Naturalistic audiovisual illusions reveal the cortical sites involved in the multisensory processing of speech

Megevand, Pierre Bastien; Thézé, Raphaël; Mehta, Ashesh

How to cite

MEGEVAND, Pierre Bastien, THÉZÉ, Raphaël, MEHTA, Ashesh. Naturalistic audiovisual illusions reveal the cortical sites involved in the multisensory processing of speech. In: European journal of neuroscience, 2025, vol. 61, n° 5, p. e70043. doi: 10.1111/ejn.70043

This publication URL: https://archive-ouverte.unige.ch/unige:183576
Publication DOI: 10.1111/ejn.70043

This is the accepted version of an article accepted for publication in European Journal of Neuroscience (doi: 10.1111/ejn.70043).

Naturalistic audiovisual illusions reveal the cortical sites involved in the multisensory processing of speech

Pierre Mégevand (1,2,3), Raphaël Thézé (3), Ashesh D. Mehta (4,5)

1. Department of Clinical Neuroscience, Faculty of Medicine, University of Geneva, Switzerland
2. Division of Neurology, Geneva University Hospitals, Switzerland
3. Department of Fundamental Neuroscience, Faculty of Medicine, University of Geneva, Switzerland
4. Department of Neurosurgery, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA
5. The Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA

Corresponding author: Pierre Mégevand, Department of Fundamental Neuroscience, Centre Médical Universitaire, Rue Michel Servet 1, 1211 Genève 4, Switzerland; +41 22 379 53 88; pierre.megevand@unige.ch.

Running title
Processing audiovisual speech illusions

Word count
7094 words

Figure count
7 figures

Keywords
Audiovisual speech; multisensory processing; speech and language; comprehension; humans; intracranial electroencephalography.

Abstract

Audiovisual speech illusions are a spectacular illustration of the effect of visual cues on the perception of speech. Because they allow dissociating perception from the physical characteristics of the sensory inputs, these illusions are useful to investigate the cerebral processing of audiovisual speech. However, the meaningless, monosyllabic utterances typically used to induce illusions are far removed from natural communication through speech. We developed naturalistic speech stimuli that embed mismatched auditory and visual cues within grammatically correct sentences to induce illusory perceptions in controlled fashion. Using intracranial EEG, we confirmed that the cortical processing of audiovisual speech recruits an ensemble of areas, from auditory and visual cortices to multisensory and associative regions. Importantly, we were able to resolve which cortical areas are driven more by the auditory or the visual contents of the speech stimulus or by the eventual perceptual report. Our results suggest that higher-order sensory and associative areas, rather than early sensory cortices, are key loci for illusory perception. Naturalistic audiovisual speech illusions represent a powerful tool to dissect the specific roles of individual cortical areas in the processing of audiovisual speech.

Abbreviations

A	auditory
HFA	high-frequency activity
iEEG	intracranial electroencephalography
V	visual

1 Introduction

Speech is multisensory: in natural circumstances, the movements of the speaker's mouth, face and body are visible to their interlocutor. The visual speech cues complement the information transmitted by the auditory speech stream and can improve the comprehension of what is being said, especially when the auditory signal is corrupted by noise (Sumby & Pollack, 1954). The cortical sites underlying the multisensory processing of audiovisual speech include visual and auditory cortices, multisensory cortex in the posterior superior temporal lobes, as well as frontal cortical areas (see Beauchamp, 2016 for a review). However, perhaps due to the dynamic and multilayered nature of the audiovisual speech signal, important questions remain about the exact role of each of these sites, from the low-level processing of the auditory and visual speech cues' physical characteristics up to that of the lexical and semantic contents of speech.

Audiovisual speech illusions like the well-known McGurk effect are a striking demonstration of the influence of visual cues on the perception of speech (McGurk & Macdonald, 1976; Jiang & Bernstein, 2011). In these illusions, the presentation of mismatched auditory and visual speech cues drives perception away from the auditory stimulus (Tiippana, 2014; Thézé, Gadiri, *et al.*, 2020). Because these illusions allow dissociating perception from the physical characteristics of the sensory inputs, they represent a powerful experimental tool for investigations into the neural substrates of audiovisual speech processing. Indeed, several cognitive neuroscience studies used audiovisual speech illusions to probe these substrates (e.g. Benoit *et al.*, 2010; Nath & Beauchamp, 2012; Smith *et al.*, 2013; Erickson *et al.*, 2014; Pratt *et al.*, 2015; Tse *et al.*, 2015; Uno *et al.*, 2015; Kumar *et al.*, 2018; Proverbio *et al.*, 2018; Li *et al.*, 2021).

These studies, however, have yielded sometimes contradictory findings, in particular regarding the implication of early sensory cortices in the perception of audiovisual speech illusions. For instance, Nath and Beauchamp (2012), using functional MRI to study cortical responses to meaningless mismatched audiovisual syllables, found that activity in early auditory or visual cortex did not directly correlate with the participants' illusory perception. By contrast, Smith et al. (2013), using intracranial EEG (iEEG), observed that, when the patients' perception of these syllables was driven by the visual speech cue, rather than the auditory one, auditory cortical activity was more similar to that of the perceived (but unheard) sound than to the auditory stimulus that was actually presented. This apparent discrepancy might stem from limitations in stimulus and task design, which we detail below.

Some limitations relate to the way the McGurk effect and similar illusions are typically induced. The mismatched audiovisual stimuli usually consist of isolated syllables, which obviously depart from natural speech because they fail to convey any meaning (Van Engen *et al.*, 2022). Furthermore, the key viseme-phoneme pair is generally placed at, or very close to, the onset of the utterance, which means that preparatory articulatory gestures could disproportionally influence cortical activity and perception (Schwartz & Savariaux, 2014). A second set of issues concerns the separability of stimulus- and perception-related cortical activities: in previous work, not all possible combinations of matched and mismatched viseme-phoneme pairs were systematically presented, and perception was not systematically related to the stimuli at the single-participant, single-trial level. Thus, it was impossible to isolate which factor predominantly drove cortical activity: the physical characteristics of the stimuli's auditory and visual components; the presentation of mismatched vs. matched stimuli; the participants' perceptual report of one stimulus or the other; or any interactions between these factors.

In an attempt to lift some of these limitations, and to re-examine the role of auditory and visual cortex in the perception of audiovisual speech illusions, we designed two audiovisual speech stimulus sets where one key word embedded in a sentence could begin indifferently with a plosive or a fricative speech cue without impacting the sentence's syntactical and semantic integrity. We used intracranial EEG and a mass univariate analysis approach based on linear regression to resolve the cortical

processing of these complex, naturalistic stimuli with optimal resolution and signal-to-noise ratio at the single-trial level (Parvizi & Kastner, 2018).

2 Materials & methods

In designing the experiments as well as in their analysis and presentation, we strived to adhere to guidelines and recommendations on iEEG research (Mercier *et al.*, 2022).

2.1 Participants

Patients with drug-resistant focal epilepsy who were undergoing iEEG monitoring participated in the experiments (experiment 1: 4 patients from North Shore University Hospital, NY, USA; experiment 2: 4 patients from Geneva University Hospitals, Switzerland). Electrode implantation was determined solely on clinical grounds, without reference to the present study. All participants were fluent speakers of English (experiment 1) or French (experiment 2). The patients provided written informed consent under the guidelines of the Declaration of Helsinki, monitored by the relevant institutional review board (experiment 1: Feinstein Institutes for Medical Research IRB; experiment 2: Commission cantonale d'éthique de la recherche de la République et canton de Genève).

2.2 Stimuli and tasks

Experiment 1: the stimuli (Figure **1a**) consisted of videos showing the lower part of the face (from the nose downwards) of a male speaker uttering the following sentences: "She had to give the definition of the word 'bet' in front of her class." The key word could be either 'bet' or 'vet' (or 'wet' in a minority of trials). The videos were recorded with a smartphone camera.For mismatched stimuli, the soundtrack of a given video was paired with the image stream of another one. Several iterations of the speaker uttering the sentences were recorded, and the pairing of the mismatched stimuli was done empirically to minimize perceptible delays between the auditory and visual streams. We expected that the perception of mismatched $V_{vet}A_{bet}$ stimuli would be 'vet' in a majority of cases (i.e. driven by the visual component of the stimulus), while the perception of mismatched $V_{bet}A_{vet}$ stimuli would be 'vet' in most occurrences (i.e. driven by the auditory component; see Jiang & Bernstein, 2011; see Thézé, Gadiri, *et al.*, 2020; Thézé, Giraud, *et al.*, 2020 for similar results). The mismatched stimuli involving 'wet' were not expected to elicit any illusory percept (see Jiang & Bernstein, 2011 for similar $V_{wa}A_{ba}$ stimuli). They were used as a behavioral control and were not analyzed further here. Supporting table **S1** details the number of trials of each type that each participant was exposed to. In order to keep the total duration of the experiment reasonably short for patients (Mercier *et al.*, 2022), not all stimulus combinations were presented in equal numbers; rather, we sought to increase the numbers of mismatched over matched trials. Furthermore, we reduced the number of trials including the 'wet' key word after the first two participants.

To verify that the mismatched audiovisual stimuli would elicit illusory perceptions (defined here as perceptions that did not correspond to the auditory stimulus), we ran a pilot experiment in 5 healthy lab members. Between 18 and 21 repetitions of each stimulus were presented as detailed below. For matched stimuli, responses were almost systematically correct (a single participant answered incorrectly on a single occasion). The mismatched $V_{vet}A_{bet}$ combination, which is expected to yield the highest rate of illusory perception, elicited an illusory perception in 100%, 95%, 100%, 0% and 35% of trials (values per individual lab member), highlighting the large variability in susceptibility to audiovisual speech illusions across individuals (Nath & Beauchamp, 2012). As expected, the opposite mismatched combination ($V_{bet}A_{vet}$) elicited lower rates of illusory perception: 84%, 11%, 15%, 0% and 0% (values per individual lab member).

We presented stimuli at the patient's bedside with Presentation software (version 17.2; Neurobehavioral Systems) running on a laptop computer. Precise timing of stimulus and digital trigger

presentation with respect to the iEEG data acquisition system was verified using an oscilloscope, a microphone, and a photodiode. One trial (Figure **1b**) started with a baseline period where a fixation cross was shown at the center of screen for 500-750 ms. Then, a 4.5-second, 720-by-480-pixel, 30-frames-per-second video was displayed at the center of the screen. Finally, a response screen was shown to the participant: "What word did you hear?", with the following options for response: 'bet', 'vet' and 'wet'. The participant answered by pressing the corresponding key on the laptop's keyboard (3-alternative forced-choice task). Reaction time was not monitored. Between 46 and 71 repetitions of each stimulus were presented in randomized order. Participants could take a break every 50 trials.

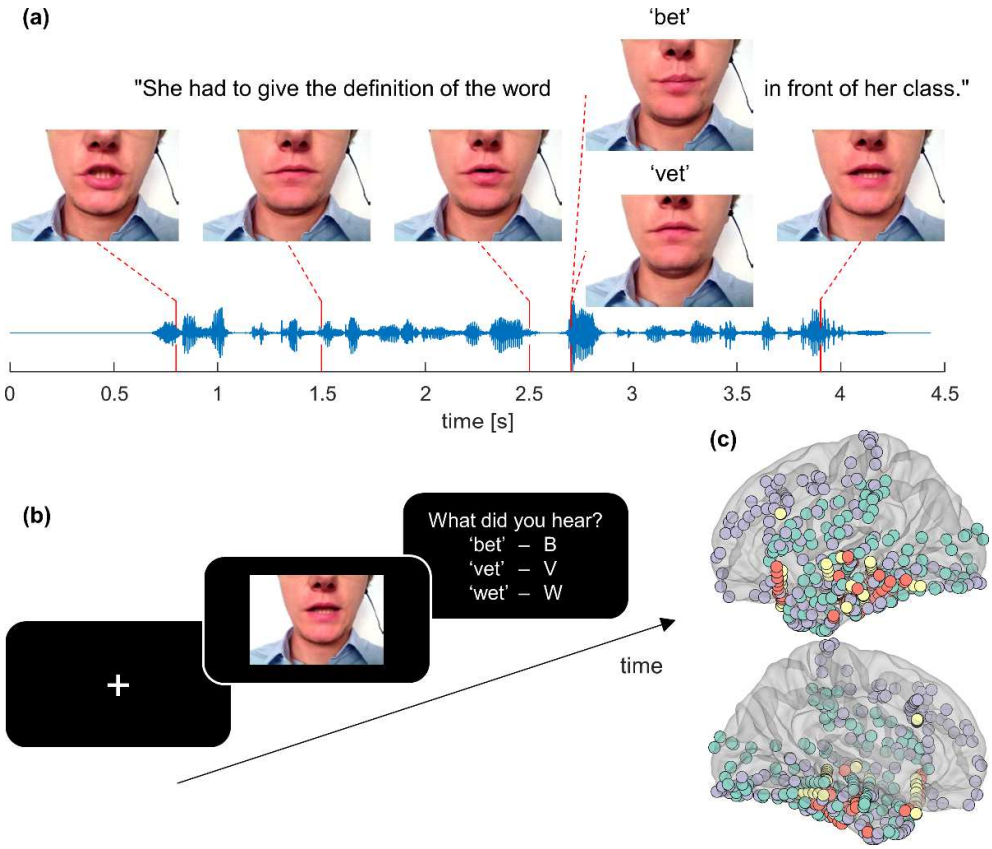


Figure 1. Experiment 1: stimuli, task and location of intracranial EEG electrodes. (a) For illustration purposes, the soundtrack of the audiovisual speech stimuli is plotted together with selected movie frames to illustrate the stimuli's salient visual cues. (b) At each trial, patients had to indicate which key word they thought they had heard (3-alternative forced-choice task). (c) iEEG electrodes (N=458) are color-coded per patient (4 patients) on lateral (top) and medial (bottom) views of the FreeSurfer average brain's left hemisphere (for the sake of simplification, right-hemisphere electrodes were flipped to the left hemisphere).

Experiment 2: the task and stimuli (Figure **5a**) are described in detail elsewhere (Thézé, Gadiri, *et al.*, 2020; Thézé, Giraud, *et al.*, 2020). Briefly, a computer-generated virtual character pronounced a sentence in French. The character's facial animations were synchronized to the soundtrack of the synthesized speech for each sentence. Toward the end of each sentence, a key word started with either "v" or "b"; either word was syntactically correct and semantically meaningful within the sentence. The viseme-phoneme pair of that word could either match or be mismatched, eliciting illusory perceptions (i.e. perceptions according to the visual, not the auditory, stimulus component) in a proportion of trials. In a pilot experiment, background auditory noise (ambient sound recorded in a café) was added to the synthesized speech soundtrack to bring perception of visual "v", auditory "b" combinations (hereafter V_vA_b) close to 50%. The level of background noise was not adjusted for each participant, meaning that

an individual participant's illusory perception rate could be far from 50%. Ten sentences were created, each in 4 combinations (2 with matched and 2 with mismatched visual and auditory components). Six virtual characters were implemented, each pronouncing each sentence combination once, so that one experiment block consisted of 240 unique trials. Two participants completed one block each, two others were able to complete 2 blocks each.

2.3 iEEG electrode localization

We localized and displayed iEEG electrodes using Voxeloc (https://github.com/HumanNeuronLab/voxeloc; Monney *et al.*, 2024) and iELVis (https://github.com/iELVis/iELVis; Groppe *et al.*, 2017). Briefly, a post-implantation high-resolution CT scan was aligned to a pre-implantation millimetric, isometric T1-weighted MRI scan using FSL, (Jenkinson *et al.*, 2012) FreeSurfer (Fischl, 2012) or NiftyReg (Modat *et al.*, 2014). Electrodes were manually identified using Voxeloc (for stereo-EEG electrodes) or BioImage Suite 3 (for subdural electrodes; Joshi *et al.*, 2011). For subdural electrodes, brain shift was compensated by projecting electrode locations back to the pre-implantation leptomeningeal surface (Dykstra *et al.*, 2012). For stereo-EEG electrodes, careful inspection of the aligned CT and MRI scans identified electrode contacts that lay outside the brain. Each iEEG electrode was attributed an anatomical location based on the individual participant's Desikan-Killiany parcellation of gyral anatomy (Desikan *et al.*, 2006). Cortical sites were deemed to be located in "early" auditory cortex if they were in the transverse temporal gyri or superior temporal gyrus; in "early" visual cortex if they were in the pericalcarine cortex, cuneus, lateral occipital gyri, or lingual gyrus; and in "late", higher-order cortical areas if they were in the supramarginal or angular gyri, inferior frontal gyrus, lateral orbitofrontal cortex, banks of the superior temporal sulcus, middle or inferior temporal gyri, fusiform gyrus, or temporal pole. To represent data from multiple participants, we brought electrode coordinates onto the FreeSurfer average brain. Because 871 of 1057 iEEG electrodes were implanted in the left hemisphere, and for the sake of simplification, we flipped right-hemisphere electrodes to the left hemisphere when displaying results aggregated over patients. Supporting figures S3-S34 display detailed results in individual patients.

2.4 iEEG recording and preprocessing

Signals were referenced to a subdermal vertex electrode, filtered, amplified and digitized using clinical systems (experiment 1: XLTEK EMU128FS or Natus Neurolink IP 256 systems, Natus Medical, digitized at 500 or 512 samples/s; experiment 2: Brain Quick LTM system, Micromed, digitized at 1024 or 2048 samples/s). We performed analyses offline using FieldTrip (Oostenveld *et al.*, 2011) and custom-made MATLAB routines. We filtered out 50- or 60-Hz line noise and its harmonics using a discrete Fourier transform filter. We inspected visually and rejected iEEG electrodes heavily contaminated with noise or abundant epileptiform activity, and rereferenced the remaining iEEG signals to average reference.

To compute the broadband high-frequency activity (HFA), which indexes local cortical activity (Crone *et al.*, 1998; Ray *et al.*, 2008; Leszczyński *et al.*, 2020), we filtered the iEEG signal of each trial between 75 and 175 Hz in 10-Hz bands (4^{th}-order Butterworth filters), computed instantaneous power using the Hilbert transform, divided power in each window by its mean over the trial, and averaged power fluctuations across bands back into a single time series. We then downsampled the HFA signal to 200 Hz. We applied baseline correction at the single-trial level by subtracting the average value of HFA before stimulus onset (experiment 1: from -300 to -100 ms; experiment 2: from -250 to 0 ms).

