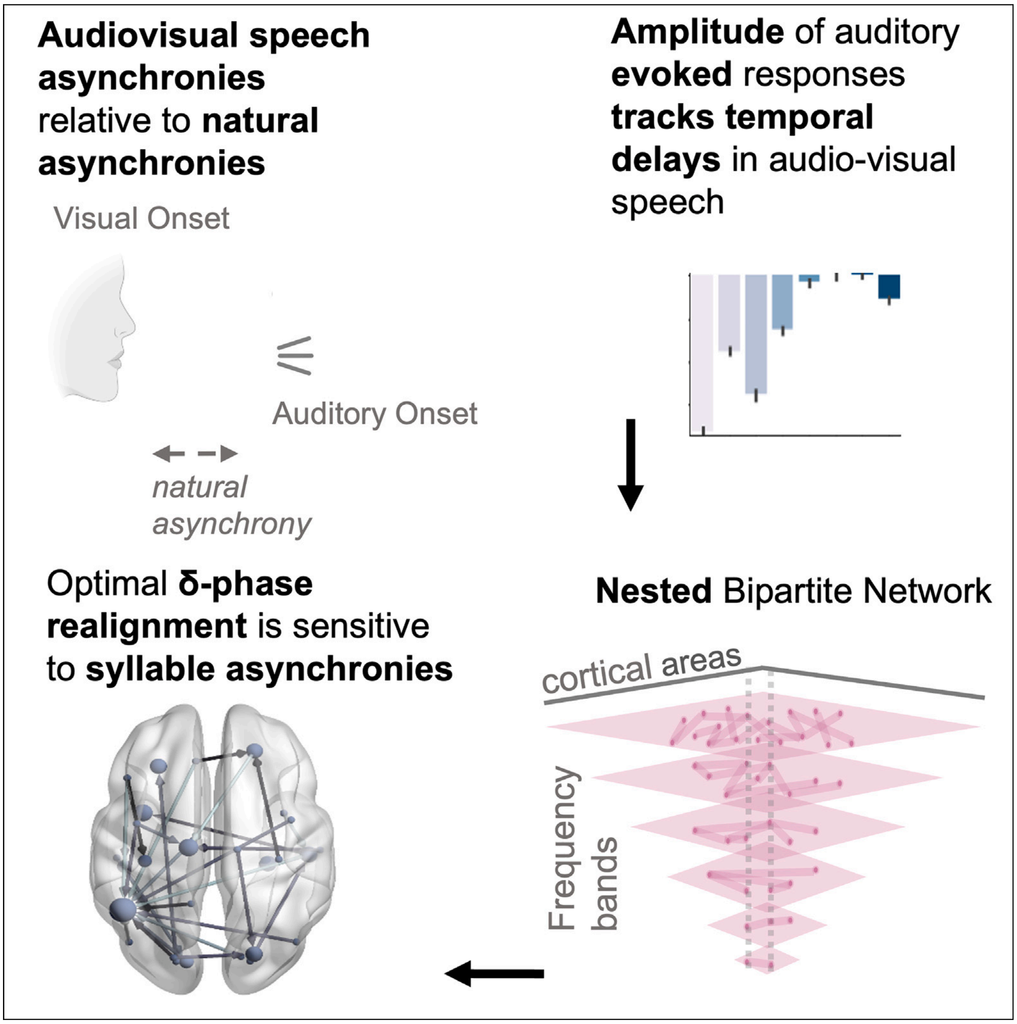
iScience

CellPress
OPEN ACCESS

Article

Hierarchically nested networks optimize the analysis of audiovisual speech


Audiovisual speech asynchronies relative to **natural asynchronies**

Amplitude of auditory **evoked** responses **tracks temporal delays** in audio-visual speech

Optimal **δ-phase realignment** is sensitive to **syllable asynchronies**

Nested Bipartite Network

Nikos Chalas, Diana Omigie, David Poeppel, Virginie van Wassenhove

nchalas@uni-muenster.de (N.C.)
virginie.van.wassenhove@gmail.com (V.v.W.)

Highlights

Brain activity is sensitive to audiovisual (AV) speech delays

Auditory evoked responses track temporal AV speech delays

Spatially synchronized nested networks track audiovisual speech asynchronies

AV temporal statistics drive top-down information transfer to auditory cortex

Chalas et al., iScience *26*, 106257
March 17, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.106257

iScience

CellPress
OPEN ACCESS

Article

Hierarchically nested networks optimize the analysis of audiovisual speech

Nikos Chalas,[1,2,6,7,*] Diana Omigie,[3] David Poeppel,[4,5] and Virginie van Wassenhove[2,*]

SUMMARY

In conversational settings, seeing the speaker's face elicits internal predictions about the upcoming acoustic utterance. Understanding how the listener's cortical dynamics tune to the temporal statistics of audiovisual (AV) speech is thus essential. Using magnetoencephalography, we explored how large-scale frequency-specific dynamics of human brain activity adapt to AV speech delays. First, we show that the amplitude of phase-locked responses parametrically decreases with natural AV speech synchrony, a pattern that is consistent with predictive coding. Second, we show that the temporal statistics of AV speech affect large-scale oscillatory networks at multiple spatial and temporal resolutions. We demonstrate a spatial nestedness of oscillatory networks during the processing of AV speech: these oscillatory hierarchies are such that high-frequency activity (beta, gamma) is contingent on the phase response of low-frequency (delta, theta) networks. Our findings suggest that the endogenous temporal multiplexing of speech processing confers adaptability within the temporal regimes that are essential for speech comprehension.

INTRODUCTION

Seeing a speaker's face facilitates the comprehension of auditory speech.[1–5] This is in part because the interlocutor's mouth movements cohere in time with the produced acoustic speech envelope, a phenomenon called temporal comodulation. The temporal comodulation of audiovisual (AV) speech signals yields statistical redundancy of speech information, which benefits multisensory integration[6] at a syllabic timescale.[3,7,8] The functional relevance of temporal comodulation for AV speech is that seeing the speaker's visual articulatory gestures entrains cortical oscillations not only in auditory[9,10] but also in motor[11] cortices. Such network-level entrainment to AV speech[12] has important implications for the optimal neural coding of multisensory speech signals and is decisive for the listener's perception.[13] For instance, electrophysiological recordings have shown that visual information can phase-reset ongoing low-frequency oscillations[14–17]: by phase-aligning ongoing low-frequency oscillations in auditory cortices, incoming visual speech inputs can temporally tune endogenous brain rhythms to the arrival of upcoming auditory speech inputs. With such a mechanism, the acoustic signal arrives at a high-excitable state in auditory cortices, which optimizes the integration of AV speech. The interplay between low-frequency oscillatory phase alignment and high-frequency responses is referred to as nestedness.

The temporal comodulation of AV speech does not rely on perfect synchrony; rather, it relies on a temporal window of integration, which is a permissible temporal delay within which AV information can be integrated. The temporal window of integration approximates the syllabic scale mentioned earlier [8,18]: visual speech (as measured by the viseme onset) tends to precede phonation with an average of 100–300 ms for syllables like [pa] or [ta],[3,19] and a large range of AV asynchronies has been described across languages.[20] The tendency for visual precedence has been argued to initiate internal predictions of upcoming auditory speech, thereby facilitating intelligibility.[7,21–29] This working hypothesis, generally consistent with multisensory causal inference models[30] and the efficiency of predictive coding, supports the notion that the more informative a signal is, the more reliable internal predictions will be. In AV speech, the more informative the visual speech, the faster and smaller the auditory evoked responses.[7,31] Additionally, in AV speech processing, temporal comodulation and temporal delays are two factors, which contribute to multisensory causal inferences.[32,33] Evidence for temporal comodulation and delay computations can be decoded from magnetoencephalography (MEG) when using simple sequences of AV stimuli.[34] Due to their importance in AV speech processing, we here questioned whether desynchronizing AV speech signals would also alter the temporal tuning of the speech processing network.

