**INTER-FACULTY MASTER PROGRAM**

**on NETWORKS and COMPLEXITY**

**SCHOOL of MATHEMATICS**

**SCHOOL of BIOLOGY**

**SCHOOL of GEOLOGY**

**SCHOOL of ECONOMICS**

**ARISTOTLE UNIVERSITY of THESSALONIKI**

**Master Thesis**

**Title:**

# Εγκεφαλικά δίκτυα στη χρονική (α)συμφωνία του οπτικοακουστικού λόγου.

# Brain Networks during audiovisual speech (a)synchronies.

**Nikolas Chalas**

**SUPERVISOR:** Stefanos Sgardelis, Professor, AUTH

**CO-SUPERVISOR:** Virginie van Wassenhove, Researcher, CEA

**CO-SUPERVISOR:** Evangelos Paraskevopoulos, Postdoctoral Researcher, AUTH

**Thessaloniki, December 2019**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ**
**στα ΔΙΚΤΥΑ και ΠΟΛΥΠΛΟΚΟΤΗΤΑ**
**ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ**
**ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ**
**ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ**
**ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

**ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ**

## Μεταπτυχιακή διπλωματική Εργασία

**Τίτλος Εργασίας:**

# Εγκεφαλικά δίκτυα στη χρονική (α)συμφωνία του οπτικοακουστικού λόγου.

# Brain Networks during audiovisual speech (a)synchronies.

**Nikolas Chalas**

**ΕΠΙΒΛΕΠΩΝ:** Στέφανος Σγαρδέλης, Καθηγητής, ΑΠΘ

**ΣΥΝ-ΕΠΙΒΛΕΠΩΝ:** Virginie van Wassenhove, Ερευνήτρια, CEA

**ΣΥΝ-ΕΠΙΒΛΕΠΩΝ:** Ευάγγελος Παρασκευόπουλος, Μεταδιδακτορικός ερευνητής, ΑΠΘ

Εγκρίθηκε από την τριμελή εξεταστική τη 17η Δεκεμβρίου 2019

| | | |
|---|---|---|
| …… | …… | …… |
| Σ. Σγαρδέλης<br>Καθηγητής Α.Π.Θ. | Ε. Αντωνοπούλου<br>Αν. Καθηγήτρια Α.Π.Θ. | Ε. Παρασκευόπουλος<br>Μεταδιδακτορικός ερευνητής Α.Π.Θ |

**Θεσσαλονίκη , Δεκέμβριος 2019**

……..

Νικόλας Χαλάς
Πτυχιούχος Βιολόγος Α.Π.Θ.

# Table of Contents

## Abstract

Speech encapsulates visual and acoustic signals that are received asynchronously but perceived as simultaneous by the listener. In the neuropsychological level, during audiovisual speech, the preceding visual speech may serve as an initial prediction, encompassing '*what'* and '*when'* the sound will occur. The current study sought to determine this mechanism. For this scope, magnetoencephalographic (MEG) data were acquired from human participants during the presentation of systematically a-synchronized audiovisual speech syllables. At first, the visual modulation was confirmed in the event-related auditory response, corroborating previous findings reporting a reduced amplitude and latency of the auditory response during naturally occurring audiovisual asynchronies. Furthermore, the neural distribution across audiovisual asynchronies matched the asymmetric temporal integration window. More importantly, it is shown that audiovisual speech asynchronies are processed in the brain with cortical entrainment. This neural synchronization matched the temporal scale of syllabic rate was not found to be distracted with various asynchronies deviating the natural audiovisual statistics, within the superior temporal gyrus. Finally, hierarchically nested oscillatory networks are found to underpin the various asynchronies, depicting the predictive hierarchies within the cortex through different oscillatory regimes.

**Keywords:** audiovisual speech; cortical entrainment; magnetoencephalography; cortical networks; multisensory integration; predictive coding

## Introduction

Paying attention to the speaker's lips enhances auditory perception (Erber, 1975; Grant & Seitz, 2000; Macleod & Summerfield, 1987; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2006; Sumby & Pollack, 1954). From an evolutionary standpoint, the capacity to synthesize the senses provides a substantial survival value, as combining cues from different sensory modalities, allows the nervous system to grasp – otherwise inconceivable - environmental signals (Stein, Stanford, & Rowland, 2014). Consequently, it has been argued that neocortical operations are multisensory (Ghazanfar & Schroeder, 2006), underpinned by a dynamic interplay of distributed cortical regions (Driver & Noesselt, 2008). Accordingly, when cross-modal and causally-linked signals reach our senses, the brain integrates them into a coherent percept, segregating irrelevant noise, while implementing a weighting principle according to the most -or least- reliable input (Cao, Summerfield, Park, Giordano, & Kayser, 2019; Kayser & Shams, 2015; Körding et al., 2007).

Regarding human communication, face-to-face interaction is inherently multisensory (Ghazanfar & Takahashi, 2014; Holler & Levinson, 2019; van Wassenhove, 2013). In a seminal study of audiovisual integration in speech comprehension, McGurk and MacDonald (1976) reported an illusion delineating the influence of visual speech in auditory perception. In this study, when an auditory syllable [b_] was presented with a dubbed visual articulation of [g_] it led to a *fused* auditory perception of [d_] or [g_]. Interestingly, when an auditory syllable [g_] was dubbed with a visual articulation of [b_] it led to a *combined* auditory [bg_], illustrating that audio-visual interactions depend on the informational content of each sensory modality (van Wassenhove, 2013). Furthermore, subsequent studies aimed at unraveling the effect of temporal concordance in audiovisual speech perception (Conrey & Pisoni, 2006; Maier, Di Luca, & Noppeney, 2011; Massaro, Cohen, & Smeele, 1996; Munhall et al., 1996; van Wassenhove, Grant, & Poeppel, 2007). A major finding was that audiovisual speech -in contrast to non-speech stimuli- exhibited a temporal tolerance in audiovisual asynchronies matching the average syllabic duration (~250 ms) and this *temporal window of integration* is asymmetric so that visual leads are better tolerated than auditory leads for a synchronous percept (van Wassenhove et al., 2007).

Humans, like other organisms, face a series of multimodal stimuli, varying in structural (Simoncelli & Olshausen, 2001) and temporal (Pollack, 2001) patterns, sampled in different time-

scales (Kiebel, Daunizeau, & Friston, 2008; Poeppel, 2003). It has been long proposed that the sensory systems seek to reduce the redundancy of incoming information for the sake of efficient neural coding (Attneave, 1954; Barlow, 1961; MacKay, 1956), a trait with an evolutionary value (Simoncelli, 2003). Hence, the brain is fine-tuned to efficiently encode the natural statistics of sensory inputs (Chandrasekaran et al. 2009; Fiser and Aslin 2001; Lungarella and Sporns 2006), refining unconscious inferences. Furthermore, one influential theory supposes that the brain is a hierarchically inferential system in which feedback predictions from higher-order areas (top-down), are compared with feedforward (bottom-up) sensory evidence with this comparison resulting in a residual-error (Friston, 2005). At the neurophysiological level, these processes can be orchestrated by the interaction of slow (top-down) and fast (bottom-up) intrinsic cortical rhythmic activity [namely cortical oscillations: delta (1 - 3 Hz), theta (3 - 7 Hz), alpha (8 – 12 Hz), beta (12 – 30 Hz) and gamma ( >30 Hz)] (Arnal & Giraud, 2012). Likewise, the interplay between low-frequency and high-frequency oscillations underpin speech comprehension (Giraud & Poeppel, 2012; Gross, Hoogenboom, et al., 2013; Kösem & van Wassenhove, 2017) and audiovisual speech perception (Arnal, Wyart, & Giraud, 2011) through the decoding of incoming information at multiple time-scales (Donhauser & Baillet, 2019; Poeppel, 2003).

Due to the nature of speech production in natural conversational settings (Fitch, 2000; Ghazanfar & Takahashi, 2014), mouth movements are correlated with auditory speech. Specifically, the visual speech precedes phonation for an average of 100-300 ms in case of consonant-vowel sequences ([pa], [ta]) (Chandrasekaran et al. 2009; Grant and Seitz 2000) whereas a range of asynchronies is found in more ecologically plausible series of syllables (Schwartz & Savariaux, 2014). This raised the hypothesis that the visual precedence could serve as an initial prediction for the upcoming auditory input. Numerous studies addressed this question, investigating the neurophysiological influence of visual speech on auditory evoked responses [namely event-related-potentials (ERPs)] (Bernstein et al. 2008; Besle et al. 2004; Irwin et al. 2018; Jääskeläinen et al. 2004; Karas et al. 2019; Pilling 2009; Simon and Wallace 2018; Wassenhove, Grant, and Poeppel 2005; and a recent meta-analysis: Baart 2016;). The central finding condenses into a reduced amplitude and earlier latency of the auditory evoked responses with concurrent visual speech. This effect was proportional to the reliability of the visual signal so that the more predictive the visual speech, the faster and the smaller the auditory evoked response (Arnal et al., 2011; Wassenhove et al., 2005).

