JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Left motor delta oscillations reflect asynchrony detection in multisensory speech perception

https://doi.org/10.1523/JNEUROSCI.2965-20.2022

Cite as: J. Neurosci 2022; 10.1523/JNEUROSCI.2965-20.2022

Received: 25 November 2020
Revised: 12 January 2022
Accepted: 14 January 2022

Left motor delta oscillations reflect asynchrony detection in multisensory speech perception

Emmanuel Biau [1,2], Benjamin G. Schultz [2], Thomas C. Gunter [3] and Sonja A. Kotz [2,3].

[1]Department of Psychology, University of Liverpool, L69 7ZA, Liverpool, United Kingdom.

[2]Basic and Applied NeuroDynamics Laboratory, Department of Neuropsychology and Psychopharmacology, University of Maastricht, 6200 MD, Maastricht, Netherlands.

[3]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103, Leipzig, Germany.

Corresponding authors:

Dr. Emmanuel Biau

Address: Department of Psychology, University of Liverpool, L69 7ZA, Liverpool, United Kingdom.

Email : e.biau@liverpool.ac.uk

Prof. Dr. Sonja A. Kotz

Address: Basic and Applied NeuroDynamics Laboratory, Department of Neuropsychology and Psychopharmacology, University of Maastricht, 6200 MD, Maastricht, Netherlands.

Email: sonja.kotz@maastrichtuniversity.nl

ABSTRACT

During multisensory speech perception, slow delta oscillations (~1 - 3 Hz) in the listener's brain synchronize with the speech signal, likely engaging in speech signal decomposition. Notable fluctuations in the speech amplitude envelope, resounding speaker prosody, temporally align with articulatory and body gestures and both provide complementary sensations that temporally structure speech. Further, delta oscillations in the left motor cortex seem to align with speech and musical beats, suggesting their possible role in the temporal structuring of (quasi)-rhythmic stimulation. We extended the role of delta oscillations to audio-visual asynchrony detection as a test case of the temporal analysis of multisensory prosody fluctuations in speech. We recorded EEG responses in an audio-visual asynchrony detection task while participants watched videos of a speaker. We filtered the speech signal to remove verbal content and examined how visual and auditory prosodic features temporally (mis-)align. Results confirm (i) that participants accurately detected audio-visual asynchrony, and (ii) increased delta power in the left motor cortex in response to audio-visual asynchrony. The difference of delta power between asynchronous and synchronous conditions predicted behavioural performance, and (iii) decreased delta-beta coupling in the left motor cortex when listeners could not accurately map visual and auditory prosodies. Finally, both behavioural and neurophysiological evidence was altered when a speaker's face was degraded by a visual mask. Together, these findings suggest that motor delta oscillations support asynchrony detection of multisensory prosodic fluctuation in speech.

KEYWORDS

Audio-visual asynchrony; multisensory speech; delta oscillations; prosody; motor cortex.

SIGNIFICANCE STATEMENT

Speech perception is facilitated by regular prosodic fluctuations that temporally structure the auditory signal. Auditory speech processing involves the left motor cortex and associated delta oscillations. However, visual prosody (i.e., a speaker's body movements) complements

auditory prosody, and it is unclear how the brain temporally analyses different prosodic features in multisensory speech perception. We combined an audio-visual asynchrony detection task with electroencephalographic recordings to investigate how delta oscillations support the temporal analysis of multisensory speech. Results confirmed that asynchrony detection of visual and auditory prosodies leads to increased delta power in left motor cortex and correlates with performance. We conclude that delta oscillations are invoked in an effort to resolve denoted temporal asynchrony in multisensory speech perception.

INTRODUCTION

Speaker prosody displays perceptible fluctuations in the speech amplitude envelope, allowing a listener to segment and parse incoming speech (Ghitza, 2017). While not isochronous, prosody imposes a temporal structure with regular alterations of strong and weak accentuated cues occurring at ~1 - 3 Hz delta rate (Ding et al., 2016; Doelling et al., 2014; Ghitza, 2017; Pell & Davis, 2012). Populations of neurons in visual and auditory cortices synchronize their firing responses with the onsets of predictable events, structuring sensory signals at a delta rate. Such “neural entrainment” reflects the early stages in sensory processing by which neural oscillations might track temporally relevant signal features. The neural representation of such sensory features must temporally align, and there is evidence that delta oscillations play a role in unisensory as well as multisensory integration (Giraud & Poeppel, 2012; Keitel et al., 2017; Kösem & van Wassenhove, 2017; Meyer, Sun & Martin, 2019). For example, using a temporal order judgement task, Kösem, Gramfort and van Wassenhove (2014) showed that the phase shifts of entrained delta oscillations in the auditory cortex linearly mapped participants' perception of audio-visual simultaneity. Other studies described an interaction of delta oscillations in visual and auditory cortices for audio-visual speech (Crosse et al., 2016; Mercier et al., 2015). Crosse, Liberto, and Lalor (2016) reported that speech envelope tracking in the auditory cortex improved through visual information, particularly in the delta range. Beyond segmentation, prosody presents in visual and auditory information and facilitate synchronization of multimodal information in social interaction (Esteve-Gibert & Guellaï, 2018; Kotz, Ravignani & Fitch, 2018). The term “visual prosody” encompasses communicative gestures (i.e., hand, head, face, and body movements) whose prominent phase temporally coincides with acoustic prosodic features such as intonational phrases, pitch accents, and boundary tones (Biau et al., 2016; Chandrasekaran et al., 2009; Munhall et al., 2004; Wagner et al., 2014). For example, listeners rely on the successful temporal analysis of gestures and sounds in speech perception (Cherry, 1953; Obermeier, Dolk & Gunter, 2012; Sumby & Pollack, 1954). Together this raises the following questions: How does the brain temporally align multiple dynamic prosodies in multisensory speech perception?

The present study investigated whether delta oscillations respond to manipulation of temporal alignment in multisensory speech (i.e., dynamic non-verbal visual and auditory prosodies). We refer to temporal alignment as the mechanism by which the brain attempts to integrate the quasi-rhythmic structure of visual and auditory prosodies in multisensory speech. Delta activity in the motor cortex has been associated with the temporal analysis of rhythmic stimuli as its phase aligns with the onsets of predictable events (Morillon et al., 2019; Morillon & Schroeder, 2015; Saleh et al., 2010). In speech, Keitel and colleagues (2018) showed that left motor delta activity tracked temporally predictable slow phrasal features in auditory sentences and predicted successful speech comprehension. This suggests that this region responds to perceptually relevant regularities in the signal to improve comprehension. Keitel et al. (2018) also found delta-beta cross-frequency coupling in the left motor region, in line with previous research, showing that motor beta oscillations respond to the temporal alignment of rhythmic auditory tones or visual cues (Fujioka, Ross & Trainor, 2015; Saleh et al., 2010). These findings led to the hypothesis that delta oscillations are involved in the temporal analysis of speech by mediating top-down control through cross-frequency coupling with beta activity (Arnal, 2012; Arnal, Doelling & Poeppel, 2015; Morillon and Baillet, 2017). In other words, delta activity could reflect how the brain gathers and temporally analyses different sensory inputs in left motor cortex and generates predictions to improve (multisensory) signal processing. Finally, the left motor cortex, including the left inferior frontal gyrus, is involved in gestures and speech integration (Biau et al., 2016; Park et al., 2016; Zhao et al., 2018).

We propose that visual and auditory prosodic features encoded in the visual and auditory sensory cortices provide two representations of the speech signal, and their (un-)successful temporal alignment may recruit the left motor cortex during speech perception. To test this hypothesis, we manipulated the temporal structure of filtered multisensory speech, including whole body or masked head movements. Participants performed an audio-visual synchrony detection task and watched small video clips of a single speaker engaged in a conversation. We also recorded their electroencephalogram (EEG). Firstly, we tested behaviourally how successfully listeners temporally align visual and auditory prosodic features in multisensory speech. We then analysed modulations of delta oscillations in response to audio-visual asynchrony to find out whether and to which degree they index (un-)successful temporal alignment in multisensory speech perception. Thirdly, we tested whether delta-beta coupling in the left motor cortex predicts multisensory (a-)synchrony detection in speech perception.

METHODS

Participants

We recruited twenty-six native Dutch speakers (mean age = 22.24, SD = 4.24; 15 females) at Maastricht University, who received €10 for participating in the experiment after giving informed consent. All participants were right-handed and had normal or corrected-to-normal vision and hearing. The protocol of the study was approved by the Research Ethical Committee of Maastricht University. Data from three participants were removed from the final analysis due to technical problems.

Stimuli

Short videos were extracted from a longer video recording used in a previous study (Gunter & Weinbrenner, 2017). The videos depicted a female actor and an experimenter (both German native speakers) engaged in a question-answer conversation. The actor sat on a chair, moved freely, and was visible from her knees up to the top of her head. Relevant segments containing the actor' answers separate from the experimenter were selected to create the current stimulus set (N = 54). Each of the 54 segments was 10 seconds long (600 frames at 60 frames per second; FPS). The audio track was extracted to be low pass filtered with Hann band windowing procedure (from 0 Hz to 400 Hz; 20 Hz smoothing) using Praat (Boersma & Weenink, 2015). In doing so, we altered speech intelligibility removing verbal content while keeping the prosodic contour of the signal. Peak frequencies were extracted from the audio and video files through Fourier transformations that calculated the frequency at which the peak amplitude occurred within a range of 0.5Hz to 8Hz. For videos, the average magnitude of grayscale pixel changes between consecutive frames was used to determine the frequency of movement and gesture (see Table 1).

[Insert Table 1]

Table 1. Summary statistics of peak frequencies obtained for video and audio signals using a Fourier transformation. Video signals: Full (Head + Body), Head only (either original head information or head-masked) and Body only (lower body part without head information). Audio signals: Audio only. The Audio and Body only measures (*) are consistent across the no-mask and head-mask conditions. Frequencies are shown for the no-mask and head-mask videos for the masked area (Head only) and all pixels (Full video).

We applied two visual manipulations to each of the 54 speech segments: (1) The presence or absence of a visual mask (no mask, head-mask), and (2) the original temporal alignment of the audio-visual information or a temporal shift of the audio signal relative to the video onset (synchronous, asynchronous). In the no-mask condition, the speaker's body and face were fully visible. In the head-mask condition, the head of the speaker was blurred to degrade visual prosody conveyed by the speaker's lips. The mask was created by applying a low-pass Gaussian filter on the upper third of the original video containing the speaker's face, attenuating a high frequency signal. This manipulation removed fine-grained facial expressions from the video while slow gestures remained intact (see Figure 1). In the synchronous condition, the original temporal alignment between visual and auditory onsets was intact. To create an asynchronous condition, we inserted a delay between the visual and auditory onsets by shifting the sound onset by +400 ms relative to the video onset (i.e., 24 frames). This manipulation maintained the natural order of visual information preceding auditory information in the synchronous condition in ecologically valid contexts but with a longer duration. In the current experiment, audio onsets did not precede corresponding video onsets. This lag duration was based on a time-window of multisensory integration established in previous studies (Biau et al., 2016; Biau & Soto-Faraco, 2013; Jessen & Kotz, 2015; Obermeier & Gunter, 2014). A 400 ms lag ensured that the delay between video and audio onsets was long enough for participants to detect audio-visual asynchrony at a success rate of approximatively 80% ± 5%. This delay was established in a pilot experiment with different participants (n = 16). Results confirmed that participants detected both synchrony and asynchrony between visual and auditory information in the audio-visual stimuli similarly (correct response rates in the synchronous condition: 0.79 ± 0.11 and asynchronous condition: 0.78 ± 0.12; t(1,15) = -0.26; *p* = 0.797; two-tailed; *Cohen's d* = 0.07). This was done to ensure we retained enough correct response trials in both conditions for further EEG analyses. Further, a central white fixation cross was displayed in each video to allow participants to focus their gaze on a central cue while attending audio-visual stimuli. Altogether, this created four conditions: Head-Mask Synchronous (HMS), Head-Mask Asynchronous (HMA), No-Mask Synchronous (NMS), and No-Mask Asynchronous (NMA) (see Figure 1A). 18 additional video clips, in which the central white fixation-cross turned red, were used as fillers, counterbalanced across conditions (colour change onset jittered between 5 and 9 seconds after the video onset; ~ 8 % of total stimuli, not included in the final data analysis). We used the fillers in a memory test to focus the participants' attention on the videos during the experiment. Finally, audio files were recombined with corresponding video files in each condition. Videos were edited using Adobe Premiere Pro CS3 and exported using the following parameters: Pixel resolution 1920 × 1080, 60 FPS compressor Indeo video 5.10, AVI format, audio sample rate 48 kHz, 16 bits Mono.