[1]Institute for Biomagnetism and Biosignal Analysis, University of Münster, P.C., 48149 Münster, Germany

[2]CEA, DRF/Joliot, NeuroSpin, INSERM, Cognitive Neuroimaging Unit; CNRS; Université Paris-Saclay, 91191 Gif/Yvette, France

[3]Department of Psychology, Goldsmiths University London, London, UK

[4]Department of Psychology, New York University, New York, NY 10003, USA

[5]Ernst Struengmann Institute for Neuroscience, 60528 Frankfurt am Main, Frankfurt, Germany

[6]School of Biology, Faculty of Sciences, Aristotle University of Thessaloniki, P.C., 54124 Thessaloniki, Greece

[7]Lead contact

*Correspondence: nchalas@uni-muenster.de (N.C.), virginie.van.wassenhove@gmail.com (V.v.W.)

https://doi.org/10.1016/j.isci.2023.106257

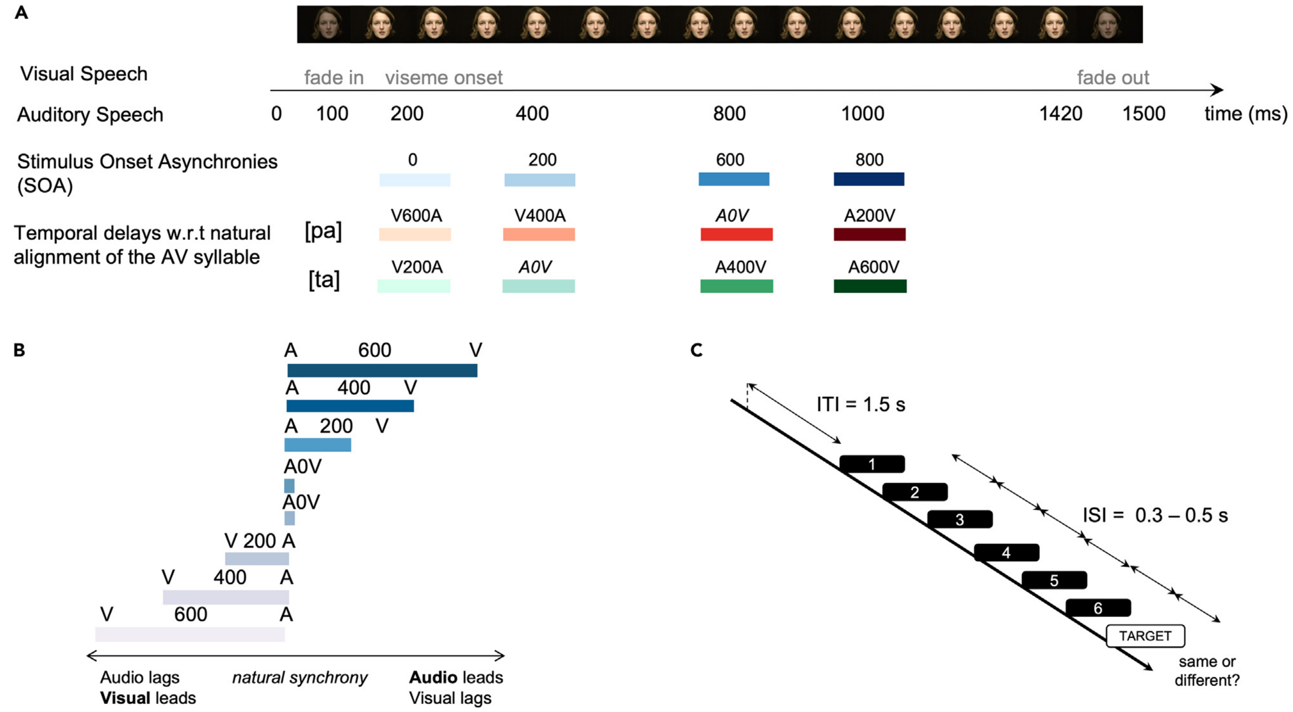


Figure 1. Stimuli and experimental design

(A and B) Temporal alignment of audiovisual (AV) syllables. All videos lasted for 1320 ms (75 frames; frame rate of 50 Hz). The first five frames faded into the static face; the last four frames faded out from the static face. The viseme onset was defined as the first frame containing a detectable speech movement, and this frame was placed 200 ms after the onset of the video for both the syllables [pa] and [ta]. The auditory speech could be played 0, 200, 600, or 800 ms later than the viseme onset, which is defined as the stimulus onset asynchronies (blue). By SOA, we thus, refer to the physical temporal delay between the viseme onset and the auditory speech onset. In the original recordings, the/a/in [pa] occurred 600 ms after the viseme onset and the/a/in [ta] occurred 200 ms after the viseme onset; these delays define the natural temporal alignment for each syllable, which we define as A0V. Each AV speech delay can be defined as a function of different temporal references and timelines: (i) the physical SOA (blues in A), (ii) the auditory delay with respect to the natural AV alignment (reds for [pa] and greens for [ta], bottom lines in A), or (iii) the visual delay (blues) with respect to natural AV alignment in an auditory speech timeline in (B). The desynchronized [pa] (A400V) matched the natural asynchrony of [ta] (A0V). Conversely, the desynchronized [ta] (V400A) matched the natural synchrony of [pa] (A0V). The set of AV speech stimuli could be directly compared as a function of their SOAs (A) or as a function of their natural speech asynchronies with reference to A0V (B).

(C) One trial lasted 14 s and consisted of a sequence of six identical AV stimuli chosen among the two possible syllables ([ta] or [pa]) and their four possible SOAs. A seventh stimulus was the target stimulus chosen among four possible stimuli: synchronized (A0V) or desynchronized [pa] (A400V) or [ta] (V400A). The inter-stimulus interval (ISI) was 400 ms. The inter-trial intervals were 1.5 s. All analyses uniquely focus on the brain responses elicited by the sequence of six syllables (responses to the target were not analyzed).

For this, we recorded participants with MEG while they were watching and listening to sequences of six identical AV syllables [pa] or [ta]. AV syllables were either temporally aligned (natural AV speech statistics; A0V) or temporally misaligned (the auditory speech signal was either advanced or delayed with respect to the original video). Each sequence of six syllables (Figure 1) was followed by a target AV syllable that matched or mismatched the preceding sequence of syllables in identity or in timing (Figure 1C). To keep participants attentive, they were required to report whether the AV target differed from the standards; however, we solely focus on the brain responses to the presentation of identical syllables in our analysis and detail our reasoning below.

First, we sought to replicate previous work showing that the auditory evoked response is sensitive to the timing of AV speech[7,21,28] and that a decreased amplitude occurs within a particular temporal window of integration.[7,8,21,28] For this, we quantified the evoked responses to desynchronized AV speech, both as a function of the stimulus onset asynchronies (Figure 1A; SOA) and as a function of the temporal distance to the natural statistics of the syllables (Figure 1B; distance to A0V). These two temporal dimensions are crucial for our work: the analyses using SOA imply that we are mostly looking at the bottom-up and feed-forward analysis of incoming AV speech signals. Conversely, the distance to A0V analysis explores a new working hypothesis according to which the distribution of AV speech asynchronies for a given syllable would be part of the internal representation of that syllable. As such, the distance to the internal distribution (A0V) could be hypothesized and tested with incoming AV speech inputs, which constitutes a more top-down or feedback analysis. Second, we characterized the sensitivity of the oscillatory functional networks to the timing of AV speech. For this, we tested the temporal multiplexing hypothesis from a network perspective and explored the implication of hierarchically nested cortical networks in AV speech processing. Our approach is original in that we used a nested bipartite network (Figure S1), which provides direct indexing of nestedness in oscillatory networks. Although a substantial body of work has described temporal multiplexing in the processing of auditory speech[35] surprisingly, little work has been dedicated to characterizing temporal multiplexing in AV speech, perhaps owing to the difficulty of tackling two sensory modalities with distinct temporal integrative properties.

RESULTS

The amplitude of auditory-evoked responses tracks temporal delays in audiovisual speech

Herein, we asked how the brain tracks physical asynchronies (*i.e.*, the ''when'') irrespective of and, as a function of, the natural AV speech timing of the syllable (*i.e.*, the ''what''). For this, we performed linear regression analyses with two types of regressors using the amplitude of the auditory-evoked responses as the main factor. The first analysis used SOAs as a regressor, which is measured as the physical temporal delay between the auditory speech onset and the viseme onset (Figure 1A, blues). Using the SOAs as a regressor asks whether physical AV asynchronies are encoded in brain responses in a feedforward manner. The second analysis used the temporal delays as distances to the natural AV speech synchrony (distance to A0V; Figure 1B). This second analysis asks whether asynchronies in reference to a possible internal representation of the natural statistics of AV speech signals (a ''signed prediction error'') are coded in brain response.

First, we performed a subject-wise linear regression analysis of single-trial MEG responses locked to the auditory speech onset for each sensor and time sample. We used the SOA as the unique regressor (4 values: 0, 200, 400, and 800 ms; Figure 1A). This analysis resulted in an average time course of beta values (estimated coefficients of subject-wise linear regressions) for each sensor, each latency, and each individual. Subsequently, we performed a group-level spatiotemporal cluster permutation one-sample t-test of the individuals' beta time courses to identify clusters (sensors x time samples) showing beta values significantly different from zero in latency and in topography. Significant clusters indicate that, at the group level, MEG-evoked activity is significantly ranked according to the regressor (for a similar approach, see[36]).

We found two clusters indicating beta values significantly different from zero at the group level at early latencies (20 magnetometers: [47 ms: 246 ms], $p < 0.01$; 22 magnetometers: [15 ms: 436 ms], $p < 0.01$, Figure 2A). A significant cluster indicates that the amplitude of the evoked response is graded as a function of the SOA: the larger the beta value, the stronger the gradation. Our analysis shows a consistent parametric decrease in phase-locked auditory responses when the SOA increases (Figure 2A). The estimated sources for this parametric tracking were bilateral, encompassing the auditory cortices and the middle temporal sulci (Figure 2C). The graded response as a function of SOA demonstrates that the brain parametrically tracks the physical asynchronies of AV speech syllables.

