Sound and Movement

Komeilipoor, N.

2015

Chapter 6

The timing in superior temporal areas predicts (mis-)perception of speech in the presence of sensorimotor stimuli

The perception of a speech sound is affected by looking at facial motion. Incongruence between sound and watching somebody articulating may yield a bias toward the visual percept, coined the McGurk effect. We tested the degree to which silent articulation of a syllable also induces such a bias and searched for its neural correlates using EEG. Listeners were instructed to identify the auditory syllables /pa/ and /ta/ while silently articulating congruent/incongruent syllables or observing videos of a speaker's face articulating them. As baseline we included an auditory only condition without competing visual or sensorimotor input. As expected perception of sound was diminished when incongruent syllables were observed and when silently articulated albeit to lesser degree. This was accompanied by significant amplitude modulations in the beta frequency band in right superior temporal areas. There, the event-related beta activities during congruent conditions were phase locked to responses evoked during the auditory only condition. This implies that proper temporal alignment of different input streams in right superior temporal areas is mandatory for both audiovisual and audiomotor speech integration.

Submitted for publication: Komeilipoor, N., Cesari, P., & Daffertshofer, A. (under review)

Introduction

The brain receives a continuous stream of information from different sensory modalities. Proper integration of input is essential for accurate perception. Seminal for this is the perception of speech sound, which is clearly affected by observation of facial motion, referred to as the McGurk effect (McGurk and MacDonald, 1976). This has inspired many researchers investigating multisensory integration (Tiippana, 2014). The perception of a sound syllable can also be affected by tactile stimulation (Gick and Derrick, 2009, Ito et al., 2009) and, here particularly interesting, by silent articulation of speech sounds (Sams et al., 2005, Mochida et al., 2013, Sato et al., 2013).

Neuroimaging revealed an involvement of the superior temporal sulcus/gyrus (STS/STG) in the McGurk effect (Calvert et al., 2000, Jones and Callan, 2003, Sekiyama et al., 2003, Bernstein et al., 2008, Irwin et al., 2011, Nath and Beauchamp, 2012, Szycik et al., 2012, Erickson et al., 2014). What precisely happens in this area to accomplish multisensory integration, however, is still unclear.

Here we capitalize on the competition between auditory and visual input as well as between auditory and sensorimotor inputs to probe how cortical oscillations contribute to multisensory integration. We adopted a protocol recently introduced by Mochida et al. (2013), in which listeners are instructed to identify the auditory syllables while silently articulating congruent/incongruent syllables, or observing videos of a speaker's face articulating congruent/incongruent syllables. Cortical activity was monitored using electro-encephalography (EEG).

Consistent with the McGurk effect (McGurk and MacDonald, 1976), we expected, when dubbing the acoustic syllable /pa/ onto the visual presentation of articulatory gestures of /ta/, subjects to typically misperceive the sound. We also expected a similar result when subjects themselves silently articulated an incongruent syllable (Sams et al., 2005, Mochida et al., 2013, Sato et al., 2013). Furthermore, we expected conventional source localization of EEG to reveal STS/STG as areas discriminating between proper and improper perception, in support with the aforementioned imaging studies. Finally, we hypothesized the phase dynamics in STS/STG to be essential for multisensory integration, as we believe that temporal alignment of distinct sensory streams is key to their integration.

Methods

Subjects

Twelve volunteers (mean age 26.1 years, 5 females) participated after giving their written informed consent. All were right handed and had normal hearing and normal or corrected-to-normal vision.

Protocol

The experimental protocol has been adopted from a recent study by Mochida *et al.* (2013). The ethics committee of the Faculty of Human Movement Sciences, VU University Amsterdam had approved it prior to conduction.

Task

Participants were asked to identify the syllables /pa/ and /ta/ that they heard under the following subtask conditions: silently articulating congruent/incongruent syllables (*motor condition*), observing videos of a speaker's face articulating congruent/incongruent syllables (*visual condition*), and a condition without a subtask (*baseline condition* or *auditory only*); see Figure 1 for overview. In the motor condition, participants were instructed to articulate the syllables with as little vocalization as possible while moving the lips and tongue as much as possible.