Apparatus

The audio files were presented through EEG-compatible air tubes (ER3C Tubal Insert Earphones, Etymotic Research). Videos were presented on a 27-inch Iiyama G-MASTER (GB2760HSU-B1) TN display with a 1ms response time, a refresh rate of 144Hz, and a native resolution of 1920 x 1080 pixels connected to the stimulus presentation computer (Intel i7-6700 CPU @ 3.40 GHz, 32 GB, running 64-bit Windows 7, NVIDIA GeForce FTX 1080 GTX GPU). Stimuli were presented using a custom MATLAB script (MATLAB and Statistics Toolbox Release 2015b, The MathWorks, Inc., Natick, Massachusetts, United States) that called VideoLAN Client (VLC; VideoLAN Client, 2017; http://www.videolan.org/) to play the videos. EEG data were collected using BrainVision Recorder (Brain Products, GmbH, 2017) software on an Intel Xeon E5-1650 PC (3.5 GHz, 32GB RAM) running Windows 7. Video onsets were synchronized to EEG data using the Schultz Cigarette Burn Toolbox (Schultz, Biau, & Kotz, 2020).

Procedure

Participants were seated approximately 60 cm apart from a monitor in a sound attenuated booth while videos were displayed on a computer screen. Participants watched 234 videos organised in nine blocks of 26 randomised trials (i.e., 6 stimuli per condition + 2 fillers). The task was a two-alternative forced choice synchrony detection task (Figure 1B). Participants attended both the audio and video stimuli. Each trial began with a central white fixation cross (jittered duration 500 +/- 250 ms) followed by the stimulus. After the video ended, participants decided whether the audio and the video signals were synchronous or asynchronous by pressing the “1” or “2” key on the keyboard without time pressure (counterbalanced across participants). Additionally, participants were asked to count internally the number of times they observed a red cross in a video clip and reported it at the end of the experiment. This secondary task ensured that the participants carefully attended both visual and auditory information during the experiment. Further, we chose a relatively easy task, ensuring that performance in the audio-visual synchrony detection task was not affected. Filler trials were not included in behavioural and EEG analyses but the total number of reported red crosses served to check that attention was maintained throughout the experiment. Before the experiment, participants received five practice trials where they were presented with one example of each condition to ensure they understood the instructions. At the end of the experiment, participants were asked if they could identify the speaker’s language and to report it.

[Insert Figure 1]

Figure 1. Experimental procedure of the audio-visual asynchrony detection task. (A) The four experimental conditions. For each item, the audio signal was the same across all four versions. Visual information was manipulated by the presence or absence of a mask (no-mask or head-mask). Video and sound were either temporally aligned in the synchronous conditions (NMS, HMS), or temporally misaligned by 400ms in the asynchronous conditions (NMA, HMA). (B) Example of one trial timeline. (C) Distribution of the electrodes covering the motor region of interest (ROI; blue circles) and the control region of non-interest in the visual area (RONI; red circles).

EEG recording and pre-processing

Electrophysiological data were recorded at 1000 Hz with 128 active electrodes (ActiCap, Brain Vision Recorder, Brain Products) according to the 10-20 international standard, and impedances were kept below 10 kΩ. The ground electrode was located at AFz, and the reference electrode was placed at the right mastoid (TP10).

Offline EEG pre-processing: EEG data were pre-processed offline using Fieldtrip (Oostenveld et al., 2011) and SPM8 toolboxes (Wellcome Trust Centre for Neuroimaging). Continuous EEG signals were bandpass filtered (standard non-causal two-pass Butterworth filters) between 0.1 Hz and 100 Hz and bandstop filtered (48-52 Hz and 98-102 Hz) to remove line noise at 50 and 100 Hz. Data were epoched from 1000 ms before stimulus onset to 11000 ms after stimulus onset. Trials and channels with artefacts were excluded by visual inspection before applying an independent component analysis (ICA) to remove components related to ocular artefacts. Excluded channels were then interpolated using the method of triangulation of nearest. After re-referencing the data to an average reference, the remaining trials with artefacts were manually rejected by a final visual inspection (on average, 13.57 ± 8.32 trials across conditions per participant).

EEG data analyses at the scalp level

Time-frequency analysis was applied to each electrode using a Morlet wavelet (width: 5 cycles, from 1 to 40 Hz with 1 Hz step and 20 ms time steps) and frequency analyses were performed for each trial prior to averaging across trials in the four conditions. The power was normalised relative to a pre-stimulus baseline (-700 to -200 ms with respect to stimulus onset) to determine increases or decreases of power dependent on the conditions. The peak frequency analysis applied to the video and audio signals of the audio-visual clips revealed that prosodic features conveyed activity mainly between 2 to 3 Hz (see Table 1), which determined our frequency band of interest. In the present study, oscillatory delta activity was assessed by means of power information, that is by taking the average power across the 2-to-3 Hz frequencies (power and peak frequency values) and investigating its modulations as a function of the audio-visual speech analysis. Further, as the spontaneous speech signal is not isochronous, phase information likely would be too noisy to extract meaningful information. This is the reason why we did not investigate phase modulations here. As entrainment necessitates several cycles from recurrent stimulations to build up (Doelling et al. 2014; Thut et al., 2011; Zoefel et al. 2018) and the slower frequency in our band of interest was 2Hz (corresponding to a period of 500 ms), we defined a time window of interest from + 3 to + 9 seconds after stimulus onset. This time window ensured that neural activity sufficiently entrained to the temporal structure of the stimuli, and that the responses evoked by the stimulus onset-offsets did not influence the results. In the identified regions of interest and non-interest (see Results section), normalised mean power across pool electrodes in the 2-3 Hz frequency band was computed for the four conditions and exported for further statistical analyses.

EEG data analyses at the source level

Source localisation: We used the Montreal Neurological Institute (MNI) MRI template and a template volume conduction model from Fieldtrip. The 128 electrode positions on the volunteer's head were defined by using a Polhemus FASTRAK device (Colchester), recorded with the Brainstorm toolbox implemented in MATLAB (Tadel et al., 2011), and realigned to the template head model using Fieldtrip. The template volume conduction model and the electrode template were used to prepare the source models. Leadfields were computed based on scalp potentials and source activity was reconstructed applying a linearly constrained minimum variance (LCMV) beamforming approach implemented in Fieldtrip (van Veen et al., 1997; Wang et al., 2018). Source analyses were run on potential data (i.e., average referenced) and time-series data were reconstructed in 2020 virtual electrodes for each participant. Time-frequency analysis was computed at each of 2020 virtual sources with the exact same approach to scalp level analyses. The maximum voxel activation regions were defined by using the automated anatomical labelling atlas (AAL).

Phase-amplitude coupling (PAC) between delta and beta oscillations: We applied a modulation index (MI; Tort et al., 2010) analysis in the time-window of interest to quantify delta-beta PAC in the significant cluster revealed by source localisation in the contrast NMA- NMS (i.e. difference of delta power in the NMA condition minus NMS condition). Firstly, the power spectrum (1 - 30 Hz) was estimated across all grids of the significant cluster and trials by applying a 1/f correction time-frequency decomposition method with wavelet for each participant (Griffith et al., 2019). Fractal activity was attenuated by subtracting the linear fit of $1/f$ characteristic from the data to isolate oscillatory components before extracting the power peaks. For each epoch, the spectral power was first calculated by applying a constant time-frequency decomposition method with 5-cycles wavelet across all frequencies (from 1 to 30 Hz). This ensured that a single slope was computed and subsequently subtracted from the signal. This step generated two vectors: one vector contained the values of each wavelet frequency A, while the other vector contained the power spectrum for each electrode-sample pair B. Both vectors were then put into log-space to provide a linear line in order to get the slope and intercept of the $1/f$ curve. The linear equation Ax = B was resolved using least-squares regression, where x is an unknown constant describing the curvature of the $1/f$ characteristic. The $1/f$ fit Ax was then subtracted from the log-transformed power spectrum B. Peaks of $1/f$–corrected absolute power were then identified in the delta (1-3 Hz) and beta (20-30 Hz) bands of interest for each trial. The most prominent power spectrum peaks in the delta and beta bands were then extracted and saved as the individual delta and beta peaks. Across participants, the mean delta peak was at 2.1 Hz and the mean beta peak was at 24.16 Hz. To obtain an equal number of correct and incorrect trials across conditions, the same number of trials between all conditions was determined by taking 80% of the smallest number of available trials across all the conditions ($\text{NMS}_{\text{correct}}$, $\text{NMS}_{\text{incorrect}}$, $\text{NMA}_{\text{correct}}$, $\text{NMA}_{\text{incorrect}}$, $\text{HMS}_{\text{correct}}$, $\text{HMS}_{\text{incorrect}}$, $\text{HMA}_{\text{correct}}$ and $\text{HMA}_{\text{incorrect}}$; average minimum number of trials: 6.61 ± 3.37). The 80% subsampling was done to ensure that some participants were not overrepresented in the resampling procedure due to using 100% of their available data, as well as to vary the set of trials in the condition determining the minimum number of trials across iterations (Keitel et al., 2018). Subsampled trials were concatenated, and the operation was repeated for 50 iterations in each condition to provide enough random trials to compute the PAC (i.e., 50 trials per condition per participant). The grids of interest were identified during source localisation (see Results section) and correspond to the grids at which the difference of 2-3 Hz delta power between the condition NMA minus NMS was significant (i.e., contrast NMA-NMS; number of significant grids = 92). Second, the time-series of each grid source of the left motor cluster were duplicated and filtered separately: the first time-series was filtered around the theta peak (± 0.5 Hz) and the second time-series was filtered around the beta peak (± 5 Hz). Third, the Hilbert transform was applied to the delta and beta filtered time-series to extract the phase of the former and the power of the latter. Fourth, beta power was binned into 12 equidistant bins of 30° according to the delta phase. The binning was computed for each trial and grid source separately. The MI was computed by comparing the observed distribution to a uniform distribution for each trial and grid. The PAC was then averaged across the left motor grids and 50 iterations in each condition. Finally, we investigated whether the delta-beta coupling was specifically localised in the region of interest, identified by the source localisation analysis (i.e., left motor area), or extend to further regions in the brain. We compared the delta-beta PAC between masks in a whole brain PAC analysis (no-mask and head-mask, correct trials only as the ROI analysis did not established a relationship between PAC and behavioural performance). The difference of trial numbers between conditions was balanced by taking 80% of the smallest sample of available correct trials between all the four conditions ($\text{NMS}_{\text{correct}}$, $\text{NMA}_{\text{correct}}$, $\text{HMS}_{\text{correct}}$ and $\text{HMA}_{\text{correct}}$; average minimum number of trials: 18.78 ± 5.77). Subsampled trials were concatenated, and the operation was repeated for 40 iterations (to circumvent computational resource limits reached by concatenated epoch lengths). The delta-beta PAC was then averaged across all iterations at each grid (n = 2020) and conditions across participants.

Experimental design and statistical analysis

Audio-visual asynchrony detection task: The experiment used a full within-subject design. The effect of asynchrony and its interaction with the head-mask in audio-visual speech perception was assessed by means of the d' sensitivity index (Macmillan & Kaplan, 1985). To calculate the d' index, the hit trials (i.e., "yes" responses in synchronous conditions NMS and HMS) and false alarm trials (FA, i.e., "yes" responses in the asynchronous conditions NMA and HMA) were computed for each participant. The d' scores for asynchrony detection in the no-mask and head-mask conditions were calculated for each participant as follows: d' = Z (Hit_{rate}) – Z (FA_{rate}). The *d'* index allows considering response bias by comparing hits and false alarms to assess whether participants actually discriminated synchrony and asynchrony. Additionally, the decision criterion *c* was computed as follows: c = 0.5 x (Hit_{rate} – FA_{rate}) / 2 to determine the decision shift between no-mask and head-mask conditions. Further, the mean correct response rates were computed for each participant (hit and correct rejection trials, respectively from the synchronous and asynchronous conditions). Finally, the mean reaction times of the correct trials were computed for each participant (hit and correct rejection trials; comprised between mean reaction times ± two standard deviations range). The effects of masking the speaker's face and audio-visual asynchrony on correct response rates and reaction times were assessed using two-way repeated-measure ANOVAs with the factors mask (no-mask, head-mask), asynchrony (synchronous, asynchronous), and the interaction between mask and asynchrony, using SPSS (IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.). In the case of significant interactions, *post-hoc t*-tests were Bonferroni-corrected. To test whether participants' sensitivity to asynchrony was dependent on information conveyed by head and facial movements, the *d'* and *c* criterion in the no-mask and head-mask conditions were individually tested against zero by means of one-sample t-tests. Further, the difference of *d'* between the no-mask and head-mask conditions was assessed applying a paired-samples *t*-test and the effect size was defined using Cohen's *d*.