2.5 Data analysis

Because of the time-varying and multisensory nature of the stimuli, we used linear regression with contrast coding to disentangle cortical responses to their auditory and visual components (Cohen *et al.*, 2003; Smith & Kutas, 2015). Statistical testing was performed directly on the β coefficients of the linear regression. For both experiments, we built an analytical strategy that would reduce the complex sets of conditions ("v" vs. "b" auditory and visual stimulus components, matched vs. mismatched

components, and participants' perception reported as "v" vs. "b") into simpler 2-by-2 designs, allowing a straightforward interpretation of the interaction term. For that purpose, we carefully selected a subset out of the total number of trials that each participant was exposed to (see below).

Experiment 1: because the rate of illusory perception for mismatched stimuli tended to approach 100% for one mismatched combination (e.g. $V_{vet}A_{bet}$) and 0% for the other one (e.g. $V_{bet}A_{vet}$) in each given participant, we were not able to disentangle the effects of mismatched stimuli from those of reported perception. We retained for further analysis: the matched trials where perception matched the physical stimuli; the mismatched trials of the combination eliciting the highest rate of illusory perception that did elicit the illusory perception; and the mismatched trials of the combination eliciting the lowest rate of illusory perception that did not elicit the illusory perception. We were thus able to analyze data following a 2-by-2 design, with the identity of the visual and auditory stimulus components as factors. The numbers of trials that were retained for analysis (Supporting table S2) averages 213 and ranges from 181 to 228. Per-condition trial numbers average 53 and range from 21 (in a single participant and condition) to 69. These numbers are large enough to enable our regression analysis (Mercier *et al.*, 2022).

At each site and each time point of each trial, we modelled the HFA signal as a function of the physical characteristics of the auditory (A) and visual (V) components of the stimuli:

$$HFA = \beta_0 + \beta_1 A * \beta_2 V$$

For three participants, who had a high rate of illusory perception in response to the $V_{vet}A_{bet}$ mismatched stimulus, factor A (auditory) was set to +0.5 for A_{bet} and -0.5 for A_{vet}, and factor V (visual) was set to +0.5 for V_{vet} and -0.5 for V_{bet}. For the fourth participant, who had a high rate of illusory perception to $V_{bet}A_{vet}$, A was set to +0.5 for A_{vet} and -0.5 for A_{bet}, and V to +0.5 for V_{bet} and -0.5 for V_{vet}. Thus:

- β_0 represented the average response to all stimuli;
- β_1, the response difference due to differing auditory stimulus components;
- β_2, the response difference due to differing visual stimulus components;
- β_3, the interaction term between β_1 and β_2, the response difference due to mismatches between the auditory and visual components of the stimuli.

We analyzed β coefficients between 2.5 and 3.5 s after stimulus onset to focus on cortical processing of the key word, whose auditory component started at 2.66 for 'vet' or 2.68 s for 'bet'. We statistically assessed each site's activity by taking the maximum of the absolute value of the t statistic of each β coefficient during that window, and the latency of its occurrence. In order to threshold observed values of that maximum absolute t statistic, we simulated its null distribution 10'000 times and selected the 95^{th} percentile of that null distribution as a threshold (t=±3.7157, corresponding to a two-tailed p-value of $1.28*10^{-4}$).

Early vs. late cortical sites: to determine whether cortical sites sensitive to stimulus physical characteristics were distinct from those sensitive to mismatched stimuli, we performed a Pearson's correlation analysis of their β_1 or β_2 (whichever was largest) vs. β_3 coefficients. To investigate whether significant cortical sites were more frequently found in early sensory cortex vs. later, higher-order areas, we simulated the distribution of cortical sites' anatomical locations under the null hypothesis. For that purpose, we drew 10000 times at random the same number of cortical sites that were observed to have significant β_1, β_2 or β_3 coefficients out of the entire pool of cortical sites in experiment 1. We then computed the probability that the observed numbers of cortical sites influenced by stimulus physical characteristics or by mismatched stimuli in early or later cortical areas were more extreme than under H_0. Finally, we compared the repartition of cortical sites responsive to stimulus physical characteristics vs. mismatched stimuli in early vs. non-early sensory cortex using Fisher's exact test.

Experiment 2: to examine the interactions between the physical components of mismatched stimuli and the participants' perception, we designed a 2-by-2 analysis with stimulus (V_vA_b vs. V_vA_b) and perception ("v" vs. "b") as factors. When selecting a subset of trials from the entire experiment, we balanced trial numbers for each combination of factors in order to circumvent the issues that can affect the interpretation of the interaction term in linear regression analyses with unbalanced designs (Landsheer & Wittenboer, 2015; Smith & Kutas, 2015). For instance, participant 4 perceived V_vA_b trials as "b"-leading words (no illusion) in 44 out of 60 trials and as "v"-leading words (illusion) in 16, and V_bA_v trials as "v"-leading words (no illusion) in 10 out of 60 trials and as "b"-leading words (illusion) in 50. In this case, we selected 10 V_vA_b trials perceived "b", 10 V_vA_b trials perceived "v", the 10 V_bA_v trials perceived "v" and 10 V_bA_v trials perceived "b". Individual trials were selected so as to be as close as possible in the trial sequence to those of the least-common combination. 20, 22, 24 and 10 quadruplets of trials were selected for each participant, respectively.

At each site and each time point of each trial, we modelled the HFA signal as a function of the stimuli's physical characteristics and of patients' trial-by-trial perception:

$$HFA = \beta_0 + \beta_1 stim * \beta_2 per$$

Factor stim (stimulus) was set to +0.5 for V_bA_v trials and -0.5 for V_vA_b trials, and factor per (perception) was set to +0.5 for trials where a "v"-leading word was perceived and -0.5 for trials where a "b"-leading word was perceived. Thus:

- β_0 represented the average response to all stimuli;
- β_1, the response difference due to differing physical properties of the stimuli (V_bA_v vs. V_vA_b);
- β_2, the response difference due to differing perception ("v" vs. "b");
- β_3, the interaction term between β_1 and β_2, the interaction between the effects of mismatched stimuli and perception. In other words, β_3 represented differences in cortical responses as a function of the combination of a particular stimulus type (V_vA_b vs. V_bA_v) and its perception ("v" vs. "b"). Thus, β_3 directly indexed the effect of the audiovisual speech illusion.

We analyzed β coefficients between -0.25 and +0.75 s relative to the onset of the key word. We statistically assessed each site's activity by taking the maximum of the absolute value of the t statistic of each β coefficient during that window, and the latency of its occurrence. In order to threshold observed values of that maximum absolute t statistic, we arbitrarily applied an alpha level of $5*10^{-4}$. We selected that arbitrary value because no observation survived the simulation-based, more stringent correction for multiple comparisons over time points as applied in experiment 1.

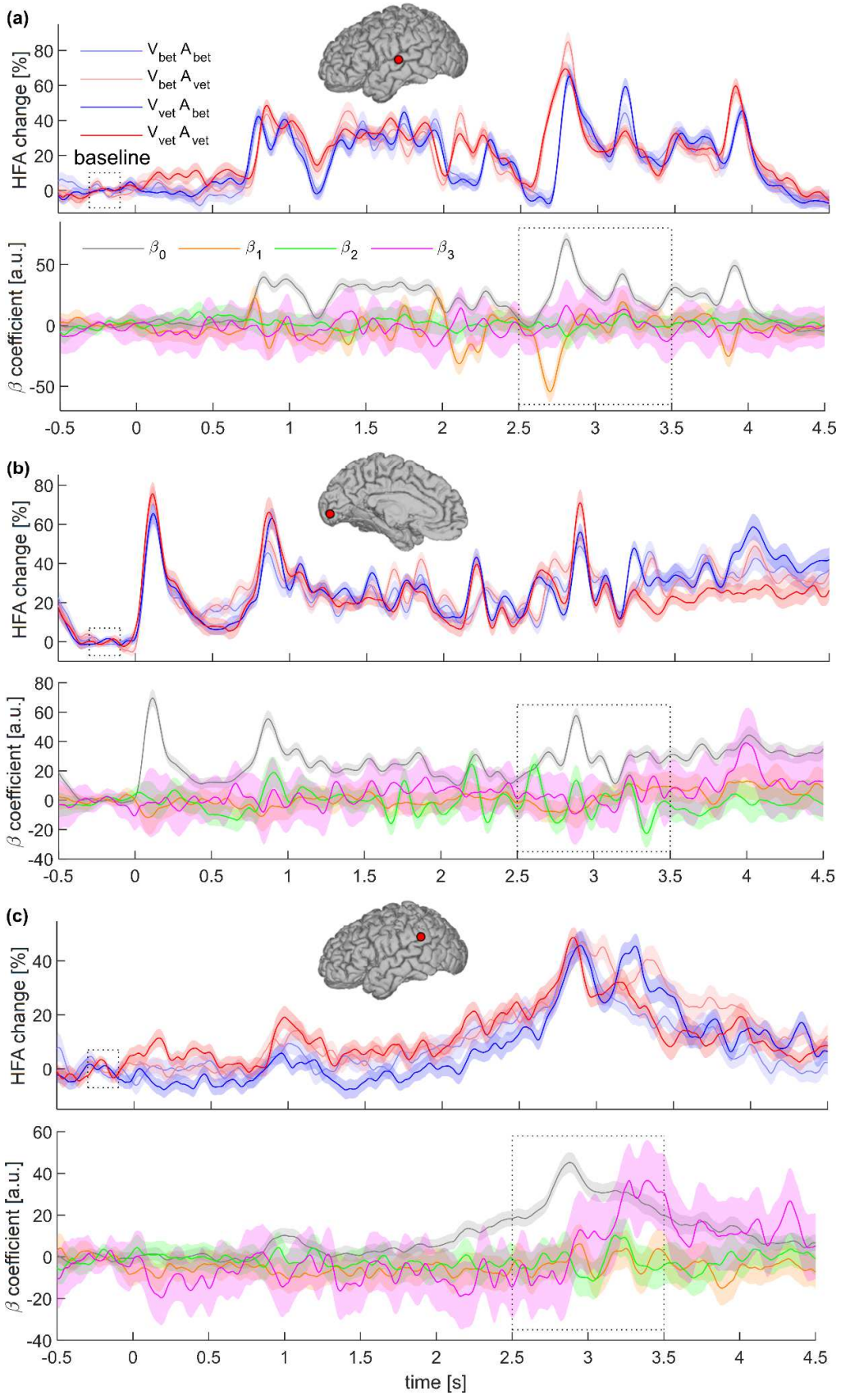

(a)
HFA change [%]
V_{bet} A_{bet}
V_{bet} A_{vet}
V_{vet} A_{bet}
V_{vet} A_{vet}
baseline
β coefficient [a.u.]
β_0 β_1 β_2 β_3
(b)
HFA change [%]
β coefficient [a.u.]
(c)
HFA change [%]
β coefficient [a.u.]
time [s]

Figure 2. Experiment 1: representative cortical responses. For all panels, HFA responses to the 4 stimuli (solid lines: average; shaded areas: standard error of the mean) are shown over the β coefficients of the linear regression (solid lines: average; shaded areas: standard error of the mean). β_0 encodes the unweighted mean HFA response to all stimuli; β_1, response differences due to differences in the auditory component of the stimuli; β_2, response differences due to differences in the visual component of the stimuli; and β_3, the interaction between the effects of the auditory and visual components. Insets show the anatomical location of each cortical site (all from the same patient's left hemisphere) (a) Auditory-driven response in the superior temporal gyrus (see inset). Notice, especially between 2.5 and 3 seconds, how responses diverge as a function of the auditory content of the stimuli (red hues vs. blue hues), which translates into non-zero values of β_1 (orange trace). (b) Visual-driven response in pericalcarine cortex. Responses diverge as a function of the stimuli's visual content (pale hues vs. dark hues), which is represented by non-zero values for β_2 (green trace). (c) Interaction response in the supramarginal gyrus. Notice how β_3 (pink trace) departs from 0 between 3 and 3.5 seconds.

3 Results

In a first attempt to induce audiovisual speech illusions using naturalistic stimuli, we produced videos of a human speaker who pronounced the following sentence: "She had to give the definition of the word 'bet' (or 'vet') in front of her class" (Experiment 1: Figure **1a,b**). The key word played no syntactical role in the sentence, but carried semantic information and respected lexical constraints. Mismatched stimuli elicited robust illusory perception in all four participants (Supporting table **S2**). Our analysis strategy aimed to separate the influence of the stimuli's visual and auditory components on cortical responses (see Figure **1c** for the cortical sites' locations). We focused our analysis on the broadband high-frequency activity (HFA), which indexes local neuronal activity (Crone *et al.*, 1998; Ray *et al.*, 2008; Leszczyński *et al.*, 2020). Using regression analysis, we modeled the influence of the auditory and visual stimuli on cortical responses, as well as their interaction.

Figure **2** shows exemplary HFA responses to the four possible combinations of auditory and visual stimuli, and the time courses of the regression coefficients. Responses in the superior temporal gyrus (auditory cortex) were strongly driven by the auditory component of the stimulus, irrespective of the visual component (Figure **2a**, notice the difference between the red vs. blue hues). This influence of the auditory stimulus is captured by the β_1 regressor, especially around the occurrence of the key word 2.7 s into the stimulus. Conversely, responses in pericalcarine cortex (visual cortex) were strongly influenced by the visual component of the stimulus (Figure **2b**, pale vs. dark hues), as indexed by the β_2 regressor. Finally, responses in the supramarginal gyrus were sensitive to the interaction between the auditory and visual components of the stimulus, as indexed by the β_3 regressor (Figure **2c**).

We quantified the impact of the stimuli's auditory and visual components on cortical responses by taking the absolute maximum value of each regressor during a 1-second window that spanned the occurrence of the key word (Figure **3** shows all significant cortical sites across participants on a template brain; see Supporting figures **S1**, where all cortical sites are plotted, and **S3-S18** for detailed results in individual participants). Overall, multiple cortical sites spanning all lobes of both cerebral hemispheres responded to the audiovisual speech stimuli (Figure **3a**). The influence of the stimuli's auditory component was strongest in the superior temporal cortex, supramarginal gyrus, and inferior and middle frontal cortex (Figure **3b**), while the impact of the visual component was mostly visible on occipital sites as well as in inferior frontal cortex (Figure **3c**). Response modulations as a function of the interaction of the auditory and visual stimuli (matched vs. mismatched) were most evident in the superior temporal and supramarginal cortex, occipital cortex, as well as orbitofrontal cortex (Figure **3d**).

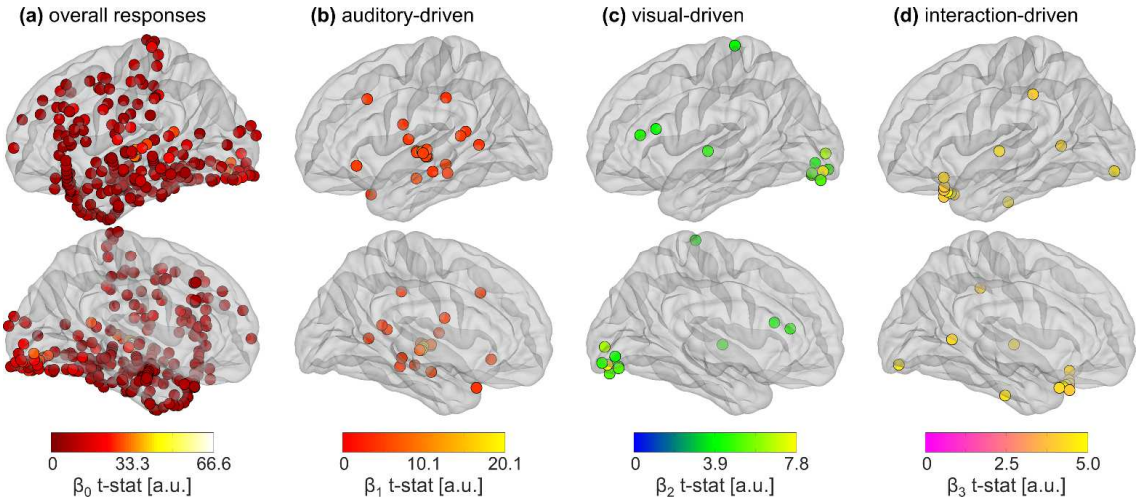


Figure 3. Experiment 1: significant HFA responses. Lateral (top) and medial (bottom) views of the FreeSurfer average brain's left hemisphere are shown (for the sake of simplification, right-hemisphere sites were flipped to the left hemisphere). Sites with significant task-related HFA response modulations are plotted in solid colors and circled in black. (a) Unweighted mean HFA responses to the task, as indexed by correlation coefficient β_0. 254 out of 458 cortical sites were significantly activated by the task. (b) HFA modulation due to the auditory component of the stimuli, as encoded by β_1 (24 sites). (c) HFA modulation due to the stimuli's visual component (β_2; 11 sites). (d) HFA modulation as a function of the interaction between the stimuli's auditory and visual components (β_3; 12 sites).

In order to determine whether there were systematic differences in the location of cortical sites responsive to the stimuli's physical characteristics (as indexed by β_1 and β_2) vs. their mismatch (as indexed by β_3), we first established that there was no correlation between the respective β coefficients (Figure **4a**; Pearson's r=-0.265, p=0.086), confirming that most sites were either influenced by the stimuli or by their mismatch, but not by both. Then, we asked whether there were more cortical sites responsive to stimulus physical characteristics in early sensory cortices vs. late, higher-order areas (see Materials & methods). We found that such sites were more numerous than expected in early cortex, and less than expected in later areas (Figure **4b**). By contrast, the repartition of cortical sites influenced by mismatched stimuli in early vs. late areas did not differ from chance. Finally, we found that the distribution of sites influenced by stimuli vs. their mismatch in early vs. non-early cortex differed significantly (20 "stimulus" sites in early cortex vs. 14 in non-early cortex, and 2 "mismatch" sites in early cortex vs. 10 in non-early cortex; p=0.0182, Fisher's exact test). Overall, these analyses show that responses to the stimuli's physical characteristics were more concentrated in early sensory cortices, which was not the case for responses to mismatched stimuli.