Stimuli

Stimuli had been produced by a Dutch male speaker. We recorded conventional videos at 50 Hz frame rate. Audio signals were digitized at a rate of 44.1 kHz. They were delivered at a level of 60 dB through paired speakers placed in front of the participants (distance 55 cm to the participant's torso) and separated by approximately 30 cm. We superimpose white noise to the syllables (signal-to-noise ratio of 5 dB). Beginning and end of the noise were faded in and out, respectively (.5s duration). Syllables were preceded by four clicks (.67s inter-click interval) to provide a cue for silently articulating a syllable in the motor condition.

The auditory syllables were paired to visual and motor components yielding four different combinations: (i) congruent /pa/ (visual/motor /pa/ auditory /pa/), (ii) congruent /ta/ (visual/motor /ta/ auditory /ta/), (iii) incongruent stimuli (visual/motor /pa/ auditory /ta/), and (iv) the converse incongruent stimuli (visual/motor /ta/ auditory /pa/).

In the *motor condition*, English characters representing /pa/ or /ta/ were presented on a front display (LCD monitor, frame rate 60 Hz, about 55 cm in front of the participant's nasion) until participants pressed the space bar of a computer keyboard to start the trial. They were asked to silently articulate the indicated syllable in time with the clicks and the onset of the syllable while watching a still frame of the video.

For the *visual condition*, a video of the speaker's face articulating either /pa/ or /ta/ was presented on the front display). Prior to video presentation, the initial frame of the video was presented from the onset of white noise until the onset of the syllable.

In the *baseline* (*auditory only*) *condition*, participants were asked to listen to the auditory stimuli while watching a still frame of the video.

Procedures

The experimental session consisted of a familiarization phase followed by a test phase. In the first, participants performed one baseline condition block followed by one motor condition block (see below for block definition). In the latter, the subjects performed three sets of motor and visual condition blocks, with the order of the two blocks randomized within each set and counterbalanced for each participant. After these six blocks, they also performed a baseline block, and a resting state recording both before and after every block; the latter will be reported elsewhere. After each stimulus presentation, the participants were asked to select which of the syllables they perceived among four possible alternatives, /pa/, /ta/, /ka/, or 'etc' displayed on the screen. They provided their responses using the computer keyboard. The next trial was initiated 10s after the participants entered their response.

Motor and visual condition blocks consisted of 60 trials in which each of the four different combinations (2 auditory syllables × 2 subtask syllables) were performed 15 times. A baseline condition block consisted of 30 trials in which each of the two auditory syllables were presented 15 times. The order of trials per block was randomized, the order blocks was randomized across participants.

We recorded EEG using a 64-channel amplifier (Refa, TMSi, Enschede, The Netherlands; Ag/AgCl electrodes mounted in an elastic cap and two mastoid electrodes) and sampled the signals at a rate of 1024 Hz. We co-registered EMG activity from left and right masseter and digastric muscles using Ag/AgCl surface electrodes (sampling 1 kHz, 16 channel Porti amplifier, TMSi, Enschede The Netherlands).

All subsequent analyses were realized using Matlab 2014a (The Mathworks, Natwick, MA) including the open-source fieldtrip toolbox (fieldtrip.fcdonders.nl).