EEG data at the scalp level: EEG data of correct trials at the scalp level were statistically analysed (NMA$_{correct}$, NMS$_{correct}$, HMA$_{correct}$ and HMS$_{correct}$). We first tested if delta power responses were modulated and dependent on the participants' sensitivity to audio-visual asynchrony in multisensory speech perception. The differences of mean power between the two contrasts NMA-NMS and HMA-HMS (NMA-NMS: difference of power NMA minus NMS; HMA-HMS: difference of power HMA minus HMS) at the electrode level were statistically assessed by applying dependent *t*-tests using Monte-Carlo cluster-based permutation tests (Maris & Oostenveld, 2007) with an alpha cluster-forming threshold set at 0.05, three minimum neighbour channels, 5000 iterations, and cluster selection based on maximum size. Cluster-based permutation statistics were applied for the time window of interest in the delta 2-3 Hz band across all the electrodes. Further, to test whether changes in fronto-central delta oscillations reflect the temporal analysis of multisensory speech rather than purely sensory-driven response activity, we performed the same tests on the theta band (4 - 8 Hz), which tracks the syllabic structure of speech (Giraud & Poeppel, 2012). We expected to find modulations of delta but not theta oscillations for audio-visual asynchrony in the region of interest if motor delta responses reflect temporal analysis. In the identified regions of interest and non-interest (see Results section), normalised mean power across pooled electrodes in the 2-3 Hz delta and 4-8 Hz theta frequency bands was computed for the four conditions and exported. This step allowed confirming that delta oscillations responded to audio-visual speech perception independently from the conditions, with an increase of power as compared to the pre-onset baseline (i.e., positive values meaning an increase of power, while negative values meaning a decrease of power in audio-visual speech perception). Statistical differences of power in relevant contrasts were assessed by means of two-way repeated-measure ANOVAs.

EEG data source localisation: We tested whether delta power responses at the source level depend on the participants' sensitivity to audio-visual asynchrony in multisensory speech perception. Differences in delta power for the two contrasts NMA-NMS (difference of power NMA minus NMS) and HMA-HMS (difference of power HMA minus HMS) were assessed by applying dependent *t*-tests using Monte-Carlo cluster-based permutation tests at source level as performed for the scalp level analysis. For visualisation of the source localisation results, the power differences in the two contrasts were grand averaged across participants, and the grand average power differences were interpolated to the MNI MRI template for visualization. Only voxels surpassing the statistical significance threshold are depicted in both contrasts (significant *t*-values at alpha = 0.05, multiple comparison cluster-corrected).

Delta-beta PAC: Cross-frequency analyses were performed to investigate whether left motor delta-beta PAC is associated with the successful detection of audio-visual speech asynchrony, dependent on whether listeners were able to match visual and auditory prosodies (no-mask conditions) or not (head-mask conditions). First, statistical differences of mean PAC across conditions in the region of interest were assessed applying a three-way repeated-measure ANOVA with the factors mask (no-mask and head-mask), asynchrony (synchronous and asynchronous), and correctness (correct and incorrect trials). Second, statistical differences of whole brain delta-beta PAC were assessed by applying dependent *t*-tests using Monte-Carlo cluster-based permutation tests as described above (whereas *t*-tests were one-tailed here as we had a strong hypothesis about delta-beta PAC modulation directionality based on results at region of interest level).

Correlations between performance in synchrony detection and delta oscillations in the identified left motor cluster: We examined the relationship between neural activity and sensitivity to audio-visual asynchrony in multisensory stimuli. By means of Pearson correlations, we tested whether the difference of delta power in the left motor cortex [Δ_{power} = delta $\text{power}_{\text{asynchronous}}$ - delta $\text{power}_{\text{synchronous}}$] predicted differences in correct responses [Δ_{CR} = $\text{CR}_{\text{asynchronous}}$ - $\text{CR}_{\text{synchronous}}$] in the no-mask and head-mask conditions. The purpose of this analysis was to link the increase of left motor delta power in response to audio-visual asynchrony and the participants' sensitivity to the temporal analysis of multisensory speech. A positive correlation between the two variables would establish that an increase of delta power predicts improved asynchrony detection when audio-visual stimuli are asynchronous. Increased sensitivity to asynchrony corresponding with increased delta power would support our hypothesis on the role of left motor cortex in the temporal analysis of audio-visual speech. For each participant, we computed the 2-3 Hz power at the grids sources from the significant cluster established in the NMA-NMS contrast source analysis (i.e., significant grids situated in the left central and frontal gyrus areas of interest). Power was averaged across grids in the four conditions separately ($NMS_{correct}$, $NMA_{correct}$, $HMS_{correct}$ and $HMA_{correct}$), and we calculated the mean difference (Δ_{Power}) separately in the no-mask (NMA-NMS) and head-mask (HMA-HMS) contrasts to obtain two delta power values per participant. Similarly, the difference of correct response rates (Δ_{CR}) was calculated in the no-mask and head-mask contrasts, resulting in two behaviour values per participant. The statistical relationship between behaviour (Δ_{CR}) and delta power (Δ_{Power}) was assessed applying Pearson correlation tests.

Difference of delta power between correct and incorrect trials across conditions: We tested whether correctness (correct vs incorrect trials) predicted delta power differences in the left motor cortex across conditions (NMA, NMS, HMA and HMS). To circumvent the unbalanced number of trials between correct and incorrect trials within conditions (which was expected according to our experimental procedure targeting ~75-85% of accuracy), we performed permutations tests on the difference of delta power $trials_{correct}$ - $trials_{incorrect}$ between the original data and 5000 permuted data as follows: First, delta power (2-3Hz) in the time-window of interest was computed at source level for all trials and conditions ($NMS_{correct}$, $NMS_{incorrect}$, $NMA_{correct}$, $NMA_{incorrect}$, $HMS_{correct}$, $HMS_{incorrect}$ $HMA_{correct}$ and $HMA_{incorrect}$). Second, correct and incorrect labels were randomly shuffled across trials in each condition. Third, for each iteration two equal samples of shuffled correct and incorrect trials were generated by taking the smallest number of available trials in each condition (i.e. between the original number of correct and incorrect trials). Fourth, the mean delta power from the left motor cluster identified in the source localisation step was computed separately for the shuffled correct and incorrect trials in each condition. Then, the mean difference of delta power $trials_{correct}$ - $trials_{incorrect}$ was computed for each iteration in the NMA, NMS, HMA and HMS conditions. Fifth, in each condition a one-sample t-test against zero (two-tailed) was performed on the difference of delta power $trials_{correct}$ - $trials_{incorrect}$ from the original data to determine the original effect size ($t\text{-value}_{original}$), as well as from every permutated data set (i.e. 5000 $t\text{-values}_{permut}$). Finally, the 5000 t-values from the t-tests were ranked and the *p*-value in each condition was calculated as

p = [(number of absolute t-values$_{\text{permut}}$ +1) > (absolute t-value$_{\text{original}}$ +1)]/(number of permutations +1).

Distance between delta peak frequencies in the stimulus and delta peak frequencies induced in the left motor cortex: We wanted to confirm that modulation of delta oscillations in left motor cortex does not reflect mere stimulation frequencies, i.e., purely sensory entrainment, but reflects process-driven temporal analysis of visual and auditory prosodies: The rationale was that in the former case, one would assume that tracking the dominant stimulus oscillation would entrain neural delta responses in the exact same frequency. In the latter case, an increase in neural delta activity would reflect the temporal analysis of sensory-specific oscillations independent of their respective frequencies. If true, there should be no direct mapping of the frequency of the delta power maxima in left motor cortex, and the frequency of dominant delta activity conveyed by the multisensory stimuli (see Table 1). To test this assumption, we probed the absolute distance (i.e., absolute difference) between the distribution of peak frequencies of the delta power induced in stimulus perception and the delta peak signal frequencies in the corresponding video clips (Full, Head only, Body only and Audio only signals; see Table 1). Firstly, we determined the individual delta peak (1-3 Hz) of each participant in every trial (correct trials only: NMS$_{\text{correct}}$, NMA$_{\text{correct}}$, HMS$_{\text{correct}}$ and HMA$_{\text{correct}}$) as described previously (see methodology in the previous phase-amplitude coupling section). Secondly, for each trial we calculated the absolute distance between the delta peak frequency of the neural power and every signal of the stimulus presented in the corresponding trial. This step resulted in four absolute distance scores per trial per participant. We averaged the absolute distance scores across participants for each stimulus. Finally, absolute difference scores were sorted by conditions (NMS, NMA, HMS and HMA), and stimulus signals (Full, Head only, Body only and Audio only). To assess statistically the distance between the peak frequencies of delta power and stimuli, we tested the mean of each score distribution against zero with a one-sample t-test (one-tailed). P-values were corrected for multiple comparisons by applying a Bonferroni correction (α = 0.05/total number of comparisons). A one-way repeated-measures ANOVA assessed the statistical difference between the multiple cases of absolute difference (16 in total = two masks x two synchronies x four signals). Similarly, we tested the consistency of the delta power frequency maxima in left motor cortex across all trials. This should confirm that any observed variations in delta activity reflect a difference in amplitude modulation on the power of the same delta activity rather than different oscillations across conditions. We computed the delta peak frequency of the neural power of every participant for each stimulus in all four conditions ($NMS_{correct}$, $NMA_{correct}$, $HMS_{correct}$ and $HMA_{correct}$). To statistically assess the consistency of activity across all trials and independent of all conditions, we averaged the EEG delta peak frequency across participants for each stimulus in all four conditions separately (i.e., 54 scores per condition). We then applied a two-way repeated-measure ANOVA with the factors mask (no-mask and head-mask) and synchrony (synchronous and asynchronous).

RESULTS

Participants reported 18.26 ± SD = 1.51 red crosses (out of 18) at the end of the experiment. Additionally, they correctly identified the speaker's native language (they all responded "German"), although they could not report any semantic content. These results confirmed that participants correctly paid attention to both the audio and video signals.

Listeners successfully temporally analysed visual and auditory prosodic features to denote audio-visual asynchrony in multisensory speech perception.

[Insert Figure 2]

Figure 2. Behavioural performances in the asynchrony detection task. (A) Average *d'* scores and correct response rates (± standard error of the mean; grey dots represent individual averages; n = 23). (B) Reaction times of correct responses across conditions (± standard error of the mean; grey dots represent individual averages). Significant contrasts are marked by stars ($p < 0.05$).

D' scores are reported in Figure 2A (left panel). To test whether participants perceived audio-visual asynchrony in both the no-mask and head-mask conditions, we preformed two independent one-sample *t*-tests. Results showed that the mean *d'* was significantly greater than zero in the no-mask and head-mask conditions, confirming that participants were sensitive to audio-visual asynchrony in both cases (no-mask: $t(1,22) = 10.25$; $p < 0.001$; *Cohen's d* = 3.04; head-mask: $t(1,22) = 8.07$; $p < 0.001$; *Cohen's d* = 2.38). A paired-samples *t*-test comparing the *d'* between the no-mask and the head-mask conditions tested the hypothesis that participants detected asynchrony better in the no-mask conditions. Results confirmed that it was indeed the case ($t(1,22) = 6.96$; $p = < 0.001$, two-tailed; *Cohen's d* = 1.46). To assess whether participants tended to respond "synchrony" more often (i.e., a liberal response bias) independently from their actual sensitivity to audio-visual synchrony, and whether this response bias differed when a head-mask was present, we performed two independent one-sample t-tests on the mean criterion c in the no-mask and head-mask conditions. Results revealed that the mean *c* criterion was significantly more negative in the head-mask conditions (-0.53 ± 0.28; $t(1,22) = -9.01$; $p < 0.001$; *Cohen's d* = 2.68) but not different from zero in the no-mask conditions (0.11 ± 0.40; $t(1,22) = 1.32$; $p = 0.1$; *Cohen's d* = 0.39). This confirmed that when the speaker's face was head-masked, participants were significantly more biased toward responding "synchrony" than in the no-mask conditions (i.e., a liberal response bias).

