012). What can multisensory processing tell us about the functional organization of auditory cortex?, in: *The Neural Bases of Multisensory Processes*, M. M. Murray and M. T. Wallace (Eds), pp. 31–48. CRC Press/Taylor and Francis, Boca Raton, FL, USA.

Bressler, S. L. and Kelso, J. A. S. (2001). Cortical coordination dynamics and cognition, *Trends Cogn. Sci.* **5**, 26–36.

Bressler, S. L. and Richter, C. G. (2014). Interareal oscillatory synchronization in top-down neocortical processing, *Curr. Opin. Neurobiol.* **31C**, 62–66.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C. and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures, *Neuroreport* **14**(17), 2213–2218. DOI:10.1097/01.wnr.0000095492.38740.8f.

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

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain, *Nat. Rev. Neurosci.* **3**, 215–229.

Engel, A. K. and Fries, P. (2010). Beta-band oscillations — signalling the status quo? *Curr. Opin. Neurobiol.* **20**, 156–165.

Green, K. P., Kuhl, P. K., Meltzoff, A. N. and Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect, *Percept. Psychophys.* **50**, 524–536.

Guggisberg, A. G., Honma, S. M., Findlay, A. M., Dalal, S. S., Kirsch, H. E., Berger, M. S. and Nagarajan, S. S. (2008). Mapping functional connectivity in patients with brain lesions, *Ann. Neurol.* **63**, 193–203.

Hanslmayr, S., Gross, J., Klimesch, W. and Shapiro, K. L. (2011). The role of α oscillations in temporal attention, *Brain Res. Rev.* **67**, 331–343.

Hasson, U., Skipper, J. I., Nusbaum, H. C. and Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation, *Neuron* **56**, 1116–1126.

Helfer, K. S. (1997). Auditory and auditory–visual perception of clear and conversational speech, *J. Speech Lang. Hear. Res.* **40**, 432–443.

Herrmann, C. S. and Knight, R. T. (2001). Mechanisms of human attention: event-related potentials and oscillations, *Neurosci. Biobehav. Rev.* **25**, 465–476.

Hipp, J. F., Engel, A. K. and Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception, *Neuron* **69**, 387–396.

Horwitz, B. (2005). Integrating neuroscientific data across spatiotemporal scales, *C. R. Biol.* **328**, 109–118.

Jain, A., Bansal, R., Kumar, A. and Singh, K. D. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students, *Int. J. Appl. Basic Med. Res.* **5**, 124–127.

Jones, J. A. and Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect, *Neuroreport* **14**, 1129–1133.

Kaiser, J. (2004). Hearing lips: gamma-band activity during audiovisual speech perception, *Cereb. Cortex* **15**, 646–653.

Kaiser, J. and Lutzenberger, W. (2005). Human gamma-band activity: a window to cognitive processing, *Neuroreport* **16**, 207–211.

Kaiser, J., Hertrich, I., Ackermann, H. and Lutzenberger, W. (2006). Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli, *NeuroImage* **30**, 1376–1382.

Keil, J., Muller, N., Ihssen, N. and Weisz, N. (2012). On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states, *Cereb. Cortex* **22**, 221–231.

Kelly, S. P., Dockree, P., Reilly, R. B. and Robertson, I. H. (2003). EEG alpha power and coherence time courses in a sustained attention task, in: *First International IEEE EMBS Conference on Neural Engineering, 2003. Conference Proceedings*, pp. 83–86.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, *Brain Res. Rev.* **29**, 169–195.

Klimesch, W. (2012). Controlled access to stored information, *Trends Cogn. Sci.* **16**, 606–617.

Kumar, G. V., Halder, T., Jaiswal, A. K., Mukherjee, A., Roy, D. and Banerjee, A. (2016). Large scale functional brain networks underlying temporal integration of audio–visual speech perception: an EEG study, *Front. Psychol.* **7**, 1558. DOI:10.3389/fpsyg.2016.01558.

Luria, A. R. (1995). *Higher Cortical Functions in Man*. Springer, Boston, MA, USA.

Maris, E., Schoffelen, J.-M. and Fries, P. (2007). Nonparametric statistical testing of coherence differences, *J. Neurosci. Meth.* **163**, 161–175.