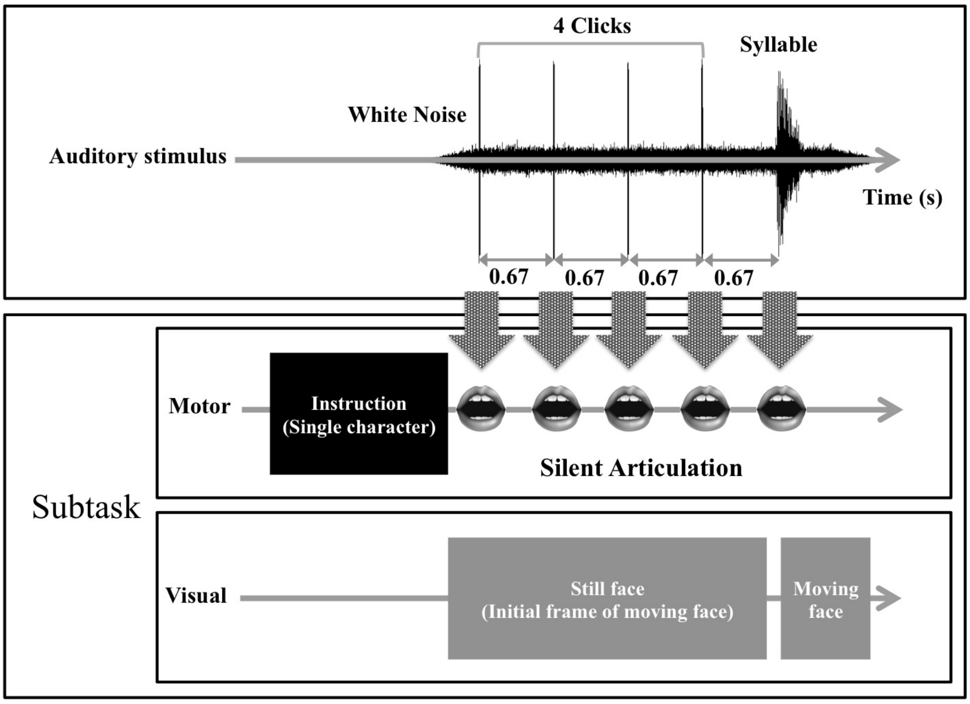

Figure. 1. *Scheme of the experimental protocol* (modified from Mochida et al., 2013). The stimulus was preceded by four clicks at .67s intervals which provided the subjects with a cue to silently articulate a syllable under the motor condition in which the syllables to be articulated by the participants were presented visually using English characters, which disappeared at the second click. Participants were asked to silently articulate the indicated syllable in time with the clicks and the onset of the syllable, while watching a still video frame. In the visual condition, videos of a speaker's face producing the syllables were presented. The videos were synchronized with the auditory stimulus. The initial frame of each video was presented from the noise onset to the stimulus onset.

Data analysis

Behavioral data

For each auditory stimulus and condition, 45 responses were collected from which the error-response rates were determined as a measure of syllable intelligibility. Subsequently, rates were categorized and averaged over trials for every participant according to the combination of visual and motor tasks with the auditory stimuli as (sets of) congruent or incongruent stimuli. The mean error-response rate of the baseline condition was subtracted from those of the congruent and incongruent stimuli sets. The resulting, unbiased rates were compared using a two-way repeated-measures analysis of variance (ANOVA) with *condition* (motor/visual), and *stimulus* (congruent/incongruent) as within-subjects factors. Post-hoc comparisons were performed via t-tests applying a

Bonferroni correction for multiple comparisons when required. A partial-eta-squared statistic served to estimate effect size.

Electrophysiological data

We considered the interval between the third and forth (last) click onset (~ pre) and the interval between the last click onset and the end of syllable presentation (~ post) intervals of interest. In detail, we selected epochs of 1.34s (± .67s) around last click onset (= speech onset). Equivalent to the behavioral responses, trials were categorized into two stimuli categories: congruent and incongruent.

Preprocessing. All signals were filtered with a 50 Hz notch filter (2^{nd}-oder bi-directional Butterworth) to remove power-line artifacts. We reduced movement artifacts by high-pass filtering (2^{nd}-order bi-directional Butterworth, cut-off frequency 3 Hz for EEG and 10 Hz for EMG signals). Since EMG signals were collected in a bipolar montage, we full-wave rectified them using the modulus of the corresponding analytic signal also referred to as Hilbert-amplitude. Subsequently, we removed movement and muscle artifacts by combining EEG and EMG and applying independent component analysis detailed as *Supplementary Material*.

Source localization. We concentrated on neural activity at the beta frequency band (15-30 Hz) using beamformers based on dynamic imaging of coherent sources (DICS; Gross et al., 2001); the analysis of other frequency bands can be found as *Supplementary Material.* We note that beta band oscillations are prospectively modulated during the McGurk illusion (Keil et al., 2012) and, of course, also during motor tasks (Van Wijk et al., 2009, Houweling et al., 2010). The beta band is hence primary target to assess the role of phase entrainment in multisensory processing.