The mean correct response rates across conditions are depicted in Figure 2A (right panel). NMS: 0.78 ± 0.09; NMA: 0.80 ± 0.13; HMS: 0.82 ± 0.11; HMA: 0.48 ± 0.14. To test whether the presence of the head-mask affected participants' perception of audio-visual asynchrony, we performed a two-way repeated-measure ANOVA with the main factors mask and asynchrony on accuracy. Results confirmed a significant interaction between the mask and asynchrony ($F(1,22) = 82.04$; p < 0.001; $\eta_p^2 = 0.789$). Bonferroni-corrected pairwise comparisons showed that performance decreased significantly only in the asynchronous condition of the head-mask conditions (HMA) but not in the three other conditions (NMS, NMA and HMS; no significant difference between them). These results show that the synchrony between visual and auditory stimulus information predicted participants' performance differently, dependent on the presence or absence of the head-mask. The test also revealed a significant main effect of mask ($F(1, 22) = 115.22$, $p < 0.001$; $\eta_p^2 = 0.84$) and asynchrony ($F(1, 22) = 34.52$, $p < 0.001$; $\eta_p^2 = 0.61$) for correct response rates.

Reaction times across conditions are reported in Figure 2B. Similarly, a two-way repeated-measure ANOVA with the main factors mask and asynchrony was performed on the reaction times. Results revealed a significant main effect of mask on reaction times ($F(1, 22) = 16.50$, $p < 0.01$; $\eta_p^2 = 0.43$). No significant effect of asynchrony ($F(1, 22) = 0.67$, $p = 0.42$; $\eta_p^2 = 0.03$) or an interaction between mask and asynchrony was found ($F(1, 22) = 2.32$, $p = 0.14$; $\eta_p^2 = 0.1$). These results show that accurate responses were faster when the face of the speaker was not masked compared to head masked.

Together, the behavioural results support our hypothesis that participants can successfully temporally analyse slow auditory and visual prosodic features in an audio-visual asynchrony detection task. This sensitivity to audio-visual (a)synchrony was altered by the amount of available visual information: On the one hand, the temporal analysis of visual and auditory prosodic information did not change the participants' sensitivity to audio-visual asynchrony in the no-mask conditions. Therefore, the no-mask conditions represent a case of successful temporal analysis in audio-visual speech perception. On the other hand, participants were both slower and less accurate in detecting audio-visual asynchrony in the head-mask conditions, which represents the case of less successful temporal analysis of audio-visual speech perception. Response accuracy in HMS did not differ from the no-mask conditions (although participants were slower in responding correctly), whereas it decreased to chance-level in the asynchronous head-mask condition (HMA). Consequently, the visual mask affected participants' sensitivity to audio-visual asynchrony in both HMS and HMA conditions to a different degree, likely due to the delay between visual and auditory stimulus onsets.

Delta oscillations in the left motor cortex denote asynchrony between the visual and auditory prosodies in multisensory speech perception.

We then addressed whether delta oscillations in the left motor cortex relate to the temporal analysis of multisensory information, and whether responses depend on the amount of visual information available. First, a cluster-based permutation tests revealed a significant increase in delta power (2-3 Hz) in response to the audio-visual asynchrony when the speaker's face was visible (no-mask: NMA-NMS) but not when it was masked (head-mask: HMA-HMS) (NMA-NMS: $p < 0.001$, cluster statistic = 117.23; HMA-HMS: No positive cluster; multiple comparisons are cluster-corrected). No significant negative clusters were found in both contrasts. Importantly, the topography of the significant delta cluster in the no-mask contrast showed a main fronto-central response when video and audio signals were asynchronous, in line with the expected source localization of delta in the motor region (Figure 3B; Puzzo et al., 2010; Stegemöller et al., 2017). To assess the potential interaction of visual information and audio-visual asynchrony detection in this motor region of interest, we defined a set of electrodes as the region of interest (ROI) representative of the delta response topography: F1, Fz, F2, FFC3h, FFC1h, FFC2h, FFC4h, FC3, FC1, FCz, FC2, FC4, FCC3h, FCC1h, FCC2h, fCC4h, C1, Cz and C2 (Figure 1C). The mean delta power across the electrodes of the ROI was computed separately in the four conditions and confirmed an increase of induced delta activity compared to the pre-stimulus baseline (NMS: 0.64 ± 0.17; NMA: 0.74 ± 0.15; HMS: 0.70 ± 0.16 and HMA: 0.68 ± 0.20; see Figure 3A and 3C). A two-way repeated-measure ANOVA revealed a significant interaction between the factors mask and asynchrony for delta power ($F(1, 22) = 5.78$, $p = 0.03$; $\eta_p^2 = 0.21$). Bonferroni-corrected pairwise comparisons showed that in the no-mask contrast, delta power was significantly greater in the asynchronous (NMA) than synchronous (NMS) condition ($p = 0.02$), whereas asynchrony did not affect delta power responses in the head mask contrast ($p > 0.5$). No further pairwise comparison was significant in the post hoc tests. The significant interaction established that the detection of temporal (a)synchrony of visual and auditory information modulated increases in delta power differently and dependent on the availability of visual information (i.e., no-mask versus head-mask).

[Insert Figure 3]

Figure 3. Delta responses to audio-visual asynchrony at the scalp level. (A) Time-frequency spectra of the mean power differences in the motor ROI between asynchronous and synchronous conditions in the no-mask (NMA-NMS; left) and head-mask (HMA-HMS; right) contrasts. The white dashed lines correspond to the onset of the video and the window of interest is marked by the pink dashed rectangles. (B) Topographical distribution of the difference of 2-3 Hz delta power in the time-window of interest, in the no-mask (NMA-NMS; top) and head-mask (HMA-HMS; bottom) contrasts. The pink dots display electrodes with

significant *t*-values (alpha threshold = 0.05). (C) Delta power across the electrodes of interest in the four conditions (2-3 Hz band). Significant contrasts are marked by stars ($p < 0.05$).

Secondly, to separate the influence of audio-visual speech (a)synchrony perception from sensory processing, delta responses were also examined in a control visual region of non-interest (RONI; Figure 1C). The region of non-interest was located in the occipital cortex where we did not expect higher audio-visual speech analysis to take place as visual information was identical between synchronous and asynchronous conditions within mask contrasts (RONI electrodes: PPO1h, PPO2h, PO3, POz, PO4, POO1, POO2, POO9h, O1, Oz, O2, POO10h, Ol1h, Ol2h, O9 and O10). We compared the effect of audio-visual asynchrony between the identified motor region (ROI) and the visual sensory area (RONI) to confirm that delta response modulations did not reflect signal processing only (Figure 4A). The mean differences of 2-3Hz delta power (NMA-NMS and HMA-HMS) were computed in the regions of interest and non-interest at the same time-window (Figure 4B; ROI: NMA-NMS = 0.1 ± 0.09; HMA-HMS = -0.03± 0.19; RONI: NMA-NMS = 0.05 ± 0.10; HMA-HMS = 0.01 ± 0.24). A two-way repeated-measures ANOVA with the mean factors region (ROI or RONI) and mask (no-mask or head-mask) was performed to assess whether the responses of delta oscillations to asynchrony reflected multisensory speech analysis or purely signal processing taking place in sensory areas (i.e., visual occipital areas). Results revealed a significant interaction between region and mask ($F(1, 22) = 5.75$, $p = 0.025$; $\eta_p^2 = 0.21$). First, Bonferroni-corrected pairwise comparisons showed that in the no-mask contrast the delta power difference NMA-NMS (but not HMA-HMS) was significantly greater in the region of interest than in the region of non-interest (respectively p = 0.025 and p = 0.572). Only in the region of interest the difference of power NMA-NMS was significantly greater than HMA-HMS (respectively $p = 0.019$ and $p = 0.113$). No main effect of mask ($F(1, 22) = 0.25$, $p = 0.622$; $\eta_p^2 = 0.21$) or region ($F(1, 22) = 2.18$, $p = 0.154$; $\eta_p^2 = 0.09$) was found.

[Insert Figure 4]

Figure 4. Comparisons between the motor region of interest (ROI) and the visual region of non-interest (RONI). (A) TFRs of the difference of spectrum in the no-mask contrast (NMA-NMS) in the ROI and RONI. (B) The mean differences of 2-3Hz delta power (NMA-NMS and HMA-HMS) were computed in the regions of interest and non-interest. Significant contrasts are marked by stars ($p < 0.05$).

Thirdly, the mean power in the 4 - 8 Hz band was computed in the four conditions separately from the ROI electrodes and confirmed an increase of theta activity compared to the pre-stimulus onset baseline (NMS: 0.86 ± 0.25; NMA: 0.85 ± 0.18; HMS: 0.83 ± 0.16 and HMA: 0.81 ± 0.24). A two-way repeated-measure ANOVA revealed no significant main effect of mask (F(1, 22) = 2.77, p = 0.11; η_p^2 = 0.11), asynchrony (F(1, 22) = 0.27, p = 0.606; η_p^2 = 0.01) or interaction between the factors mask and asynchrony (F(1, 22) = 0.05, p = 0.825; $\eta_p^2 < 0.01$) on theta power in the region of interest. Further, the cluster-based permutation tests revealed no significant modulation of theta power by audio-visual asynchrony in any of the mask contrasts (NMA-NMS: no significant cluster; HMA-HMS: no significant cluster; multiple comparisons are cluster-corrected). These results confirmed that audio-visual asynchrony detection modulated delta power over the expected fronto-central region. Further, the delta power response was attenuated when listeners were less able to integrate visual and auditory prosodies (i.e., in the head-mask as compared to the no-mask conditions). This result suggests that increased delta activity in left motor cortex plays a role in audio-visual asynchrony detection as it only increased in asynchronous but not synchronous multisensory speech perception. Therefore, delta activity might be associated with the brain’s effort to resolve mismatches between visual and auditory prosodies in the temporal analysis of multisensory speech.

Next, we analysed the source localisation of the delta power modulations observed when video and audio signals were presented in asynchrony in both no-mask and head-mask contrasts. Cluster-based permutation t-tests between synchronous and asynchronous conditions at the source level revealed that asynchrony significantly increased delta oscillation responses when the head of the speaker was visible (NMA-NMS: p = 0.042; cluster statistic = 233.02) but not when it was head-masked (HMA-HMS: p = 0.27; cluster statistic = 38.27). The projections of the significant t-values on the brain's surface showed an increase of delta power originating mainly in the left precentral region and the left inferior frontal gyrus (Figure 5A). The source results support the topographies of the delta power modulations observed at the scalp level, which revealed fronto-central differences in the no-mask contrast only (Figure 3B). Similar to the scalp level analysis, we computed the mean 2-3Hz power across the significant grids in all four conditions in the time-window of interest. Power was normalised relative to the pre-stimulus baseline to determine an increase of delta power during stimulus presentation in all four conditions. Four one-sample *t*-tests against zero confirmed a significant increase of delta power in response to audio-visual speech perception in all four conditions (respectively NMS: 0.69 ± 0.15; $t(1,22) = 22.25$; $p < 0.001$; *Cohen's d* = 4.64; NMA: 0.76 ± 0.17; $t(1,22) = 21.81$; $p < 0.001$; *Cohen's d* = 4.55; HMS: 0.70 ± 0.17; $t(1,22) = 19.84$; $p < 0.001$; *Cohen's d* = 4.14 and HMA: 0.69 ± 0.20; $t(1,22) = 16.74$; $p < 0.001$; *Cohen's d* = 3.49). We then performed a two-way repeated-measure ANOVA with the main factors mask and asynchrony on mean power as in the scalp level analysis, but the test did not reveal any significant effects (mask: $F(1, 22) = 1.42$, $p = 0.25$; $\eta_p^2 = 0.061$; asynchrony: $F(1, 22) = 1.75$, $p = 0.20$; $\eta_p^2 = 0.074$; mask*asynchrony: $F(1, 22) = 1.94$, $p = 0.18$; $\eta_p^2 = 0.081$). Further, we tested whether the modulation of delta responses in the left motor areas by audio-visual asynchrony predicted detection performance in the no-mask and head-mask conditions (Figure 5B). Pearson correlations revealed a positive correlation between the correct response rate differences (Δ_{CR}) and delta power differences (Δ_{Power}) in the no-mask contrast (NMA-NMS: $r = 0.36$; $p = 0.046$, one-tailed) but not in the head-mask contrast HMA minus HMS (HMA-HMS: $r = 0.04$; $p = 0.43$, one-tailed). These results confirmed that when participants perceived asynchrony between video and audio signals (no-mask conditions), the difference in delta power between asynchronous and synchronous conditions predicted detection accuracy. This was not the case when participants were less able to detect temporal alignment between visual and auditory information (head-mask conditions).