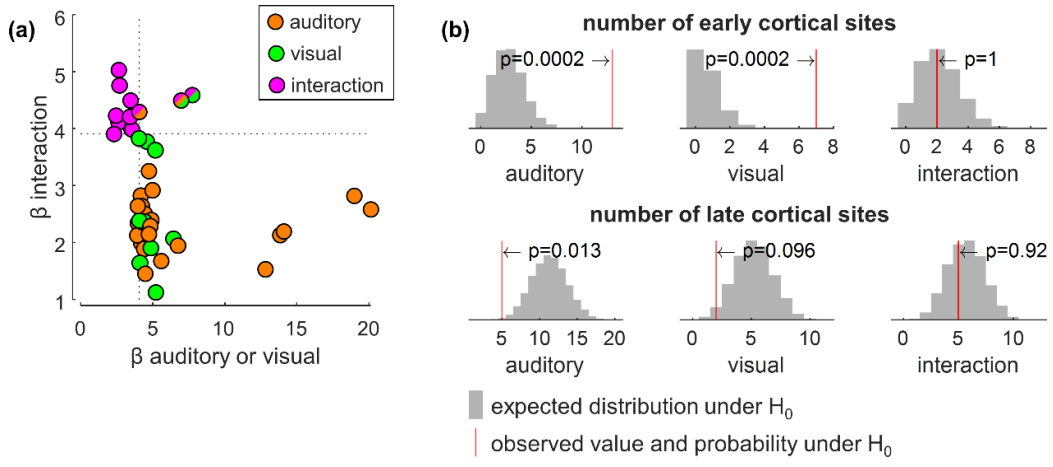


Figure 4. Experiment 1: early vs. late cortical sites. (a) Scatter plot of the β coefficients for all 43 significant sites. The x axis shows the β coefficient for either the auditory or the visual stimulus components (whichever was largest), while the y axis shows the β coefficient for the interaction (mismatched stimuli). Sites that displayed significant influence of both the stimulus' physical components and their interaction are shown with 2 or 3 colors. The dashed lines indicate the lowest β values that reached statistical significance on either axis. There is no positive correlation between the influence of stimulus physical characteristics and that of the interaction. (b) Expected (gray bars) vs. observed (red lines) numbers of early (top) and late (bottom) cortical sites influenced by the stimulus' physical components and their interaction. The number of responses to stimulus physical characteristics is larger than expected by chance in early sensory cortices, and smaller than expected by chance in later cortical areas.

In experiment 1, since one of the mismatched audiovisual stimuli ($V_{vet}A_{bet}$ in most cases) induced an illusory perception on a large majority of trials, while the reverse mismatch ($V_{bet}A_{vet}$) did not induce an illusory perception on most trials, we cannot disentangle to which degree cortical responses are reflecting the occurrence of mismatched stimuli vs. the participants' perceptual report. In order to dissociate these two factors, it was necessary to design another stimulus set, which would not induce illusory perception systematically, but only in a fraction of trials. For that purpose, we used 3D animated virtual characters and speech synthesis to create a novel stimulus set, where words that could begin indifferently with a "b" or a "v" were seamlessly embedded into full sentences in French (Experiment 2: Figure **5a**; Thézé, Gadiri, *et al.*, 2020). Although the rate of illusory perception to mismatched stimuli varied from participant to participant, each one experienced illusory perception to both types of mismatched stimuli in a fraction of trials. We were thus able to orthogonalize the physical characteristics of the stimuli (V_vA_b vs. V_bA_v) from the participants' perception ("v" vs. "b"-leading word) in our analysis of cortical responses (see Figure **5b** for the location of recording sites).

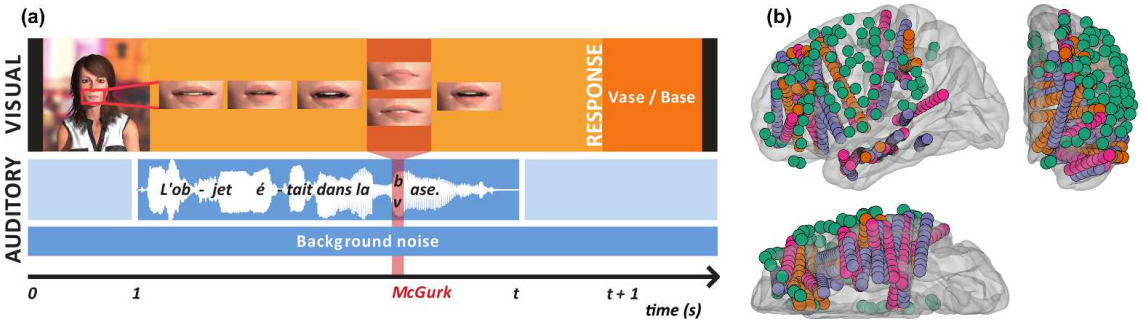


Figure 5. Experiment 2: stimuli, task and location of intracranial EEG electrodes. (a) Timeline of an example trial (from Thézé, Giraud, *et al.*, 2020). An animated computer-generated avatar pronounces a sentence among background noise. At a key viseme-phoneme pair, the visual and auditory stimuli can match or be mismatched, eliciting illusory perceptions (i.e. perceptions according to the visual, not the auditory, stimulus component) in a proportion of trials. The French-language example sentence translates to "The object was in the base / mud". (b) iEEG electrodes (N=576) are color-coded per patient (4 patients) on lateral (top left), ventral (bottom) and frontal (right) views of the FreeSurfer average brain's left hemisphere (for the sake of simplification, right-hemisphere electrodes were flipped to the left hemisphere).

As in experiment 1, multiple cortical sites in the temporal, parietal, frontal and insular lobes of both cerebral hemispheres responded to these complex audiovisual speech stimuli (Figure **6a**). The effect of the stimuli's physical characteristics was observed mostly in the inferior frontal and orbitofrontal cortices as well as the middle temporal gyrus (Figure **6b**), while the influence of the participants' reported perception was visible in inferior and middle frontal cortex and the anterior temporal lobe (Figure **6c**). The interaction between these two factors, which directly indexed the audiovisual speech

illusion (see Materials & methods), was strongest in the supramarginal gyrus, the middle temporal gyrus, rostral inferior and middle frontal gyri, as well as the insula and the cingulate gyrus (Figure 6d). As previously reported (Nath & Beauchamp, 2012), we found significant variability across participants (see Supporting figures S2, where all cortical sites are plotted, and S19-S34 for detailed results in individual participants).

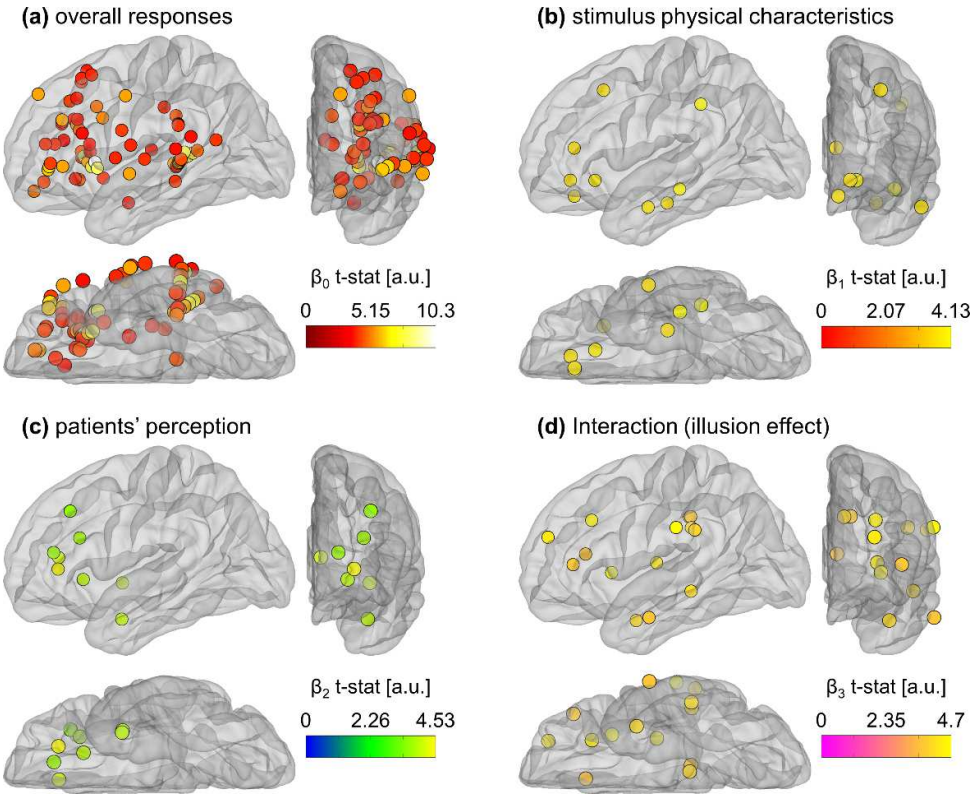


Figure 6. Experiment 2: significant HFA responses. Lateral (top left), ventral (bottom) and frontal (right) views of the FreeSurfer average brain's left hemisphere are shown (for the sake of simplification, right-hemisphere sites were flipped to the left hemisphere). Sites with significant task-related HFA response modulations are plotted in solid colors and circled in black. (a) Unweighted mean HFA responses to the task, as indexed by correlation coefficient β_0. 69 out of 576 sites were significantly activated by the task. (b) HFA modulation due to the physical characteristics of the stimuli (V_bA_v vs. V_vA_b), as encoded by β_1 (9 sites). (c) HFA modulation due to the patients' reported perception ("V" vs. "B"; β_2; 8 sites). (d) HFA modulation as a function of the interaction between the stimuli's physical characteristics and the patients' perception (β_3; 14 sites).

The fine temporal resolution of iEEG allowed us to examine when the effect of this interaction between mismatched stimuli and the participants' perception was maximally influencing cortical activity (Figure 7a). We found a tendency for earlier effects in the supramarginal gyrus, insula, and posterior cingulate cortex, whereas temporal and frontal cortical sites were influenced later. However, statistical analysis failed to disclose significant latency differences across lobes (Figure 7b; 1-way ANOVA: $F(10,3)=1.54$, $p=0.265$), probably because of the low number of observations and the coarse anatomical grouping. Taken together, our results indicate that the perception of audiovisual speech illusions recruits a widespread ensemble of cortical sites that includes relatively early sensory cortices, higher-order sensory and multisensory cortical areas, as well as prefrontal cortex on both cerebral hemispheres.

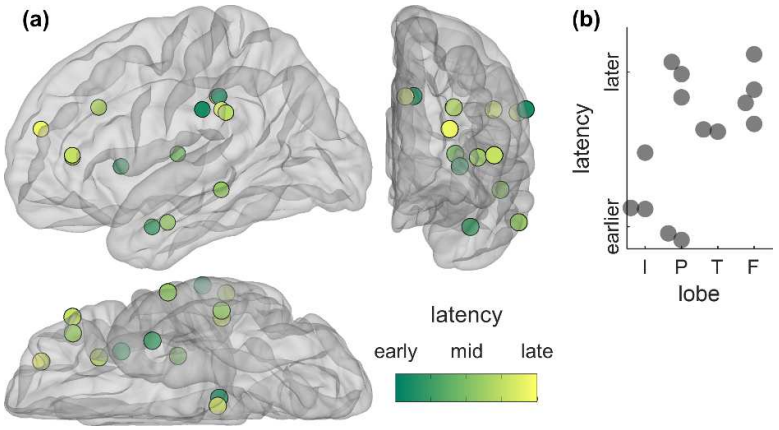


Figure 7. Experiment 2: latency of "illusion effect". (a) Latency of maximal influence on cortical responses of the interaction between the stimuli's physical characteristics and the participants' reported perception. Only sites that were considered significant in Figure 6d are shown here. (b) Latencies grouped by lobe (I: insula; P: parietal; T: temporal; F: frontal). Although latencies seemed to be shorter in the insula and part of the parietal lobe, and longer in frontal cortex, statistical analysis failed to disclose significant latency differences across lobes.

4 Discussion

Ever since their discovery, audiovisual speech illusions have fascinated neuroscientists and the general public alike (MacDonald, 2018). Leveraging these illusions to probe the neural substrates of multisensory speech perception, however, is not straightforward, because of the complexity of audiovisual speech as a sensory input. In fact, previous studies reported somewhat conflicting results regarding the key locus for their perception. Here, we combined intracranial EEG, which optimally resolves neural signals in humans, with innovative stimulus design to reveal that the perception of audiovisual speech illusions involves an ensemble of higher-order sensory and associative cortical areas in the temporal, parietal and frontal lobes. Importantly, we are able to separate the neural processing of the auditory and visual components of the multisensory speech stream, as well the neural activity that relates to reported perception. Furthermore, our naturalistic stimuli alleviate most of the concerns formulated against the use of audiovisual speech illusions to investigate the mechanisms underlying multisensory speech perception.

Audiovisual speech illusions are typically induced by presenting mismatched streams of simple utterances in the form of meaningless syllables. While such stimuli robustly and seamlessly induce illusory perceptions, they are not an ideal approach to examine multisensory speech processing, because they violate the phonological, lexical, syntactical and semantic constraints of natural communication through speech (Van Engen *et al.*, 2022). Our naturalistic stimuli, which take advantage of the fact that a number of words (in English or in French) that begin either with a "b" or a "v" can be used interchangeably in a sentence without upsetting its grammatical soundness, fully respect these constraints (Thézé, Gadiri, *et al.*, 2020). The one remaining difference between audiovisual speech illusions and natural audiovisual speech, the incongruence between the auditory and visual speech cues, is a necessary feature of our experimental design and analysis plan. An additional issue of traditional stimuli is that the mismatched auditory and visual cues are presented at, or very soon after, the onset of the utterance. Consequently, preparatory articulatory movements are particularly conspicuous, which might exaggerate the effects of visual cues on speech perception (Schwartz & Savariaux, 2014). One way to alleviate this concern is to use stimuli of the "aba/aga" type, where the initial vowel (which is identical for all stimuli) ensures that preparatory articulatory movements are identical across stimuli (Keil *et al.*, 2012). These utterances, however, remain meaningless. Our

approach to stimulus design takes care of this concern by placing the mismatched audiovisual speech cues towards the end of the sentence, with the added advantage that all our stimuli are syntactically correct and meaningful sentences.

The aforementioned issues with traditional stimuli might explain some of the discrepancies in the results of previous studies that probed the multisensory processing of speech with audiovisual illusions. Another limitation of prior work pertains to the degree to which the influences of the stimuli's physical characteristics could be dissociated from those of the elicited percept. For instance, Smith et al. (2013) used iEEG to examine cortical responses to mismatched $A_{Ba}V_{Va}$ stimuli, which induced an illusory perception as "va", and contrasted them to cortical responses to matched $A_{Ba}V_{Ba}$ or $A_{Va}V_{Va}$ stimuli. They found that auditory cortical responses to $A_{Ba}V_{Va}$ were more similar to those to $A_{Va}V_{Va}$ than to $A_{Ba}V_{Ba}$. However, cortical responses to the reverse mismatched stimulus ($A_{Ba}V_{Va}$, which was not expected to elicit an illusory perception) were not examined. Thus, whether auditory cortex was driven more by the physical characteristics of the stimulus or the patients' perception remains unresolved.

Even when all combinations of matched and mismatched auditory and visual stimulus components are presented, an additional confound stems from the very robustness with which some mismatched combinations induce illusory perceptions. Indeed, if a participant's perception systematically corresponds to the visual component of a mismatched stimulus and to the auditory component of the reverse combination, then it becomes impossible to fully dissociate stimulus- from perception-related cortical activity. Our own findings are, in part, affected by this confound (see the results of our experiment 1). In order to fully disentangle stimulus- from response-related cortical activity, we designed a stimulus set that induced illusory perception in a fraction of trials, in a quasi-bistable fashion (Thézé, Gadiri, *et al.*, 2020; Thézé, Giraud, *et al.*, 2020).

Our results indicate that, while very early sensory cortical areas are obviously crucial for the perception of audiovisual speech, they are not the major locus for the perception of audiovisual speech illusions, because their activity relates to the physical properties of the stimuli much more than to the participants' reported perception. In that respect, our findings are in agreement with those of Nath and Beauchamp (2012) rather than Smith et al. (2013). Cortical sites in inferior frontal and orbitofrontal cortex react specifically to mismatched audiovisual speech stimuli, but they seem to do so regardless of which sensory component drove perception (see also Nath & Beauchamp, 2012). We found cortical sites that tracked both the nature of the speech stimuli and their perception in the supramarginal gyrus, anterior temporal lobe, lateral prefrontal cortex, as well as the insula and cingulate cortex. Thus, we delineate a group of higher-order sensory and associative cortical areas that underlie the perception of audiovisual speech illusions.

Our study suffers from several important limitations. Because the cortical regions targeted by iEEG electrodes are determined based on clinical grounds, and because epilepsy does not affect all regions with equal frequency, our coverage of the cortical mantle is incomplete, particularly so for superior parietal cortex in both experiments and for the occipital lobe in experiment 2. The superior temporal sulcus, a key locus for the perception of audiovisual speech illusions (see Beauchamp, 2016 for a review), is not sampled in this study, as it is only rarely targeted by iEEG electrodes for the following reasons: it is inaccessible to subdural electrodes, which lie on the arachnoid membrane and cannot explore sulci; and it is unlikely to be explored much by stereo-EEG electrodes. Indeed, stereo-EEG electrode trajectory planning follows a number of rules designed to minimize risk to the patient (Vakharia *et al.*, 2018): the entry point is on a gyral crown, and trajectories do not cross sulcal pial boundaries, in order to avoid pial blood vessels; and the anchoring bolt (and hence the electrode trajectory) is orthogonal to the skull in order to ensure safe anchoring and an accurate trajectory. Because of these rules, and given that the superior temporal sulcus itself runs approximately orthogonally to the skull, it is only exceptionally grazed by electrodes en route to a deeper target

(although see Nourski *et al.*, 2021 for an example of superior temporal sulcus responses to auditory speech).

The coverage problem is compounded by our low participant numbers. This is due both to the rarity of iEEG as a method and to the complexity of our experiments, which we reserved to patients who were able to focus their attention for a long period of time. Importantly, because the majority of iEEG electrodes in our sample were implanted in the left hemisphere, we cannot explore in much detail inter-hemispheric differences in the processing of audiovisual speech. Furthermore, we used iEEG to establish correlations between sensory inputs and perceptual reports, but we are unable to determine which cortical sites are causally implicated in the perception of audiovisual speech illusions, as opposed to merely correlating with this perception. Perturbational approaches, for instance with focal electrical or magnetic stimulation of cortical tissue (Beauchamp *et al.*, 2010; Keller *et al.*, 2017; Murakami *et al.*, 2018), allow exploring questions of causality more directly.

Author contributions

PM: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, supervision, visualization, writing – original draft preparation. RT: conceptualization, formal analysis, investigation, methodology, visualization, writing – review & editing. ADM: funding acquisition, investigation, supervision, writing – review & editing.

Acknowledgements

PM was supported by the following career grants from the Swiss National Science Foundation: 148388, 167836 and 194507. We thank the patients who kindly accepted to participate in this study, Silvia Marchesotti for helpful discussions on the analysis strategy, the members of the Mehta, Schroeder, and Giraud labs for stimulating discussions on multisensory cortical processing, and the physicians and personnel of the Neurology and Neurosurgery departments of North Shore University Hospital and Geneva University Hospitals for their professional support.