Cross-spectral density matrices of all conditions were determined using a multi-tapering method with center frequency of 22.5 ± 7.5 Hz for a time period of .67s before and after syllable onset. A volume conduction model was derived from the MNI template brain resulting in an anatomically realistic 3-shell model. The lead-field matrix was estimated using the boundary element method (BEM) for each grid point in the brain. A grid with 5 mm resolution was normalized onto a standard Montreal Neurological Institute (MNI) brain in order to calculate group statistics and for illustrative purposes. To establish significance across participants, we used a non-parametric permutation t-statistic (Monte Carlo method; 1000 iterations; $\alpha < .05$, (Oostenveld et al., 2010). By this we identified voxels with statistically significant activity contrasts pre and post stimulus presentation.

To determine the time course of activity of the most significant areas we employed linearly constrained minimum variance (LCMV) beamformers (Van Veen et al., 1997). Epochs of pre and post intervals were band-pass filtered in beta frequency band with a second-order bi-directional Butterworth filter and projected onto the location determined via DICS. The resulting time series were aligned at stimulus onsets per subtask conditions (45 signals for visual and motor, 15 signals for auditory alone) and averaged. We used the post-stimulus intervals to assess the phase relation between the beta band event-related activities after congruent/incongruent stimuli relative to baseline.

Phase dynamics. We defined the phase at each source via the analytical signal, i.e. we computed the instantaneous Hilbert-phase of the beta-band filtered LCMV beamformed EEG. The degree of phase synchrony in the motor and video conditions was estimated as the phase-locking value (PLV, Mormann et al., 2000) of the difference between the respective phases and the phase in the baseline condition. If the phase in the two conditions is not altered relative to baseline, then the corresponding PLV will be equal to 1, otherwise it will be smaller bounded by 0. The baseline was the *auditory only* condition, in which auditory processing was ‘optimal’. A large PLV hence implies visual or sensorimotor input streams to be entrained to the auditory one.

PLVs underwent the same statistical assessment as the error-response rates, i.e. a two-way repeated-measures analysis of variance (ANOVA) with *condition* (motor/visual), and *stimulus* (congruent/incongruent) as within-subjects factors and a post-hoc t-tests with Bonferroni correction when required (partial-eta-squared as effect size).

Results

Behavioral data

For the error-response rates in motor and visual conditions we found a significant main effect for *condition* ($F_{1,11}$=48.318, *p*<.001, η^2=.815). Rates were larger for the visual than the motor subtask. We also found a significant main effect for *stimulus* ($F_{1,11}$=81.022, *p*<.001, η^2=.88), where rates were larger for incongruent than congruent stimuli. The interaction between *condition* and *stimulus* was also significant ($F_{1,11}$=17.343, *p*<.005, η^2=.612); for both the visual and motor condition the response rate was larger for incongruent compared to congruent stimuli (*p*<.001 and *p*<.005, respectively). For the incongruent stimuli the rate was significantly larger in the visual than in the motor condition (*p*<.001); cf. Figure 2.

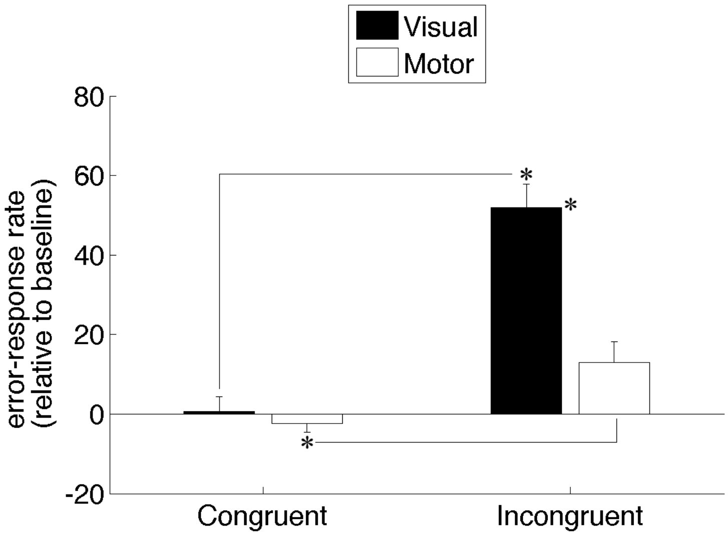

Figure 2. *Effect of congruent and incongruent stimuli on syllable intelligibility*. The error-response rate relative to baseline across all 12 participants for both congruent and incongruent stimuli during the two conditions (visual and motor). Error bars denote standard errors. Significant comparisons between conditions are highlighted with an asterisk (* *p*<005).