A feasible neural mechanism explaining the aforementioned predictive visual modulation originates from invasive electrophysiological recordings and illustrates a phase-(re)setting of ongoing low-frequency oscillations to the presence of visual cue (Kayser, Petkov, & Logothetis, 2008; Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007; Mégevand et al., 2019; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). This way auditory inputs arrive at a high-excitability state, optimizing their processing. Furthermore, Electro- and Magneto-encephalographic (M/EEG) measurements have shown that audio-visual speech entrains cortical oscillations in the auditory cortex (Luo, Liu, & Poeppel, 2010; Power, Mead, Barnes, & Goswami, 2012). Park et al. (2016) extended these results showing that the motor cortex tracks lip-movements and thus guides this entrainment, in line with the Motor Theory of Speech Perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Moreover, it has been shown that the phase of theta oscillation before the presentation determined the speech percept of an ambiguous syllable (ten Oever & Sack, 2015). Although, as attention affects speech tracking (Ding, Chatterjee, & Simon, 2014; O'Sullivan et al., 2015), it is unclear if asynchronies deviating the naturally occurring audiovisual speech asynchronies (Chandrasekaran et al., 2009; Schwartz & Savariaux, 2014) would yield a phase-resetting of auditory response, as attentional resources dedicated to the audition could reduce the bias originating from visual modality (Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010).

In the current study, we seek to further disentangle the neuropsychological effect of audio-visual speech asynchronies in audiovisual speech perception. For this reason, we acquired magnetoencephalographic (MEG) data during the presentation of ecologically-valid audiovisual syllables ([pa] and [ta]) while we systematically deviated the natural visual-to-auditory asynchronies (stimulus onset asynchronies; SOAs). At first, in the ERP analysis, we expect to replicate previous work (Simon & Wallace, 2018), in which an asymmetric temporal window of integration will be depicted across the amplitude of the ERP responses, with the maximum suppression occurring at natural audio-visual speech asynchronies. Then, we scope to determine the functional oscillatory networks at different oscillatory regimes (delta, theta, alpha, beta, gamma, and high gamma) underpinning the identification of audiovisual speech asynchronies, deviating from the naturally occurred ones, while estimating phase modulations for low-frequency oscillations (delta, theta).

## Material & Methods

### *Participants*

Fourteen volunteers took part in the study. All had normal, corrected-to-normal vision, and normal hearing. All participants provided written informed consent prior to taking part in the experiment, in accordance with the Declaration of Helsinki (2008) and the Ethics Committee on Human Research at NeuroSpin (Gif-sur-Yvette, France).

### *Stimuli*

The experiment was written in MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) with PsychToolbox (version 3.0.11). Audiovisual, visual and audio stimuli were extracted from videos of a female speaker pronouncing the syllables [ta] and [pa]. For each syllable, a video lasting 1320 ms (at a digitization rate of 50 Hz) was transformed into 66 frames (1 frame = 20 ms). The acoustic content was extracted as a wav file at 44.1 kHz. Video and sound editing was carried out in Adobe Premiere Pro and Audacity, respectively.

To reduce the visual onset and offset responses, which would inevitably be caused by the sudden appearance and disappearance of the high-contrast faces on a black background, we modified the contrast of the first frame to produce five fade-in frames. The last frame was similarly treated to produce four fade-out frames. A single audiovisual (AV) stimulus consisted of 75 frames in total. In both [pa] and [ta] videos, the onset of the first visible mouth movement occurred at the $10^{th}$ frame *i.e.* 200 ms after the video onset.

In the AV trials, the auditory speech signals were presented at 4 variable timings relative to the naturally measured audiovisual asynchronies (**Figure 1a**, natural audiovisual asynchrony indicated in red). The natural asynchrony for the [ta] syllable was 200 ms so that the auditory utterance occurred 200 ms after the visual speech onset. The natural asynchrony for the recorded [pa] syllable was 600 ms so that the auditory utterance occurred 600 ms after the visual speech onset. Throughout the text, we will consider the natural asynchronies as A0V.

Both [pa] and [ta] were physically desynchronized with respect to their natural asynchronies by the same delays so that the acoustic signal and the first visual movement were

desynchronized by 200, 400, 600 and 800 ms (**Figure 1**). For the [pa] syllable, these four physical timing conditions corresponded to the acoustic speech preceding by 600 ms, 400 ms, 0 ms and lagging by 200 ms the natural speech asynchrony. These asynchronies are mapped to the natural synchrony and will be referred to as A600V, A400V, A0V, and V200A. For the [ta] syllable, the same four possible physical timing conditions corresponded to the acoustic signal preceding by 200 ms, 0 ms and lagging by 400 and 600 ms the natural speech asynchrony. These [ta] asynchronies correspond to A200V, A0V, V400A, and V600A.

Each trial comprised a sequence of 6 identical syllables (standards) presented with an average inter-stimulus interval (ISI) of 400 ms. ISIs were selected from a uniform distribution between 300 and 500 ms. Each sequence ended with the presentation of one of the four possible target syllables: A0V and V400A for [ta], and A0V and A400V for [pa]. The subsequent trial started 1500 ms after the participant's response (**Figure 1b**).

The experiment consisted of three types of blocks. In the first type of block (AV), only audiovisual stimuli were presented. We call a trial a sequence composed of 6 successive and identical syllables (standards) followed by a target syllable. There were a total of 6 AV blocks, each comprising 4 identical trials of each of the 8 possible syllables. This resulted in a grand total of 24 trials per syllable (6 presentations x 24 trials = 144 presentations) across the 6 experimental AV blocks. In the second type of block only auditory [pa] trials were presented: 24 trials of the auditory [pa] syllable were presented with the same regularity as the AV stimuli but with no visual input (black screen). In the third type of block, only visual [pa] syllables were presented: 24 trials of the visual [pa] syllable were presented with the same regularity as the in the other blocks. As in the AV blocks, each trial in the audio alone and video alone blocks ended with the presentation of a target stimulus which was identical or different from the previously presented sequences. Only AV blocks were analyzed, as the audio and visual-only blocks served as distractors in the experimental design, enabling participants to rest their ears and eyes during the MEG acquisition.

All audiovisual asynchronies and trigger timing were carefully checked with a photodiode, microphone, and oscilloscope measurements to insure minimal temporal variance across presentations and trigger-stimulus delay steadiness for the entire time of the MEG acquisition (all below 5 ms).

***Procedure***

Participants were seated comfortably in an upright position under the MEG dewar located in a magnetically shielded room. Visual stimuli were presented on a projection screen located about 1 m away from the participant and auditory stimuli were presented via ear-plugs (Etymotic Research Inc., USA). The experiment started with the presentation of a practice run in which two AV trials were presented. The instructions on the screen requested participants to attend to the presented stimuli. Participants were instructed to use the response box and to answer by button press whether the final target stimulus was the same or different from the preceding stimuli in the trial. When it was clear that they understood the instructions, the first block was presented. Halfway through each block, listeners were offered the opportunity to take a short break, which they could terminate with the press of a button. In total, participants were recorded for 8 blocks (6 AV blocks, 1 audio alone, and 1 visual alone block). For each participant, the visual and audio alone block alternated between the 4th and 8th positions.

***MEG Recordings***

Data were recorded using the Elekta Neuromag Vector View 306 MEG system (Neuromag Elekta LTD, Helsinki system), which comprises 306 sensors (102 magnetometers, 204 orthogonal planar gradiometers). Seven electrodes were used to record electrocardiographic (3 electrodes, ECG), and vertical and horizontal electrooculographic (4 electrodes, EOG) signals. A 3-dimensional Fastrak digitizer (Polhemus, USA) was used to digitize the position of three fiducial head landmarks (Nasal and Pre-auricular points). Four head-position coils were used as indicators of head position in the MEG helmet to help with the coregistration with MRI data. The sampling rate for the MEG acquisition was set to 1 kHz with a band-pass filter of 0.03 to 330 Hz.

