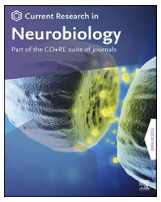
Current Research in Neurobiology 2 (2021) 100014

Contents lists available at ScienceDirect

Current Research in Neurobiology

journal homepage: www.editorialmanager.com/crneur/Default.aspx



Auditory detection is modulated by theta phase of silent lip movements

Emmanuel Biau [a,b,*], Danying Wang [a,b], Hyojin Park [a,b], Ole Jensen [a,b], Simon Hanslmayr [a,b,c]

[a] School of Psychology, University of Birmingham, Edgbaston, Birmingham, UK
[b] Centre for Human Brain Health, University of Birmingham, Birmingham, UK
[c] Centre for Cognitive Neuroimaging, Institute for Neuroscience and Psychology, University of Glasgow, Glasgow, UK

ARTICLE INFO

Keywords:
Lip movements
Theta oscillations
Entrainment
Auditory processing

ABSTRACT

Audiovisual speech perception relies, among other things, on our expertise to map a speaker's lip movements with speech sounds. This multimodal matching is facilitated by salient syllable features that align lip movements and acoustic envelope signals in the 4–8 Hz theta band. Although non-exclusive, the predominance of theta rhythms in speech processing has been firmly established by studies showing that neural oscillations track the acoustic envelope in the primary auditory cortex. Equivalently, theta oscillations in the visual cortex entrain to lip movements, and the auditory cortex is recruited during silent speech perception. These findings suggest that neuronal theta oscillations may play a functional role in organising information flow across visual and auditory sensory areas. We presented silent speech movies while participants performed a pure tone detection task to test whether entrainment to lip movements directs the auditory system and drives behavioural outcomes. We showed that auditory detection varied depending on the ongoing theta phase conveyed by lip movements in the movies. In a complementary experiment presenting the same movies while recording participants' electro-encephalogram (EEG), we found that silent lip movements entrained neural oscillations in the visual and auditory cortices with the visual phase leading the auditory phase. These results support the idea that the visual cortex entrained by lip movements filtered the sensitivity of the auditory cortex via theta phase synchronization.

Introduction

When hearing gets difficult, people often visually focus on their interlocutors' mouth to match lip movements with sounds and to improve speech perception. Mouth opening indeed shares common features with auditory speech envelope, which temporally synchronize on dominant 4–8 Hz theta rhythms imposed by syllables (Park et al., 2016; Chandrasekaran et al., 2009; Luo and Poeppel, 2007). The present study focuses on theta activity conveyed by moving lips because the speaker's mouth provides a direct source of visual speech information matching sounds. In the brain, neural oscillations from the auditory cortex track the auditory envelope structure during speech perception, suggesting that this "entrainment" reflects signal analysis (Keitel et al., 2018; Peelle and Davis, 2012; Gross et al., 2013; Giraud and Poeppel, 2012). Although the term entrainment is currently under debate (Meyer et al., 2019; Obleser and Kayser, 2019; Haegens and Zion Golumbic, 2018; Rimmele et al., 2018), here we use it to describe neural patterns tracking salient features conveyed in speech signals which occur at theta frequency (4–8 Hz).

Previous studies demonstrated that the perception of moving lips entrains oscillations in the visual cortex and modulates activity in the auditory regions, although not limited to the theta band (Bourguignon et al., 2020; Crosse et al., 2019). Further, information specific to lip movements is represented not only in the visual cortex but also in the auditory cortex (Park et al., 2018). An fMRI study reported overlapping activations in the bilateral primary auditory cortex during the perception of isolated words presented either visually (i.e., silent moving lips) or aurally (Calvert et al., 1997). More recently, a study used intracranial recordings in epileptic patients to investigate the neural responses evoked by the perception of lip movements in the auditory cortex during the presentation of syllables in uni- or multimodal conditions (Besle et al., 2008). They reported activations in response to silent lip movements in the visual cortex followed by similar responses in the secondary auditory cortex, suggesting crossmodal activation via direct feedforward processes. However, all these results beg the question of whether visual perception of lip movements modulates the auditory cortex in a functional way. In other words, do purely visually induced theta speech rhythms impose time windows that render the auditory cortex more sensitive to inputs in a phasic manner? If the answer to this question is yes, then visually focusing on your interlocutor's mouth when you have trouble understanding them would indeed be an effective filter modulator to increase auditory sensitivity.