Electrophysiological data

Source localization. DICS beamformer analysis revealed contributing areas in temporal, occipital, and parietal cortices. Most consistent sources were located around right STG, which will be highlighted in what follows. An overview of the other sources (and other frequency bands) is provided as *Supplementary Material*.

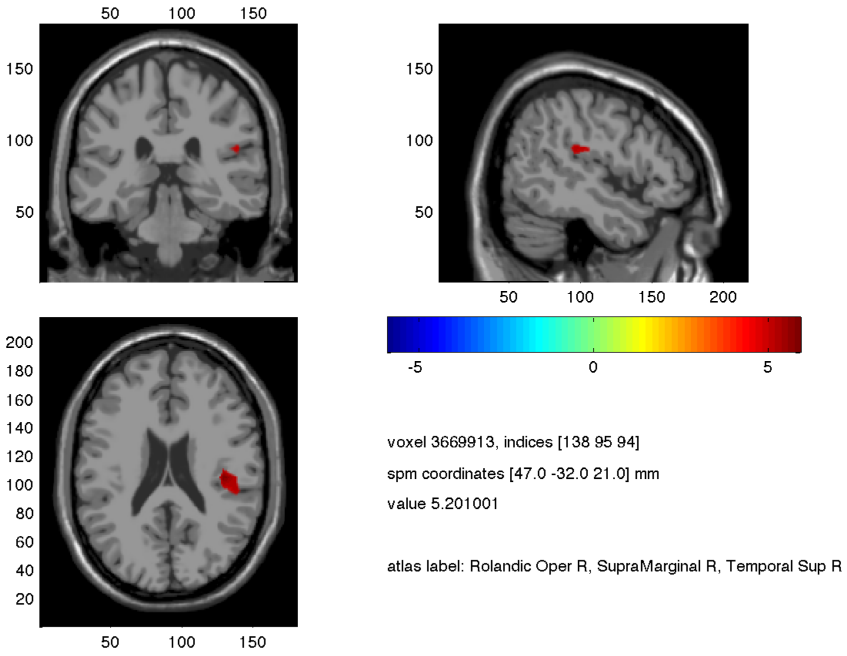
50
100
150
150
100
50
150
100
50
50
100
150
200
-5
0
5
200
180
160
140
120
100
80
60
40
20
50
100
150
voxel 3669913, indices [138 95 94]
spm coordinates [47.0 -32.0 21.0] mm
value 5.201001
atlas label: Rolandic Oper R, SupraMarginal R, Temporal Sup R

Figure. 3. *Example of DICS source projection.* Right Rolandic operculum, right superior temporal gyrus and right supramarginal gyrus were identified as possible generators of the beta-band changes. The red areas represent cortical tissue displaying a significant difference between the time periods of .67s before versus after syllable onset. The image was transformed to MNI template space and overlaid on the template structural image. The peak coherence was observed at MNI coordinates (47, −32, 21).

Right STG, right supramarginal gyrus, and right Rolandic operculum turned out to be the three major sources contrasting conditions from baseline in the beta band; an example is given in Figure 3. This underscores the role of STG in (mis-)perception of sounds (e.g., Beauchamp et al., 2010). We selected the averaged maximum values of seven voxel locations (MNI coordinates: 47, -31, 16).

Phase dynamics. The PLV of the event-related beta activities in visual and motor conditions relative to the auditory only condition revealed a significant main effect for *stimulus* ($F_{1,11}$=6.3, p<.05, η^2=.364) with PLV being significantly larger for congruent than for incongruent stimuli. The interaction between *condition* and *stimulus* was not significant ($F_{1,11}$=1.643, p>.05, η^2=.13); see Figure 4.