Ethics statement

All experiments described here followed the guidelines of the Declaration of Helsinki, complied with applicable laws, and were duly authorized by the relevant institutional review board (experiment 1: Feinstein Institutes for Medical Research Institutional review board; experiment 2: Commission cantonale d'éthique de la recherche de la République et canton de Genève). The patients provided written informed consent.

Conflict of interest statement

The authors report no conflict of interest pertaining to the present study.

Data availability statement

The data and code that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions: raw data include personally identifiable information from human patients.

References

Beauchamp, M.S. (2016) Audiovisual Speech Integration: Neural Substrates and Behavior. In Hickok, G. & Small, S.L. (eds), *Neurobiology of Language*. Academic Press, San Diego, pp. 515–526.

Beauchamp, M.S., Nath, A.R., & Pasalar, S. (2010) fMRI-Guided Transcranial Magnetic Stimulation Reveals That the Superior Temporal Sulcus Is a Cortical Locus of the McGurk Effect. *J. Neurosci.*, **30**, 2414–2417.

Benoit, M.M., Raij, T., Lin, F.H., Jääskeläinen, I.P., & Stufflebeam, S. (2010) Primary and multisensory cortical activity is correlated with audiovisual percepts. *Hum. Brain Mapp.*, **31**, 526–538.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. edn. Lawrence Erlbaum Associates, New York.

Crone, N.E., Miglioretti, D.L., Gordon, B., & Lesser, R.P. (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain J. Neurol.*, **121**, 2301–2315.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., & Killiany, R.J. (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, **31**, 968–980.

Dykstra, A.R., Chan, A.M., Quinn, B.T., Zepeda, R., Keller, C.J., Cormier, J., Madsen, J.R., Eskandar, E.N., & Cash, S.S. (2012) Individualized localization and cortical surface-based registration of intracranial electrodes. *NeuroImage*, **59**, 3563–3570.

Erickson, L.C., Zielinski, B.A., Zielinski, J.E.V., Liu, G., Turkeltaub, P.E., Leaver, A.M., & Rauschecker, J.P. (2014) Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Front. Psychol.*, **5**.

Fischl, B. (2012) FreeSurfer. *NeuroImage*, **62**, 774–781.

Groppe, D.M., Bickel, S., Dykstra, A.R., Wang, X., Mégevand, P., Mercier, M.R., Lado, F.A., Mehta, A.D., & Honey, C.J. (2017) iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *J. Neurosci. Methods*, **281**, 40–48.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., & Smith, S.M. (2012) FSL. *NeuroImage*, **62**, 782–790.

Jiang, J. & Bernstein, L.E. (2011) Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol. Hum. Percept. Perform.*, **37**, 1193–1209.

Joshi, A., Scheinost, D., Okuda, H., Belhachemi, D., Murphy, I., Staib, L.H., & Papademetris, X. (2011) Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics*, **9**, 69–84.

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

Keller, C.J., Davidesco, I., Megevand, P., Lado, F.A., Malach, R., & Mehta, A.D. (2017) Tuning face perception with electrical stimulation of the fusiform gyrus. *Hum. Brain Mapp.*, **38**, 2830–2842.

Kumar, G.V., Kumar, N., Roy, D., & Banerjee, A. (2018) Segregation and Integration of Cortical Information Processing Underlying Cross-Modal Perception. *Multisensory Res.*, **31**, 481–500.

Landsheer, J.A. & Wittenboer, G. van den (2015) Unbalanced 2 x 2 Factorial Designs and the Interaction Effect: A Troublesome Combination. *PLOS ONE*, **10**, e0121412.

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A.Y., Tal, I., Haegens, S., Melloni, L., Knight, R.T., & Schroeder, C.E. (2020) Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. *Sci. Adv.*, **6**, eabb0977.

Li, L., Li, R., Huang, X., Shen, F., Wang, H., Wang, X., Deng, C., Wang, C., Yang, J., Zhang, L., Li, J., Zou, T., & Chen, H. (2021) Motor Circuit and Superior Temporal Sulcus Activities Linked to Individual Differences in Multisensory Speech Perception. *Brain Topogr.*, **34**, 779–792.

MacDonald, J. (2018) Hearing Lips and Seeing Voices: the Origins and Development of the 'McGurk Effect' and Reflections on Audio–Visual Speech Perception Over the Last 40 Years. *Multisensory Res.*, **31**, 7–18.

McGurk, H. & Macdonald, J. (1976) Hearing lips and seeing voices. *Nature*, **264**, 691–811.
Mercier, M.R., Dubarry, A.-S., Tadel, F., Avanzini, P., Axmacher, N., Cellier, D., Vecchio, M.D., Hamilton, L.S., Hermes, D., Kahana, M.J., Knight, R.T., Llorens, A., Megevand, P., Melloni, L., Miller, K.J., Piai, V., Puce, A., Ramsey, N.F., Schwiedrzik, C.M., Smith, S.E., Stolk, A., Swann, N.C., Vansteensel, M.J., Voytek, B., Wang, L., Lachaux, J.-P., & Oostenveld, R. (2022) Advances in human intracranial electroencephalography research, guidelines and good practices. *NeuroImage*, **260**, 119438.
Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., & Ourselin, S. (2014) Global image registration using a symmetric block-matching approach. *J. Med. Imaging*, **1**, 024003.
Monney, J., Dallaire, S.E., Stoutah, L., Fanda, L., & Mégevand, P. (2024) Voxeloc: A time-saving graphical user interface for localizing and visualizing stereo-EEG electrodes. *J. Neurosci. Methods*, **407**, 110154.
Murakami, T., Abe, M., Wiratman, W., Fujiwara, J., Okamoto, M., Mizuochi-Endo, T., Iwabuchi, T., Makuuchi, M., Yamashita, A., Tiksnadi, A., Chang, F.-Y., Kubo, H., Matsuda, N., Kobayashi, S., Eifuku, S., & Ugawa, Y. (2018) The Motor Network Reduces Multisensory Illusory Perception. *J. Neurosci.*, **38**, 9679–9688.
Nath, A.R. & Beauchamp, M.S. (2012) A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, Neuroergonomics: The human brain in action and at work, **59**, 781–787.
Nourski, K.V., Steinschneider, M., Rhone, A.E., Kovach, C.K., Banks, M.I., Krause, B.M., Kawasaki, H., & Howard, M.A., III (2021) Electrophysiology of the Human Superior Temporal Sulcus during Speech Processing. *Cereb. Cortex*, **31**, 1131–1148.
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011) FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Comput. Intell. Neurosci.*, **2011**, 1–9.
Parvizi, J. & Kastner, S. (2018) Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.*, **21**, 474–483.
Pratt, H., Bleich, N., & Mittelman, N. (2015) Spatio-temporal distribution of brain activity associated with audio-visually congruent and incongruent speech and the McGurk Effect. *Brain Behav.*, **5**, e00407.
Proverbio, A.M., Raso, G., & Zani, A. (2018) Electrophysiological Indexes of Incongruent Audiovisual Phonemic Processing: Unraveling the McGurk Effect. *Neuroscience*, **385**, 215–226.
Ray, S., Crone, N.E., Niebur, E., Franaszczuk, P.J., & Hsiao, S.S. (2008) Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci. Off. J. Soc. Neurosci.*, **28**, 11526–11536.
Schwartz, J.-L. & Savariaux, C. (2014) No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. *PLoS Comput. Biol.*, **10**, e1003743.
Smith, E., Duede, S., Hanrahan, S., Davis, T., House, P., & Greger, B. (2013) Seeing Is Believing: Neural Representations of Visual Stimuli in Human Auditory Cortex Correlate with Illusory Auditory Perceptions. *PLOS ONE*, **8**, e73148.
Smith, N.J. & Kutas, M. (2015) Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, **52**, 157–168.
Sumby, W.H. & Pollack, I. (1954) Visual Contribution to Speech Intelligibility in Noise. *J. Acoust. Soc. Am.*, **26**, 212–215.
Thézé, R., Gadiri, M.A., Albert, L., Provost, A., Giraud, A.L., & Mégevand, P. (2020) Animated virtual characters to explore audio-visual speech in controlled and naturalistic environments. *Sci. Rep.*, **10**, 1–12.
Thézé, R., Giraud, A.L., & Mégevand, P. (2020) The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. *Sci. Adv.*, **6**, eabc6348.
Tiippana, K. (2014) What is the McGurk effect? *Front. Psychol.*, **5**, 725.

Tse, C.-Y., Gratton, G., Garnsey, S.M., Novak, M.A., & Fabiani, M. (2015) Read My Lips: Brain Dynamics Associated with Audiovisual Integration and Deviance Detection. *J. Cogn. Neurosci.*, **27**, 1723–1737.

Uno, T., Kawai, K., Sakai, K., Wakebe, T., Ibaraki, T., Kunii, N., Matsuo, T., & Saito, N. (2015) Dissociated Roles of the Inferior Frontal Gyrus and Superior Temporal Sulcus in Audiovisual Processing: Top-Down and Bottom-Up Mismatch Detection. *PLOS ONE*, **10**, e0122580.

Vakharia, V.N., Duncan, J.S., Witt, J.-A., Elger, C.E., Staba, R., & Engel Jr, J. (2018) Getting the best outcomes from epilepsy surgery. *Ann. Neurol.*, **83**, 676–690.

Van Engen, K.J., Dey, A., Sommers, M.S., & Peelle, J.E. (2022) Audiovisual speech perception: Moving beyond McGurk. *J. Acoust. Soc. Am.*, **152**, 3216–3225.

Supporting information for

Naturalistic audiovisual illusions reveal the cortical sites involved in the multisensory processing of speech

Pierre Mégevand, Raphaël Thézé, Ashesh D. Mehta

Participant	$V_{bet}A_{bet}$	$V_{bet}A_{vet}$	$V_{vet}A_{bet}$	$V_{vet}A_{vet}$	$V_{wet}A_{bet}$	$A_{wet}A_{wet}$	Total
1	51	60	71	51	71	5	309
2	51	60	71	46	71	5	304
3	50	70	70	50	5	5	250
4	50	70	70	50	5	5	250

Supporting table S1. Experiment 1: numbers of trials of each type for each participant.

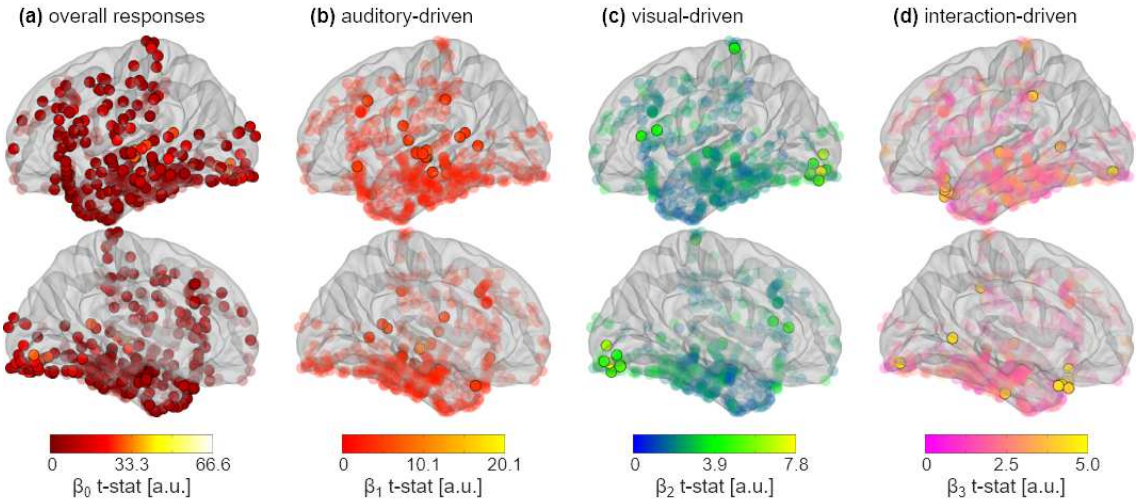
Participant 1		Auditory	
		Bet	Vet
Visual	Bet	51/51 "Bet"	60/60 "Vet"
	Vet	**66/71** **"Vet"**	51/51 "Vet"

Participant 2		Auditory	
		Bet	Vet
Visual	Bet	48/51 "Bet"	21/70 "Vet"
	Vet	**68/71** **"Vet"**	44/46 "Vet"

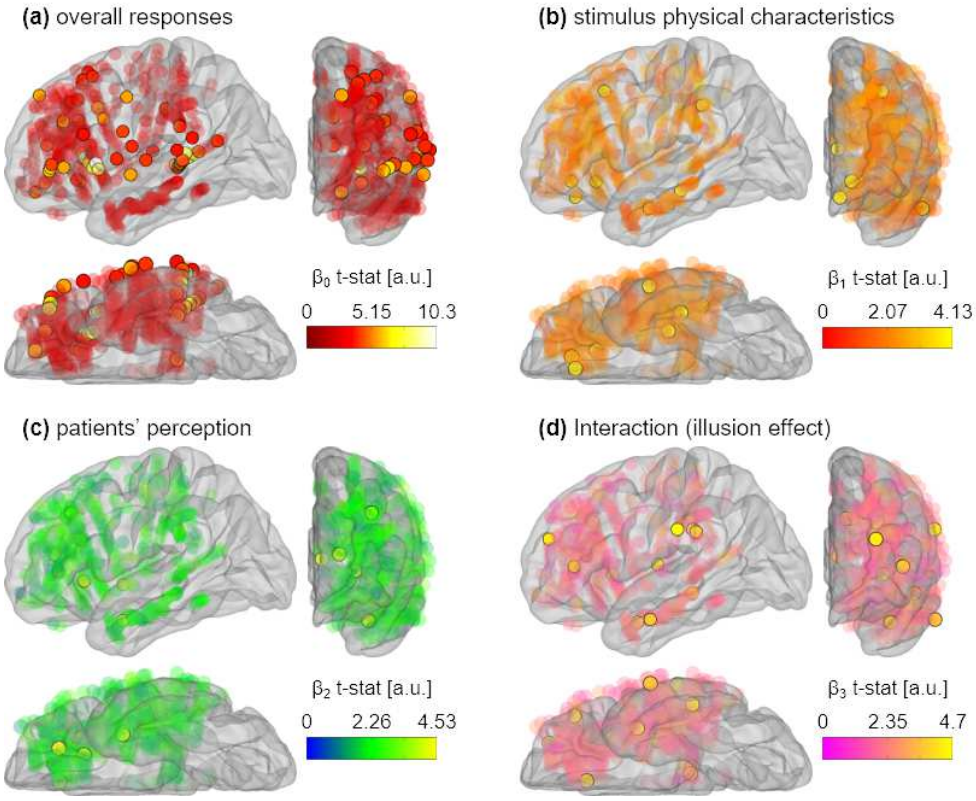
Participant 3		Auditory	
		Bet	Vet
Visual	Bet	50/50 "Bet"	**49/70** **"Bet"**
	Vet	69/70 "Bet"	49/50 "Vet"

Participant 4		Auditory	
		Bet	Vet
Visual	Bet	50/50 "Bet"	60/70 "Vet"
	Vet	**68/70** **"Vet"**	50/50 "Vet"

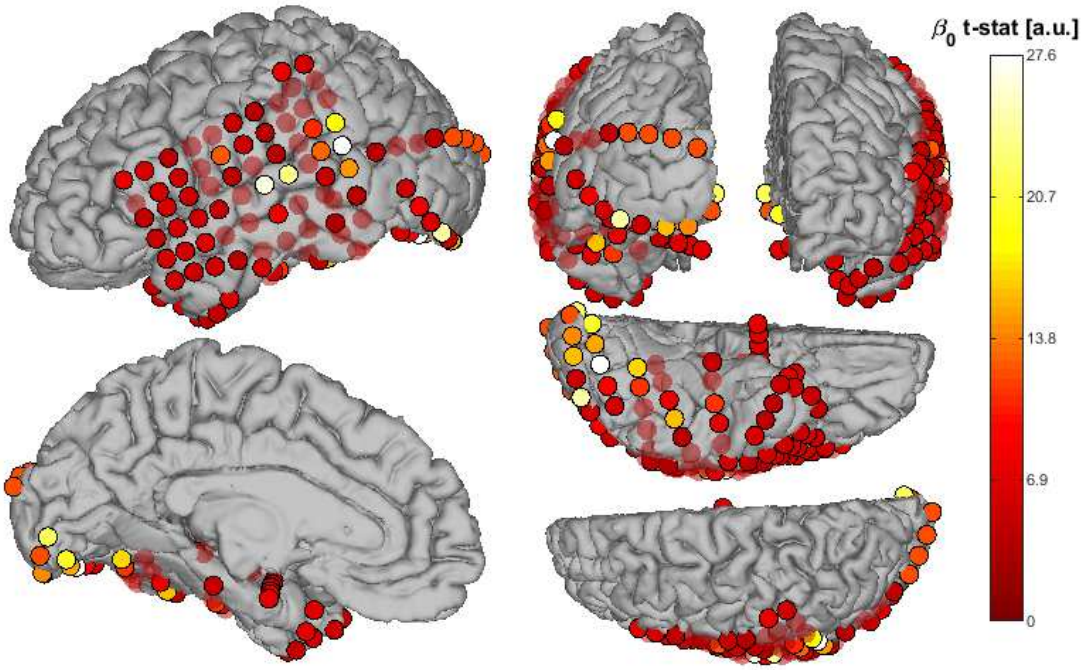
Supporting table S2. Experiment 1: behavioral results. For each participant, the mismatched audiovisual stimulus combination that yielded the highest rate of illusory perception is highlighted in bold. Only trials in the numerator were retained for further analysis (see Materials & methods).