Second, we tested the hypothesis that the natural statistics of AV speech help generate an adequate temporal prediction of the acoustic inputs. For this, we asked whether the temporal distance between the viseme onset and the natural AV speech timing (A0V) could be measured from brain activity (Figure 2B). We predicted that the larger the deviance from the natural statistics A0V, the larger the residual errors in the evoked responses would be (in a predictive coding framework) and thus, the larger the amplitude of the auditory-evoked response. To test this, we used the same regression analysis approach, but this time, we used the temporal distances from natural A0V as a regressor (Figure 1B; 8 values: −600, −400, −200{query; is it minus or En dash}, 0, 200, 400, and 600 ms). This analysis resulted in three clusters (50 magnetometers, [- 169 ms: - 7 ms], $p < 0.01$; 25 magnetometers, [141 ms: 454 ms], $p < 0.01$; 37 magnetometers, [8 ms: 378 ms], $p < 0.01$) in which the beta values were significantly different from zero at the population-level. This observation indicates that the temporal distance from the natural AV speech asynchrony (A0V) was coded for both syllables (Figure 2B). The source estimates for this effect were bilateral and mostly located in the posterior part of the superior temporal sulci (Figure 2D).

Last, the same single-trial regression analysis using the temporal distances from natural A0V as a regressor (8 values: −600, −400, −200, 0, 200, 400, and 600 ms) was separately applied to the auditory-evoked

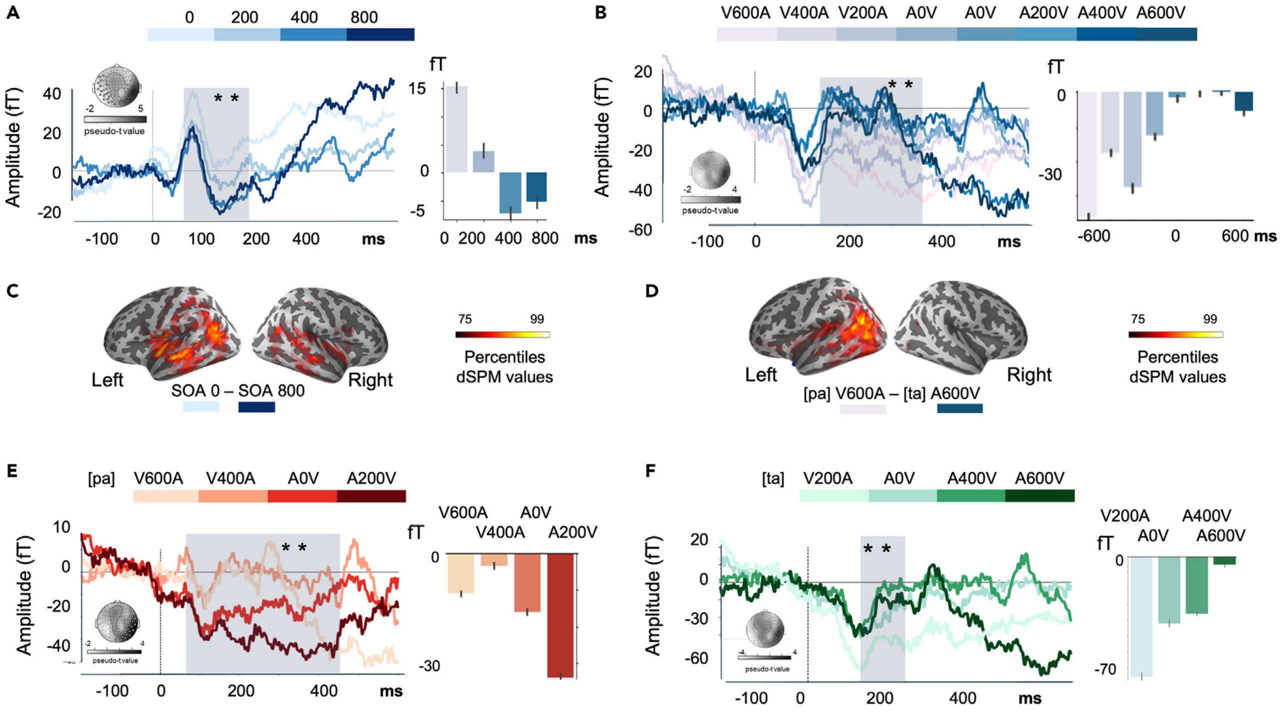


Figure 2. Parametric tracking of temporal delays in AV speech by phase-locked responses

(A) The amplitude of the auditory-evoked responses parametrically varied as a function of the AV speech asynchronies (shaded areas) measured as a function of stimulus-onset-asynchronies SOA (the left panel; Figure 2A) and (B) as a function of the temporal distance to the natural AV speech synchrony A0V (the right panels; Figure 2B).

(C) Grand-averaged source estimates of the contrast SOA = 0 ms with SOA = 800-ms combining brain responses to [pa] and [ta].

(D) Grand-averaged source estimates contrasting V600A ([pa]) vs. A600V ([ta]). In both source estimates, bilateral regions encompassed auditory cortices, posterior occipito temporal, and middle temporal cortices (right panel).

(E) Auditory-evoked responses elicited by the presentation of different asynchronies for [pa].

(F) Auditory-evoked responses elicited by the presentation of different asynchronies for[ta]. For both syllables, a significant gradation of amplitude was shown to be sensitive to the temporal distance between the natural AV speech asynchrony (A0V), and the AV speech asynchronies. $^{**}p < 0.01$. Topographies indicate significant sensors and pseudo-t values. For the bar plots, data are presented as mean ± SEM.

responses elicited by each syllable ([pa] and [ta]; Figures 2E and 2F, respectively). Although, we expected to replicate similar findings, the differences in visemic saliency of the [pa] and [ta] syllables were predicted to affect the timing sensitivity of the auditory-evoked responses, with a stronger timing sensitivity for [pa] than for [ta]. We found two clusters showing population-level beta significantly different from zero for [pa] (22 magnetometers, [36 ms: 428 ms], $p < 0.01$; 27 magnetometers, [84 ms: 184 ms], $p < 0.01$; Figure 2E), and for [ta] (12 magnetometers, [141 ms: 244 ms], $p < 0.01$; 12 magnetometers, [153 ms : 222 ms], $p < 0.01$; Figure 2F). The temporal extent of the clusters elicited by AV speech asynchronies was much broader for [pa] ([36 ms: 428 ms]) than for [ta] ([141 ms: 244 ms]).

Altogether, we found systematic, bilateral modulations of the auditory-evoked responses as a function of AV speech asynchronies, either measured as a function of SOA or as a distance to the natural AV speech synchrony. In all analyses, we consistently observed that the earlier the visual speech onset, the larger the amplitude of the auditory-evoked responses, and the closer to the auditory onset, the more suppressed the amplitude of the auditory-evoked response. The linear regression approach allowed capturing at once the gradation in the amplitude of the auditory-evoked responses. The significant impact of the temporal delay confirms prior findings suggesting that visemic information predicts and guides acoustic analysis, although cluster analysis limits a robust interpretation of latency findings.[37] Importantly, our approach reveals the coexistence of at least two temporal statistics in the cortical analysis of AV speech: those fed forward by multisensory evidence as demonstrated by the parametric changes in evoked responses elicited by SOAs and those fed endogenously, possibly by an internal predictive speech model, which were captured by the sensitivity to the temporal distances to natural AV speech temporal statistics (A0V) akin to a signed prediction error.

Next, we explored the dynamic changes of the AV speech network in response to AV speech asynchronies and turned to a novel functional connectivity analysis of human brain activity, which allowed establishing the existence of temporal multiplexing in functional networks, and the sensitivity of oscillatory nestedness to AV speech asynchronies seen in phase-locked (evoked) responses.

Spatially synchronized nested oscillatory networks are sensitive to AV speech asynchronies

To identify the oscillatory networks sensitive to AV speech asynchronies, we analyzed brain activity elicited by AV speech syllables in source space (see functional connectivity analysis; Supp. Figure 1). Several studies in AV speech have reported an extensive network contributing to the analysis of AV speech signals engaging various networks and operating at distinct frequency regimes[6,12,38,39]).

As in previous studies, we assessed the full functional network as a function of the oscillatory regime (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]) for 46 combined sets of cortical parcels according to the human connectome project atlas.[40] For the network analysis, we did not distinguish between syllable identity and focused our statistical model on distance to A0V as described in Figure 1B: a 2 × 4 mixed model ANOVA at the network level with within-subject factors of order (2: auditory or visual leads) and distance from A0V (4: 0, 200, 400, and 600 ms) revealed an effect of speech asynchronies (measured as the distance from A0V) in all oscillatory networks ($p < 0.05$, FDR corrected; Figure S1). When considering the descriptive statistics of each network, we observed that low-frequency networks in the δ and θ ranges were spatially more widespread than their counterparts in the higher-frequency regimes. This observation enriched the mechanistic predictions of temporal multiplexing[35] at the network level: a global network-level regulation at slow frequency may temporally fine-tune nested local networks at high frequency. To test this novel working hypothesis and estimate the degree of nestedness in the oscillatory networks shown to be sensitive to AV speech asynchronies, we used a bipartite network approach.