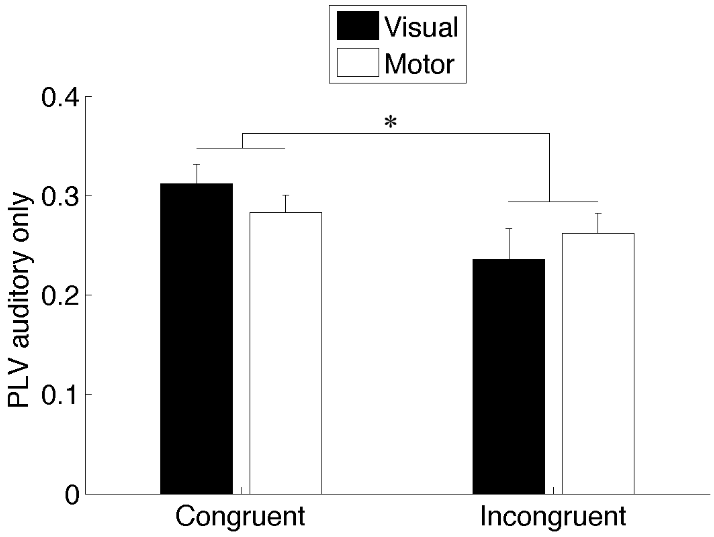

Figure 4. *PLV in the right STG between event-related activities during visual and motor conditions and baseline across all 12 participants for both congruent and incongruent stimuli*. Error bars denote standard errors. Significant comparisons between conditions are highlighted with an asterisk (* p<.05).

Discussion

We used EEG to identify neural correlates of the influence of observing and/or silently articulating of congruent/incongruent syllables on auditory perception. We found that (a) perception of auditory syllables is degraded when the subjects observed and, to a lesser degree, when silently articulated incongruent syllables; (b) beta band activity in STG discriminated congruent from incongruent stimuli and – given result (a) – proper and improper perception; (c) as hypothesized, in STG the event-related beta modulations during congruent conditions (i.e. properly perceived syllables) were phase locked to the responses evoked during the auditory only condition.

Different studies evidenced involvement of the motor system in speech and gestural perception (Fadiga et al., 2002, Komeilipoor et al., 2014). But does that imply that the motor system itself processes perception? The behavioral results (a) do confirm that speech motor control contributes to listening ability, which supports the general idea of a close link between speech production and perception (Sams et al., 2005, Mochida et al., 2013, Sato et al., 2013). Our EEG source results (b), however, do hint at a major involvement of auditory rather than the motor system in audio-articulatory interaction. The importance of STG in the McGurk illusion, in particular, and audiovisual speech integration, in general, has already been underscored in several imaging studies

(Calvert et al., 2000, Jones and Callan, 2003, Sekiyama et al., 2003, Bernstein et al., 2008, Irwin et al., 2011, Nath and Beauchamp, 2012, Szycik et al., 2012, Erickson et al., 2014). It has also been shown that fMRI-guided TMS delivered over STG yields a significant reduction in the perception of the McGurk illusion (Beauchamp et al., 2010). A recent MEG study showed that the perception of the McGurk illusion is preceded by high beta activity in STG (Keil et al., 2012) meaning that not only the overall activity and excitability in STG is important but also its dynamics. Our results corroborate this suggestion, implying that especially right STG plays a central role not only for audiovisual but also for audiomotor speech convergence. Apparently both types of multisensory integrations rely on the same neural mechanism.

In line with (Keil et al., 2012), the beta frequency band was most informative when pinpointing STG. More importantly, however, our finding (c), i.e. the pronounced phase locking between event-related activities during congruent conditions and the auditory-only baseline suggests that proper timing between the input through different modalities is mandatory for their proper integration. If stimuli are incongruent, it appears that additional processing of (one of) the individual inputs yields a beta-desynchronization in STG and, by this, a misperception of sound. Speculating about more detail is beyond the scope of the current study, as using sole 64-channel EEG comes with certain limitations. Future studies should address the involvement of superior temporal auditory regions with more spatial resolution but we suggest employing the (modified) by Mochida *et al*. (2013) as it allows for tackling not only the integration of auditory and visual input but also that of motor information.

Supplementary material

Artifact removal. We identify which EEG components were contaminated by EMG, by combining all z-scored signals (64·EEG + 6·EMG = 70 channels) and conducted an independent component (IC) analysis. Independent component analysis (ICA) was carried out in the Fieldtrip toolbox using the logistic infomax ICA algorithm of Bell and Sejnowski (1995).