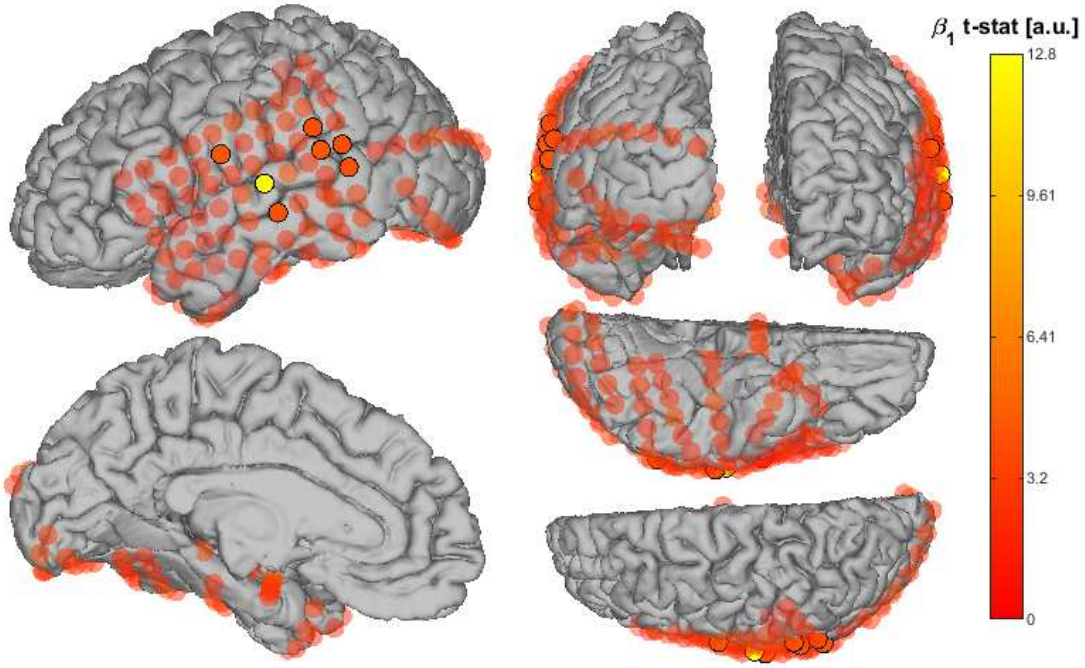
Supporting figure S1. Experiment 1: HFA responses. Lateral (top) and medial (bottom) views of the FreeSurfer average brain's left hemisphere are shown (for the sake of simplification, right-hemisphere sites were flipped to the left hemisphere). Sites with significant task-related HFA response modulations are plotted in solid colors and circled in black; sites that failed to reach significance are plotted in transparent colors. (a) Unweighted mean HFA responses to the task, as indexed by correlation coefficient β_0. 254 out of 458 cortical sites were significantly activated by the task. (b) HFA modulation due to the auditory component of the stimuli, as encoded by β_1 (24 sites). (c) HFA modulation due to the stimuli's visual component (β_2; 11 sites). (d) HFA modulation as a function of the interaction between the stimuli's auditory and visual components (β_3; 12 sites).



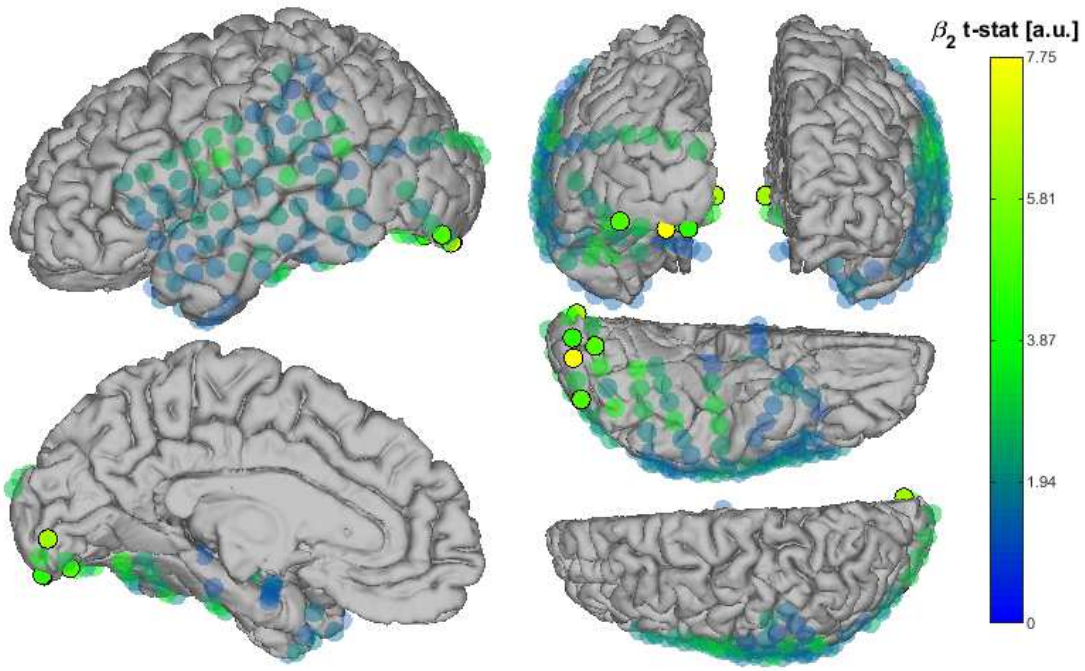
Supporting figure S2. Experiment 2: HFA responses. Lateral (top left), ventral (bottom) and frontal (right) views of the FreeSurfer average brain's left hemisphere are shown (for the sake of simplification, right-hemisphere sites were flipped to the left hemisphere). Sites with significant task-related HFA response modulations are plotted in solid colors and circled in black; sites that failed to reach significance are plotted in transparent colors. (a) Unweighted mean HFA responses to the task, as indexed by correlation coefficient β_0. 69 out of 576 sites were significantly activated by the task. (b) HFA modulation due to the physical characteristics of the stimuli (VbAv vs. VvAb), as encoded by β_1 (9 sites). (c) HFA modulation due to the patients' reported perception ("V" vs. "B"; β_2; 8 sites). (d) HFA modulation as a function of the interaction between the stimuli's physical characteristics and the patients' perception (β_3; 14 sites).



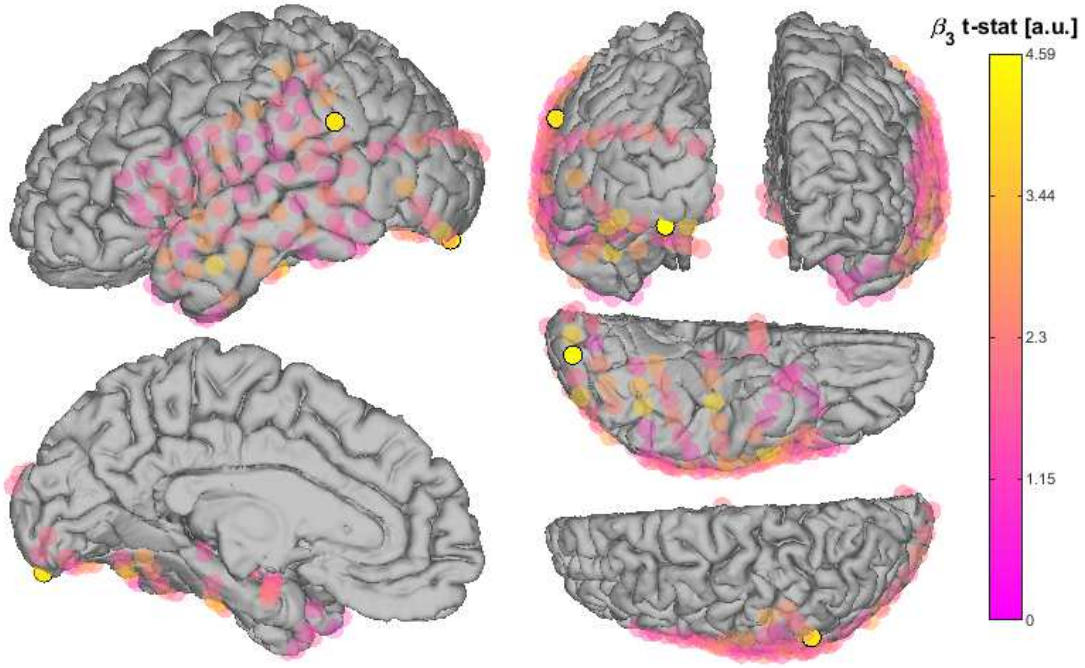
Supporting figure S3. Experiment 1, participant 1, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 141 intracranial EEG electrodes sampling the left cerebral hemisphere.



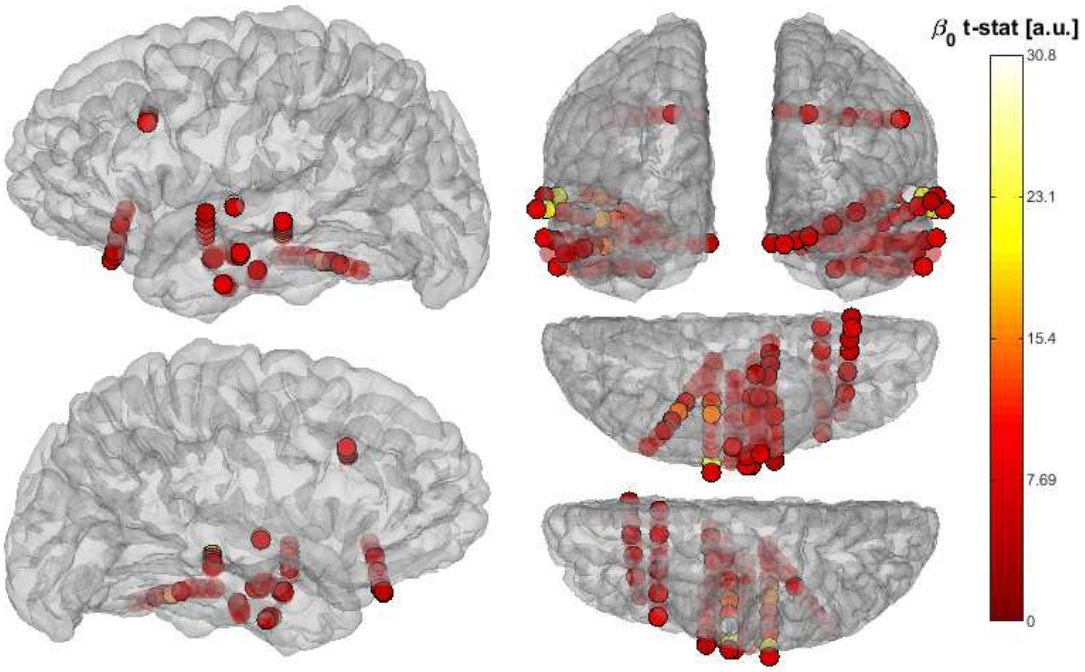
Supporting figure S4. Experiment 1, participant 1, absolute maximum value of the t-statistic for regressor β_1 (influence of auditory component).



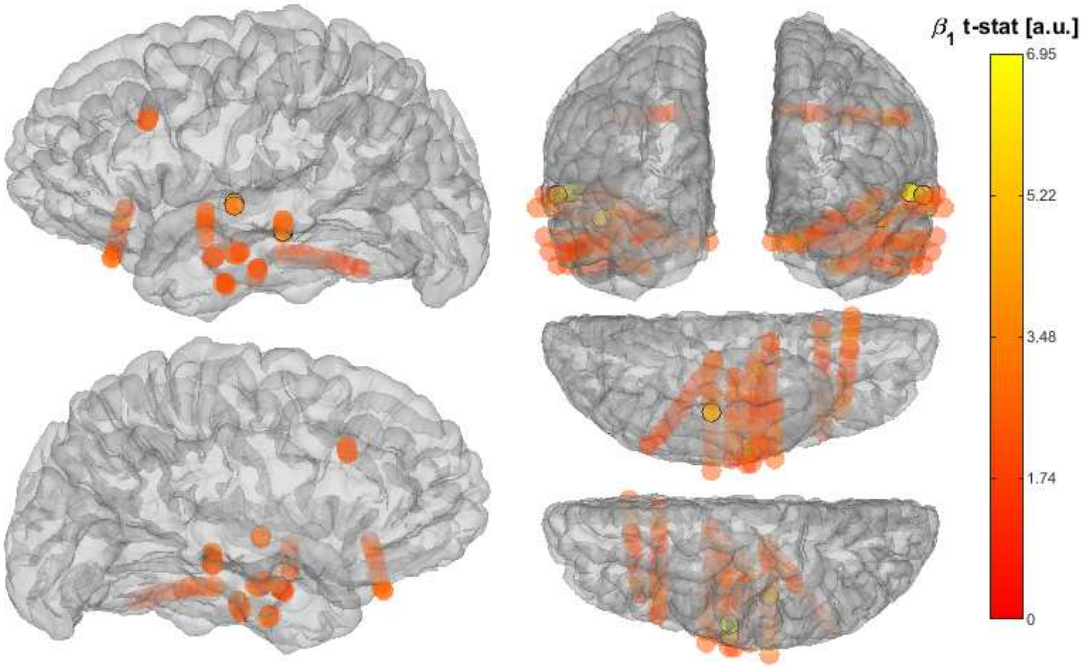
Supporting figure S5. Experiment 1, participant 1, absolute maximum value of the t-statistic for regressor β_2 (influence of visual component).



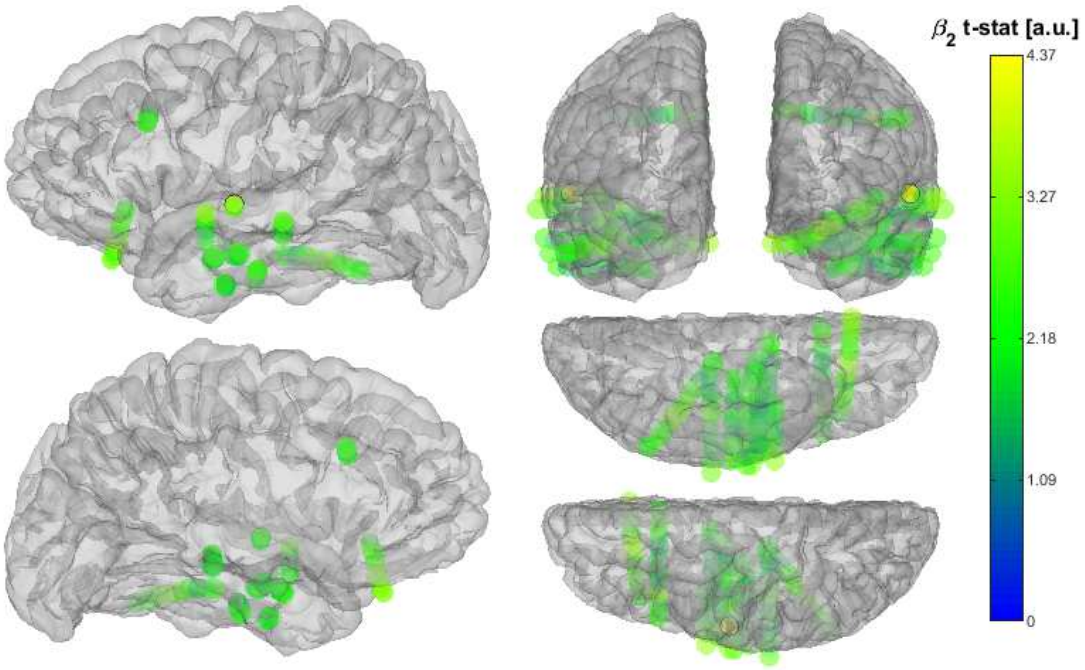
Supporting figure S6. Experiment 1, participant 1, absolute maximum value of the t-statistic for regressor β_3 (interaction between auditory and visual components).



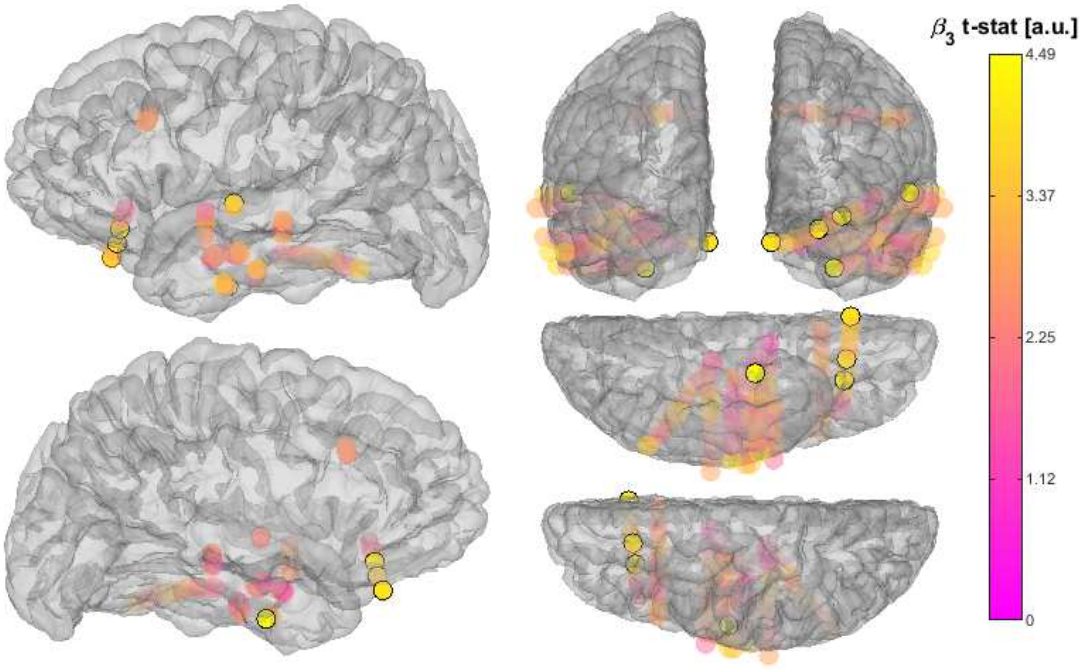
Supporting figure S7. Experiment 1, participant 2, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 88 intracranial EEG electrodes sampling the left cerebral hemisphere.



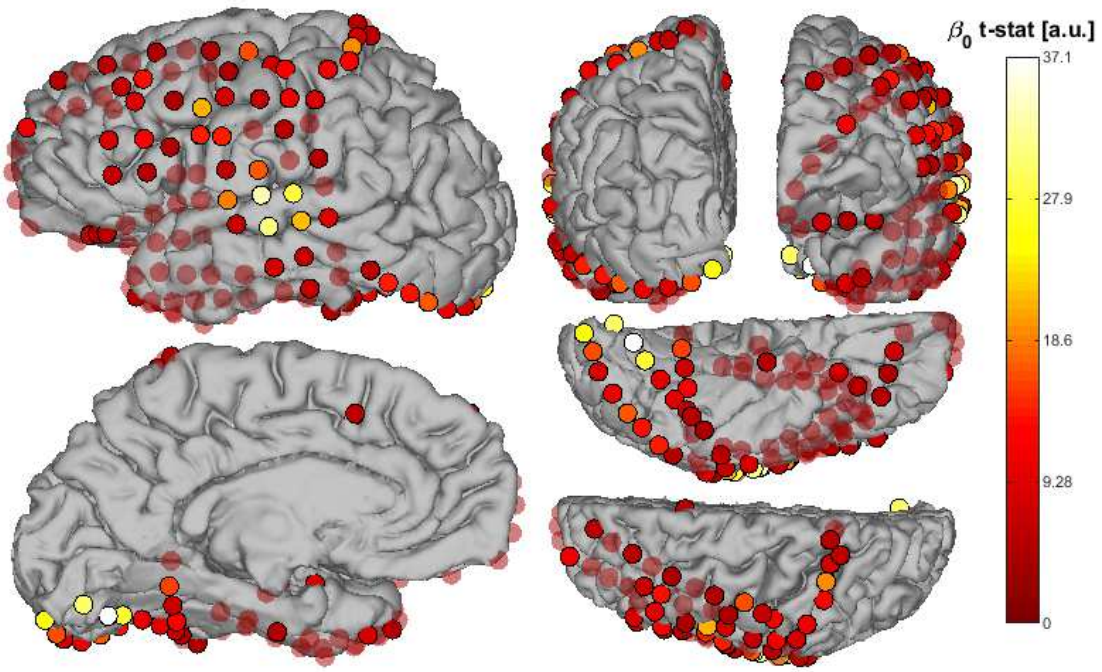
Supporting figure S8. Experiment 1, participant 2, absolute maximum value of the t-statistic for regressor β_1 (influence of auditory component).