Bipartite networks (Figure 3) are widely used in other research fields[41] but have not been exploited in neurosciences. We briefly illustrate a typical usage: in community ecology, the interactions of plants and pollinators form a nested bipartite (plants x pollinators) network[42,43] 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 contributes to a generalist-specialist balance while distinguishing mutualistic from antagonistic interactions.[44,45] By analogy, in brain networks, we assume that the generalists are large-scale (spatial dimension) and slow frequency (temporal dimension), whereas the specialists are small-scale operating in high frequency. A bipartite network analysis provides a direct assessment of the interactions across spatiotemporal scales in networks, which is central to the question of nestedness in temporal multiplexing we are addressing here. To the best of our knowledge, this is the first application of bipartite networks to the description of nested oscillatory functional networks in human brain activity. To provide intuition on what to expect, we provide an illustration of a nested bipartite network (Figure 3A) with a random network (Figure 3B).

To generate the bipartite adjacency matrix, we calculated the node strength i.e., the sum of incoming and outgoing connections of cortical labels for each significant frequency band-specific network (Figure S1). Thus, the node strength serves as a proxy for the contribution of each brain label in a frequency-specific network. We then reconstructed a 46 (node strengths per cortical area) by 6 (frequencies) matrix (Figure S2A) which represented the bipartite network (Figure S2B). To estimate the degree of nestedness, we computed the matrix's temperature[46]: a matrix's temperature increases as it deviates from perfect nestedness. Perfect nestedness is defined when a boundary line can separate connections (black squares) in the matrix's column from the absence of connection (white squares; Figure 3A top panel). By definition, the temperature of a matrix is a number between 0 and 100, but the size and the structure of each matrix actually define the temperature values that the matrix can take. Thus, the probability of obtaining the temperature of the matrix by chance (compared to null models) is more informative, rather than the temperature of the matrix per se.

We found that the bipartite network in AV speech processing was significantly nested when compared to 1,000 surrogate networks with the same characteristics (temperature = 7.99; $p = 0.01$; Figures 3B and 3C). Figure 3D illustrate the hierarchical clustering of the adjacent nodes and frequency bands generating a cluster map according to the node strength that conformed to the specialist/generalist dichotomy: crucially, and as predicted by the temporal multiplexing hypothesis, we found that higher-frequency

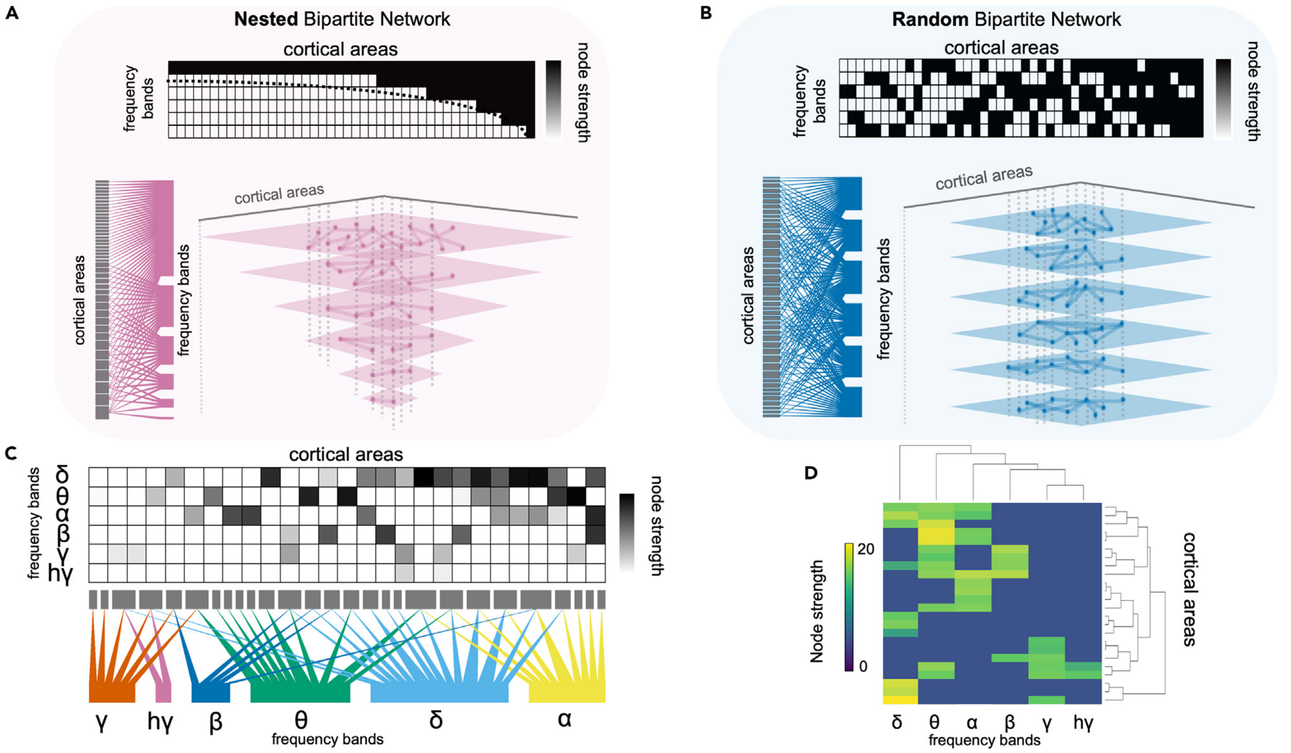


Figure 3. Nested oscillatory networks sensitive to AV speech asynchronies

(A) Ideal nested bipartite network (pink). The connections (black squares) indicate the significant activation of a node in the oscillatory network. In this example, the top node in the group *frequency bands* is connected to all nodes of the group *cortical areas*. Going downward, each frequency band connects with a subset of nodes in cortical areas in a hierarchical manner.

(B) Random bipartite network (blue). In this example, connections are drawn randomly with the Erdős-Rényi model (probability of connection = 0.5). Contrary to the ideal nestedness of (A), a random bipartite network does not show systematic node patterns between cortical areas and oscillatory regimes.

(C) Nested functional networks in AV speech processing. The outcomes of our analysis showed a significant network nestedness. The darker the square, the more common nodes are shared between brain areas. The split of shared connectivity among oscillatory regimes is provided in the bottom panel. The overall patterning indicates nestedness in the oscillatory networks with notable larger networks for low-frequency oscillations.

(D) Hierarchical clustering of the frequency regimes and cortical areas within the bipartite speech network. Higher-frequency oscillations (γ and high-γ) are spatially nested within lower-frequency oscillations (δ and θ). For the hierarchical clustering, distances were calculated with the Minkowski method.

oscillatory regimes (γ and hγ) were spatially nested within slower-frequency oscillatory networks (δ and θ; see Figure S4 for the cortical labels).

We performed the same analysis, this time using the SOAs instead of the distance to A0V and thus, considered the identity of the syllable as a factor following the description in Figure 1A. We applied a 2 × 4 ANOVA with the factors of the syllable (2: [pa] and [ta]) and SOA (4: 0 , 200 , 400 , and 600 ms). In this case, the corresponding bipartite network was also significantly nested as a main effect of SOAs (temperature value = 11.1, p = 0.01, Figure S5A) with a partly different clustering spatiotemporal pattern from the one calculated in terms of natural asynchronies (Figure S5B).

Optimal δ-phase realignment is sensitive to syllable asynchronies

Having established the presence of spatiotemporal multiplexing and nestedness in the oscillatory networks sensitive to AV speech asynchronies, we then asked whether causal pairwise interactions of phase information in low-frequency networks were linked to the synchronized activity of low-frequency and large-scale networks (in particular, δ and θ). With this analysis, we sought to identify cortical areas modulating the synchronized activity of low-frequency networks, which we previously found to modulate high-frequency networks in cortical space.

To establish the directionality of nestedness in oscillatory networks, we used phase transfer entropy (pTE;[47]). pTE is a measure of directed connectivity among oscillatory networks that quantifies the direction

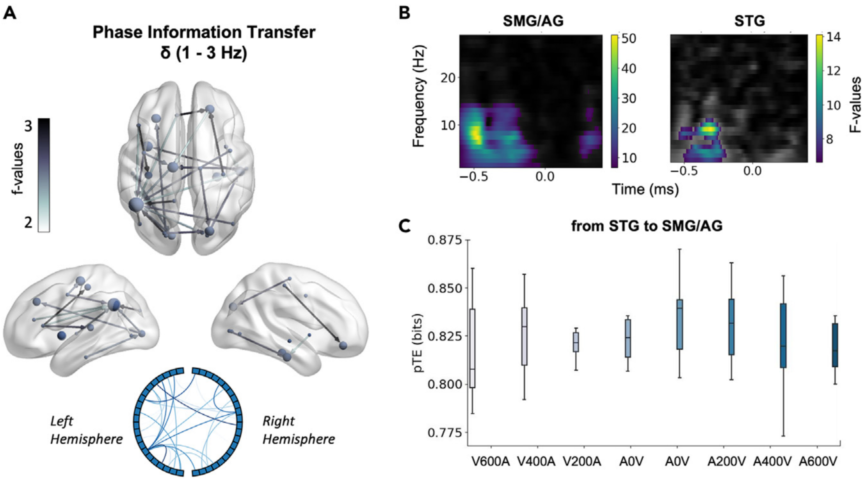


Figure 4. Phase interactions on oscillatory activity sensitive to asynchronized AV speech

(A) Causal phase interactions elicited by asynchronized AV speech in the δ oscillatory network (estimated as distance from A0V). The phase-time series were estimated using Hilbert-transform on single-trial source estimates and pairwise phase interactions were estimated using phase transfer entropy (pTE; $p < 0.05$; FDR corrected). The color scale indicates t-values. The size of the nodes corresponds to the node strength in the network. The arrows indicate the directionality of phase-information transfer.