* Corresponding author. School of Psychology, University of Birmingham, B15 2TT, Birmingham, UK.
E-mail address: e.biau@bham.ac.uk (E. Biau).

https://doi.org/10.1016/j.crneur.2021.100014
Received 21 November 2020; Received in revised form 12 May 2021; Accepted 19 May 2021

Results

Entrainment to lip movements during silent speech drives behavioural performance

To address this question, we adapted an auditory tone detection paradigm in which a continuous white noise was presented simultaneously with silent movies displaying speakers engaged in conversations (Fig. 1A and see Material and Methods; Movie1 and Sound1 for examples). Participants were instructed to press a key as fast and accurate as possible every time they detected a pure tone (1 kHz, 100 ms) embedded in the white noise at individual threshold (determined with a calibration task). In the condition of interest, there were two target tones: the first tone occurred randomly in the first half of the trial (0–2.5 s after trial onset) and the second tone occurred randomly in the second half of the trial (2.5–5 s). Importantly, the random generation of tone onsets equivalently sampled the phases of visual lips activity by chance in the first and second half windows, and guaranteed sufficient trials per bin (Fig. 1B and see Material and Methods for random-distribution statistical assessment). Two additional conditions containing zero or one single tone were introduced to estimate the false alarm rates (FA) and to reduce the predictability of the second tone by the occurrence of the first one. The three conditions were counterbalanced and randomised across six blocks of 50 trials (100 trials per condition).

To test the first hypothesis of visual entrainment affecting auditory processing, a random movie was displayed with the sound in each trial. Participants were asked to attend carefully to the silent movies centred on the speakers' nose, albeit non-informative to perform the tone detection task. Crucially, the videos were preselected such that lip movements occurred in the 4–8 Hz theta range. We determined at which theta frequency the vertical mouth apertures and auditory speech envelope showed significant dependencies in the original clips by using mutual information method (see Material and Methods section). This paradigm allowed us to link directly the onset of detected tones with the phase of the ongoing theta activity conveyed by the lip movements. As neural entrainment increases over time (Thut et al., 2011; Hanslmayr et al., 2019), we compared behavioural performance between the early and late time-windows (containing respectively the first and second tones).

We addressed whether visual entrainment predicted detection performances depending on the tone position in the two tones condition by adopting two different analytic approaches. In a first approach, we assumed that participants entrained at different preferred phases, and we tested for a phase locking within participants after realigning individual phases on their preferred phase (Fig. 2; see Material and Methods). Such approach allowed detecting the presence of a phasic modulation independently from the individual mean phase at the two separate tones. Second, we compared the amplitude of the phasic modulation between correctly detected (hit trials) and missed tones (miss trials) to test how visual entrainment may account for auditory detection performances at the first and second tones. We hypothesized that visual entrainment shapes auditory perception by tuning auditory activity more consistently towards a more optimal state for when the tones' onsets occur, and increasing their detection. In contrast, we expected inconsistent tuning of the auditory system at the onsets of the missed tones reflected by a weaker phasic modulation. One might argue that a similar phase modulation in the miss trials exists as for hit trials, where the auditory system

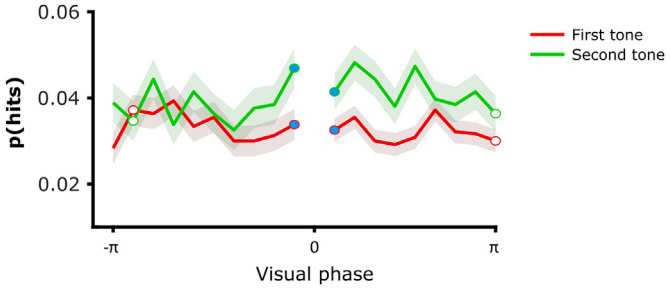


Fig. 2. Phasic modulation along the realigned phase in the two tones condition. Probability of correctly detected tones p(hits) = hits/(hits + misses) along the visual phase realigned on the preferred bin (0° bin, not plotted) in the two tones condition. The red line depicts p(hits) at the first tone and the green line depicts p(hits) at the second tone (mean ± standard deviation). The phase modulation was estimated by subtracting the average response of the two bins adjacent to the bin opposite to the preferred phase (white dots) from the average response of the two bins adjacent to the preferred phase (blue dots).