[Insert Figure 5]

Figure 5. Delta oscillation responses to audio-visual asynchrony at the source level for no-mask and head-mask contrasts. (A) Contrast NMA – NMS projected onto the brain's surface (significance *t*-values; cluster-corrected at alpha threshold = 0.05). The maximum voxel MNI coordinates is located left precentrally [-50 19 40] but significant activation was also found in

the left inferior frontal gyrus (pars triangularis; maximum voxel MNI coordinates [-30 31 0]). No significant difference was found when the head of the speaker was masked (HMA – HMS contrast; not represented). (B) Scatterplots of audio-visual asynchrony detection performance and delta power in the significant cluster region (left motor cortex). The difference of delta power in the left motor cluster (Δ_{Power}; x-axis; z scores) correlated with the difference of audio-visual asynchrony detection (Δ_{CR}; y-axis; z scores) between asynchronous and synchronous conditions only when the face of the speaker was visible, and participants could integrate video and audio onsets (no-mask conditions). (C) Average delta power differences between correct and incorrect trials from the significant left motor cluster in the four conditions NMS, NMA, HMS and HMA (± standard error of the mean; grey dots represent individual averages; $n = 23$; outliers not represented). Significant differences from zero are marked by stars ($p < 0.05$). (D) Left panel: Peak frequency correspondence between delta activity carried in the video clips and delta power responses induced in the left motor cluster. The bars represent the mean absolute distance between the delta peak frequencies in the stimulus and the peak frequencies of neural delta power induced during the corresponding trial (± standard error of the mean). Peak frequency matching was assessed for the synchronous and asynchronous conditions in the two mask conditions (no-mask and head-mask), and the different signal types of each stimulus: Full, Head only, Body only and Audio only (see Table 1). The mean of the absolute difference scores were significantly greater than zero in all conditions and for all the stimulus signals. (D) Right panel: Consistency of delta peaks across the ordered stimuli in all four conditions. The upper panel displays the mean delta peak frequencies in the left motor cortex across all participants (± standard error of the mean) for each stimulus in the no-mask conditions (black squares: NMS; orange squares: NMA). The lower panel displays the mean EEG delta peak frequencies across all participants (± standard error of the mean) for each stimulus in the head-mask conditions (black squares: HMS; orange squares: HMA). The variations in delta responses observed across all conditions reflect a difference of power amplitude modulation on the same oscillatory activity.

We tested whether correctness predicted delta power response modulations in the significant cluster identified in the left motor cluster (Figure 5C). Permutation tests revealed that delta power from the left motor cortex was significantly greater in the correct trials as compared to incorrect trials in the asynchronous conditions no-mask (NMA: *Cohen's d* = 0.552; p = 0.002) and head-mask (HMA: *Cohen's d* = 0.401; p = 0.011). In contrast, no significant difference of delta power between correct and incorrect trials were found in the synchronous conditions (NMS: *Cohen's d* = 0.28; p = 0.429; HMS: *Cohen's d* = 0.232; p = 0.332). These results showed that increases of delta power in the left motor cortex predicted sensitivity to audio-visual alignment in the asynchronous conditions but not in the synchronous conditions. Further, we aimed to control that delta responses induced in the left motor cortex during audio-visual speech perception did not reflect purely stimulus driven entrainment to the delta activity carried in the visual or auditory signals of the video clips (Figure 5D left panel). The mean peak of delta power in the left motor cortex in the four conditions was respectively for NMS: 2.06 ± 0.13 Hz; NMA: 2.07 ± 0.18 Hz; HMS: 2.03 ± 0.17 Hz and HMA: 2.06 ± 0.21 Hz. To statistically assess the distance between the peak frequencies of left motor delta power and stimulus delta activity, we tested the mean of each score distribution against zero with a one-sample t-test (one-tailed). Results revealed that the mean absolute distance was significantly greater than zero in all conditions ($p < p_{corrected}$). Further, a one-way repeated-measure ANOVA tested the difference of absolute distance between all conditions. Results revealed a significant effect of condition ($F(15,848)$ = 2.995; p < 0.001; η_p^2 = 0.05). However, Bonferroni-corrected pairwise comparisons revealed only a single marginal tendency for a difference between the Full$_{\text{no-mask synchronous}}$ and Head$_{\text{head-mask synchrony}}$ absolute distances (p = 0.09; 1st and 11th bars on Figure 5D). These results confirm that the delta responses induced in the left motor cortex significantly deviated from stimulus-related delta frequency, thus did not just reflect entrainment. Finally, concerning the consistency of neural delta power across trials in all conditions, results revealed no significant main effect of mask or an interaction between mask and asynchrony for motor delta peak frequency (mask: $F(1,53)$ = 1.84; p = 0.181; η_p^2 = 0.034; asynchrony: $F(1,53)$ = 1.11; p = 0.741; η_p^2 = 0.002; interaction between mask*asynchrony: $F(1,53)$ = 1.33; p = 0.717; η_p^2 = 0.003; see Figure 5D right panel). These results confirm that any observed variation in motor delta activity between the experimental conditions cannot be explained as mere stimulus frequency.

Delta-beta PAC reflects sensitivity to audio-visual temporal asynchrony in speech perception but is not limited to the left motor cortex.

Finally, we assessed whether delta-beta PAC modulations in the left motor area reflect sensitivity to audio-visual asynchrony in speech perception. First, a three-way repeated-measure ANOVA (main factors: mask, asynchrony and correctness) revealed a main effect of mask on delta-beta PAC with delta-beta phase-coupling being significantly greater in the no-mask than in the head-mask conditions (F(2,22) = 4.72; p = 0.041; η_p^2 = 0.18; see Figure 6A). No further significant main effects or interactions were found. These results show greater left motor cortex delta-beta PAC when participants were more sensitive to asynchronous audio-visual speech in the no-mask conditions than when they were less able to match visual and auditory prosodic features (head-mask conditions). Nevertheless, we cannot fully discard that delta-beta PAC also increased in the head-mask condition as compared to a control baseline condition (e.g., visual or auditory only condition). Secondly, we investigated whether the delta-beta PAC difference between no-mask and head-mask conditions was restricted to the left motor areas. As accuracy and asynchrony did not affect delta-beta PAC in the cluster of interest, we selected only correct trials for the delta-beta PAC analysis at the whole brain level and combined synchronous and asynchronous trials within no-mask and head-mask conditions (i.e., NMCs: NMA+ NMS; HMCs: HMA + HMS). The cluster-based permutation tests revealed one significant positive cluster peaking in the superior motor area and in the left middle temporal lobe (although not exclusively; see Figure 6B), confirming that delta-beta PAC was significantly larger in the no-mask (NMCs) compared to the head-mask (HMCs) case (NMCs - HMCs : p = 0.043, cluster statistic = 216.69).

[Insert Figure 6]

Figure 6: Phase-amplitude coupling between delta and beta oscillations. (A) PAC analysis in the left motor cluster. The figure represents the modulation of delta-beta PAC in a significant cluster, dependent on the mask and audio-visual asynchrony. Significance is indicated by an asterisk (p < 0.05, Bonferroni-corrected). Delta-beta PAC from the left motor cortex was greater in the no-mask than the head-mask conditions but did not discriminate between correct and

incorrect trials. Significant contrasts are marked by stars (p < 0.05). (B) Delta-beta PAC difference between no-mask (NMA+ NMS) and head-mask (HMA + HMS) case in the whole brain. Results revealed significant maximum differences located in the superior motor area (MNI coordinates [0 11 50]) and in the left middle temporal lobe (MNI coordinates [-50 -1 -20]).

In summary, the EEG results mirrored the behavioural results as modulations in left motor delta power reflect the successful detection of audio-visual asynchrony when participants were able to see face and visible articulators (no-mask conditions), but not in the head-mask conditions. An increase in left motor delta power only predicted differential sensitivity to audio-visual asynchrony in the no-mask conditions and related to correctly perceiving asynchronous audio-visual speech. Importantly, a control analysis confirmed that variations in left motor delta activity reflect an amplitude difference based on the same oscillatory activity across all stimuli rather than oscillation differences of stimulation *per se*. Lastly, delta-beta PAC in the left motor cortex was greater when listeners detected audio-visual asynchrony more accurately during speech perception (i.e., no-mask as compared to head-mask conditions). Nevertheless, this result did not exclude that delta-beta PAC also increased in the head-mask conditions, but to a lesser extent.

DISCUSSION

The present study investigated the role of motor delta oscillations during the temporal analysis of multisensory prosodic features in speech perception. The behavioural results of the audio-visual asynchrony detection task confirmed that listeners processed both prosodies in multisensory speech perception when sufficient visual information was available. At the brain level, the perception of audio-visual asynchrony induced an increase in left motor delta activity (extending to the inferior frontal gyrus). Further, the difference of delta power between asynchronous and synchronous conditions predicted participants' sensitivity of audio-visual asynchrony. In contrast, participants were less able to discriminate audio-visual information when a speaker's facial information was masked. This is evident in the absence of difference in delta activity between asynchronous and synchronous conditions. Finally, delta-beta PAC in the left motor cortex was significantly greater when listeners were more accurate in perceiving asynchrony between visual and auditory information during multisensory speech perception (no-mask vs. head-mask conditions). Altogether, the current results indicate that the delta timescale provides a flexible framework to synchronise a listener’s brain activity with multisensory speech input. Thus, motor delta activity seems to play a role in detecting temporal mismatches between visual and auditory prosodies and is as a sensitive measure of (un-)successful temporal analysis in multisensory speech perception.

Behaviourally, the results of the asynchrony detection task confirm our first hypothesis, that is, listeners temporal analyse prosodic events in multisensory speech perception. This finding was expected as visual information complements auditory information and often improves speech perception (Sumby & Pollack, 1954; van Wassenhove et al., 2005). Speaker’s articulatory movements and gestures temporally aligned with acoustic prosodic cues, providing listeners with a reliable temporal structure of the speech signal in the delta range (Biau et al., 2016; Esteve-Gibert & Guellaï, 2018; Wagner et al., 2014). Participants likely use these salient prosodic events as landmarks to align them into a coherent multisensory speech percept. The results suggest that successful temporal analysis can focus the listeners’ attention within brief time-windows containing complementary multisensory prosodic events. This is in line with the theory of dynamic attending, stating that non-random external stimulation drives periodic attention allocation towards critical events (Large & Jones, 1999). Noteworthy, the differences of performance between the no-mask and head-mask conditions indicate that participants likely relied on complementary information conveyed by the speaker’s head, face, and fine articulatory gestures to achieve the integration of the visual prosodic signal (Cross, Butler, & Lalor, 2015). Of note, when the speaker’s face was masked, participants’ response accuracy decreased significantly while they remained sensitive to the temporal alignment of the audio-visual signals in the synchronous condition (HMS). Our results suggest that participants adopted a liberal guessing strategy and tended to respond “synchronous” more often in the head-mask conditions (i.e., negative *c* criterion and decrease of d’ as compared to the no-mask conditions). Therefore, we assume that if participants were not sensitive at all to audio-visual temporal alignment, such a bias would have only increased and led to responding “synchronous” even more systematically. Consequently, correct response rates would have significantly increased in the HMS as compared to the NMS and NMA conditions. Our results show that this is not the case as HMS accuracy was equivalent to NMS and NMA accuracy. Rather, performance in the HMA condition decreased to chance level. While somewhat speculative, a comparable delta power increase in the HMS and NMA conditions at the scalp level (Figure 3C) may reflect a similar increased effort to reach comparable accuracy levels when integrating a blurred visual signal with an auditory signal. Such an increased effort was not observed in the HMA condition, potentially due to the larger delay between visual and auditory signal onsets that prevented participants to align them. In other words, increased delta power reflects an analytic effort and explains comparable delta power patterns in the NMA and HMS conditions. Lastly, although we applied a unique head-mask to obscure visual facial prosody, this is technically a gradual masking approach because blurring the face did not prevent the participants from using other available visual prosodic information (e.g., head nods, upper parts of the speaker’s body, and hand gestures). Nevertheless, future studies could adopt a more fine-grained gradual masking approach by using different levels of visual degradation, masking different effectors (e.g., mouth, head, hands, and breathing) to examine which movements best carry information needed for successful temporal analysis in multisensory speech perception.