(B) Main effects of AV speech asynchronies on inter-trial coherence (ITC) in SMG/AG (Left; 1–9 Hz; −500 ms to −186 ms) and STG (Right; 1–9.5 Hz; −390 ms to −200 ms). Statistically significant clusters were drawn at $p < 0.05$.

(C) pTE values from STG to SMG/AG (measured as bits) across AV asynchronies. Data are presented as mean ± SEM. SMG/AG: supramarginal gyrus/angular gyrus; STG: superior temporal gyrus.

of information transfer: the magnitude of pTE between two phase-time series indicates the extent to which the phase of one time series can predict the phase of another time series. Herein, we used pTE to characterize information transfer in the identified nested oscillatory networks and investigated how information may be regulated at different time scales. Specifically, we asked whether global low-frequency networks drove local high-frequency networks as predicted by temporal multiplexing.

We calculated the pTE using single-trial source estimates locked to the auditory speech onset, and the functional cortical labels of each oscillatory network described above (Figure S1). The pTE was computed per participant and per AV asynchronies. The pTE values of each brain region (network node) were submitted to a 2 × 4 mixed model ANOVA at the network level with within-subject factors of order (2: auditory or visual leading) and distance from A0V (4: 0, 200, 400, and 600 ms). We found no significant effects on pTE in the θ network, whereas pTE in the δ network was significantly affected by distance to A0V ($p < 0.05$, FDR corrected; Figure 4A). The significant changes of pTE as a function of distance to A0V suggest that the global δ network is sensitive to AV speech asynchronies in a manner that affects phase information transfer. This δ effect showed a significant directed transfer of information in cortical phase space in a network consisting of 36 connections (Figure 4A). Within the δ network, the left supramarginal gyrus/angular gyrus (SMG/AG) formed a central hub as revealed by a node strength of 34.6 (mean = 6.8; SD = 9.39). The phase response in the SMG/AG hub was principally predicted by frontal and motor cortices, i.e., the SMG/AG received phase information from frontal (the left inferior frontal cortex and the right orbital and frontal cortex) and motor (left and right premotor cortex) regions. Additionally, the left inferior frontal cortex contributed to the modulation of the δ phase response in superior temporal gyrus (STG). In turn, the δ network also showed a significant feedforward pTE to the central SMG/AG hub.

Our observations give a central and convergent role to the SMG/AG region in orchestrating the analysis of AV speech asynchronies relative to the distance to A0V through possible changes in neural excitability of the δ-band in regions often reported in the AV speech processing, namely: frontal cortices, motor, and auditory regions.[48]

We repeated the same analysis, this time as a function of SOAs. For this, we applied a 2 × 4 ANOVA with the factors of syllable ([pa] and [ta]) and SOA (0 , 200 , 400 , and 600 ms). Similar to the analysis on distance to A0V, we observed the significant main effect of SOAs on pTE between nodes including SMG/AG, STG, motor, frontal, and visual cortices (Figure S5C). Interestingly; however, SMG/AG did not stand out as a central hub of the network, and the delta network was overall sparser. This indicates that the SMG/AG as a hub in the phase information transfer network during temporal integration of AV speech is specific to temporal distances from natural asynchronies, possibly indicative of predictive processing mechanism.

To better understand how global phase interactions in δ-band affect local inter-trial phase synchronization during the processing of AV speech asynchronies, we computed inter-trial-coherence (ITC) from single-trial source estimates in the SMG/AG and in the STG. ITCs from each region were submitted to 2 × 4 mixed model ANOVA at the network level with within-subject factors of order (2: auditory or visual leads) and distance from A0V (4: 0, 200, 400, and 600 ms). This analysis confirmed an effect of distance to A0V on ITC (including δ) preceding the auditory speech onset in both SMG/AG and STG regions (Figure 4B). To illustrate the causal interaction of the global δ pTE network within SMG/AG and STG activity, we plot the phase information transfer (pTE values; measured as bits) across audiovisual asynchronies (as distances from A0V). It is evident that the pTE increases when syllables are presented in the vicinity of the A0V and decreases away from it (also see Figure S6A for information transfer from IFG → SMG/AG). Thus, we conclude that as temporal asynchronies deviate away from natural speech asynchronies (A0V), the less phase interaction between SMG/AG and STG in the δ-band.

Next, we computed the Hilbert-transformed single-trial δ responses and plotted the δ-band phase responses (Figure S6B). In the SMG/AG regions and for the syllable [pa], the δ-phase responses differed between the natural and the extreme asynchrony A0V and A600V (Figure S6B; left panel). No differences were found for [ta] (Figure S6B; bottom panel). These results show that for [pa] but not for [ta], the excitability of δ activity in SMG/AG is differentially modulated prior to the arrival of the sound as a function of speech asynchrony. This differential modulation in δ activity was not observed in STG, where we did observe phase-time series modulated by a cycle duration of around 200–300 ms (Figure S6B).

Taken together, we found causal phase interactions in the δ network sensitive to AV speech asynchronies, with SMG/AG having a central role in the orchestration of the interactions. Both SMG/AG and STG exhibited significant broad-band phase-synchronization with a main effect of SOA on activity prior to acoustic onset. However, SMG/AG showed a selective optimal excitability δ-phase response before the onset of a sound solely for syllable [pa]. These results confirm the importance of δ-phase in possibly realigning the networks when faced with salient AV speech asynchronies.

DISCUSSION

In the present study first, we sought to determine the neurophysiological responses sensitive to the alignment of AV speech in time. We replicate and extend previous work[7,21,28] showing the suppression of auditory-evoked responses to the presentation of natural AV speech: gradual changes in the amplitude of the auditory-evoked responses were found as a function of the degree of temporal deviance from the natural AV synchrony (A0V) of each syllable. Although those temporal deviances are not ecologically valid, they constitute a useful experimental manipulation to understand the mechanisms for AV speech integration. Second, we characterized oscillatory networks sensitive to AV speech timing in several frequency regimes. The coupling of oscillatory activity coordinates neural activity[49] and can take different forms.[49–54] We demonstrate that the representation of AV speech asynchronies modulates hierarchically nested long-range oscillatory cortical networks, which is consistent with long-range communication.[55] Using a bipartite network analysis, we show that AV speech timing is mediated by spatially nested hierarchical oscillatory networks and that high-frequency activity (γ and high-γ) is nested within low-frequency (δ and θ) regimes. We further report that frontal and motor cortices causally modulate the excitability of SMG/AG within the δ network. SMG/AG appears as a central incoming hub. Last, we demonstrate optimal δ-phase alignment of STG, but not of SMG/AG, to the precedence of visual speech, irrespective of the incoming temporal concordance between auditory and visual speech inputs.

Tracking audiovisual speech timing in the evoked (phase-locked) activity

We show an early modulation of auditory-evoked responses as a function of AV speech asynchrony. Those internal predictions are shaped by the temporal statistics of sensory inputs is a domain-general observation[56] also reported in AV speech,[57] which indicates that AV speech integration may be grounded in the interaction of speech-specific and domain-general multisensory processing.[58] The distribution of brain-evoked responses was asymmetric around natural AV speech asynchrony (A0V), which matches the typical profile of the temporal window of integration in AV speech previously reported in behavioral and computational studies[8,59–62] as well as neurophysiological work.[21] Specifically, the modulation of the auditory-evoked responses by visual speech lasted longer (∼400 ms) when the viseme was more informative ([pa] than [ta]). This timing also corresponds to low-frequency rhythms (∼3 Hz) that match the temporal comodulation (2–6 Hz) of mouth movements and sound envelopes reported for consonant-vowel syllables.[19] This observation fits with previous predictions: in naturalistic AV speech, the low-frequency components track visual and auditory information[9] and may explain the long temporal integration windows reported in continuous AV speech processing.[63]

Also in this syllabic time range, the θ network implicating the dorsal stream was a significant contributor to the bipartite networks and oscillatory nestedness. However, speech asynchrony did not significantly affect the causal interactions with the θ network. This observation is intriguing considering that slower local-phase responses showed 200–300 temporal modulations (consistent with[13]). Additionally, the initial evoked response (phase-locked) analysis suggests that cortical responses are sensitive to two kinds of temporal alignments (physical asynchronies and asynchronies with respect to natural speech asynchronies). The coexistence of internal representations of the natural temporal statistics (internal templates of the natural asynchronies for AV speech) may be reflected in the stability of the θ network, whereas incoming AV speech statistics may largely be monitored by the δ network.