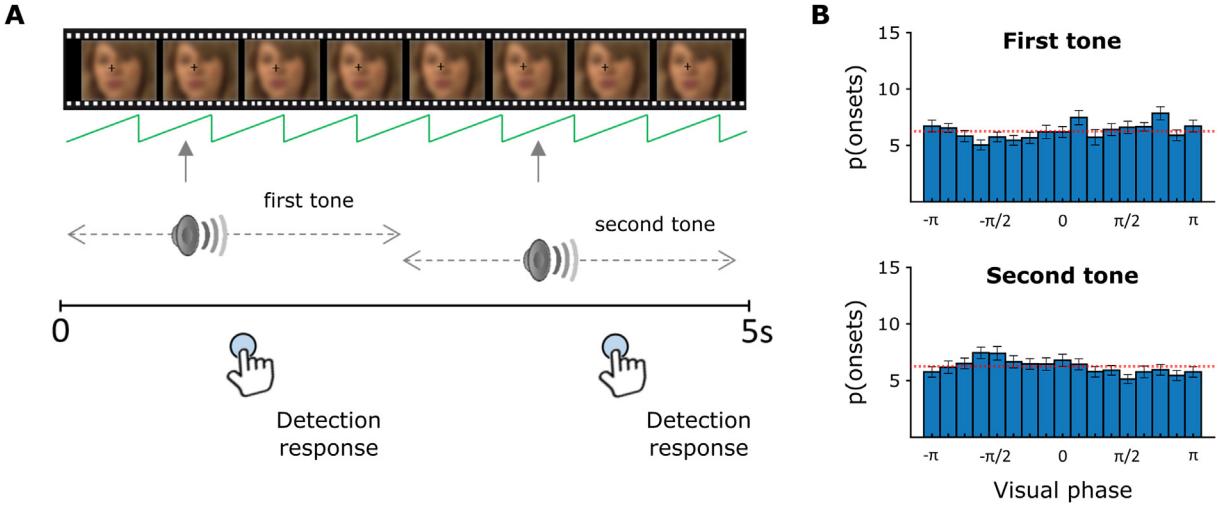


Fig. 1. Experimental Paradigm of the Tone Detection Task (TDT). (A) In each trial, continuous white noise and a silent movie were presented together for 5 s. A first pure tone occurred randomly in the first half of the trial while a second tone occurred randomly in the second half of the trial. Participants were instructed to respond as fast and accurately as possible whenever they detected a tone. In the one tone condition, the white noise track contained only one tone that occurred randomly between the two halves of the trial. In the zero tone condition, the sound of the trial contained only white noise (*N.B.* The face of the speaker has been blurred only in the figure for anonymity purpose). (B) Distribution of tone onsets along the visual phase in the first (top) and second (bottom) windows of the main condition of interest. The x-axis represents the visual phase binned in equidistant bins (from -π to +π; π/8 step). The y-axis depicts the proportion of tones occurring in each bins across participants (in proportion from total tones ± standard error of the mean). The red dashed lines represent the theoretical proportion of tone onsets expected in each bin for a random uniform distribution (n bins = 16; 6.25% per bin).