***MEG preprocessing***

Data were preprocessed with the MNE-python toolbox (Gramfort et al., 2014) in accordance with accepted guidelines for MEG research (Gross, Baillet, et al., 2013). Noisy MEG sensors were identified manually and interpolated during Signal Space Separation (SSS). The head position recorded at the beginning of each block was used to transform the signal to a standard head position by aligning head position across trials. Artifacts generated from eye blinks and

heartbeats were isolated and removed automatically from raw MEG signals using Independent Component Analysis (ICA), matching their activity with EOG and ECG signals. Prior to epoching, data were low-passed filtered at 120 Hz and downsampled at 500 Hz. Cortical surfaces, inner and outer skull surfaces were reconstructed from individual MRI with Freesurfer (http://surfer.nmr.mgh.harvard.edu) and individual binary element models (BEM) were calculated. Cortical surfaces were extracted from FreeSurfer and projected to 5120 vertices with 4.9mm spacing per hemisphere. The inverse solution was computed using the dynamic statistical parametric mapping (dSPM; Dale et al., 2000) inverse operator with a loose orientation constraint (loose = 0.2, depth = 0.8) and with a source covariance matrix estimated from the baseline activity.

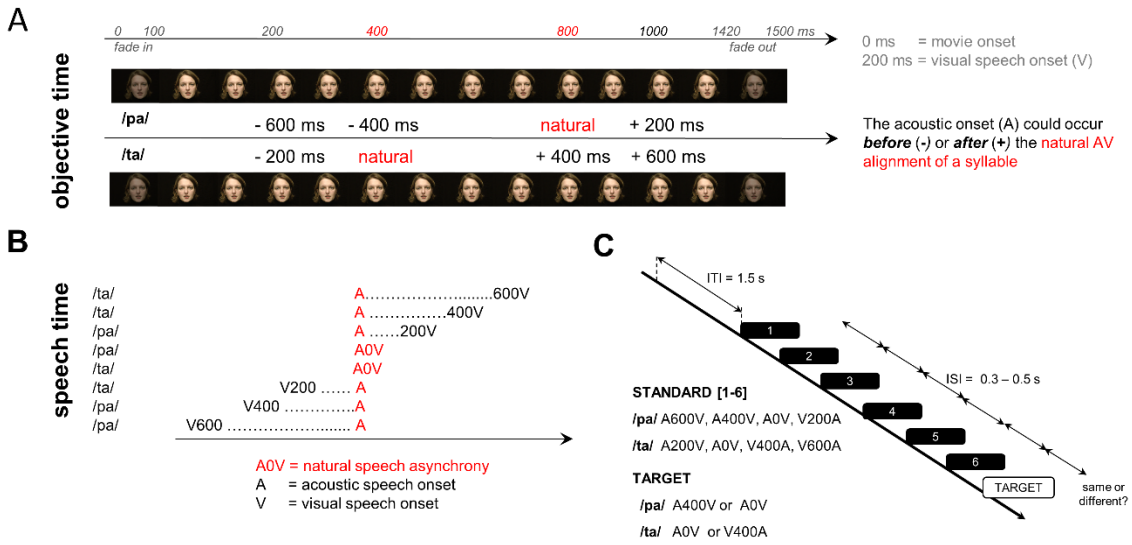


**Figure 1: Stimuli and experimental design. Panel A**: All stimuli were natural audiovisual speech syllables [pa] or /ta/. Audiovisual (AV) stimuli were naturally synchronized (red, « A0V ») or desynchronized in time. In the naturally synchronized conditions, the plosive "p" of the /pa/ syllable occurred 600 ms after the first visible visual gesture; in the /ta/ syllable, the plosive "t" occurred 200 ms after the first visible facial gesture. **Panel B:** Four possible AV asynchronies were then constructed for each syllable yielding 8 experimental asynchronies aligned to naturally occurring AV speech asynchronies. **Panel C**: One trial lasted 14 s and consisted of the sequential presentation of six identical AV stimuli chosen among the two possible syllables (/ta/ or /pa/) and the four possible asynchronies of each syllable. The seventh and last stimulus was the target stimulus. Importantly, the physical AV timing of the desynchronized /pa/ matched the natural asynchrony of /ta/ and, conversely, the physical AV timing of the desynchronized /ta/ (V400A) matched the natural synchrony of /pa/. The inter-stimulus-interval (ISI) was 400 ms; the inter-trial-intervals were 1.5 s.

### *Event-Related-Fields (sensor space)*

The recorded data were separated into epochs ranging from - 600 ms to 600 ms after each auditory event (auditory epochs), and from -200 to 1600 ms after the first frame of the visual input (visual epochs). Baseline correction was applied using the -100 ms to 0 ms interval. Epochs containing signals exceeding 7 fT/cm for gradiometers and 7000 fT for magnetometers were considered as artifacts and rejected from the analysis. Evoked responses were generated individually, averaging from an equal number of epochs across all AV blocks for each experimental condition, ensuring an equal signal to noise ratio for each condition. The analysis was restricted to magnetometers, for simplicity of topographical interpretation.

To investigate the effect of audiovisual speech asynchronies subject-wise linear regressions between the amplitude of the evoked response for each trial, time-point, and sensor were estimated. Linear regressions were estimated separately for estimating the linear effect of the audiovisual

speech asynchronies respective and irrespective of the syllable. Afterwards, the corresponding regression coefficients (β values) were submitted to a cluster spatiotemporal one-sample t-test corrected for multiple comparison to identify the significant clusters of sensors and latencies.

***Functional connectivity analyses***

*Synchronized oscillatory activity in delta, theta, alpha, beta, gamma, and high gamma band*

First, we were interested in estimating the pairwise phase synchronizations between cortical labels across different frequency bands. For this, we calculated the weighted phase-lag index (wPLI; Vinck et al. 2011). wPLI has been found to be a robust estimator of phase synchronization exhibiting increased sensitivity of detecting phase-synchronization and changes compared to other phase-synchronization metrics (Niso et al., 2013). Visual epochs were used for the estimation of wPLI in single-trial source estimates within experimental conditions and across participants. Frequency-bands of interest were: delta (δ: 1 – 3 Hz), theta (θ: 4 – 7 Hz), alpha (α: 8 – 12 Hz), beta (β: 13 – 30 Hz), gamma (γ: 30 – 40 Hz) and high gamma (hγ: 40 – 120 Hz). Thus, 46 x 46 adjacency matrices were created for each condition and each frequency band, each one depicting a functional network.

To investigate the effect of audiovisual asynchronies in different oscillatory networks, we proceeded with a network-level statistical analysis using the Network-Based-Statistics toolbox (Zalesky, Fornito, & Bullmore, 2010). Specifically, a 2 x 4 mixed-model ANOVA with two within-subject factors syllable ([pa] and [ta]) and asynchrony (SOA: 0, 200, 600, and 800 ms) was designed for assessing the main effect of asynchrony. The significance level was set to $p < 0.05$, corrected for multiple comparisons with FDR correction (Benjamini & Yekutieli, 2001).

Within each statistically significant oscillatory network (δ, θ, α, β, γ, and hγ) we calculated the node strength, namely the sum of the weights (*i.e*, F-values) of connections (or edges) which each node carried. Subsequently, a 46 x 6 adjacency matrix was constructed (46 node strengths for each frequency band) and treated as a bipartite network. The node strength of a node within a network depicts the sum of incoming and outgoing connections that each node contributes to the network. Accordingly, nodes in bipartite networks are divided into two groups (*i.e*, node strength and frequency band in our case) and connections are drawn only between those. The temperature of the matrix was calculated as a metric of bipartite nestedness with the function *nestedtemp* of

vegan package in R (Oksanen et al., 2013; R Core Team, 2018). Nestedness in bipartite networks is a measure originating from community ecology, quantifying the functional structure within interacting systems (*i.e.*, in our case between cortical labels and frequency bands).

*Directed-phase information transfer in the delta band*

Second, we proceeded with the estimation of Phase Transfer Entropy (Franchini & Hillebrand, 2017; Lobier, Siebenhühner, Palva, & Palva, 2014) between the single-trial activity of 46 functional cortical labels. As we were interested in the phase modulation of the induced activity, we used auditory epochs and subtracted evoked activity. Transfer Entropy (Schreiber, 2000) quantifies causal statistical dependencies between two signals. More specifically, TE measures to what extent knowledge of signal *X* can reduce uncertainty (as quantified by Shannon's Entropy) in predicting the future of signal *Y*, beyond the degree to which *Y* predicts its own future. Conceptually, TE aligns with Granger Causality (Granger, 1969), comparing conditional nonlinear probabilities distributions via Kullback-Leibler divergence for the estimation of statistical dependencies. Furthermore, pTE extends this notion into phase time-series, evaluating the influence of signal's *X* phase on the signal's *Y* phase and thus, detecting phase information transfer. Hilbert-transformation was applied for the extraction of the instantaneous phases and Kullback-Leibler divergence was measured from phase distributions with a number of bins adjusted according to Scott (1979) while time-delay was set as the average pairwise time needed for a sign-flip. Single-trial 46 x 46 adjacency matrices were averaged per condition individually, resulting in 8 adjacency matrices per participant. Thereafter, the statistical analysis was similar to the aforementioned.