The EEG results confirmed an increase in motor delta activity in response to audio-visual asynchrony detection, extending the role of delta activity to the temporal analysis of multisensory prosodies. Previous literature associated delta oscillations in the motor cortex with the perception of auditory rhythmic stimulation (Keitel et al., 2018; Morillon et al., 2019; Morillon & Schroeder, 2015). The present results extend these findings to the temporal analysis of non-isochronous events that act as punctual “snap fasteners” streaming visual and auditory signals within relevant time-windows. As long as they provide the brain with sufficient time for the temporal analysis of multiple sensory inputs, salient prosodic features do not have to be perfectly regular to trigger delta motor responses. The present EEG results corroborate this hypothesis in three ways: First, we did not observe different delta responses in auditory and visual cortices when audio-visual stimuli were synchronous. This would have reflected low-level feature tracking in early sensory processing (Cross, Butler & Lalor, 2015; Ghitza, 2017; Gross et al., 2013; Mai, Minett & Wang, 2016). Further, a control analysis confirmed that delta responses in left motor cortex did not simply reflect stimulus entrainment as they significantly differed from the frequency of the audio-visual stimuli. Next, audio-visual asynchrony would likely decrease pure entrainment by making signal tracking more difficult than when different channels of the same input are processed synchronously. Further, we found no theta activity in response to audio-visual asynchrony at the scalp level that would have indicated an effect driven specifically by the rate of the prosodic features (e.g., lip movements). Additionally, differences in left motor cortex delta power only predicted accuracy in the no-mask contrast. Moreover, delta power increased more for accurate than inaccurate responses in the asynchronous conditions, independent of the presence or absence of a head-mask. Lastly, participants perceived audio-visual synchrony less accurately when the speaker's facial information was blurred. This was shown in weaker delta motor responses and that synchronous and asynchronous conditions displayed not differences in delta power. Together, these results confirm that left motor delta oscillations might reflect the successful detection of audio-visual asynchrony, likely linked to the temporal analysis of multisensory speech.

Importantly, the responses found in the left inferior frontal gyrus align well with previous research that established a role in cross-modal information integration between gestures and speech (Park et al., 2018; Willems, Ozyürek & Hagoort, 2009; Zhao et al., 2018). Here, participants perceived information carried by two modalities, and integrated gestures' kinematics with auditory envelope modulations to perform an asynchrony detection task. Further investigations will need to address whether the response modulations in the left IFG were specific to the temporal integration of gesture and speech or could be reproduced using moving dots following gestures' dynamics (Biau et al., 2016; Holle et al., 2012). In contrast, we found no differential activation in further brain regions associated with multisensory speech integration such as the left posterior superior temporal sulcus (Marstaller & Burianová, 2014). Here, delta oscillations did not reflect multisensory integration *per se* but the temporal alignment of multisensory information. It is worth noting that the present study focused on the temporal analysis of visual and auditory prosodies and addressed how they (mis-)align. It could be of further interest to look more closely into the temporal dynamics between sensory areas in multisensory speech perception. For instance, comparing delta phase offsets between synchronous and asynchronous conditions could help to understand whether the synchronisation of delta oscillations between visual and auditory areas predicts delta responses in the left motor cortex. Future studies may overcome the limitations of the current study to perform source reconstruction analysis (e.g., including visual and auditory only conditions as localizers), and address the role of delta synchronisation between sensory areas in multisensory speech perception.

Finally, delta-beta coupling in the left motor cortex was larger in the no-mask conditions, when listeners perceived audio-visual temporal alignment in both directions (i.e., synchronous or asynchronous). Although somewhat speculative, delta-beta coupling might take place after proper temporal analysis of visual and auditory prosodic features has occurred and might support top-down predictions (e.g., auditory-motor coupling). For instance, Park et al. (2015) showed that the left frontal areas modulated the phase of delta oscillations in the left auditory cortex by means of top-down control in speech perception. Reciprocally, delta-beta PAC in the auditory cortex respond to the modulations of rhythmic regularity in auditory speech perception (Chang, Bosnyak and Trailor, 2019). Further, Keitel et al. (2018) reported that delta-beta PAC in the left motor cortex predicted behavioural performance in speech comprehension. Future research will need to unravel whether delta-beta coupling provides a ubiquitous means of cross-regional communication to align temporally different dynamic input in sensory cortices (Arnal, 2012; Fujioka, Ross & Trainor, 2015; Morillon et al., 2019). For example, Fontolan et al. (2014) reported that delta-beta coupling in the associative auditory cortex modulated the phase of gamma activity related to phonological processing in the primary auditory cortex in auditory sentence perception (Giraud & Poeppel, 2012). Alternatively, delta-beta PAC may drive the periodicity of attention to critical time-windows containing relevant accentuated speech information, which fits with the dynamic attention theory (Large & Jones, 1999). It is important to note that delta-beta PAC may increase in the head-mask conditions as well, but simply to a lesser extent than in the no-mask conditions. If true, top-down predictions may be generated during multisensory speech perception even when participants were less successful in detecting audio-visual temporal (mis)alignment.

We propose that motor delta oscillations mirror the successful detection of asynchronous multisensory prosodies, encoded separately in auditory and visual sensory cortices. The slow timescale of delta (1-3Hz) may also offer the brain some flexibility to create a coherent multisensory percept despite the natural delay between visual and auditory signal onsets in speech (Chandrasekaran et al., 2009). In social interactions where conditions change quickly, such a delta framework would help listeners to align speech information in a bottleneck fashion to maintain stable synchronization in speech flow (Kotz, Ravignani & Fitch, 2018). When two dynamic events cannot be integrated in a critical delta time-window due to their temporal offsets, any effort to resolve such an audio-visual mismatch increases and shows in amplified motor delta activity. At a certain point, i.e., when onsets of visual and auditory prosody onsets mismatch (~ 400ms), delta power reaches a critical threshold, leading to the successful detection of audio-visual asynchrony in speech. Further investigations will need to address whether this with other timescales present in both the speech signal and brain oscillations. For instance, we cannot fully discard that the prosodic contour in our stimuli still contained a syllable structure embedded in it (e.g., at onsets and stress peaks). Further, lip movements and auditory envelope convey syllabic information occurring at a theta rate (4-8 Hz) providing other robust temporal information in the speech signal during face-to-face conversations (Chandrasekaran et al., 2009; Giraud & Poeppel, 2012). Therefore, delta and theta activities may actually couple to strengthen speaker-listener synchronization in social communicative interactions.

CONCLUSION

Our findings show that left motor delta oscillations play a role in audio-visual asynchrony detection of visual and auditory prosodies, and by extension contribute to the successful temporal analysis of multisensory speech. We propose that a critical delta time window allows for the (un-)successful temporal alignment of dynamic prosodic features, conveyed by distinct sensory modalities in speech perception.

ACKNOWLEDGMENT

This research was supported by a postdoctoral fellowship from the European Union's Horizon 2020 research and innovation program, under the Marie Sklodowska-Curie grant agreement (No. 707727).

RESOURCE SHARING

Consent for sharing data at the level of the individual participant was received. Data for individual participants and associated scripts will be made available upon publication of the manuscript. Further information or requests should be directed to the corresponding authors.

REFERENCES

Arnal, L. H. (2012). Predicting "When" Using the Motor System's Beta-Band Oscillations. *Frontiers in Human Neuroscience*, *6*, 225

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta-Beta Coupled Oscillations Underlie Temporal Prediction Accuracy. *Cerebral Cortex*, *25*(9), 3077–3085

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152

Biau, E., Fernandez, L. M., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio-visual alignment are modulated by the communicative nature of the stimuli. *NeuroImage,132*, 129–137

Boersma, P., and Weenink, D. (2015). Praat: Doing Phonetics by Computer. Version 5.4.17

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Chang, A., Bosnyak, D. J., & Trainor, L. J. (2019). Rhythmicity facilitates pitch discrimination: Differential roles of low and high frequency neural oscillations. *NeuroImage*, *198*, 31–43. https://doi.org/10.1016/j.neuroimage.2019.05.007

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. The Journal of the Acoustical Society of America. 25 (5): 975–79

Crosse M.J., Butler J.S., Lalor E.C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. The Journal of Neuroscience 35:14195–14204

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *The Journal of Neuroscience*, *36*(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85 Pt 2*, 761–768

Esteve-Gibert N., & Guellaï B. (2018). Prosody in the Auditory and Visual Domains: A Developmental Perspective. *Frontiers in Psychology*, 9:338. doi:10.3389/fpsyg.2018.00338

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., and Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat. Commun.* 5:4694. doi: 10.1038/ncomms5694

Fujioka, T., Ross, B., & Trainor, L. J. (2015). Beta-Band Oscillations Represent Auditory Beat and Its Metrical Hierarchy in Perception and Imagery. *Journal of Neuroscience*, *35*(45), 15187–15198

Ghitza, O. (2017). Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*, *32*(5), 545–561. https://doi.org/10.1080/23273798.2016.1232419

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517

Griffiths, B. J., Parish, G., Roux, F., Michelmann, S., Plas, M. Van Der, Kolibius, D., & Hanslmayr, S. (2019). Directional coupling of slow and fast hippocampal gamma with neocortical alpha / beta oscillations in human episodic memory. *Proceedings of the National Academy of Sciences*, 1–9. https://doi.org/10.1073/pnas.1914180116

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752

Gunter, T. C., & Douglas Weinbrenner, J. E. (2017). When to Take a Gesture Seriously: On How We Use and Prioritize Communicative Cues. *Journal of Cognitive Neuroscience, 29*(8), 1355-1367

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in psychology*, *3*, 74. https://doi.org/10.3389/fpsyg.2012.00074

Jessen, S., & Kotz, S. A. (2015). Affect differentially modulates brain activation in uni- and multisensory body-voice perception. *Neuropsychologia*, *66*, 134–143

Keitel, A., Ince, R. A. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, *147*, 32–42. https://doi.org/10.1016/j.neuroimage.2016.11.062

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*(3), e2004473

Kösem, A., Gramfort, A., & van Wassenhove, V. (2014). Encoding of event timing in the phase of neural oscillations. *NeuroImage*, *92*, 274–284. https://doi.org/10.1016/j.neuroimage.2014.02.010

Kösem, A., & van Wassenhove, V. (2017). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*(5), 536–544

Kotz S.A., Ravignani A., Fitch W.T. The Evolution of Rhythm Processing. *Trends Cogn Sci*. 2018;22(10):896-910. doi:10.1016/j.tics.2018.08.002

Large, E.W., & Jones, M.R. (1999). The dynamics of attending: how people track time-varying events. *Psychol. Rev.* 106, 119

Mai, G., Minett, J. W., & Wang, W. S. Y. (2016). Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. *NeuroImage*, *133*, 516–528. https://doi.org/10.1016/j.neuroimage.2016.02.064

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190

Marstaller L, Burianov_a H.(2014). The multisensory perception of co-speech gestures - a review and meta-analysis of neuroimaging studies. *J Neurolinguistics, 30*, 69-77.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, 98, 185–199.

Mercier MR, Molholm S, Fiebelkorn IC, Butler JS, Schwartz TH, Foxe JJ. (2015). Neuro-oscillatory phase alignment drives speeded multisensory response times: an electro-corticographic investigation. J Neurosci 35: 8546–8557.