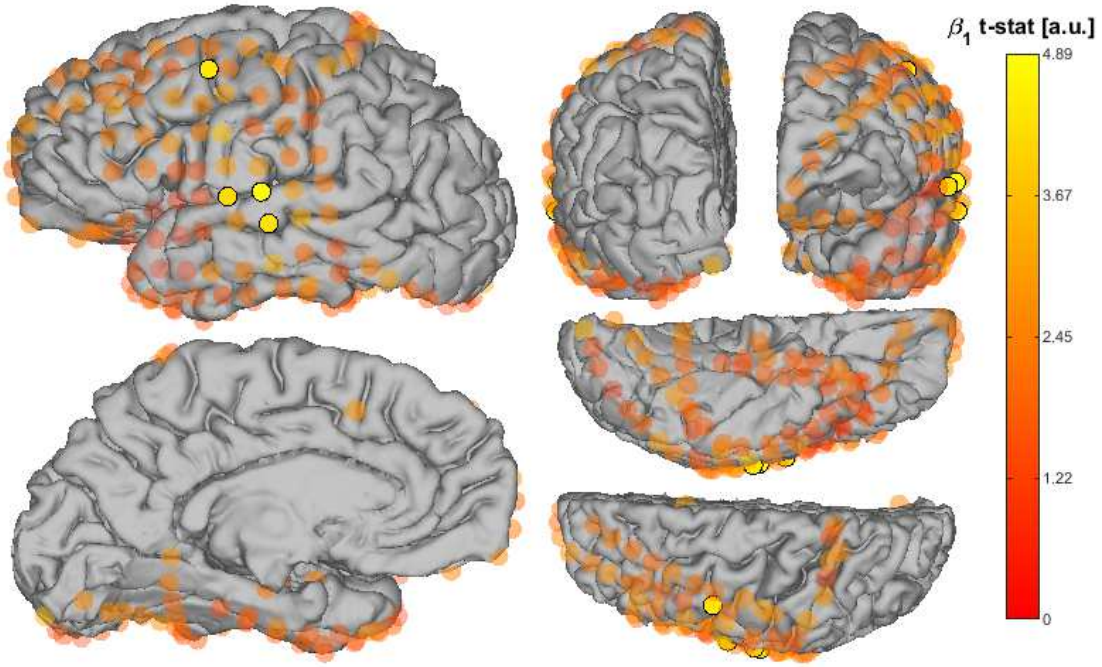
Supporting figure S9. Experiment 1, participant 2, absolute maximum value of the t-statistic for regressor β_2 (influence of visual component).



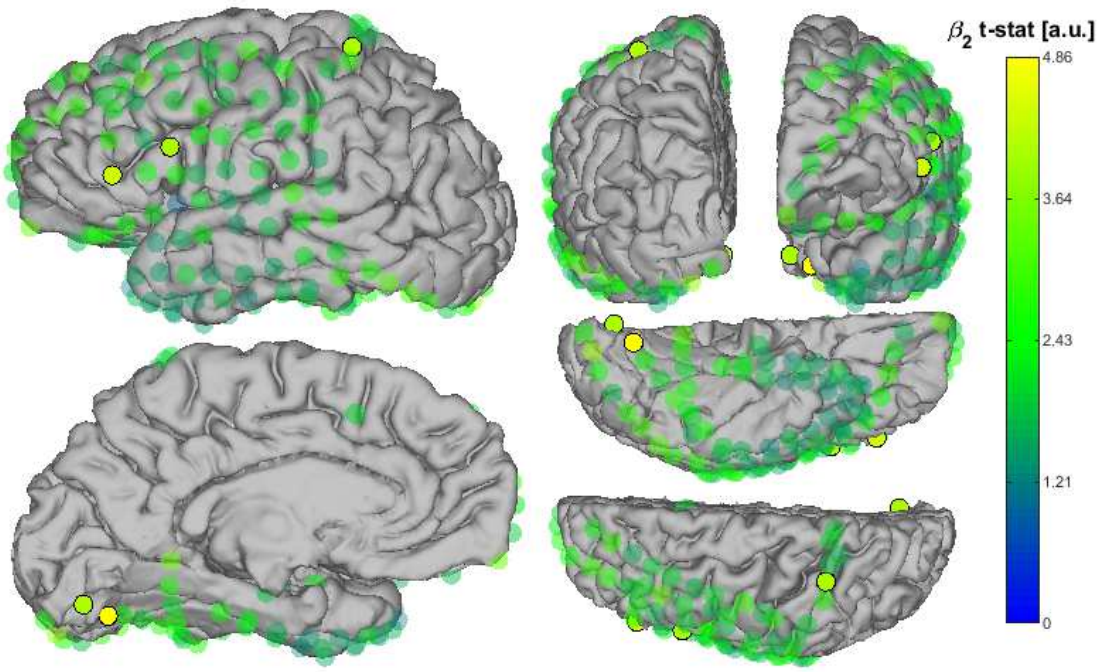
Supporting figure S10. Experiment 1, participant 2, absolute maximum value of the t-statistic for regressor β_3 (interaction between auditory and visual components).



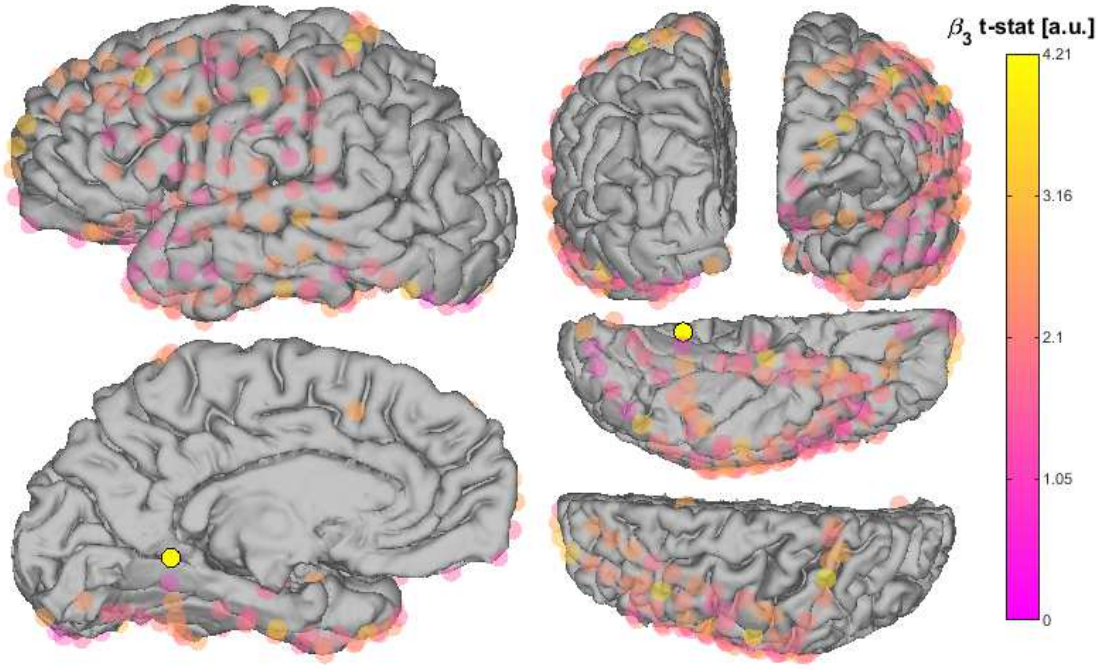
Supporting figure S11. Experiment 1, participant 3, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 156 intracranial EEG electrodes sampling the left cerebral hemisphere.



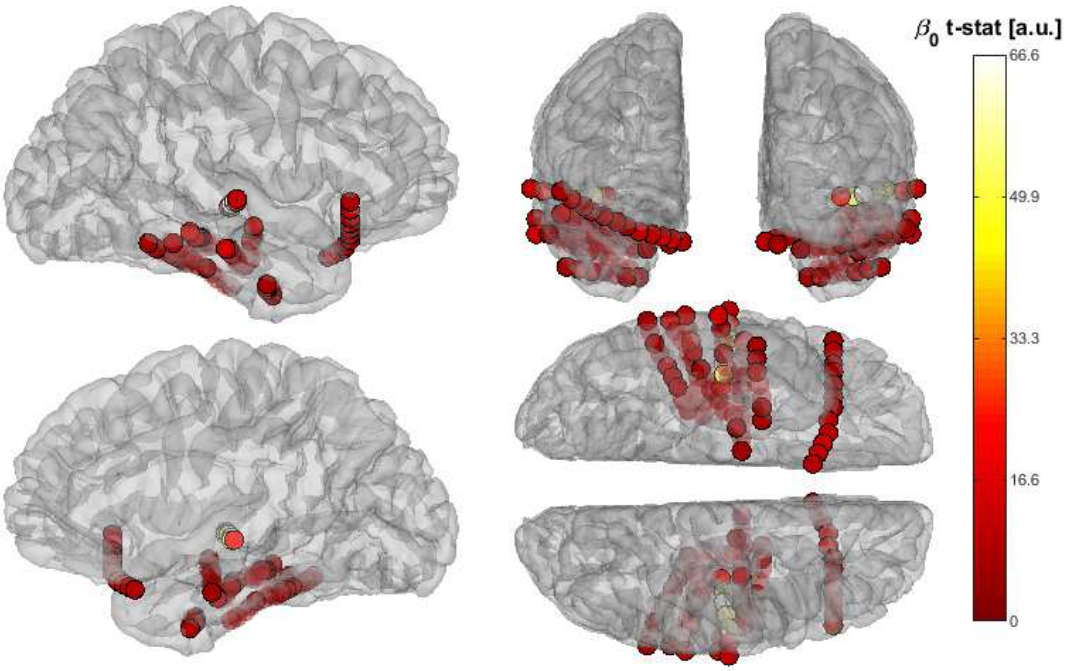
Supporting figure S12. Experiment 1, participant 3, absolute maximum value of the t-statistic for regressor β_1 (influence of auditory component).



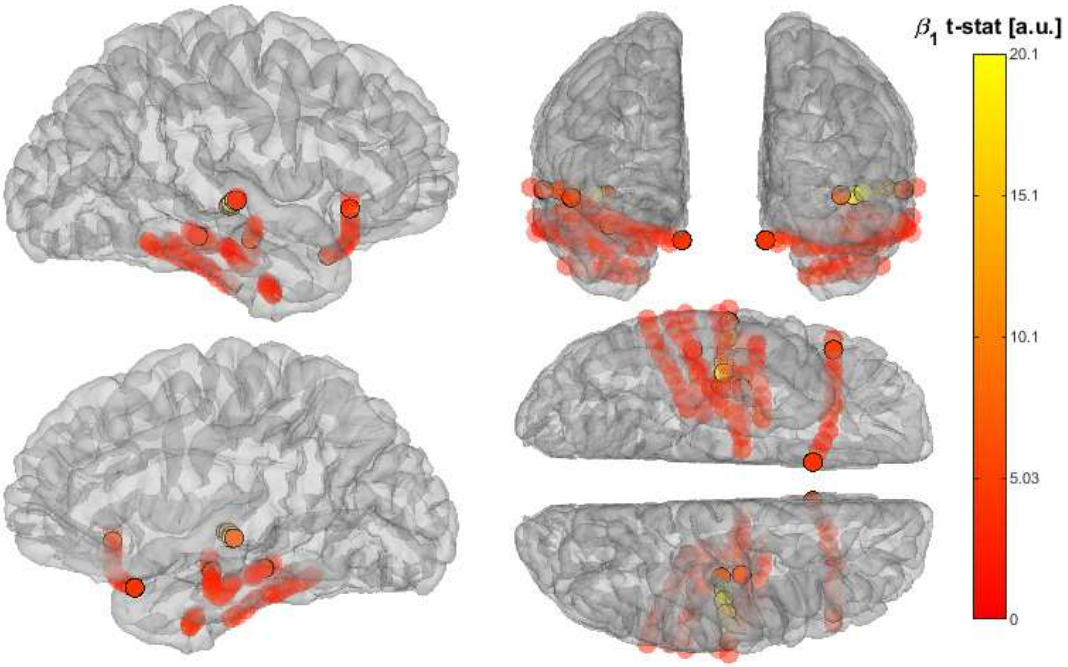
Supporting figure S13. Experiment 1, participant 3, absolute maximum value of the t-statistic for regressor β_2 (influence of visual component).



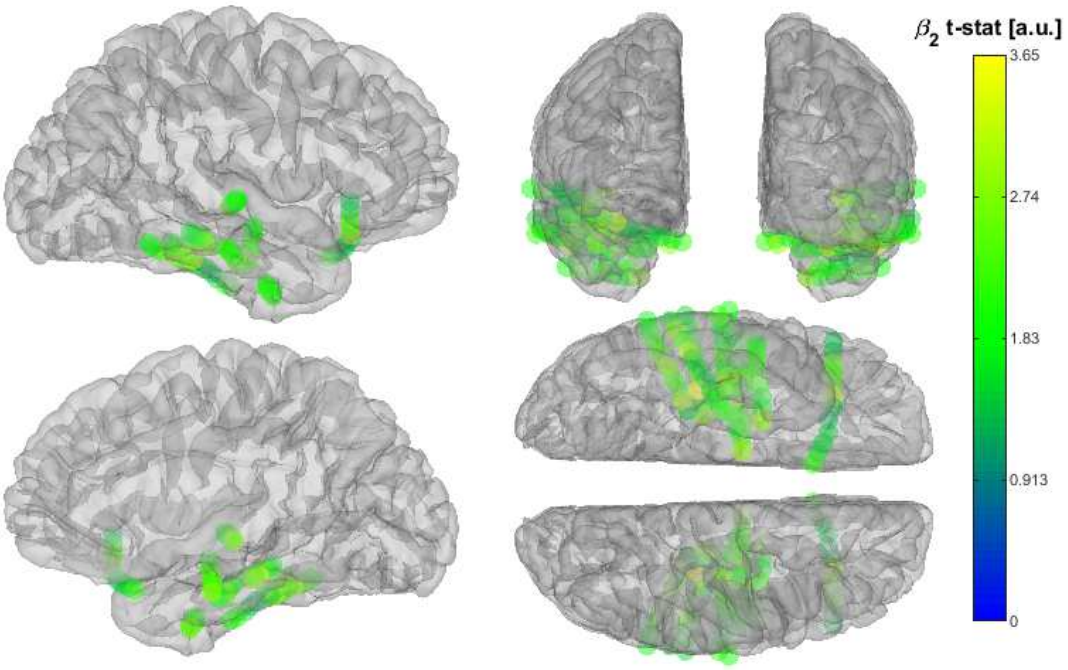
Supporting figure S14. Experiment 1, participant 3, absolute maximum value of the t-statistic for regressor β_3 (interaction between auditory and visual components).



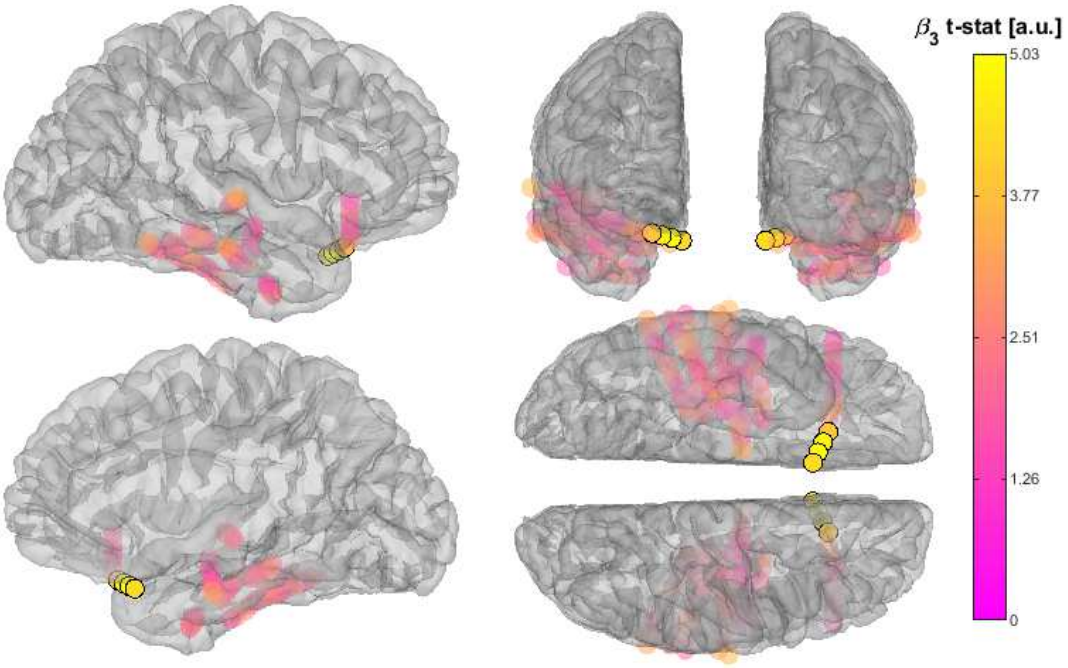
Supporting figure S15. Experiment 1, participant 4, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 73 intracranial EEG electrodes sampling the right cerebral hemisphere.