Nested oscillatory networks track audiovisual speech asynchronies

Nestedness in bipartite networks has been described in other fields,[41] with important implications for the functional understanding of the system under scrutiny.[45,64] Following an analogy with community ecology, we find that δ and θ frequency bands are "generalist" rhythms that operate over longer integrative timescales and dominate interactions with other cortical regions. Conversely, **γ** and high-**γ** regimes are "specialists" interacting with a subset of cortical areas at faster timescales and spatially within the δ and θ networks. The relationship between slow- and fast-scale rhythmic activities across cortical regions has long been speculated.[65] Oscillatory networks operating in short timescales are restricted to small neuronal spaces, whereas slow timescales implicate long-range activity.[66,67] Here, we demonstrate that for the identification of AV speech asynchronies, high-frequency oscillatory networks are confined to a cortical space regulated by slower oscillatory networks. That different oscillatory regimes facilitate the processing of incoming information at multiple temporal scales, i.e., temporal multiplexing is well described in sensory processing[55,68] and in speech perception.[69–71] We extend those findings and propose that oscillatory multiplexing is confined in cortical space by slow oscillatory rhythms. We speculate that slow activity in cortical regions (such as δ and θ) may provide the cortical space with high cortical excitability for faster activity to occur and, in our case, orchestrate feedback and feedforward operations during processing of AV speech asynchronies.

δ-band activity in supramarginal (SMG) and superior temporal gyrus (STG) during the processing of asynchronous audiovisual speech

In our results, the neural tracking of AV speech asynchronies was accomplished by a functional network of causal phase interactions in the δ range, implicating phase modulations within cortical areas. Within this hierarchical network, the SMG/AG stands as a main hub, receiving connections from frontal, parietal, and occipital areas. The SMG/AG was found when analyzing the distance to A0V but not the SOA. Previous research has implicated SMG/AG in AV speech perception.[39,72] Bernstein and colleagues reported that this region was significantly activated during congruent AV speech perception, with a temporally broad activation of the auditory-evoked response.[24] Here, we provide new evidence that this activation is driven by cortical causal interactions that modulate the phase of intrinsic low-frequency oscillations locally. Consolidating this notion, the activity of SMG/AG phase aligned during naturally asynchronous (A0V) conditions around the auditory onset. This is in line with previous reports showing that temporally and contextually congruent AV speech enhances entrainment in cortical areas ([73]; see also fMRI findings[74]).

Importantly, our control analysis testing the effect of SOA (and not the distance to the natural asynchrony) showed no involvement of the SMG/AG hub. This supports the idea that SMG/AG is specific to temporal distances from natural asynchrony—not physical asynchronies—underlines the implication of predictive processing mechanisms during temporal integration of AV speech. Altogether, our observations suggest that predictive processing in AV speech may not only generate internal hypotheses about the identity of the incoming speech inputs but also the expected delay between auditory and visual signals. This working hypothesis is complementary to the general hypothesis that neural oscillations align with external sensory inputs to improve sensitivity.[75,76] Here, the idea we put forward is that the natural delays between auditory and visual speech are a feature of speech representation that can serve the elicitation of internal hypotheses, and thus, generate prediction errors (measured as distance from the expected delay A0V). This feature would be important in that temporal delays too could be predicted, thereby optimizing the processing of signals in real-time.

The phase of δ oscillations in STG was significantly modulated by a frontal node during the presentation of asynchronous AV speech. This is consistent with previous work,[77] which identified frontal top-down, phase-modulating signals, during the presentation of natural AV speech streams. The ongoing activity in δ-band activity in STG exhibited a synchronized response to auditory speech during the presentation of various AV speech asynchronies with a cycle duration of around 200–300 ms, suggesting a functional relation in the STG between δ- and θ-band activity. The relevance of this in AV speech processing should be explored considering that our pTE analysis revealed no clear evidence of causal links with the θ network. In our analysis, AV syllables were matching in identity (temporal asynchronies were specific to a syllable). Thus, the transfer of phase information within the low-frequency δ-band cannot be distinguished from potential delta oscillations that would arise from a syllable's identity mismatch or from stimuli that are independent from language.

Limitations of the study

This study comes with limitations. One obvious drawback is that the audiovisual asynchronies used as experimental stimuli were isolated syllables, which deviate from an ecologically valid framework of natural speech. That inherently narrows our findings to the audiovisual syllables tested along with respect to their natural speech asynchronies, but still constitutes a useful experimental approach to study AV speech integration. To circumvent this issue, a future direction should include experimental designs where continuous AV speech is presented with naturally various AV delays which are in turn manipulated into an experimental design.

Conclusions

Large-scale oscillatory networks operating at multiple temporal scales are sensitive to the asynchronies of AV speech. These oscillatory networks are hierarchically structured in a bipartite network: higher-frequency activity is spatially nested within lower-frequency oscillatory activity. The effective cortical phase modulations within the low-frequency δ oscillations (1–3 Hz) revealed that SMG/AG was a central incoming hub top-down regulated by frontal regions. δ oscillations orchestrate temporal multiplexing in speech processing on the basis of the temporal statistics of incoming AV speech and regulate locally the neural synchronization of high-frequency responses. Altogether, our findings suggest that the fine-tuning of low-frequency oscillations entrained by the timing of AV speech enables the optimization of integrative processes at the finer temporal scales necessary for linguistic computations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability statements
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Participants
- METHOD DETAILS
  - ○ Stimuli
  - ○ Procedure
  - ○ MEG recordings
  - ○ MEG preprocessing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Event-related-fields (sensor space)
  - ○ Time-frequency analyses
  - ○ Functional connectivity analyses (source space)
  - ○ Directed-phase information transfer in the delta band

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106257.

ACKNOWLEDGMENTS

We thank the members of UNIACT for their logistical help on the recruitment of volunteers at NeuroSpin. We thank Manuel Vidal for his assistance in recording the original speech videos. We would like to thank Joachim Gross for insightful comments in a previous version of the manuscript and Stefanos Sgardelis for inspiring discussions. This work was supported by an European Research Council-YStG-263584 to V.v.W. and an Erasmus+ to N.C.

AUTHOR CONTRIBUTIONS

Designed research: V.v.W., D.O., and D.P.; Performed research: N.C., D.O., and V.v.W; Analyzed data: N.C., and V.v.W; Wrote the paper: N.C., and V.vW.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 1, 2022
Revised: December 22, 2022
Accepted: February 17, 2023
Published: February 20, 2023

REFERENCES

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

2. Erber, N.P. (1975). Auditory-visual perception of speech. J. Speech Hear. Disord. *40*, 481–492.

3. Grant, K.W., and Seitz, P.F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. *108*, 1197–1208.

4. MacLeod, A., and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. Br. J. Audiol. *21*, 131–141.

5. Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., and Foxe, J.J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebr. Cortex *17*, 1147–1153.

6. Giordano, B.L., Ince, R.A.A., Gross, J., Schyns, P.G., Panzeri, S., and Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. Elife *6*, e24763.

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

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

9. Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. *8*, e1000445.

10. Power, A.J., Mead, N., Barnes, L., and Goswami, U. (2012). Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. Front. Psychol. *3*, 216.

11. Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. Elife *5*, e14521.

12. Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., and Gross, J. (2018). Representational interactions during audiovisual speech entrainment: redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. PLoS Biol. *16*, e2006558.

13. Thézé, R., Giraud, A.-L., and Mégevand, P. (2020). The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. Sci. Adv. *6*, eabc6348.

14. Kayser, C., Petkov, C.I., and Logothetis, N.K. (2008). Visual modulation of neurons in auditory cortex. Cerebr. Cortex *18*, 1560–1574.

15. Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends Cognit. Sci. *12*, 106–113.

16. Lakatos, P., Chen, C.M., O'Connell, M.N., Mills, A., and Schroeder, C.E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron *53*, 279–292.

17. Mégevand, P., Mercier, M.R., Groppe, D.M., Zion Golumbic, E., Mesgarani, N., Beauchamp, M.S., Schroeder, C.E., and Mehta, A.D. (2020). Crossmodal phase reset and evoked responses provide complementary mechanisms for the influence of visual speech in auditory cortex. J. Neurosci. *40*, 8530–8542.

18. van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. Front. Psychol. *4*, 388.

19. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. PLoS Comput. Biol. *5*, e1000436.

20. Schwartz, J.-L., and Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLoS Comput. Biol. *10*, e1003743.

21. Simon, D.M., and Wallace, M.T. (2018). Integration and temporal processing of asynchronous audiovisual speech. J. Cognit. Neurosci. *30*, 319–337.