***Time-Frequency Analyses***

Auditory epochs were considered for the time-frequency analysis. For this scope, a Morlet wavelet transform was applied in single-trial source estimates from 1 to 120 Hz. For adjusting the trade-off between temporal and frequency precision, the number of cycles for the wavelets was adapted as a function of frequency (number of cycles = $f / 2$ ). Baseline correction was applied to power and inter-trial coherence (ITC) dividing the mean activation of the baseline activity (-200 to 0 ms) and taking the log. The log-transformed data were normally distributed allowing the use

of parametric statistics: the statistical analysis of power and ITC comprised a 2 x 4 ANOVA with factors of 'what' (syllable: /pa/ and /ta/) and 'when' (SOA: 0 ms, 200 ms, 400 ms, and 600 ms). Only the main effect of 'when' is reported, as neither the main effect of 'what', nor the interaction of 'when' and 'what' yielded significant results.

## Results

### *The amplitude of evoked responses linearly tracks audio-visual speech asynchronies*

First, we asked whether the brain linearly track physical asynchronies (*i.e*, the "when"), irrespective of the natural audiovisual speech timing of the syllable (*i.e,* the "what"). We proceeded with a subject-wise linear regressions between the single-trial amplitude of each sensor and time-point using the SOAs as a single regressor. Subsequently, we performed a spatiotemporal cluster permutation one-sample t-test on the beta values obtained for each sensor and time point per participant for the identification of clusters (sensors x time points) that significantly differed from zero. We found two early significant clusters (20 magnetometers, 47-246 ms, $p<0.01$; 22 magnetometers, 15-436 ms, $p<0.01$), suggesting that participants could track the objective AV asynchronies in a linear manner (Fig.2A-B, left panels).

We then asked whether participants could track the distance between the presented visual speech and the natural audiovisual speech timing. Here, our working hypothesis was that an internal speech model would rely on the natural statistics of AV speech to generate an adequate prediction of the timing of auditory occurrence for [pa] or for [ta]. Hence, we predicted that the larger the deviance from the natural audiovisual speech timing, the larger the residual errors would be in the evoked responses. To address this question, we used the same regression analysis approach, this time using a single regressor capturing the distance from natural AV speech asynchronies. We found three significant clusters (50 magnetometers, -7 - 169 ms, $p<0.01$; 25 magnetometers, 141 – 454 ms, $p<0.01$; 37 magnetometers, 8 – 378 ms, $p <0.01$) indicating that participants could successfully extract the presented audiovisual distance from the natural AV-asynchronies for both syllables (Fig.2A-B, right panels).

We then used the same regression analysis within the single-trial activity of each syllable ([pa] and [ta]) to investigate differences in sensors and latencies. We found two significant clusters for [pa] syllable (22 magnetometers, 36 – 428 ms, $p<0.01$; 27 magnetometers, 84 – 184 ms, $p<0.01$; Fig. 2C, left panel) and two significant clusters for [ta] syllable (12 magnetometers, 141 – 244 ms, $p<0.01$; 12 magnetometers, 153 – 222 ms, $p<0.01$; Fig. 2C, right panel). Asynchronies in the syllable [pa] elicited broader latency differences (36 – 428 ms) as compared to [ta] syllable (141 – 244 ms).

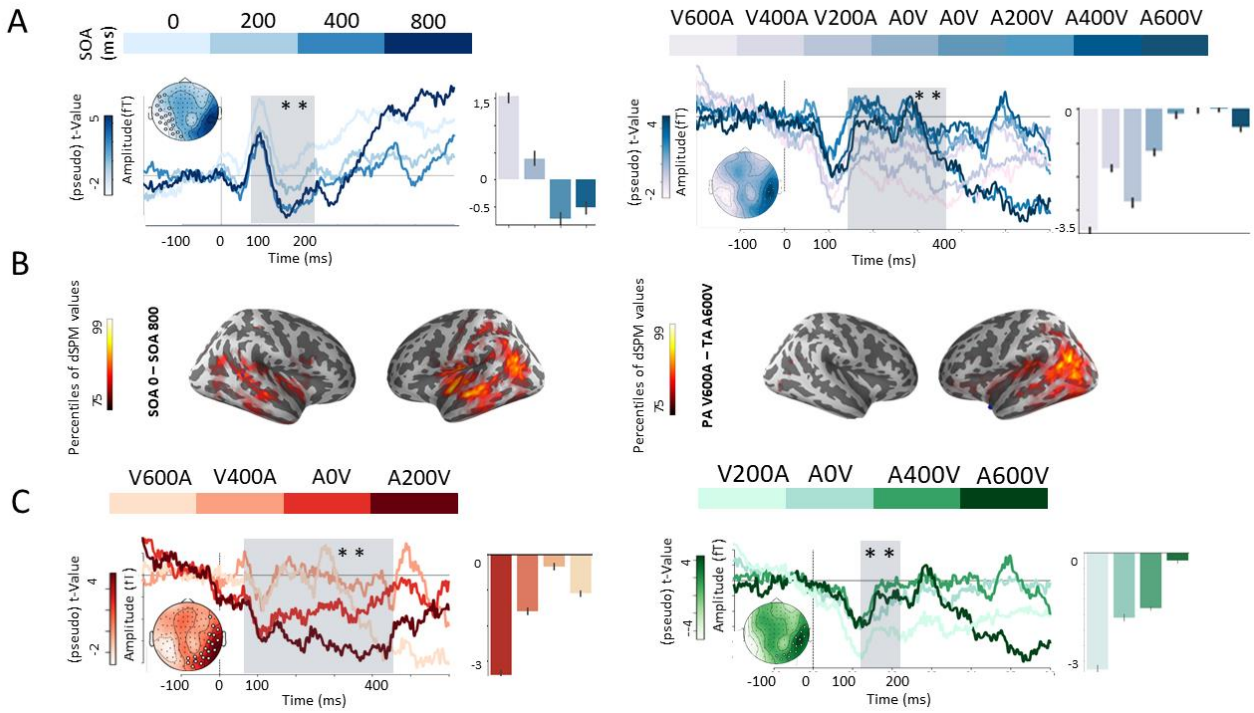


**Figure 2: Linear tracking of AV speech asynchronies post-auditory speech onset. Panel A**: The evoked responses elicited by the different four different AV asynchronies irrespective of, or as a function of the syllable (left and right panels, respectively). In both sorting, a significant graded amplitude differences (shaded areas) was found following the auditory speech onset so that the closer to natural AV speech synchrony, the smaller the amplitude of the evoked response. **Panel B**: Source reconstructed activity of the grand average evoked responses elicited by the differences between two extreme SOA conditions (left panel: SOA 0 ms and 800 ms) showing bilateral engagement of auditory cortices and a large sections of the posterior and middle Superior Temporal Sulci. The differences elicited by the contrast between PA V600A and TA A600V (right panel) thereby combining the what and when differences are localized in a large posterior temporal regions likely including visual motion areas. **Panel C**: The amplitude of the evoked responses elicited by the different AV asynchronies for the syllables [pa] (left) and [ta] (right). ** $p<0.01$ corrected.

***Synchronized oscillatory networks: nested networks underpin audiovisual speech asynchronies***

To identify possible oscillatory networks sensitive to audiovisual speech asynchronies, we performed a 2 x 4 mixed model ANOVA in the network-level with within-subject factors syllables ([pa] and [ta]) and SOA (0, 200, 400, 800 ms) assessing the main effect of the factor 'time'. The analysis revealed statistically significant networks ($p < 0.05$, FDR corrected; **Fig. 3A**) for delta (1 – 3 Hz), theta (4 – 7 Hz), alpha (8 – 12 Hz), beta (12 – 30 Hz), gamma (30 – 40 Hz), and high gamma (60 – 120 Hz) frequency bands. Specifically, the network consisted of 18 edges in the delta band, 13 edges in the theta band, 5 edges in the alpha band, 10 edges in the beta band, and 6 edges in the gamma band, and 3 edges in the high gamma band. Subsequently, in each significant oscillatory network, we calculated the node degree for each node, and we constructed a 46 x 6 matrix (46: node degrees; 6: frequency bands) which we treated as a bipartite network. The aforementioned bipartite network was significantly nested when compared to 1000 surrogate networks with the same characteristics ($t = 7.99$; $p = 0.01$). Moreover, according to the node's strength, we proceeded with the hierarchical clustering of the adjacent nodes and frequency bands, generating a cluster-map. As shown (**Fig. 3C**) we found that higher-frequency oscillatory regimes (gamma and high-gamma) were spatially nested within slower-frequency oscillatory networks (delta and theta).