Massaro, D. W. (1989). A fuzzy-logical model of categorization behavior, in: *Human Information Processing: Measures, Mechanisms, and Models*, D. Vickers and P. L. Smith (Eds), pp. 367–379. North-Holland, Amsterdam, The Netherlands.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, 691–811.

McIntosh, A. R. (2004). Contexts and catalysts: a resolution of the localization and integration of function in the brain, *Neuroinformatics* **2**, 175–182.

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

Murray, R. F., Bennett, P. J. and Sekuler, A. B. (2002). Optimal methods for calculating classification images: weighted sums, *J. Vis.* **2**, 79–104. DOI:10.1167/2.1.6.

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

Nath, A. R. and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion, *Neuroimage* **59**, 781–787.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S. and Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency, *Clin. Neurophysiol.* **115**, 2292–2307.

Nyhus, E. and Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory, *Neurosci. Biobehav. Rev.* **34**, 1023–1035.

Payne, L., Guillory, S. and Sekuler, R. (2013). Attention-modulated alpha-band oscillations protect against intrusion of irrelevant information, *J. Cogn. Neurosci.* **25**, 1463–1476.

Pitts, M. A., Padwal, J., Fennelly, D., Martínez, A. and Hillyard, S. A. (2014). Gamma band activity and the P3 reflect post-perceptual processes, not visual awareness, *NeuroImage* **101**, 337–350.

Pulvermüller, F., Lutzenberger, W., Preissl, H. and Birbaumer, N. (1995). Spectral responses in the gamma-band: physiological signs of higher cognitive processes? *Neuroreport* **6**, 2059–2064.

Railo, H., Koivisto, M. and Revonsuo, A. (2011). Tracking the processes behind conscious perception: a review of event-related potential correlates of visual consciousness, *Conscious. Cogn.* **20**, 972–983.

Rutiku, R., Martin, M., Bachmann, T. and Aru, J. (2015). Does the P300 reflect conscious perception or its consequences? *Neuroscience* **298**, 180–189.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W. and Foxe, J. J. (2007). Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion, *Neuropsychologia* **45**, 587–597.

Sauseng, P., Conci, M., Wild, B. and Geyer, T. (2015). Predictive coding in visual search as revealed by cross-frequency EEG phase synchronization, *Front. Psychol.* **6**, 1655. DOI:10.3389/fpsyg.2015.01655.

Sekiyama, K., Kanno, I., Miura, S. and Sugita, Y. (2003). Auditory–visual speech perception examined by fMRI and PET, *Neurosci. Res.* **47**, 277–287.

Shelton, J. and Kumar, G. P. (2010). Comparison between auditory and visual simple reaction times, *Neurosci. Med.* **1**, 30–32.

Sigala, R., Haufe, S., Roy, D., Dinse, H. R. and Ritter, P. (2014). The role of alpha-rhythm states in perceptual learning: insights from experiments and computational models, *Front. Comput. Neurosci.* **8**, 36. DOI:10.3389/fncom.2014.00036.

Simson, R., Vaughan, H. G. and Ritter, W. (1977). The scalp topography of potentials in auditory and visual Go/NoGo tasks, *Electroencephalogr. Clin. Neurophysiol.* **43**, 864–875.

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C. and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception, *Cereb. Cortex* **17**, 2387–2399.

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B. and James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception, *NeuroImage* **49**, 3308–3318.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Am.* **26**, 212–215.

Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind, *Front. Integr. Neurosci.* **9**, 19. DOI:10.3389/fnint.2015.00019.

Thakur, B., Mukherjee, A., Sen, A. and Banerjee, A. (2016). A dynamical framework to relate perceptual variability with multisensory information processing, *Sci. Rep.* **6**, 31280. DOI:10.1038/srep31280.

Van Wassenhove, V., Grant, K. W. and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech, *Proc. Natl Acad. Sci. USA* **102**, 1181–1186.

Van Wassenhove, V., Grant, K. W. and Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception, *Neuropsychologia* **45**, 598–607.

VanRullen, R. (2016). Perceptual cycles, *Trends Cogn. Sci.* **20**, 723–735.

Wallace, M. T., Meredith, M. A. and Stein, B. E. (1993). Converging influences from visual, auditory, and somatosensory cortices onto output neurons of the superior colliculus, *J. Neurophysiol.* **69**, 1797–1809.

Wang, X. J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition, *Physiol. Rev.* **90**, 1195–1268.