is entrained to a non-optimal phase. However, it is important to note that participants may have not detected auditory tones for a number of reasons like attentional lapses, fatigue, or distracting thoughts. Consequently, the resultant phase modulation in the miss trials embedded either the same preferred phase as for hits but with more variability across trials, or a distinct preferred phase angle reflecting a less optimal state when tones' onsets occurred and leading to missing more tones. Therefore, we expected the phase modulation to predict behavioural outcomes, with a greater phase modulation in hit trials as compared to miss trials. If true, this would be the case only during the second tone window but not during the first tone window, i.e. when visual entrainment though lip movement perception has built up sufficiently. As a first step, we tested the existence of a significant phase modulation when participants effectively detected the tones in each condition (hit trials only). We computed the probability of correctly detecting the tone p(hits) = hits/(hits + misses) as a single measure of performance. Phasic modulation was computing by subtracting the average response of the two bins adjacent to the bin opposite to the preferred phase, i.e. with the highest p(hits), from the average response of the two bins adjacent to the preferred phase (Fig. 2). The existence of a significant phase modulation at the first and second tone in each condition was assessed statistically by comparing the mean phasic modulations against zero by applying independent one-sampled T-tests (one-tailed; p-value adjusted for multiple comparisons with a Bonferroni correction). Results revealed a strong tendency for the mean distance to be significantly greater than zero at the second tone ($t(1,23) = 2.356$; $p_{adjusted} = 0.055$; Cohen's $d = 0.481$), but not at earlier first tone ($t(1,23) = -0.046$; $p_{adjusted} = 1$; Cohen's $d = -0.001$). Although the p-value corrected for multiple comparisons was slightly above $\alpha = 0.05$, this result suggests the existence of a phase modulation driving the correct detection of the second tones but not the earlier first tones in our condition of interest. Following up, we compared the phasic modulation between hit and miss trials. For each participant, the individual phase was binned and realigned on the bin yielding the greater response for hit and miss trials separately. Visual entrainment in hit and miss trials was estimated separately by computing the distance between the preferred bin and a baseline averaged across bins. The phase modulation at first and second tones was estimated by the t-value from a paired-sample T-test (one-tailed) comparing the statistical difference of visual entrainment between hit and miss trials (effect size = $t\text{-value}_{hit\text{-}miss}$; see (Zoefel et al., 2019)). To circumvent the unbalanced numbers of hit and miss trials, we tested the phase modulation in the first and second tone windows with two separate permutation tests after randomly shuffling the hit and miss labels from original data. The two permutation tests testing the original t-value against the t-values of the permuted data revealed that the phasic modulation predicted performance in the second tone window (permutations: 10,000; effect size = 2.214; $p < 0.001$), but not in the first tone window (permutations: 10,000; effect size = −0.495; $p = 0.708$). A third permutation test addressed the interaction between phase modulation and tone position by assessing the statistical difference of phase modulation between the second and first tone windows, i.e. $[t\text{-value}_{hit\text{-}miss}]_{second\ tone}$ - $[t\text{-value}_{hit\text{-}miss}]_{first\ tone}$. Results revealed that the difference of phase modulation was significantly greater in the second tone window as compared to the first tone window, confirming an interaction between phase modulation and tone position (permutations: 10,000; effect size = 0.819; $p = 0.026$). These results established the presence of a visual phase modulation that predicted tone detection performance at the second tone but not earlier at the first tone.

As participants viewed the same videos, which theoretically imposed the same modulations onto the brain activity, we hypothesized that a similar phase should emerge across participants. Accordingly, we tested visual entrainment across participants with a second analytic approach. We compared the mean phase distributions between first and second tone onsets across participants (Fig. 3A; see Material and Methods). For each participant, the corresponding phases in ongoing lip activity at detected first and second tone onsets were averaged across hit trial only. Individual mean phases were then averaged to estimate the mean phase locking across subjects to the theta signal conveyed visually in the first and second tone time-windows. Two Rayleigh's uniformity tests were performed on the first and second grand average theta phase distributions separately. For the first tone window, the Rayleigh's test did not reject the hypothesis of uniform distribution ($n = 24$; $\mu = 1.944$ rad or 111.384°; $r_{first\ tone} = 0.282$; $p = 0.148$, Bonferroni-corrected). In contrast, the Rayleigh's test revealed that mean phases were not uniformly distributed in the second tone window ($n = 24$; $\mu = -0.999$ rad or 302.763°; $r_{second\ tone} = 0.44$; $p < 0.01$, Bonferroni-corrected). Further, a permutation test was performed on the resultant vector length r difference between the first and second tones (effect size: $z\text{-value}_{hits\ second\ tone}$ - $z\text{-value}_{hits\ first\ tone}$). Permutations were computed after randomly shuffling the tone position information from the original data. This analysis addressed whether visual entrainment of correct trials in the second tone