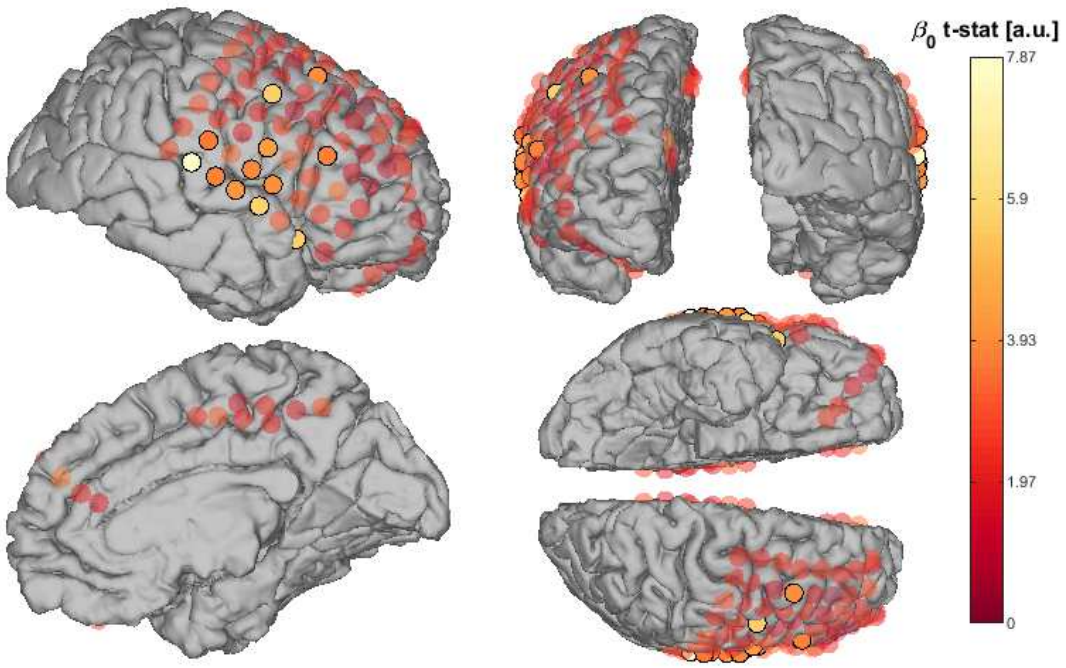
Supporting figure S16. Experiment 1, participant 4, absolute maximum value of the t-statistic for regressor β_1 (influence of auditory component).



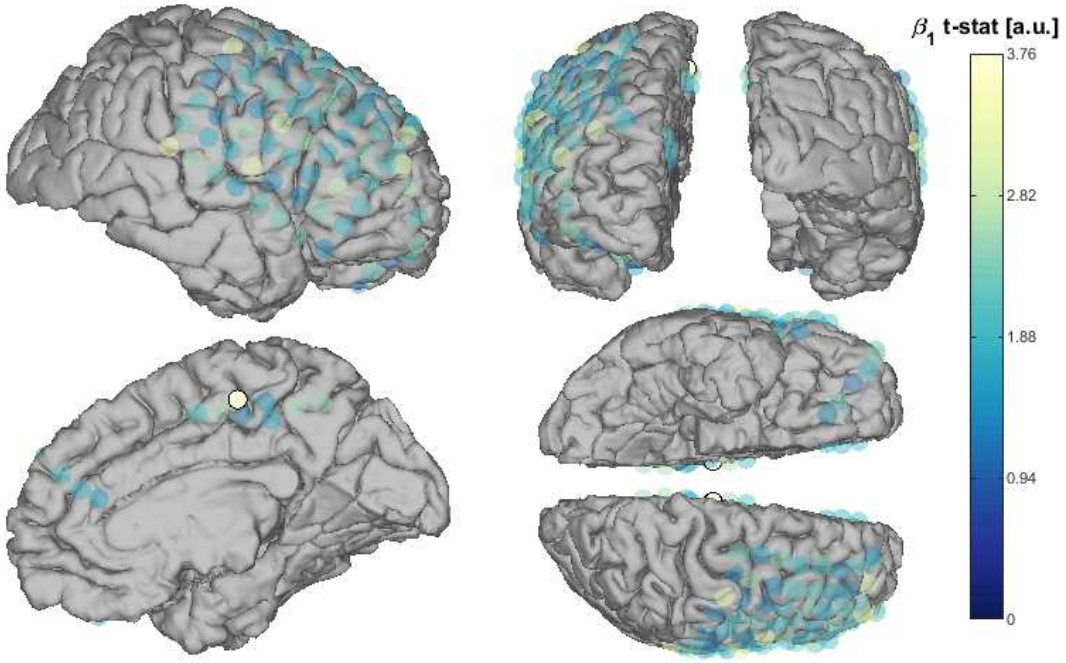
Supporting figure S17. Experiment 1, participant 4, absolute maximum value of the t-statistic for regressor β_2 (influence of visual component).



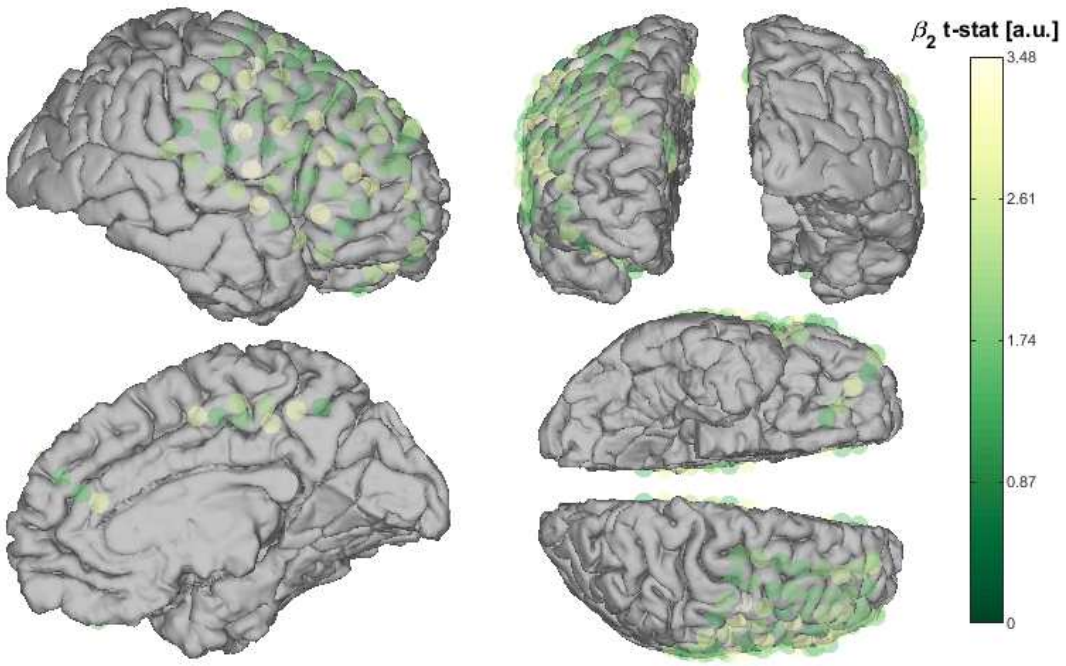
Supporting figure S18. Experiment 1, participant 4, absolute maximum value of the t-statistic for regressor β_3 (interaction between auditory and visual components).



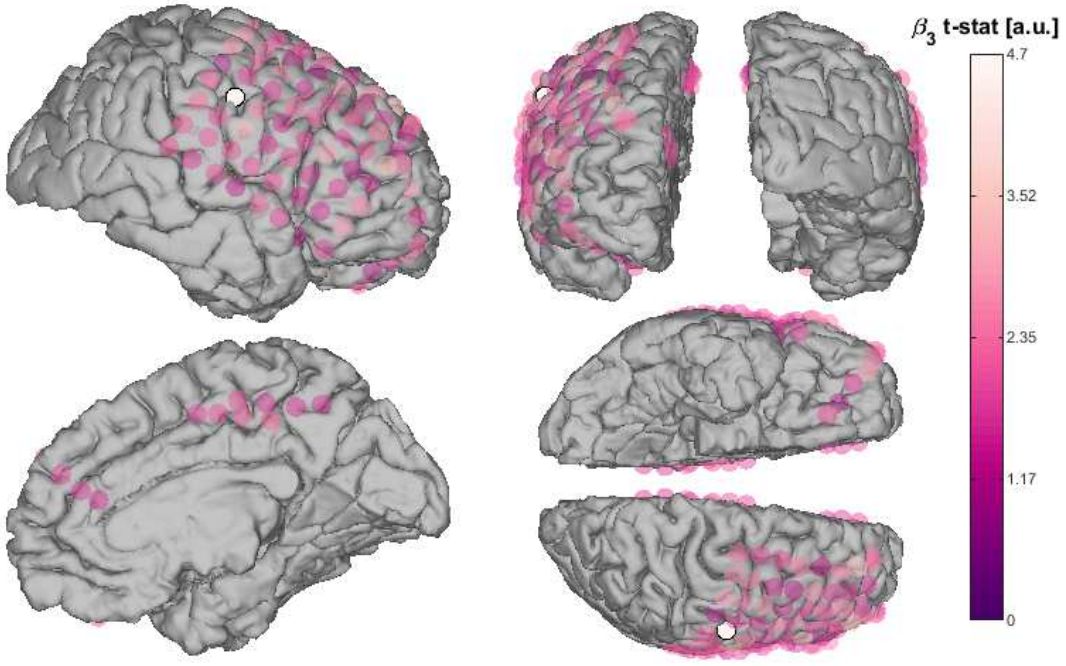
Supporting figure S19. Experiment 2, participant 1, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 83 intracranial EEG electrodes sampling the right cerebral hemisphere.



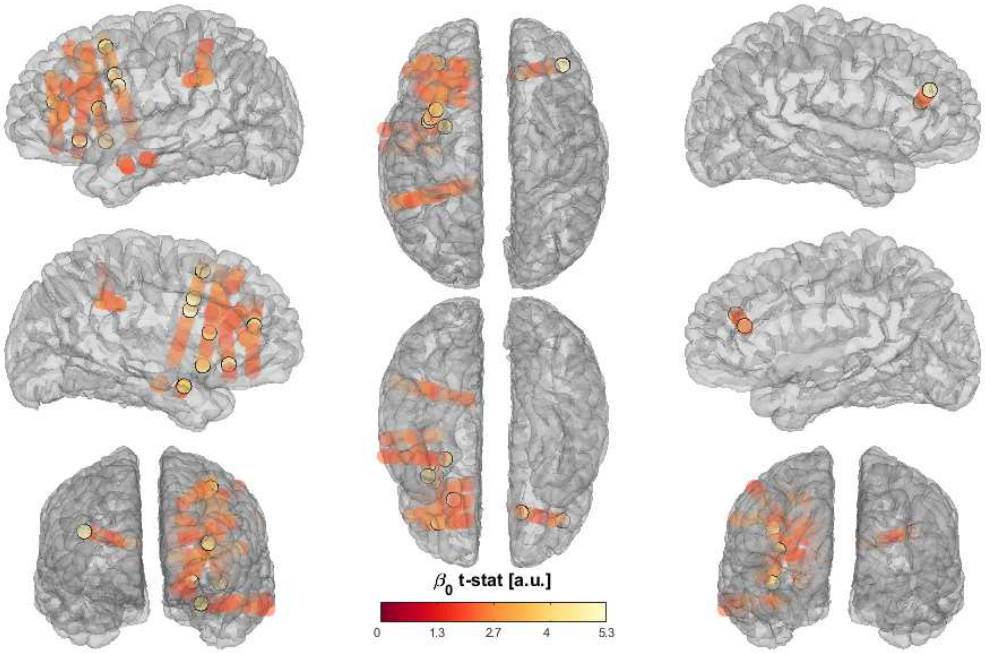
Supporting figure S20. Experiment 2, participant 1, absolute maximum value of the t-statistic for regressor β_1 (influence of stimuli's physical characteristics).



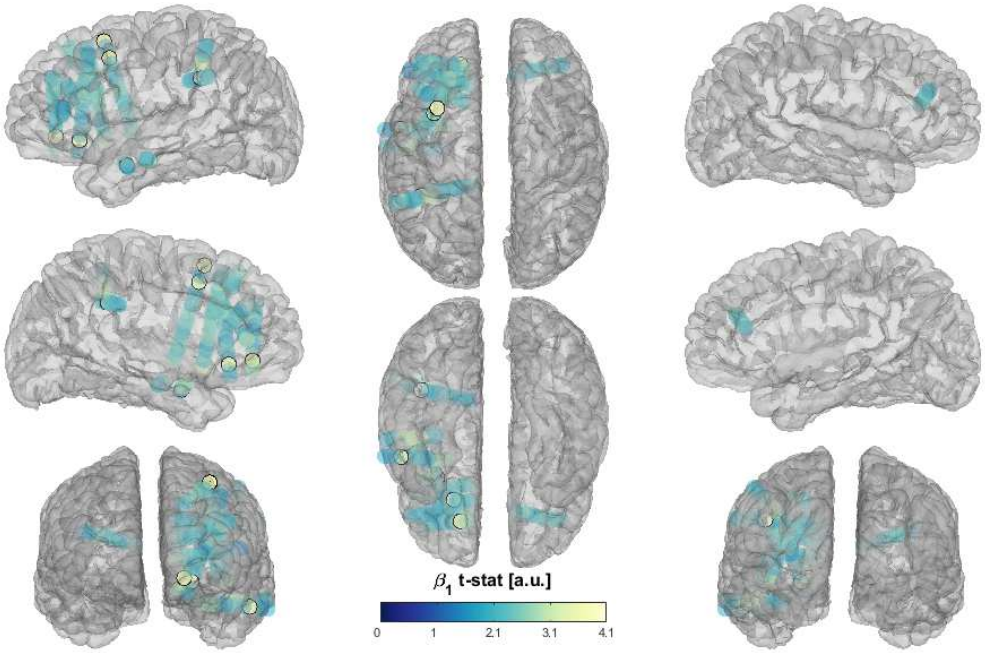
Supporting figure S21. Experiment 2, participant 1, absolute maximum value of the t-statistic for regressor β_2 (influence of participant's reported perception).



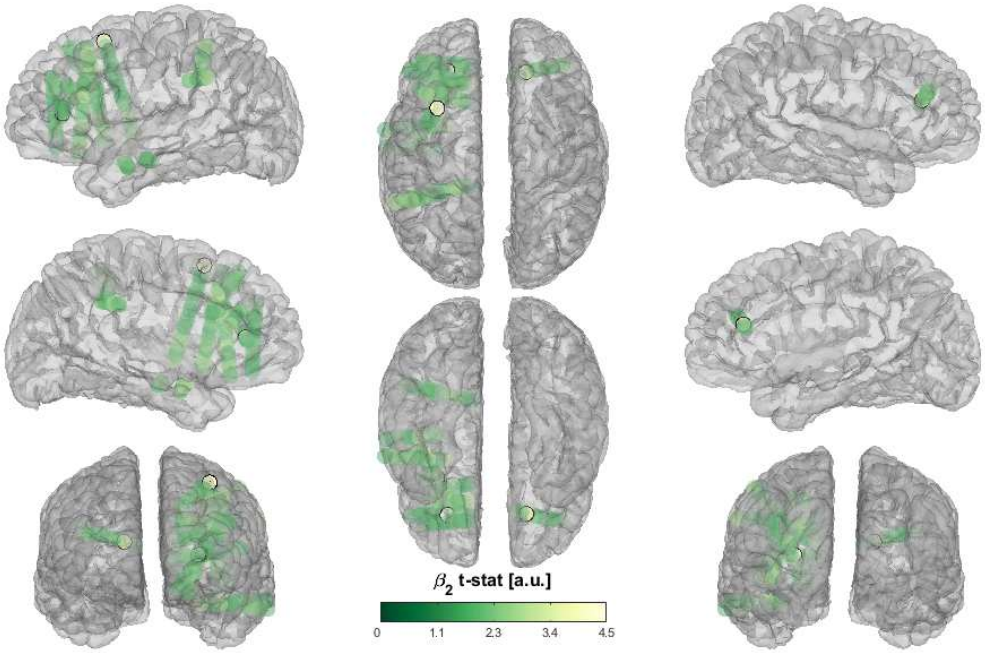
Supporting figure S22. Experiment 2, participant 1, absolute maximum value of the t-statistic for regressor β_3 (interaction of stimuli's physical characteristics and participant's reported perception).



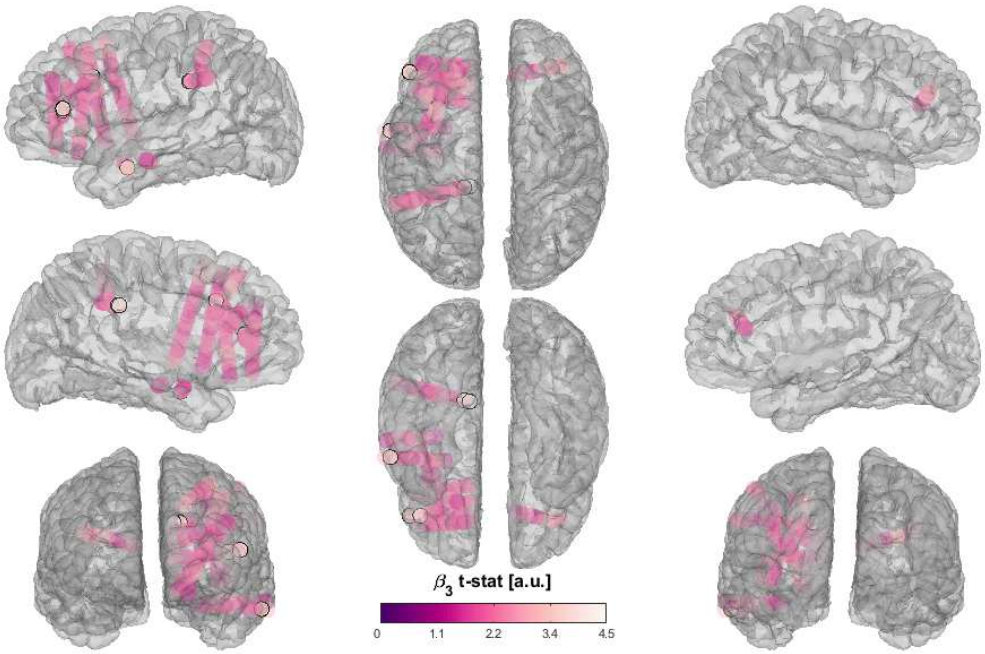
Supporting figure S23. Experiment 2, participant 2, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 151 intracranial EEG electrodes sampling the left cerebral hemisphere, 10 sampling the right hemisphere.