22. Karas, P.J., Magnotti, J.F., Metzger, B.A., Zhu, L.L., Smith, K.B., Yoshor, D., and Beauchamp, M.S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. Elife *8*, e48116.

23. Irwin, J., Avery, T., Brancazio, L., Turcios, J., Ryherd, K., and Landi, N. (2018). Electrophysiological indices of audiovisual speech perception: beyond the McGurk effect and speech in noise. Multisensory Res. *31*, 39–56.

24. Bernstein, L.E., Auer, E.T., Wagner, M., and Ponton, C.W. (2008). Spatiotemporal dynamics of audiovisual speech processing. Neuroimage *39*, 423–435.

25. Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. J. Speech Lang. Hear. Res. *52*, 1073–1081.

26. Besle, J., Fort, A., Delpuech, C., and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. Eur. J. Neurosci. *20*, 2225–2234.

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

28. Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. Psychophysiology *53*, 1295–1306.

29. Arnal, L.H., Morillon, B., Kell, C.A., and Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. J. Neurosci. *29*, 13445–13453.

30. Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429–433.

31. Arnal, L.H., Wyart, V., and Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat. Neurosci. *14*, 797–801.

32. Parise, C.V., Spence, C., and Ernst, M.O. (2012). When correlation implies causation in multisensory integration. Curr. Biol. *22*, 46–49.

33. Parise, C.V., and Ernst, M.O. (2016). Correlation detection as a general mechanism for multisensory integration. Nat. Commun. *7*, 11543.

34. Pesnot Lerousseau, J., Parise, C.V., Ernst, M.O., and van Wassenhove, V. (2021). Multisensory correlation computations in the human brain uncovered by a time-resolved encoding model. Preprint at bioRxiv. https://doi.org/10.1101/2021.01.28.428606.

35. Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. *15*, 511–517.

36. Gauthier, B., Pestke, K., and van Wassenhove, V. (2019). Building the arrow of time over time: a sequence of brain activity mapping imagined events in time and space. Cerebr. Cortex *29*, 4398–4414.

37. Sassenhagen, J., and Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology *56*, e13335.

38. Lange, J., Christian, N., and Schnitzler, A. (2013). Audio-visual congruency alters power and coherence of oscillatory activity within and between cortical areas. Neuroimage *79*, 111–120.

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

40. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. Nature *536*, 171–178.

41. Costa, L. da F., Rodrigues, F.A., Travieso, G., and Villas Boas, P.R. (2007). Characterization of complex networks: a survey of measurements. Adv. Phys. X. *56*, 167–242.

42. Petanidou, T., Kallimanis, A.S., Tzanopoulos, J., Sgardelis, S.P., and 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. Ecol. Lett. *11*, 564–575.

43. Jordano, P., Bascompte, J., and Olesen, J.M. (2002). Invariant properties in coevolutionary networks of plant-animal interactions. Ecol. Lett. *6*, 69–81.

44. Lewinsohn, T.M., Inácio Prado, P., Jordano, P., Bascompte, J., and M Olesen, J. (2006). Structure in plant-animal interaction assemblages. Oikos *113*, 174–184.

45. Corso, G., de Araujo, A.I.L., and de Almeida, A.M. (2011). Connectivity and nestedness in bipartite networks from community ecology. J. Phys, Conf. Ser. *285*, 012009.

46. Atmar, W., and Patterson, B.D. (1993). The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia *96*, 373–382.

47. Lobier, M., Siebenhühner, F., Palva, S., and Palva, J.M. (2014). Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. Neuroimage *85* (Pt 2), 853–872.

48. Biau, E., Schultz, B.G., Gunter, T.C., and Kotz, S.A. (2022). Left motor δ oscillations reflect asynchrony detection in multisensory speech perception. J. Neurosci. *42*, 2313–2326.

49. Jensen, O., and Colgin, L.L. (2007). Cross-frequency coupling between neuronal oscillations. Trends Cognit. Sci. *11*, 267–269.

50. Palva, S., and Palva, J.M. (2012). Discovering oscillatory interaction networks with M/EEG: challenges and breakthroughs. Trends Cognit. Sci. *16*, 219–230.

51. Tass, P., Rosenblum, M.G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., Schnitzler, A., and Freund, H.J. (1998). Detection ofPhase locking from noisy data: application to magnetoencephalography. Phys. Rev. Lett. *81*, 3291–3294.

52. Tort, A.B.L., Komorowski, R., Eichenbaum, H., and Kopell, N. (2010). Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. J. Neurophysiol. *104*, 1195–1210.

53. Canolty, R.T., and Knight, R.T. (2010). The functional role of cross-frequency coupling. Trends Cognit. Sci. *14*, 506–515.

54. Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. Science *313*, 1626–1628.

55. Akam, T., and Kullmann, D.M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. Nat. Rev. Neurosci. *15*, 111–122.

56. Nobre, A., Correa, A., and Coull, J. (2007). The hazards of time. Curr. Opin. Neurobiol. *17*, 465–470.

57. Baart, M., Stekelenburg, J.J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia *53*, 115–121.

58. Eskelund, K., Tuomainen, J., and Andersen, T.S. (2011). Multistage audiovisual integration of speech: dissociating identification and detection. Exp. Brain Res. *208*, 447–457.

59. Conrey, B., and Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. J. Acoust. Soc. Am. *119*, 4065–4073.

60. Maier, J.X., Di Luca, M., and Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. J. Exp. Psychol. Hum. Percept. Perform. *37*, 245–256.

61. Massaro, D.W., Cohen, M.M., and Smeele, P.M. (1996). Perception of asynchronous and conflicting visual and auditory speech. J. Acoust. Soc. Am. *100*, 1777–1786.

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

63. Crosse, M.J., Di Liberto, G.M., and Lalor, E.C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J. Neurosci. *36*, 9888–9895.

64. Mariani, M.S., Ren, Z.-M., Bascompte, J., and Tessone, C.J. (2019). Nestedness in complex networks: observation, emergence, and implications. Phys. Rep. *813*, 1–90.

65. von Stein, A., and Sarnthein, J. (2000). Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization. Int. J. Psychophysiol. *38*, 301–313.

66. Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. Science *304*, 1926–1929.

67. Csicsvari, J., Jamieson, B., Wise, K.D., and Buzsáki, G. (2003). Mechanisms of gamma

oscillations in the hippocampus of the behaving rat. Neuron *37*, 311–322.

68. Panzeri, S., Brunel, N., Logothetis, N.K., and Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. Trends Neurosci. *33*, 111–120.

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

70. Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., and Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. Elife *4*, e06213.

71. 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.

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

73. Crosse, M.J., Butler, J.S., and Lalor, E.C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. *35*, 14195–14204.

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

75. Arnal, L.H., and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. Trends Cognit. Sci. *16*, 390–398.

76. Peelle, J.E., and Sommers, M.S. (2015). Prediction and constraint in audiovisual speech perception. Cortex *68*, 169–181.

77. Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., and Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol. *25*, 1649–1653.

78. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M.S. (2014). MNE software for processing MEG and EEG data. Neuroimage *86*, 446–460.

79. Gross, J., Baillet, S., Barnes, G.R., Henson, R.N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., et al. (2013). Good practice for conducting and reporting MEG research. Neuroimage *65*, 349–363.

80. Dale, A.M., Liu, A.K., Fischl, B.R., Buckner, R.L., Belliveau, J.W., Lewine, J.D., and Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron *26*, 55–67.

81. Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., and 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*, 1548–1565.

82. Niso, G., Bruña, R., Pereda, E., Gutiérrez, R., Bajo, R., Maestú, F., and del-Pozo, F. (2013). HERMES: towards an integrated toolbox to characterize functional and effective brain connectivity. Neuroinformatics *11*, 405–434.

83. Zalesky, A., Fornito, A., and Bullmore, E.T. (2010). Network-based statistic: identifying differences in brain networks. Neuroimage *53*, 1197–1207.

84. Oksanen, A.J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P.R., Hara, R.B.O., Simpson, G.L., Solymos, P., et al. (2013). Vegan: community ecology package. R Package Version *3*, 0–291.

85. R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). https://www.R-project.org/.

86. Schreiber, T. (2000). Measuring information transfer. Phys. Rev. Lett. *85*, 461–464.

87. Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica *37*, 424.

88. Barnett, L., Barrett, A.B., and Seth, A.K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. Phys. Rev. Lett. *103*, 238701.

89. Scott, D.W. (1979). On optimal and data-based histograms. Biometrika *66*, 605–610.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Python 3.7	Anaconda	https://www.anaconda.com
MATLAB R2018a	The Mathworks	RRID:SCR_001622
R	https://www.r-project.org	RRID: SCR_001905
MNE-Python	https://github.com/mne-tools/mne-python	https://mne.tools/stable/index.html
PsychToolBox	http://psychtoolbox.org	RRID: SCR_002881
FreeSurfer	Open Source	http://surfer.nmr.mgh.harvard.edu/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Nikos Chalas (nchalas@uni-muenster.de).

Materials availability statements

This study did not generate new unique reagents.