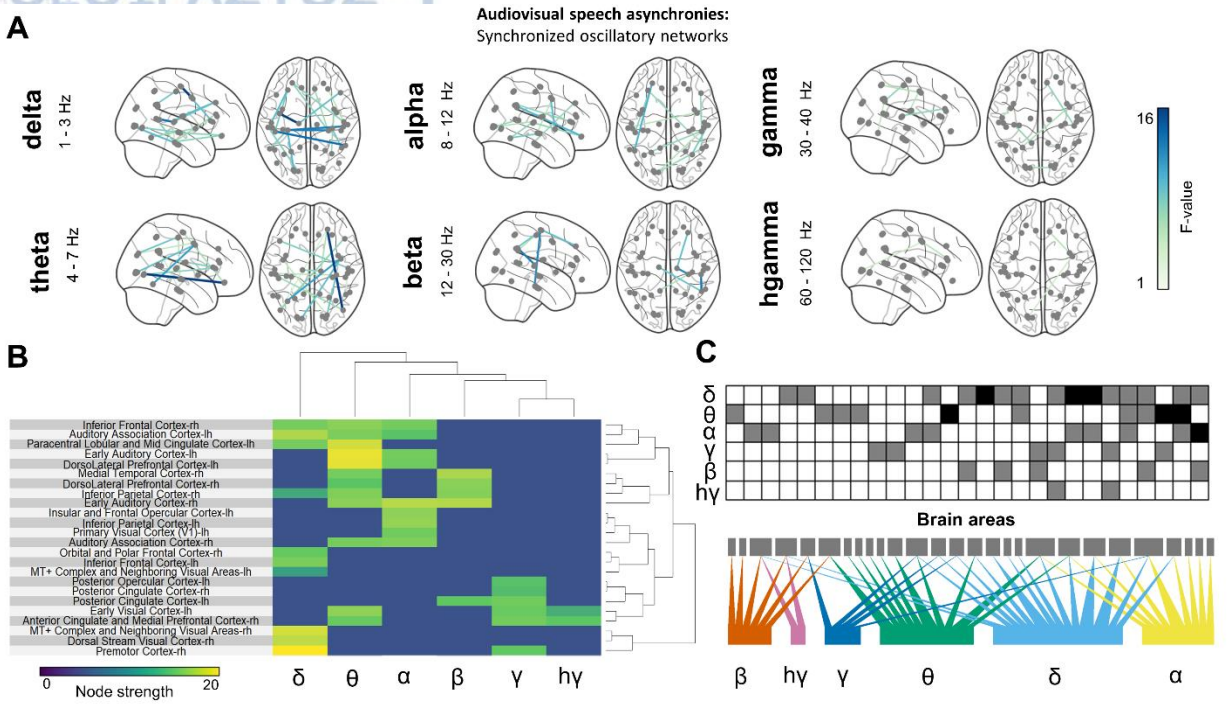


**Figure 3*:* Nested oscillatory networks underpin audiovisual speech asynchronies. Panel A**: Functional networks implicated in the identification of audiovisual speech asynchronies (combining [pa] and [ta]) in delta (1 - 3 Hz), theta (4 - 7 Hz), alpha (8 - 12 Hz), beta (12 - 30 Hz), gamma (30 - 40 Hz), and high gamma (60 - 120 Hz) regimes. The source activity was extracted from 46 functional labels and pairwise synchronization was calculated using the weighted Phase-Lag Index (wPLI). Statistically significant connections were drawn at $p < 0.05$ level (FDR corrected for multiple comparisons). **Panel B**: Hierarchical clustering of the frequency bands and cortical areas within the bipartite network. In bipartite networks, connections are drawn between two sets of nodes (brain area and frequency band) and higher-frequency oscillations (gamma and high gamma) can be seen here to be spatially nested within lower-frequency oscillations (theta and delta). **Panel C:** The functional networks in different oscillatory regimes (delta, theta, alpha, beta, gamma, and high gamma) as a bipartite network. The darker the square, the more common nodes shared between brain areas. The split of shared connectivity among oscillatory regimes is provided in the bottom panel.

### *Direction of information in the phase of delta band during audiovisual speech asynchronies*

As we observed that higher-frequency oscillatory networks were nested within the slower-frequency networks in the bipartite network (**Fig. 3**), we were further interested in quantifying the causal pairwise interactions of phase information in δ and θ bands. We calculated the phase transfer entropy between induced source estimates at the single-trial level in 46 functional cortical labels, per individual and per experimental condition. The network-level statistics were similar to the aforementioned. We focus solely on the results of the δ band considering that the θ band did not yield statistical significance. The δ band showed a significant ($p < 0.05$, FDR corrected) functional network depicting the directed-transfer of information in cortical phase space (**Fig. 4A**). The

network consisted of 36 edges and the supramarginal gyrus / angular gyrus (SMG / AG) region was the central hub within the network, as revealed by its node strength (node strength = 34.6). The Superior Temporal Gyrus appeared phase-modulated from the frontal regions. This resuls indicates the central role of the SMG / AG in the identification of desynchronized audiovisual speech through possible excitability changes in the δ band.

***Entrained activity of the left Supramarginal gyrus and Superior Temporal gyrus with asynchronous audiovisual speech***

To investigate the δ oscillatory response elicited by desynchronized audiovisual speech, we extracted the phase of induced δ activity (source estimates) on a per trial basis and averaged the epochs for each experimental condition across participants. Based on our previous analysis, the regions of interest were the SMG and the STG. As shown in **Fig. 4B** (top row), the SMG showed a significant entrained oscillatory response to the natural audiovisual asynchrony (A0V) and audiovisual desynchrony (A600V) during the presentation of the [pa] syllable; no other asynchronies (V400A, A200V) or syllable [ta] showed a significant induced oscillatory response. To the contrary, the STG (**Fig. 4B** bottom row) showed a significant δ response for all syllables and desynchronies. Thus, during the presentation of audiovisual [pa], the excitability of induced oscillatory activity can fully differentiate the natural (A0V) from the fully desynchronized (A600V) conditions and all other experimental desynchronies (V400A, A200V), indicating optimal excitability - before the onset of the sound - of the auditory cortex during the natural audiovisual speech asynchrony. This was observed for [pa] but not for [ta]

***Time-Frequency representation***

Time-Frequency responses were calculated from the induced single-trial source activity of the SMG and STG (**Fig. 4C** left panel). The statistical analysis of the TFRs for the whole auditory epochs yielded significant changes in power as a function of AV speech asynchronies in the two regions of interest. Specifically, SMG showed a significant cluster in the lower β range (14 – 18 Hz) preceding the auditory speech onset by 500 ms to 350 msn, and subsequently, in the θ range (5 – 8 Hz), 350 ms to 200 ms preceding the onset of the auditory speech. In the STG, the statistical

analysis revealed three significant clusters. One cluster was again located in the low-β range (12 – 16 Hz), at a latency spanning 500 ms to 350 ms prior to the onset of auditory speech. A second cluster located in the high-β range (25 – 27 Hz) around 500 ms to 400 ms before auditory speech onset. The third cluster in the low-β range (14 – 18 Hz) was found 100 ms to 400 ms following the presentation of the auditory syllable. All clusters were significant at the level of $p < 0.01$ (cluster-level corrected). We further compared the power within significant clusters across AV speech asynchronies (**Fig. 4C**, right panel): in the SMG, θ power was higher in the [pa] syllable during V400A asynchrony and during natural asynchrony (A0V) for the [ta] syllable. Furthermore, β power decreased as the sound was deviating away from the objective synchrony, both for syllables [pa] and [ta]. In the STG, the high-β power preceding the auditory speech onset decreased when the sound deviated away from the objective audiovisual synchrony, in contrast with the low-beta activity after the presentation of the sound, which yielded increased activity.