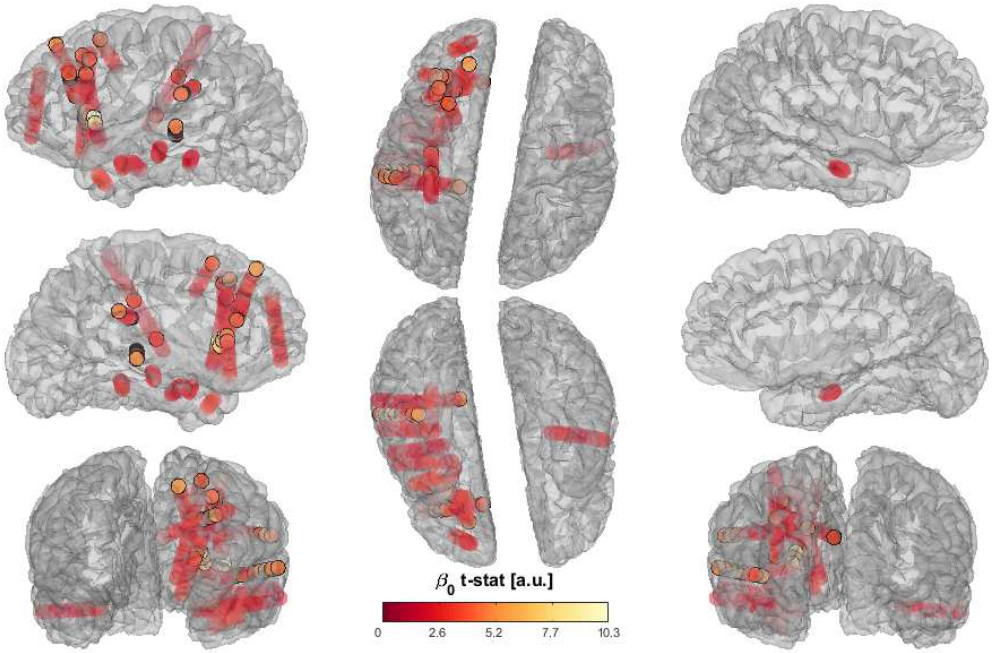
Supporting figure S24. Experiment 2, participant 2, absolute maximum value of the t-statistic for regressor β_1 (influence of stimuli's physical characteristics).



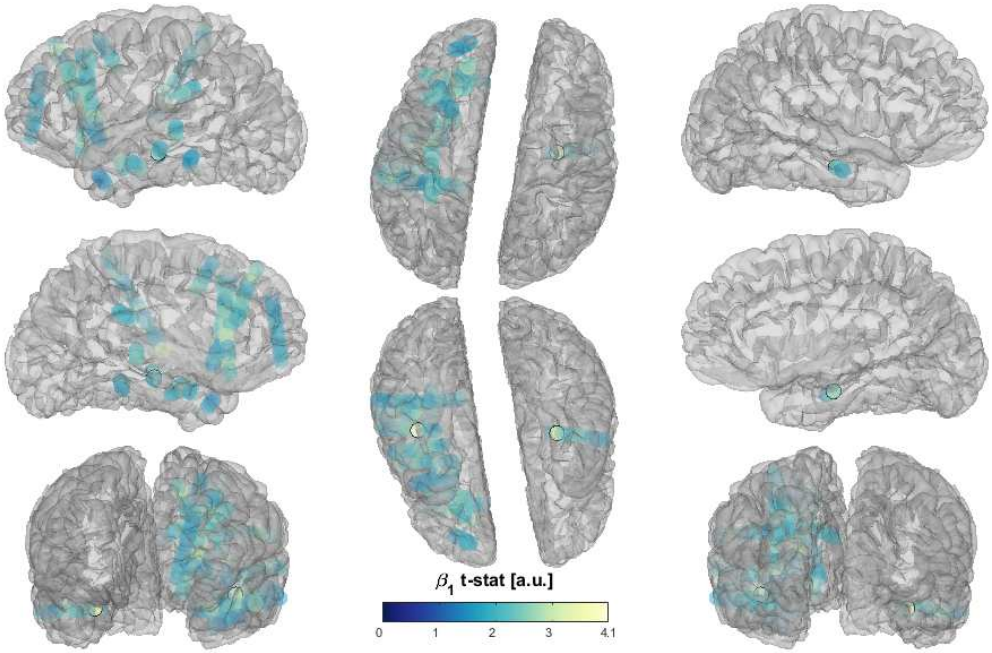
Supporting figure S25. Experiment 2, participant 2, absolute maximum value of the t-statistic for regressor β_2 (influence of participant's reported perception).



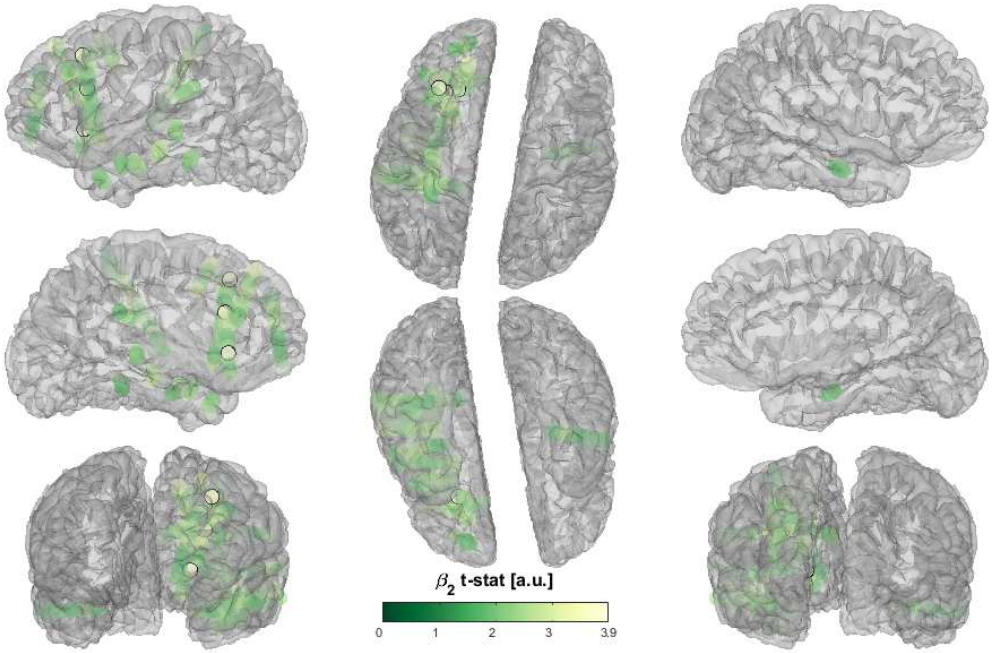
Supporting figure S26. Experiment 2, participant 2, absolute maximum value of the t-statistic for regressor β_3 (interaction of stimuli's physical characteristics and participant's reported perception).



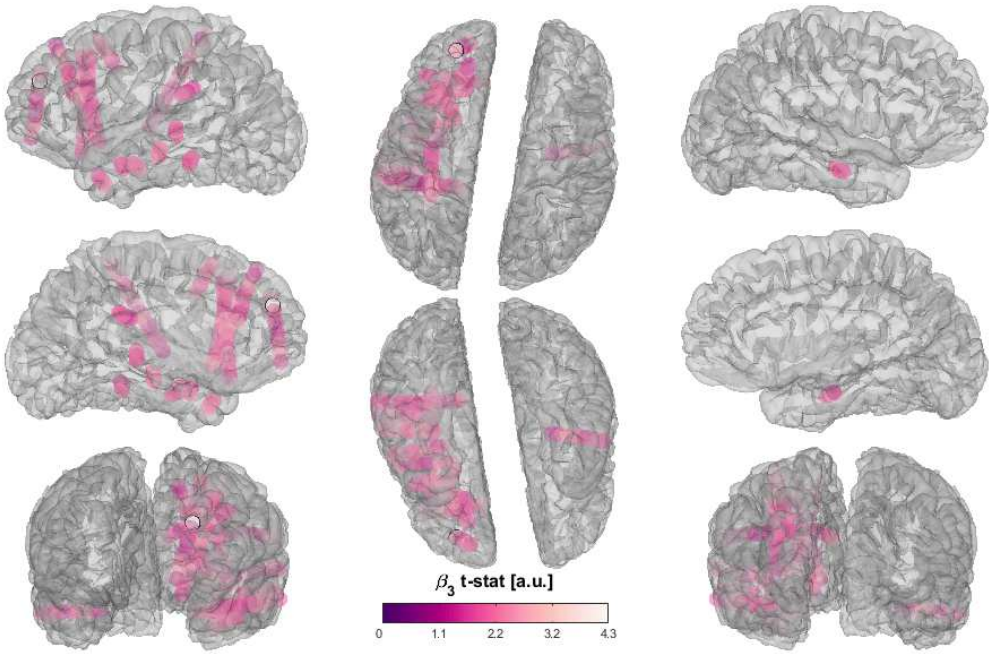
Supporting figure S27. Experiment 2, participant 3, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 166 intracranial EEG electrodes sampling the left cerebral hemisphere, 10 sampling the right hemisphere.



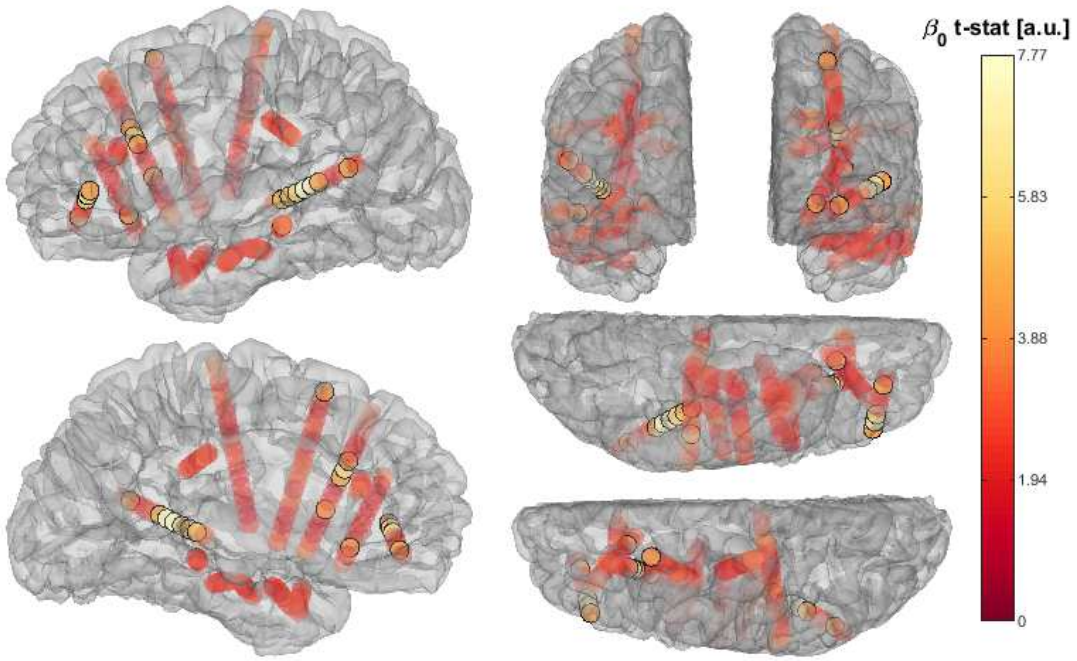
Supporting figure S28. Experiment 2, participant 3, absolute maximum value of the t-statistic for regressor β_1 (influence of stimuli's physical characteristics).



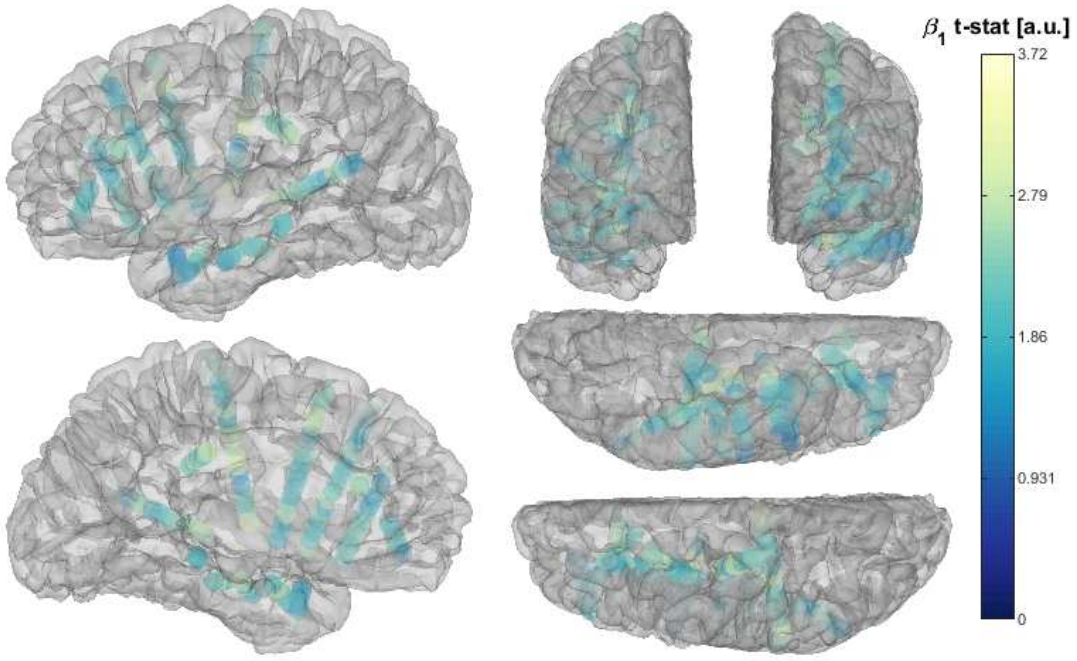
Supporting figure S29. Experiment 2, participant 3, absolute maximum value of the t-statistic for regressor β_2 (influence of participant's reported perception).



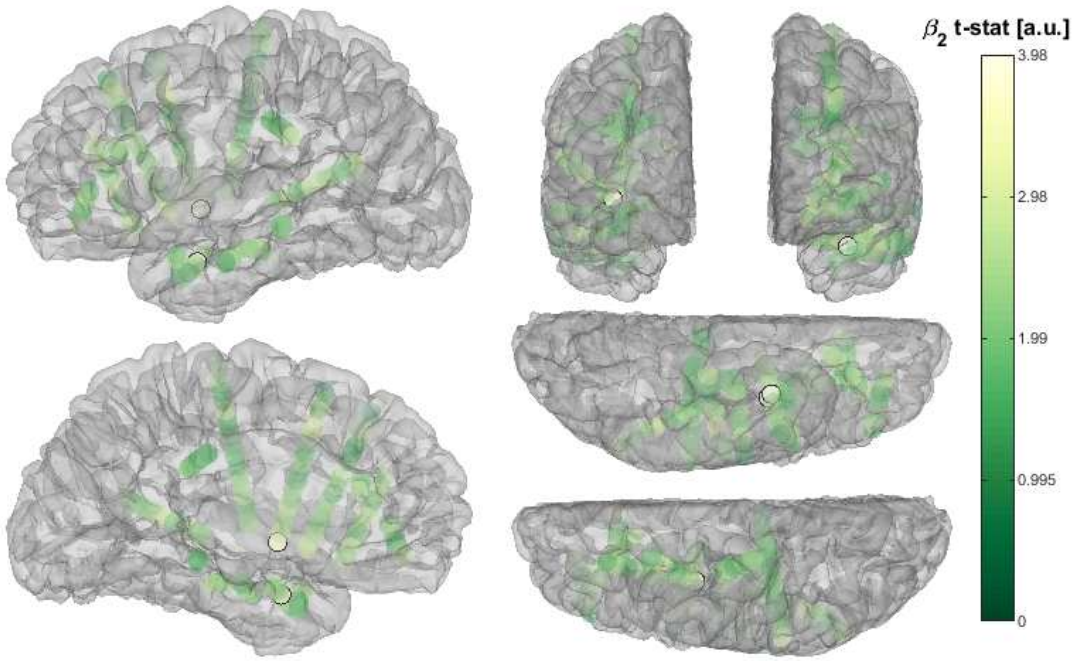
Supporting figure S30. Experiment 2, participant 3, absolute maximum value of the t-statistic for regressor β_3 (interaction of stimuli's physical characteristics and participant's reported perception).



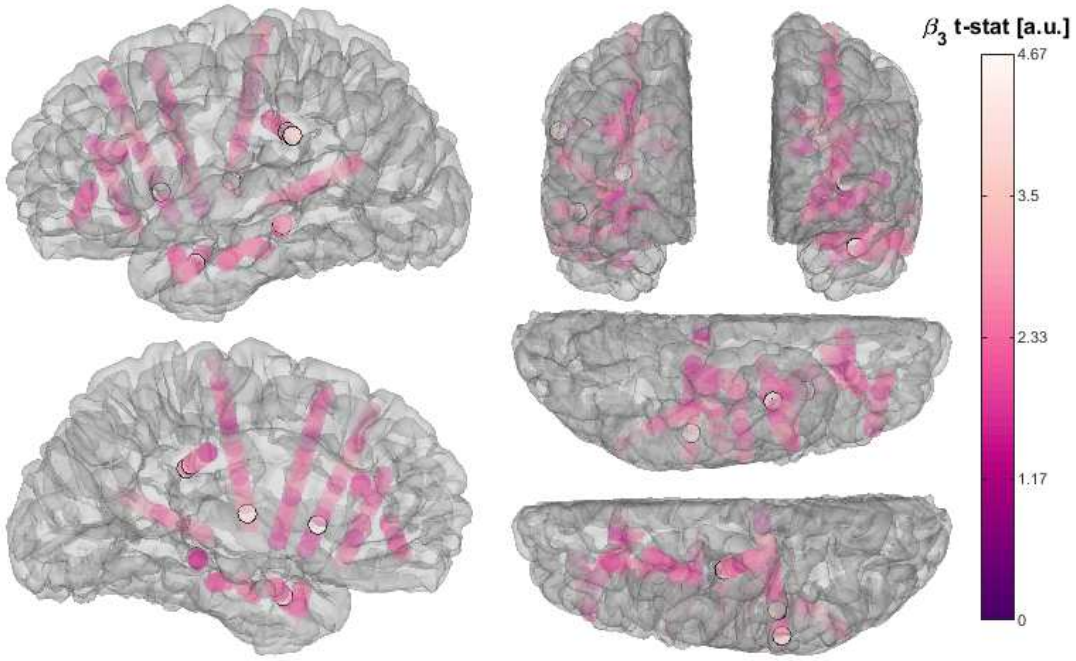
Supporting figure S31. Experiment 2, participant 4, absolute maximum value of the t-statistic for regressor β_0 (overall response to all stimuli). 156 intracranial EEG electrodes sampling the left cerebral hemisphere.



Supporting figure S32. Experiment 2, participant 4, absolute maximum value of the t-statistic for regressor β_1 (influence of stimuli's physical characteristics).



Supporting figure S33. Experiment 2, participant 4, absolute maximum value of the t-statistic for regressor β_2 (influence of participant's reported perception).



Supporting figure S34. Experiment 2, participant 4, absolute maximum value of the t-statistic for regressor β_3 (interaction of stimuli's physical characteristics and participant's reported perception).