Meyer, L., Sun, Y., & Martin, A. E. (2019). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, 1–11. https://doi.org/10.1080/23273798.2019.1693050

Morillon, B., & Schroeder, C. E. (2015). Neuronal oscillations as a mechanistic substrate of auditory temporal prediction. *Annals of the New York Academy of Sciences*, *1337*(1), 26–31. https://doi.org/10.1111/nyas.12629

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), E8913–E8921

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience and Biobehavioral Reviews*, Vol. 107, pp. 136–142. https://doi.org/10.1016/j.neubiorev.2019.09.012

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137

Obermeier, C., Dolk, T., & Gunter, T. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex, 48*, 857-870

Obermeier, C., & Gunter, T. C. (2014). Multisensory Integration: The Case of a Time Window of Gesture-Speech Integration. *Journal of Cognitive Neuroscience*, 1–16

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open-source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:156869

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Current Biology*, *25*(12), 1649–1653

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *ELife*, *5*

Park, H., Ince, R., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS biology*, *16*(8), e2006558. https://doi.org/10.1371/journal.pbio.2006558

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*, 320

Puzzo, I., Cooper, N. R., Vetter, P., & Russo, R. (2010). EEG activation differences in the pre-motor cortex and supplementary motor area between normal individuals with high and low traits of autism. *Brain Research*, *1342*, 104–110

Saleh, M., Reimer, J., Penn, R., Ojakangas, C. L., & Hatsopoulos, N. G. (2010). Fast and Slow Oscillations in Human Primary Motor Cortex Predict Oncoming Behaviorally Relevant Cues. *Neuron*, *65*(4), 461–471

Schultz, B. G., Biau, E., & Kotz, S. A. (2020). An open-source toolbox for measuring dynamic video framerates and synchronizing video stimuli with neural and behavioral responses. *Journal of Neuroscience Methods*, 108830

Stegemöller, E. L., Allen, D. P., Simuni, T., & MacKinnon, C. D. (2017). Altered premotor cortical oscillations during repetitive movement in persons with Parkinson's disease. *Behavioural Brain Research*, *317*, 141–146

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215

Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM (2011) Brainstorm: a user-friendly application for MEG/EEG analysis. Comput Intell Neurosci 2011 :879716

Thut, G., Veniero, D., Romei, V., Miniussi, C., Schyns, P., & Gross, J. (2011). Rhythmic TMS causes local entrainment of natural oscillatory signatures. *Current Biology*, *21*(14), 1176–1185. https://doi.org/10.1016/j.cub.2011.05.049

Tort, A. B. L., Komorowski, R., Eichenbaum, H., & Kopell, N. (2010). Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *Journal of Neurophysiology*, *104*(2), 1195–1210. https://doi.org/10.1152/jn.00106.2010

van Veen, B. D., van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Bio-Medical Engineering*, *44*(9), 867–880

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232

Wang, D., Clouter, A., Chen, Q., Shapiro, K. L., & Hanslmayr, S. (2018). Single-trial phase entrainment of theta oscillations in sensory regions predicts human associative memory performance. *Journal of Neuroscience*, *38*(28), 6299–6309

Willems, R. M., Ozyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage*, *47*(4), 1992–2004. https://doi.org/10.1016/j.neuroimage.2009.05.066

Zhao, W., Riggs, K., Schindler, I., & Holle, H. (2018). Transcranial Magnetic Stimulation over Left Inferior Frontal and Posterior Temporal Cortex Disrupts Gesture-Speech Integration. *The Journal of Neuroscience*, *38*(8), 1891–1900. https://doi.org/10.1523/JNEUROSCI.1748-17.2017

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. *Current Biology*, *28*(3), 401-408.e5. https://doi.org/10.1016/j.cub.2017.11.071

Movie Legends

Movie 1: Example of an audio-visual stimulus presented in the no-mask asynchronous condition (NMA). In this videoclip, video and audio information were presented in asynchrony (i.e., video onset led audio onset by 400 milliseconds), and the face of the speaker was fully visible.

Movie 2: Example of an audio-visual stimulus presented in the no-mask synchronous condition (NMS). In this videoclip, video and audio information were presented in synchrony, and the face of the speaker was fully visible.

Movie 3: Example of an audio-visual stimulus presented in the head-mask asynchronous condition (HMA). In this videoclip, video and audio information were presented in asynchrony (i.e., video onset led audio onset by 400 milliseconds), and the face of the speaker was blurred.

Movie 4: Example of an audio-visual stimulus presented in the no-mask synchronous condition (HMS). In this videoclip, video and audio information were presented in synchrony, and the face of the speaker was blurred.

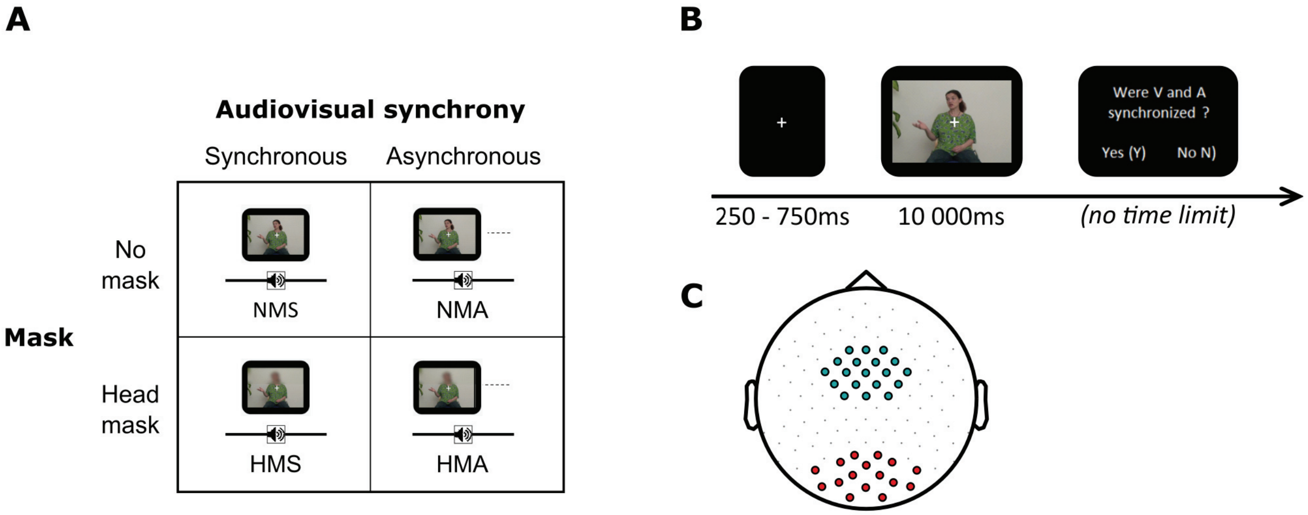
Movie1_still: Example of a frame taken from a videoclip in the no-mask synchronous condition (NMS).

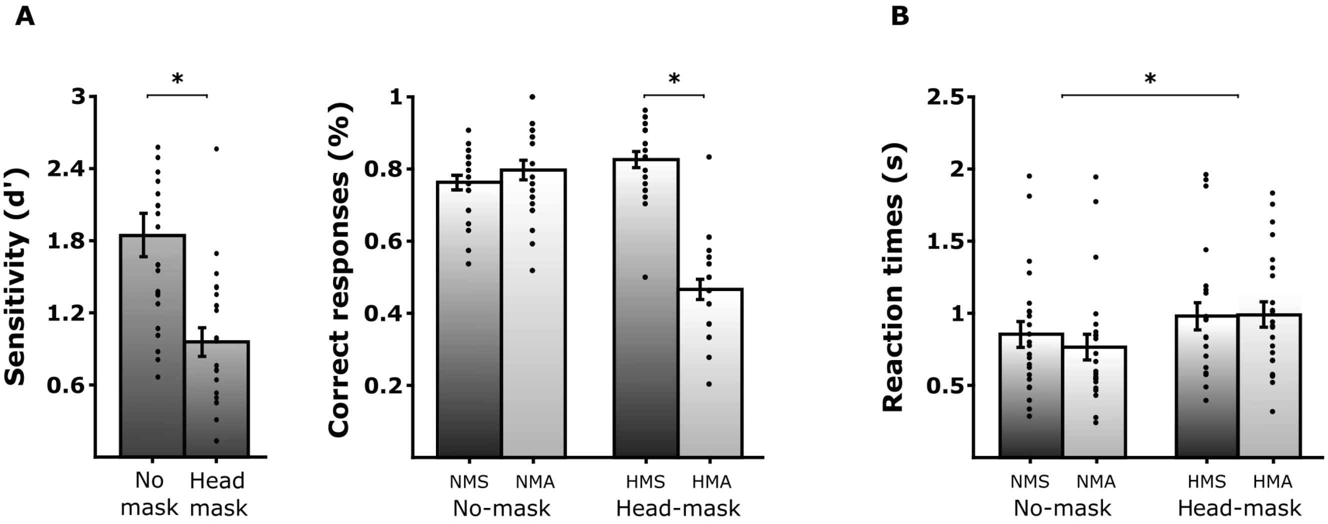
Movie2_still: Example of a frame taken from a videoclip in the no-mask asynchronous condition (NMA).

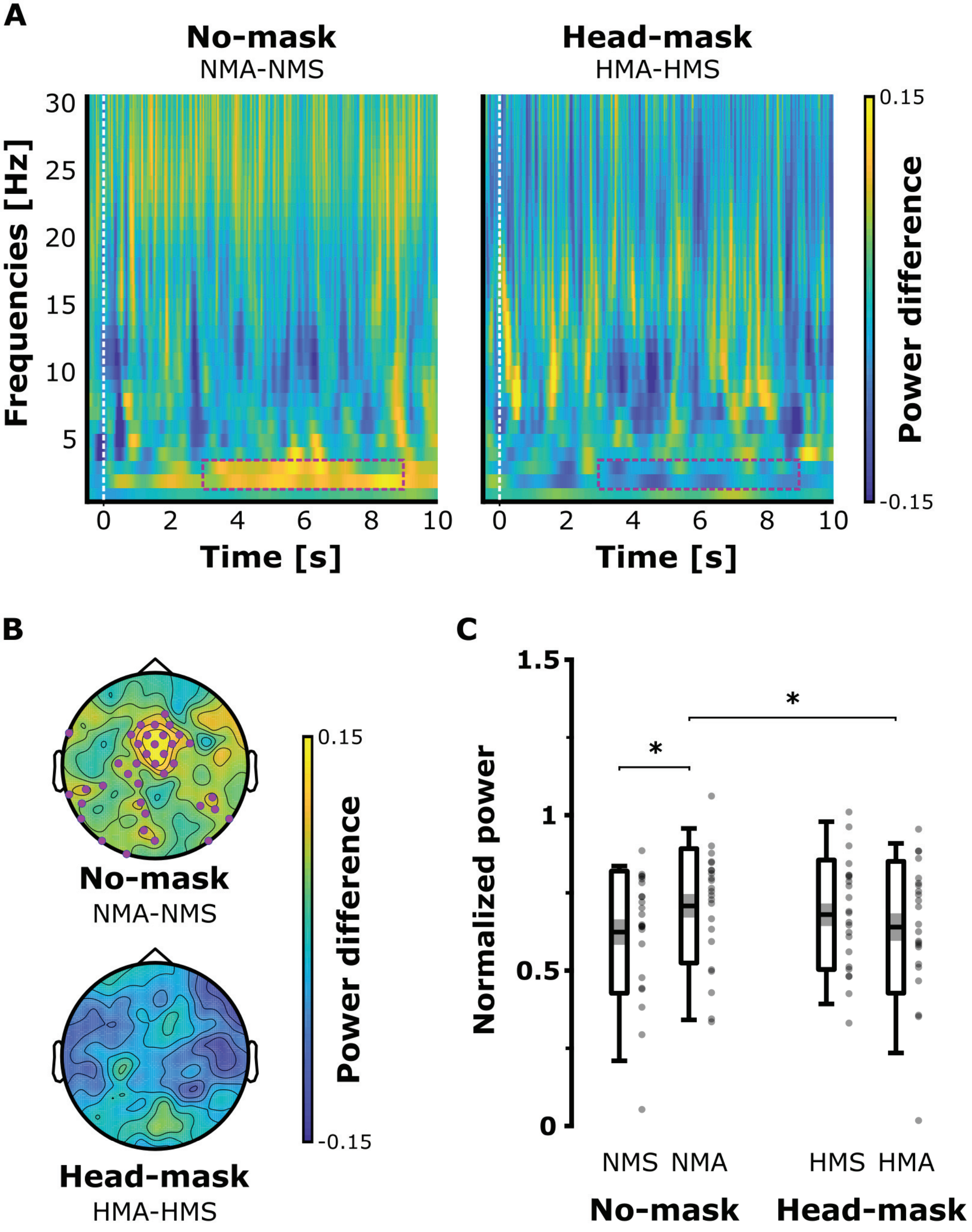
Movie3_still: Example of a frame taken from a videoclip in the head-mask synchronous condition (HMS).