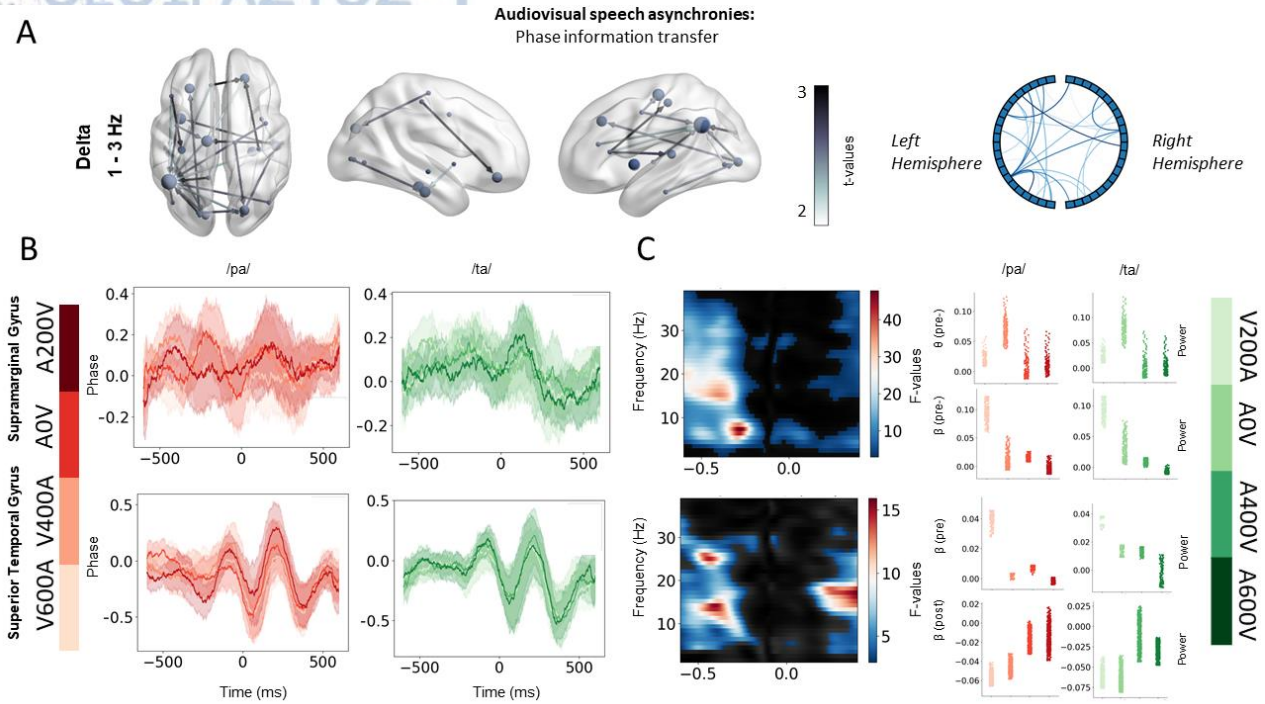


**Figure 4: Auditory induced activity of asynchronous audiovisual speech. Panel A**: Causal phase interactions of auditory speech in the delta band (1 - 3 Hz) during visual speech asynchronies (V600A, V400A, V200A, A0V, A200V, A400V, A600V) with respect to the auditory speech ([pa] and [ta] are combined). Single-trial source activity was extracted from 46 cortical functional labels and band-pass filtered in the delta (1 - 3 Hz) range. Phase-time series were estimated with Hilbert-transform and pairwise phase interactions were estimated with phase transfer entropy algorithm. Statistically significant connections were drawn at $p < 0.05$ (FDR corrected). The color scale indicates t-values. The size of the nodes depicts the node strength (i.e., the sum of weights for incoming and outgoing connections) within the network. **Panel B**: Grand average of phase-time series during visual speech asynchronies for /pa/ (left; red) and /ta/ (right; green) in left Supramarginal Gyrus (Upper) and left Superior Temporal Gyrus (Upper). **Panel C**: Main effect of AV asynchronies in Time-Frequency Representation (Left) of left Inferior Parietal Cortex (Upper) and left Auditory Cortex (Lower). Statistically significant clusters are drawn at $p < 0.05$. The power of statistically significant clusters during visual speech asynchronies for /pa/ (right; red) and /ta/ (right; green).

## Discussion

In the present study, we sought to determine the neurophysiological indices of asynchronous audiovisual speech. For this, magnetoencephalographic data were recorded from participants attending a stream of syllables ([pa] and [ta]) whose auditory input was systematically deviated from natural audiovisual asynchronies (A0V). Firstly, we've replicated previous work (Baart, 2016; Simon & Wallace, 2018; Wassenhove et al., 2005) showing suppression of the ERP signal in the precedence of visual speech. Moreover, the amplitude of the ERP response was linearly-predicted during the various audiovisual asynchronies and this modulation was evident for objective audiovisual asynchronies and for the temporal deviance from natural audiovisual asynchronies. We also show that the detection of audiovisual speech asynchronies is underpinned by spatially nested oscillatory networks, in which the gamma and high gamma synchronized activity were orderly nested throughout the cortex within delta and theta regimes. Moreover, we report that audiovisual speech asynchronies are underpinned by causal modulations in the excitability of ongoing oscillations in the delta range, within which SMG is the central –incoming-hub. Finally, we show an entrained activity of the STG to the precedence of visual speech, irrespective of a temporal concordance.

### *Tracking audiovisual speech timing in the evoked response*

We show an early modulation of auditory evoked response as a function of audiovisual speech asynchronies. This was evident when tracking objective audiovisual asynchronies (SOA 0, SOA 200, SOA 400, SOA 800) as well as asynchronies deviating the naturally occurred audiovisual asynchronies in both experimental syllables. This accounts for evidence of internalized predictions of the natural temporal statistics (Nobre, Correa, & Coull, 2007) in audiovisual speech, depicted in the auditory ERP response. The effected time-window was found around the P2 component, in which visually-induced suppressions are not confined to speech (Vroomen & Stekelenburg, 2010). Although, it has been found to dissociate speech to non-speech audiovisual incongruencies (Baart, Stekelenburg, & Vroomen, 2014), indicating that audiovisual speech integration may be grounded on the interaction of speech specific and domain-general multisensory components (Eskelund, Tuomainen, & Andersen, 2011).

As expected, the earlier the visual speech onset occurred, the less predictive, and thus the larger the subsequent auditory response was observed. The neural distribution of responses was also asymmetric, matching the typical profile of temporal window of integration found previously in behavioral (van Wassenhove et al., 2007) and neurophysiological studies (Simon & Wallace, 2018). Interestingly, the modulation of visual speech lasted longer (~ 400 ms), corresponding to a low-frequency cycle (~ 3 Hz), in the case of the highly informative viseme [pa]. These low-frequency components match the temporal correlation (2 – 6 Hz) of mouth movements and sound envelopes in CV syllables (Chandrasekaran et al. 2009). Likewise, it fits with previous findings showing that in naturalistic audiovisual speech streams the low-frequency components track visual and auditory information (Luo et al., 2010) while modulating audiovisual speech integration at longer temporal windows (Crosse, Di Liberto, & Lalor, 2016).

***Nested oscillatory networks underpin the tracking of audiovisual speech asynchronies***

The interaction of oscillatory activity in different frequency bands has been hypothesized to coordinate neural activity (Jensen & Colgin, 2007). In these temporal scales, those interactions encapsulate phase-phase coupling (J. M. Palva, Palva, & Kaila, 2005; S. Palva & Palva, 2012; Tass et al., 1998), amplitude-amplitude coupling (de Lange, Jensen, Bauer, & Toni, 2008; Helfrich, Huang, Wilson, & Knight, 2017), and phase-amplitude coupling (R. T. Canolty et al., 2006; Ryan T. Canolty & Knight, 2010; Tort, Komorowski, Eichenbaum, & Kopell, 2010). The nesting of high-frequency oscillations amplitude within the phase of low-frequency has been suggested to underpin long-range communication (Akam & Kullmann, 2014) with implications in working memory (Axmacher et al., 2010; Lisman & Jensen, 2013; Roux & Uhlhaas, 2014), speech processing (Giraud & Poeppel, 2012) and more recently, the temporal precision of information (Arnal, Doelling, & Poeppel, 2015; Grabot et al., 2019). Here, we extend those findings, showing that the identification of audiovisual speech asynchronies is supported by long-range oscillatory networks, hierarchically nested from higher- to lower-frequencies within cortical areas.

Even though the concept of bipartite networks flourishes in other research areas (Costa, Rodrigues, Travieso, & Villas Boas, 2007), to our knowledge, this is the first report of cortical oscillatory activity described as a nested bipartite network. For example, in community ecology the interactions of plants and pollinators have been proposed to form a nested bipartite (plants x

pollinators) network (Jordano, Bascompte, & Olesen, 2003; Petanidou, Kallimanis, Tzanopoulos, Sgardelis, & Pantis, 2008) in which nested structures depict the interactions between specialists (*i.e.* species that interact with few others) and generalists (*i.e.* species that interact with many others). This partitioning of interactions contribute to a generalist-specialist balance while distinguishing mutualistic from antagonistic interactions (Corso, De Araujo, & De Almeida, 2011; Lewinsohn, Inácio Prado, Jordano, Bascompte, & M. Olesen, 2006). In our paradigm, delta and theta oscillatory rhythms are the "generalists" operating on longer integrative time-scales and dominating the interactions with other cortical areas. On the contrary, gamma and high gamma regimes are interacting with fewer - and a subset of - cortical areas at faster time scales and within the delta and theta networks. Our interpretation stands within the predictive coding framework (Friston, 2005), speculating that the identification of audiovisual asynchronies is based upon bottom-up evidence (analyzed in high-frequency activity) and which refine internal predictions acting in slow-frequency oscillations in constrained cortical space.