ICs that were contaminated by EMG were discarded, and from the remaining ICs we reconstructed the EMG-free EEG. For this, we defined EMG-contamination (1) as being highly correlated with (one of) the z-scored EMG signals and the six ICs with highest correlation were selected for removal. (2) ICs with median frequency higher than the minimum (nine subjects) or the average (three subjects) of the median frequency of the EMG signals were also selected for removal.

We identified on average 26 ICs as artifacts.

Finally, the EEG reconstruction was realized by multiplying the ICs via the pseudo-inverse of the thus reduced mixing matrix (i.e. the de-mixing matrix) followed by inverting the aforementioned z-scoring.

DICS results at different frequency bands.

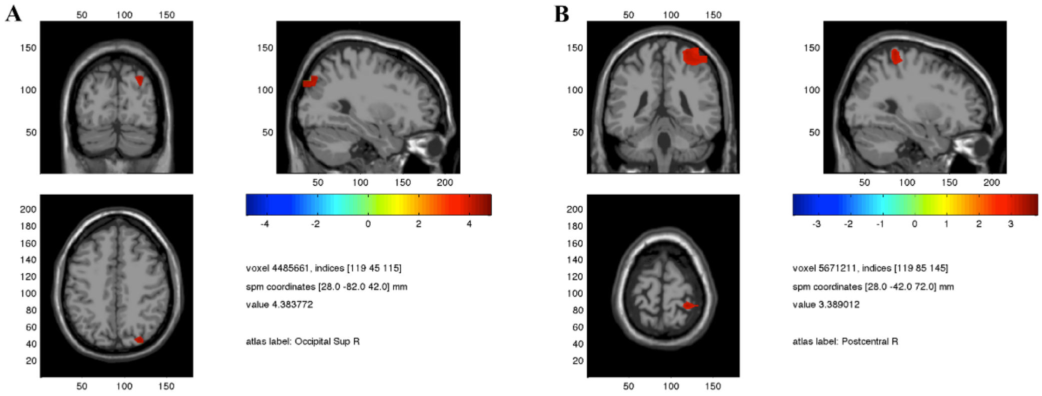
A
50 100 150
150 100 50
200 180 160 140 120 100 80 60 40 20
50 100 150 200
-4 -2 0 2 4
voxel 4485661, indices [119 45 115]
spm coordinates [28.0 -82.0 42.0] mm
value 4.383772
atlas label: Occipital Sup R
B
-3 -2 -1 0 1 2 3
voxel 5671211, indices [119 85 145]
spm coordinates [28.0 -42.0 72.0] mm
value 3.389012
atlas label: Postcentral R

Figure S1. *DICS source projections.* Panel A shows sources in visual areas (28, −82, 42), panel B in M1-S1 (28, −42, 72); see Tables S1-S3 for overview.

Table S1. Sources localized using DICS contrasting pre versus post syllable presentation periods during congruent and incongruent visual subtasks.

		Visual	
	Subtask	Congruent	Incongruent
Frequency Range	Theta (4-8 Hz)	Calcarine R	Postcentral R
	Alpha (8-12 Hz)	Rolandic oper R Heschl R Temporal sup R	Precentral R
	Beta (15-30 Hz)	Temporal sup R SupraMarginal R	Parietal Sup L
	Gamma (30-80 Hz)	Occipital sup R	Occipital sup R

Table S2. Sources localized using DICS contrasting pre versus post syllable presentation periods during congruent and incongruent motor subtasks.

		Motor	
	Subtask	Congruent	Incongruent
Frequency Range	Theta (4-8 Hz)	Calcarine R	Calcarine L
	Alpha (8-12 Hz)	Postcentral R SupraMarginal R	Temporal inf R
	Beta (15-30 Hz)	Rolandic oper R SupraMarginal R Temporal sup R	Rolandic oper R SupraMarginal R Temporal sup R
	Gamma (30-80 Hz)	Temporal sup R	Occipital sup R

Table S3. Source areas localized using DICS contrasting pre versus post syllable presentation periods during auditory alone condition.

		Auditory
	Subtask	#
Frequency Range	Theta (4-8 Hz)	Postcentral R
	Alpha (8-12 Hz)	Postcentral R
	Beta (15-30 Hz)	Rolandic oper R SupraMarginal R Temporal sup R
	Gamma (30-80 Hz)	SupraMarginal R Temporal sup R