Data and code availability

- Anonymized MEG and structural MRI data are available on demand by emailing the lead contact.
- This study did not report original code.
- Any additional information required is available from the lead contact up request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

Sixteen volunteers (9 females; mean age = 24.94; SD = 4.53) 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).

METHOD DETAILS

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 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 thus consisted of 75 frames in total. In both [pa] and [ta] videos, the onset of the first visible mouth movement occurred on the 10th frame *i.e.* 200 ms after the video onset.

In the AV trials, the auditory speech signals were presented at any of four possible timings relative to the naturally measured audiovisual asynchronies (Figure 1A, natural audiovisual asynchrony indicated in red). The natural asynchrony for the [ta] syllable being 200 ms, the auditory utterance for the natural asynchrony stimulus occurred 200 ms after the visual speech onset. The natural asynchrony for the recorded [pa] syllable being 600 ms, the auditory utterance occurred 600 ms after the visual speech onset. Throughout text, we will refer to the natural asynchronies as A0V.

Both syllables [pa] and [ta] were physically desynchronized (with respect to their natural asynchronies) with identical delays such that the acoustic signal and the first visual movement were desynchronized by 200 ms, 400 ms, 600 ms or 800 ms (Figure 1). For the [pa] syllable, these four physical timing conditions corresponded to the acoustic speech preceding the natural speech asynchrony by 600 ms, 400 ms and 0 ms, and lagging behind it by 200 ms. These asynchronies were mapped to the natural synchrony and will be referred to as A600V, A400V, A0V, and V200A. For the [ta] syllable, the same four physical timing conditions corresponded to the acoustic signal preceding the natural speech asynchrony by 200 ms and 0 ms and lagging behind it by 400 and 600 ms. We refer to these [ta] asynchronies as 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. The identification of the natural asynchrony was done for both syllables through visual inspection of the naturally articulated video recording. 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 was given (Figure 1B).

The experiment consisted of three types of blocks. In the first type of block (AV), only audiovisual stimuli were presented. A trial was composed of a sequence of 6 successive and identical syllables (standards) followed by a target syllable (Figure 1C). 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 in the other blocks. As in the AV blocks, each trial in the audio-only and video only blocks ended with the presentation of a target stimulus that was identical to 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. The audio and visual-blocks were collected but not analyzed here.

All audiovisual asynchronies and trigger timing were carefully checked with photodiode, microphone, and oscilloscope measurements to ensure 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 onto 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 guided participants to attend to the presented stimuli. Participants were instructed to use the response box to indicate by button press whether the final target stimulus was the same as 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 fourth and eighth 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 for later coregistration with MRI data. The sampling rate for MEG acquisition was set to 1 kHz with a band-pass filter of 0.03–330 Hz.

MEG preprocessing

Data were preprocessed with the MNE-python toolbox[78] in accordance with accepted guidelines for MEG research.[79] Noisy MEG sensors were identified manually and interpolated using 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), also matching their activity with EOG and ECG signals. Prior to epoching, data were low-passed filtered at 120 Hz and downsampled to 500 Hz. Cortical surfaces and 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 estimated. Cortical surfaces extracted from FreeSurfer were projected to 5120 vertices with 4.9 mm spacing per hemisphere. The inverse solution was computed using the dynamic statistical parametric mapping [dSPM;[80]] inverse operator, with a loose orientation constraint (loose = 0.2, depth = 0.8) and a source covariance matrix estimated from the baseline activity. Subsequently, the cortical surface was parceled according to HCP-MMP1.0 into combined whole-brain 46 functional labels,[40] as provided by MNE-python.

QUANTIFICATION AND STATISTICAL ANALYSIS

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). Both epochs represent the brain responses from AV syllables, with different triggering. Baseline correction, for both, was applied using the −100 ms–0 ms interval. Epochs containing signals exceeding 7 fT/cm for gradiometers and 7000 fT for magnetometers were considered artifacts and rejected from the analysis. Visual epochs stand for epochs starting at the beginning of the video (using a visual trigger) and auditory epochs stand for epochs starting at the onset of the sound (with an auditory trigger). Thus, auditory epochs were used when auditory-locked responses were of interest of the analysis (evoked analysis and phase-information transfer), whereas the longer in time visual epochs were used when auditory phase response were not of essential interest and were analyzed to account for predictive processes after the visual onset (bipartite network analysis). Evoked responses were generated individually, and were the average of an equal number of epochs across all AV blocks for each experimental condition. This ensured an equal signal to noise ratio for each condition. As described above, the syllables were repeated six times before the presentation of a target stimulus, and only the data from the repeated AV syllables were analyzed. We decided to analyze signals from each repeated syllable to increase the signal-to-noise ratio for the effective stimulus-responses. That undoubtedly leads to a concern about a repetition effect which could affect our findings. In our case though, the repetition effect would be the same in all experimental conditions (as there is an equal number of repeated syllables per AV condition), and thus it would be disregarded by the statistical comparisons we followed throughout the manuscript. The analysis was restricted to magnetometers, for simplicity of topographical interpretation.

To investigate the effect of AV speech asynchronies, subject-wise linear regressions using the amplitude of the auditory evoked responses for each trial, latency, and sensor were estimated. Separate linear regressions were performed to test the effect of AV speech asynchronies, as function of and irrespective of the syllable. This analysis resulted in time series of linear regression coefficients (β values) per individual and per sensor. The regression coefficients time series were submitted to a cluster spatiotemporal one-sample t-test (corrected for multiple comparisons) in order to identify the significant clusters of sensors and their latencies.

Time-frequency analyses

We performed a time-frequency analysis on auditory-locked single-trial source estimates using Morlet wavelet transforms (from 1 to 120 Hz). To deal with the trade-off between temporal and frequency precision, the number of wavelet cycles was estimated as a function of frequency (number of cycles = *f/2*). Baseline correction was applied to power and inter-trial coherence (ITC), by dividing by the mean activation of the pres-stimulus baseline activity (−200 to 0 ms) and taking the log. The evoked response was removed from the epoch (Figures 4B and 4C) so as to provide a description of induced power and ITC (removing the evoked component). The log-transformed data were normally distributed, allowing the use of parametric statistics: the statistical analysis of power and ITC consisted in a 2 × 4 mixed model ANOVA at the network-level with within-subject factors of order (2: AV or VA) and distance from A0V (4: 0, 200, 400, 600).

Functional connectivity analyses (source space)

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

To estimate the pairwise phase synchronizations between cortical regions across different frequency bands, we calculated the weighted phase-lag index or wPLI.[81] wPLI is a robust estimator of phase synchronization, exhibiting greater sensitivity than other phase-synchronization metrics, in detecting phase-synchronization and changes.[82] In this analysis, all visual-locked epochs were used for the estimation of wPLI in single-trial source estimates. The 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). 46 × 46 adjacency matrices were created for each SOA and each frequency band, each one depicting a functional network (Figures S1 and 3A–3C).

To investigate the effect of AV speech asynchronies in different oscillatory networks, we performed a network-level statistical analysis using the Network-Based-Statistics toolbox.[83] Specifically, a 2 × 4 mixed-model ANOVA with two within-subject factors, syllable ([pa] and [ta]) and asynchrony (SOA: 0, 200, 600, and 800 ms), was designed to assess the main effect of asynchrony on what. The significance level was set to $p < 0.05$, corrected for multiple comparisons using FDR correction.

For each statistically significant oscillatory network (δ, θ, α, β, γ, and hγ), we calculated the node strength, which is the sum of the weights (*i.e*, F-values) of connections (or edges) which each node carried. The node strength of a node within a network depicts the sum of incoming and outgoing connections that each node contributes to the network.

Subsequently, a 46 × 6 adjacency matrix was constructed (46 node strengths for each frequency band) and treated as a bipartite network. The nodes in bipartite networks were divided into two groups: node strength and frequency band in our case. The connections were drawn only between those two groups. The temperature of the matrix represents a measure of (dis)order[46] and was calculated as a metric of bipartite nestedness, using the function *nestedtemp* in the R package *vegan*.[84,85] Nestedness in bipartite networks is a measure originating from community ecology, quantifying the functional structure within interacting systems (in our case between cortical labels and frequency bands).

Directed-phase information transfer in the delta band

To illustrate the phase modulation of auditory cortex responses, we used auditory-locked epochs. We estimated the Phase Transfer Entropy[47] of single-trial MEG activity in the 46 cortical labels previously used in defining the oscillatory networks. Transfer Entropy[86] quantifies causal statistical dependencies between two signals: Transfer Entropy measures the extent to which knowledge of signal X can reduce the 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, Transfer Entropy aligns well with Granger Causality[87,88] in comparing conditional nonlinear probability distributions via Kullback-Leibler divergence for the estimation of statistical dependencies. Phase Transfer Entropy or pTE extends this notion into phase time-series, evaluating the influence of signal X's phase on the signal Y's 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.[89] The time-delay was set as the average pairwise time needed for a sign-flip. Single-trial 46 × 46 adjacency matrices were averaged per condition, resulting in an adjacency matrix for each SOA per participant. Thereafter, the statistical analysis was similar to the previous functional connectivity analysis.