***Entrainment of Supramarginal gyrus and Superior temporal gyrus during audiovisual speech asynchronies***

The functional network of causal phase-interactions in the delta range unfolded the low-frequency (1 – 3 Hz) phase-modulations within cortical areas, underpinning identifications of temporal incongruencies during audiovisual speech. Within this network, the supramarginal gyrus / angular gyrus (SMG / AG) is the main hub, with connections from frontal, parietal and occipital areas. Previous research has underlined the role of SMG / AG in audiovisual speech perception (Jones & Callan, 2003; Kaiser, Hertrich, Ackermann, & Lutzenberger, 2006). Moreover, Bernstein et al. (2008) using EEG and spatiotemporal analysis of ERPs found that SMG / AG was significantly activated during congruent audiovisual speech perception, with a temporally-broad activation within the auditory response. Here, we provide evidence that this activation is driven by cortical effective interactions, modulating the phase of intrinsic low-frequency oscillations in the area. For corroborating this notion, the SMG / AG showed entrained oscillatory activity during the presentation of objectively synchronous (A600V) and naturally asynchronous (A0V) condition around the auditory onset. This is in line with previous which utilized direct comparison of audiovisual speech streams and cortical responses (Temporal Response Function; TRF) and found

that -temporally and contextually- congruent audiovisual speech enhances entrainment in cortical areas (Crosse, Butler, & Lalor, 2015).

The ongoing-phase of the Superior temporal gyrus (STG) within the delta range was significantly modulated from the frontal region(s), during the identification of asynchronous audiovisual speech. This is in line with Park et al. (2015) who identified frontal top-down and phase-modulating signals, during the presentation of natural audiovisual speech streams. Interestingly, the ongoing activity of the SMG was found to exhibit an entrained activity of auditory speech during the presentation of various visual speech asynchronies, with a cycle of around 200-300ms. This suggests that, regarding auditory areas, audiovisual speech asynchronies didn't reduce the attentional resources dedicated to the auditory speech (Womelsdorf & Fries, 2007), disregarding thereby the bias induced by the unattended -visual- modality (Talsma et al., 2010). Furthermore, we found a significant decrease in the beta power of STG and SMG / AG before the auditory onset. Previous invasive ECoG study has underlined the role of beta oscillations in updating the content of sensory predictions (Sedley et al., 2016), whereas a decrease in the power of beta-oscillations was associated with the maintenance of predictions (Chao, Takaura, Wang, Fujii, & Dehaene, 2018). We extended those findings, corroborating the notion that beta-oscillations serve as a top-down signal, orchestrating the visual modulation in the auditory response during the perception of audiovisual speech (Arnal & Giraud, 2012; Biau & Kotz, 2018). Lastly, the difference in the latencies of beta power and neural synchronization in the STG serves as an indication of a causal effect of beta power in the entrainment of the STG, although, this hypothesis was not directly tested in the current study.

***Conclusion***

The current data demonstrate that audiovisual speech asynchronies are processed in the cortex with neural entrainment in the scale of around 250ms, matching the temporal scale of speech processing and corroborating the hypothesis that low-frequency cortical oscillations are essential for audiovisual speech integration. Furthermore, it is shown that this entrainment is underpinned by nested oscillatory networks in spatial scale and by effective cortical phase modulations within delta oscillations (1 – 3 Hz).

## References

Akam, T., & Kullmann, D. M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience*, *15*(2), 111–122. https://doi.org/10.1038/nrn3668

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta-beta coupled oscillations underlie temporal prediction accuracy. *Cerebral Cortex*, *25*(9), 3077–3085. https://doi.org/10.1093/cercor/bhu103

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801. https://doi.org/10.1038/nn.2810

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193. https://doi.org/10.1037/h0054663

Axmacher, N., Henseler, M. M., Jensen, O., Weinreich, I., Elger, C. E., & Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(7), 3228–3233. https://doi.org/10.1073/pnas.0911531107

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology*, *53*(9), 1295–1306. https://doi.org/10.1111/psyp.12683

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*(1), 115–121. https://doi.org/10.1016/j.neuropsychologia.2013.11.011

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. *Sensory Communication*, 216–234.

https://doi.org/10.7551/mitpress/9780262518420.003.0013

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under depencency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage*, *39*(1), 423–435. https://doi.org/10.1016/J.NEUROIMAGE.2007.08.035

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234. https://doi.org/10.1111/j.1460-9568.2004.03670.x

Biau, E., & Kotz, S. A. (2018). Lower Beta: A Central Coordinator of Temporal Prediction in Multimodal Speech. *Frontiers in Human Neuroscience*, *12*, 434. https://doi.org/10.3389/fnhum.2018.00434

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., … Knight, R. T. (2006). High Gamma Power Is Phase-Locked to Theta Oscillations in Human Neocortex. *Science*, *313*(5793), 1626–1628. https://doi.org/10.1126/science.1128115

Canolty, Ryan T., & Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, *14*(11), 506–515. https://doi.org/10.1016/j.tics.2010.09.001

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal Inference in the Multisensory Brain. *Neuron*, *102*. https://doi.org/10.1371/journal.pbio.1002075

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

Chao, Z. C., Takaura, K., Wang, L., Fujii, N., & Dehaene, S. (2018). Large-Scale Cortical Networks for Hierarchical Prediction and Prediction Error in the Primate Brain. *Neuron*, *100*, 1–15. https://doi.org/10.2139/ssrn.3188377

Conrey, B., & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *The Journal of the Acoustical Society of America*, *119*(6), 4065–4073. https://doi.org/10.1121/1.2195091

Corso, G., De Araujo, A. I. L., & De Almeida, A. M. (2011). Connectivity and nestedness in bipartite networks from community ecology. *Journal of Physics: Conference Series*, *285*(1), 1–6. https://doi.org/10.1088/1742-6596/285/1/012009

Costa, L. da F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, *56*(1), 167–242. https://doi.org/10.1080/00018730601170527

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

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. *Journal of Neuroscience*, *36*(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67. https://doi.org/10.1016/s0896-6273(00)81138-1

de Lange, F. P., Jensen, O., Bauer, M., & Toni, I. (2008). Interactions between posterior gamma and frontal alpha/beta oscillations during imagined actions. *Frontiers in Human Neuroscience*, *2*, 7. https://doi.org/10.3389/neuro.09.007.2008

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46. https://doi.org/10.1016/J.NEUROIMAGE.2013.10.054

Donhauser, P. W., & Baillet, S. (2019). Two Distinct Neural Timescales for Predictive Speech

Processing. *Neuron*, *0*(0). https://doi.org/10.1016/j.neuron.2019.10.019

Driver, J., & Noesselt, T. (2008). Multisensory Interplay Reveals Crossmodal Influences on 'Sensory-Specific' Brain Regions, Neural Responses, and Judgments. *Neuron*, *57*(1), 11–23. https://doi.org/10.1016/j.neuron.2007.12.013

Erber, N. P. (1975). Auditory-Visual Perception of Speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: dissociating identification and detection. *Experimental Brain Research*, *208*(3), 447–457. https://doi.org/10.1007/s00221-010-2495-9

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, *4*(7), 258–267. https://doi.org/10.1016/S1364-6613(00)01494-7

Franchini, M., & Hillebrand, A. (2017). *Phase Transfer Entropy in Matlab*. Figshare. https://doi.org/doi:10.6084/m9.figshare.3847086.v12

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285. https://doi.org/10.1016/J.TICS.2006.04.008

Ghazanfar, A. A., & Takahashi, D. Y. (2014). The evolution of speech: Vision, rhythm, cooperation. *Trends in Cognitive Sciences*, *18*(10), 543–553. https://doi.org/10.1016/j.tics.2014.06.004

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging

computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Grabot, L., Kononowicz, T. W., Dupré la Tour, T., Gramfort, A., Doyère, V., & van Wassenhove, V. (2019). The strength of alpha-beta oscillatory coupling predicts motor timing precision. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 2473–18. https://doi.org/10.1523/JNEUROSCI.2473-18.2018

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424. https://doi.org/10.2307/1912791

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197. https://doi.org/10.1121/1.1288668

Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., … Schoffelen, J.-M. (2013). Good practice for conducting and reporting MEG research. *NeuroImage*, *65*(100), 349–363. https://doi.org/10.1016/j.neuroimage.2012.10.001

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). https://doi.org/10.1371/journal.pbio.1001752

Helfrich, R. F., Huang, M., Wilson, G., & Knight, R. T. (2017). Prefrontal cortex modulates posterior alpha oscillations during top-down guided visual perception. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(35), 9457–9462. https://doi.org/10.1073/pnas.1705965114

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, 1–14. https://doi.org/10.1016/j.tics.2019.05.006

Irwin, J., Avery, T., Brancazio, L., Turcios, J., Ryherd, K., & Landi, N. (2018). Electrophysiological Indices of Audiovisual Speech Perception: Beyond the McGurk Effect and Speech in Noise. *Multisensory Research*, *31*(1–2), 39–56. https://doi.org/10.1163/22134808-00002580

Jääskeläinen, I. P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R., … Sams, M. (2004). Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *Neuroreport*, *15*(18), 2741–2744.