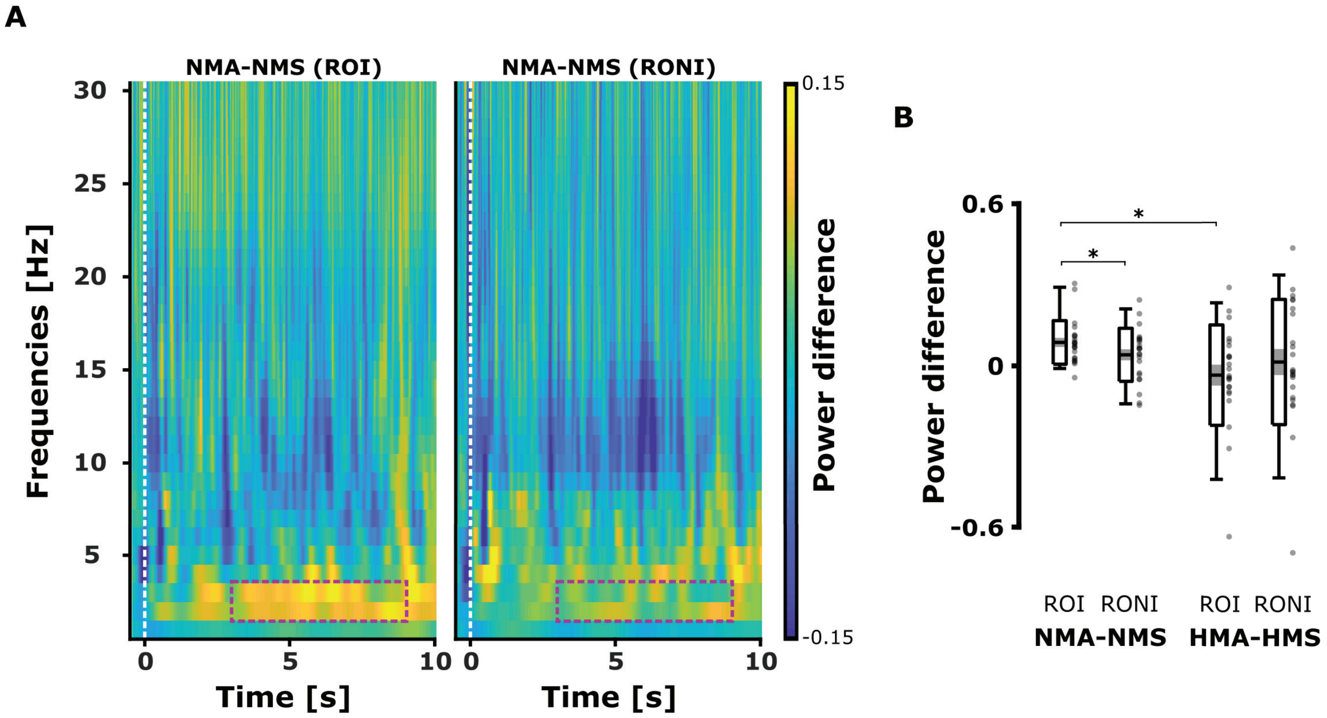
Movie4_still: Example of a frame taken from a videoclip in the head-mask asynchronous condition (HMA).

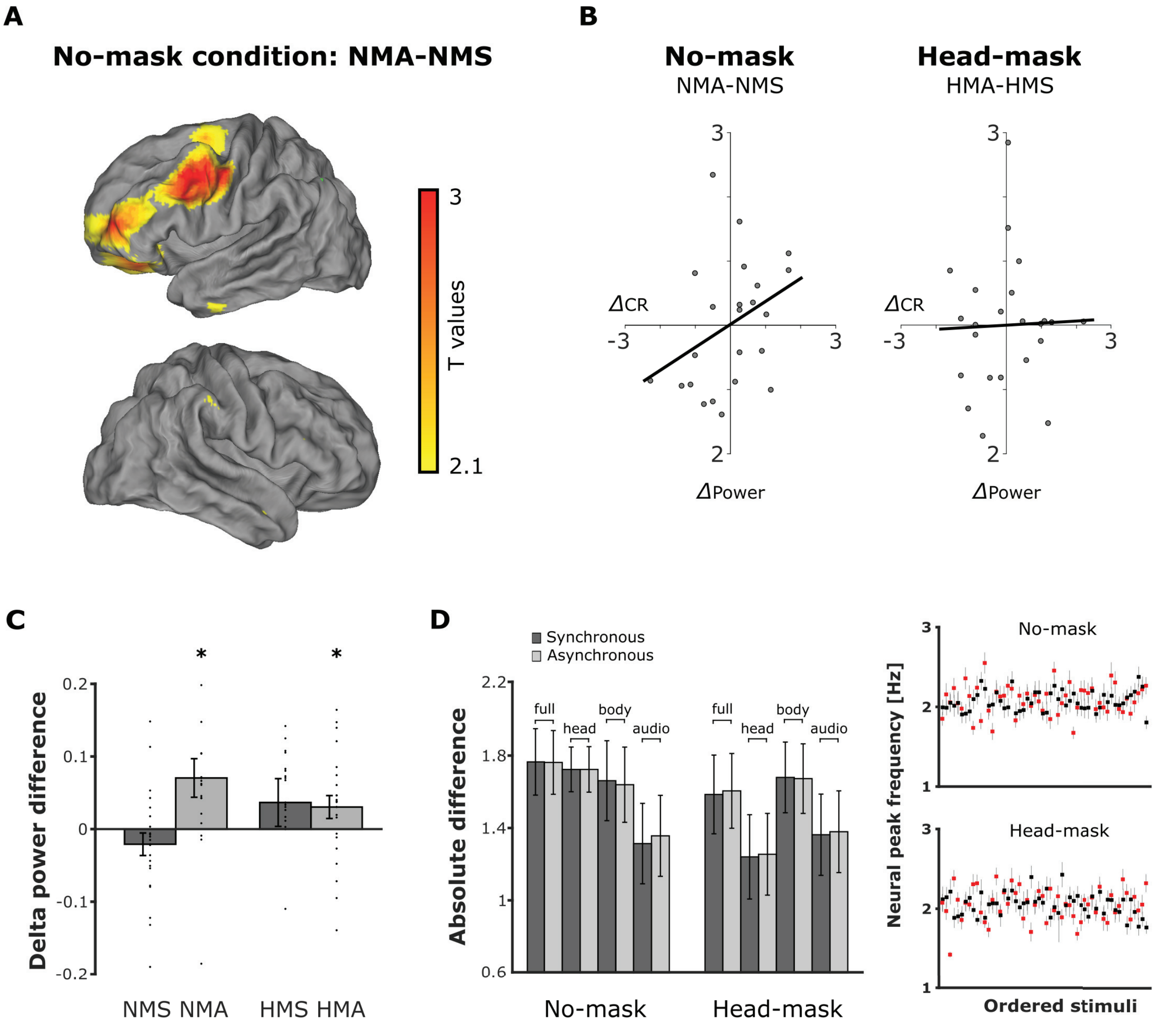
Stimuli	Mean Peak Freq.	SD Peak Freq.	Min. Peak Freq.	Max Peak Freq.
Full (No-mask + Body)	3.65	1.02	0.86	6.13
Full (Head-mask + Body)	3.10	1.45	0.86	6.13
Head only (No-mask)	3.59	0.91	1.00	3.99
Head only (Head-mask)	2.27	1.53	0.86	6.13
Body only*	3.37	1.25	0.86	6.13
Audio only*	2.74	1.44	0.86	5.86

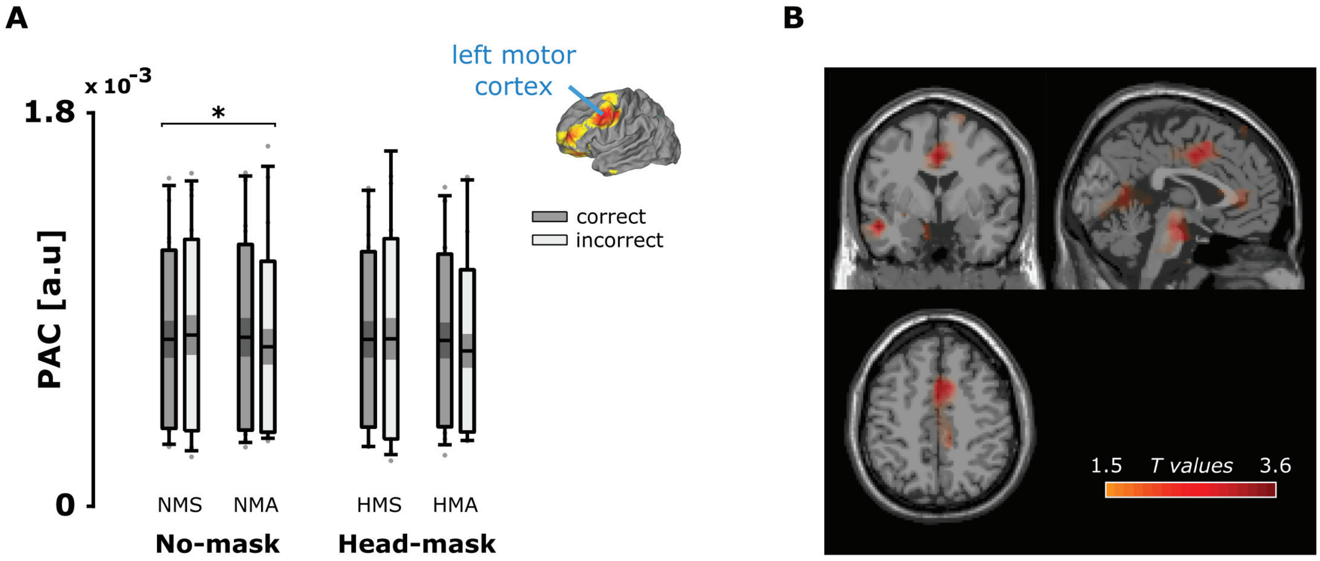

A
Audiovisual synchrony
Synchronous
Asynchronous
Mask
No mask
Head mask
NMS
NMA
HMS
HMA
B
+
Were V and A synchronized ?
Yes (Y)
No N)
250 - 750ms
10 000ms
(no time limit)
C


A
B
*
3
2.4
1.8
1.2
0.6
Sensitivity (d')
No mask
Head mask
1
0.8
0.6
0.4
0.2
Correct responses (%)
NMS
NMA
HMS
HMA
No-mask
Head-mask
2.5
2
1.5
1
0.5
Reaction times (s)


A
No-mask
NMA-NMS
Head-mask
HMA-HMS
Frequencies [Hz]
30
25
20
15
10
5
0
2
4
6
8
10
Time [s]
0.15
-0.15
Power difference
B
No-mask
NMA-NMS
Head-mask
HMA-HMS
0.15
-0.15
Power difference
C
Normalized power
1.5
1
0.5
0
*
*
NMS
NMA
HMS
HMA
No-mask
Head-mask


A
NMA-NMS (ROI)
NMA-NMS (RONI)
30
25
20
15
10
5
Frequencies [Hz]
0
5
10
Time [s]
0
5
10
Time [s]
0.15
-0.15
Power difference
B
0.6
0
-0.6
Power difference
*
*
ROI
RONI
ROI
RONI
NMA-NMS
HMA-HMS


A
No-mask condition: NMA-NMS
3
T values
2.1
B
No-mask
NMA-NMS
Head-mask
HMA-HMS
3
ΔCR
-3
3
2
ΔPower
3
ΔCR
-3
3
2
ΔPower
C
*
*
0.2
0.1
0
-0.1
-0.2
Delta power difference
NMS NMA
HMS HMA
D
Synchronous
Asynchronous
2.2
1.8
1.4
1
0.6
Absolute difference
full
head
body
audio
full
head
body
audio
No-mask
Head-mask
3
2
1
No-mask
3
2
1
Head-mask
Neural peak frequency [Hz]
Ordered stimuli


A
x 10^-3
1.8
0
PAC [a.u]
*
NMS
NMA
HMS
HMA
No-mask
Head-mask
left motor cortex
correct
incorrect
B
1.5
T values
3.6