Jensen, O., & Colgin, L. L. (2007). Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences*, *11*(7), 267–269. https://doi.org/10.1016/j.tics.2007.05.003

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport*, *14*(8), 1129–1133. https://doi.org/10.1097/01.wnr.0000074343.81633.2a

Jordano, P., Bascompte, J., & Olesen, J. M. (2003). Invariant properties in coevolutionary networks of plant-animal interactions. *Ecology Letters*, *6*(1), 69–81. https://doi.org/10.1046/j.1461-0248.2003.00403.x

Kaiser, J., Hertrich, I., Ackermann, H., & Lutzenberger, W. (2006). Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli. *NeuroImage*, *30*(4), 1376–1382. https://doi.org/10.1016/J.NEUROIMAGE.2005.10.042

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., & Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech, 1–19.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, *18*(7), 1560–1574. https://doi.org/10.1093/cercor/bhm187

Kayser, C., & Shams, L. (2015). Multisensory Causal Inference in the Brain. *PLOS Biology*, *13*(2), e1002075. https://doi.org/10.1371/journal.pbio.1002075

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain.

*PLoS Computational Biology*, *4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLoS ONE*, *2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

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. https://doi.org/10.1080/23273798.2016.1238495

Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal Oscillations and Multisensory Interaction in Primary Auditory Cortex. *Neuron*, *53*(2), 279–292. https://doi.org/10.1016/J.NEURON.2006.12.011

Lewinsohn, T. M., Inácio Prado, P., Jordano, P., Bascompte, J., & M. Olesen, J. (2006). Structure in plant-animal interaction assemblages. *Oikos*, *113*(1), 174–184. https://doi.org/10.1111/j.0030-1299.2006.14583.x

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Lisman, J. E., & Jensen, O. (2013). The Theta-Gamma Neural Code. *Neuron*, *77*(6), 1002–1016. https://doi.org/10.1016/j.neuron.2013.03.007

Lobier, M., Siebenhühner, F., Palva, S., & Palva, J. M. (2014). Phase transfer entropy: A novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *NeuroImage*, *85*, 853–872. https://doi.org/10.1016/j.neuroimage.2013.08.056

Lungarella, M., & Sporns, O. (2006). Mapping Information Flow in Sensorimotor Networks. *PLoS Computational Biology*, *2*(10), e144. https://doi.org/10.1371/journal.pcbi.0020144

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biology*, *8*(8), 25–26. https://doi.org/10.1371/journal.pbio.1000445

MacKay, D. M. (1956). Towards an Information-flow model of human behavior. *British Journal of Psychology*, *47*(1), 30–43. https://doi.org/10.1111/j.2044-8295.1956.tb00559.x

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141. https://doi.org/10.3109/03005368709077786

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 245–256. https://doi.org/10.1037/a0019952

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, *100*(3), 1777–1786. https://doi.org/10.1121/1.417342

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Mégevand, P., Mercier, M. R., Groppe, D. M., Golumbic, E. Z., Mesgarani, N., Beauchamp, M. S., … Mehta, A. D. (2019). Phase resetting in human auditory cortex to visual speech. *BioRxiv*, 405597. https://doi.org/10.1101/405597

Munhall, K. G., Gribble, P., Sacco, L., Ward, M., Thompson, P., Tohkura, Y., … Summerfield, Q. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *99658*(3), 35–362.

Niso, G., Bruña, R., Pereda, E., Gutiérrez, R., Bajo, R., & Maestú, F. (2013). HERMES : Towards an Integrated Toolbox to Characterize Functional and Effective Brain Connectivity. *Neuroinformatics*, *11*(4), 405–434. https://doi.org/10.1007/s12021-013-9186-1

Nobre, A., Correa, A., & Coull, J. (2007). The hazards of time. *Current Opinion in Neurobiology*, *17*(4), 465–470. https://doi.org/10.1016/j.conb.2007.07.006

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be

Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Oksanen, A. J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., … Szoecs, E. (2013). Vegan: Community Ecology Package. R Package Version, *3*(January), 0–291.

Palva, J. M., Palva, S., & Kaila, K. (2005). Phase synchrony among neuronal oscillations in the human cortex. *Journal of Neuroscience*, *25*(15), 3962–3972. https://doi.org/10.1523/JNEUROSCI.4250-04.2005

Palva, S., & Palva, J. M. (2012). Discovering oscillatory interaction networks with M/EEG: Challenges and breakthroughs. *Trends in Cognitive Sciences*, *16*(4), 219–229. https://doi.org/10.1016/j.tics.2012.02.004

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. https://doi.org/10.1016/j.cub.2015.04.049

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *ELife*, *5*(MAY2016), 1–17. https://doi.org/10.7554/eLife.14521

Petanidou, T., Kallimanis, A. S., Tzanopoulos, J., Sgardelis, S. P., & Pantis, J. D. (2008). Long-term observation of a pollination network: Fluctuation in species and interactions, relative invariance of network structure and implications for estimates of specialization. *Ecology Letters*, *11*(6), 564–575. https://doi.org/10.1111/j.1461-0248.2008.01170.x

Pilling, M. (2009). Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. *Journal of Speech, Language, and Hearing Research*, *52*(4), 1073–1081. https://doi.org/10.1044/1092-4388(2009/07-0276)

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Pollack, G. S. (2001). Analysis of temporal patterns of communication signals. *Current Opinion in Neurobiology*, *11*(6), 734–738. https://doi.org/10.1016/s0959-4388(01)00277-x

Power, A. J., Mead, N., Barnes, L., & Goswami, U. (2012). Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. *Frontiers in Psychology*, *3*, 216. https://doi.org/10.3389/fpsyg.2012.00216

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Roux, F., & Uhlhaas, P. J. (2014). Working memory and neural oscillations: α-γ versus θ-γ codes for distinct WM information? *Trends in Cognitive Sciences*, *18*(1), 16–25. https://doi.org/10.1016/j.tics.2013.10.010

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, *12*(3), 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Schwartz, J.-L., & Savariaux, C. (2014). No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. *PLoS Computational Biology*, *10*(7), e1003743. https://doi.org/10.1371/journal.pcbi.1003743

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610. https://doi.org/10.1093/biomet/66.3.605

Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Oya, H., Kawasaki, H., … Griffiths, T. D. (2016). Neural signatures of perceptual inference. *ELife*, *5*.

https://doi.org/10.7554/eLife.11476

Simon, D. M., & Wallace, M. T. (2018). Integration and Temporal Processing of Asynchronous Audiovisual Speech. *Journal of Cognitive Neuroscience*, *30*(3), 319–337. https://doi.org/10.1162/jocn_a_01205

Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, *13*(2), 144–149. https://doi.org/10.1016/S0959-4388(03)00047-3

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience*, *15*(8), 520–535. https://doi.org/10.1038/nrn3742

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410. https://doi.org/10.1016/j.tics.2010.06.008

Tass, P., Rosenblum, M. G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., … Freund, H. J. (1998). Detection of n:m phase locking from noisy data: Application to magnetoencephalography. *Physical Review Letters*, *81*(15), 3291–3294. https://doi.org/10.1103/PhysRevLett.81.3291

ten Oever, S., & Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the National Academy of Sciences*, *112*(52), 15833–15837. https://doi.org/10.1073/pnas.1517519112

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 Wassenhove, V. (2013). Speech through ears and eyes: Interfacing the senses with the supramodal brain. *Frontiers in Psychology*, *4*(JUL), 1–17. https://doi.org/10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

Vinck, M., Oostenveld, R., Van Wingerden, M., Battaglia, F., & Pennartz, C. M. A. (2011). An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, *55*(4), 1548–1565. https://doi.org/10.1016/j.neuroimage.2011.01.055

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–1596. https://doi.org/10.1162/jocn.2009.21308

Wassenhove, V. van, Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/PNAS.0408949102

Womelsdorf, T., & Fries, P. (2007). The role of neuronal synchronization in selective attention. *Current Opinion in Neurobiology*, *17*(2), 154–160. https://doi.org/10.1016/j.conb.2007.02.002

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage*, *53*(4), 1197–1207. https://doi.org/10.1016/j.neuroimage.2010.06.041