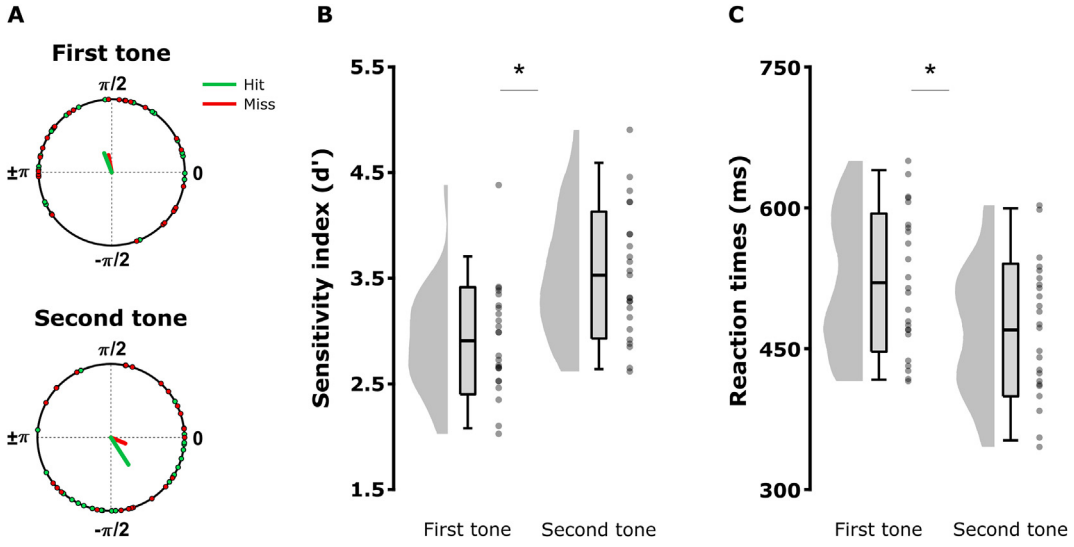


Fig. 3. Visual entrainment and tone detection performance in the two tones condition. (A) Resultant vector length r from grand average phase at the onset of first and second tones across participants (hit trials: green line; miss trials: red line). The individual mean theta phases are depicted in polar coordinates (hit trials: green circles; miss trials: red circles). (B) Mean sensitivity index (*d'*) and (C) reaction times of first and second tone hits. The graphs depict the density, the grand average (mean ± standard deviation; errors bars indicate 5th and 95th percentiles), and individual means (grey dots) for first/second tones. Significant contrasts are evidenced with stars ($p < 0.05$).

window was significantly stronger than in the first tone window, which indeed was the case (permutations: 10,000; effect size = 0.130; p = 0.021; green lines in the Fig. 3A). These results confirmed that visual entrainment across participants was greater (i.e. more consistent mean phase) for the tones detected in the second window, as compared to the ones detected earlier in the first window.

Next, we assessed whether the modulation of the visual phase computed across participants predicted tone detection performance in the first and second windows. As for the previous analytic approach, we compared hit and miss results to establish whether visual entrainment accounted for auditory detection performances differently in the first and second tone windows. If the visual system entrained more efficiently to lip movements leading to improve tones' detection in the hit trials, we expect a longer resultant vector length at the detected tones' onsets as compared to missed tones. In line with the results from the first analytic approach, we expect to find a greater vector length for the hit trials as compared to miss trials only in the second tone window when visual entrainment has built up through perception of lip movement. We applied a permutation-based approach similar to the one described previously at individual phase-locking level. Here, we tested the difference of resultant vector length between hit and miss trials across participants for the first and second tones (effect size: $[\text{z-value}_{\text{hit-miss}}]_{\text{second tone}}$ - $[\text{z-value}_{\text{hit-miss}}]_{\text{first tone}}$; green lines versus red lines in the Fig. 3A). Permutations were computed after randomly shuffling the hit and miss labels from the original data. Results revealed that the phase modulation across participants between hit and miss trials was significantly greater in the second tone window as compared to first tone window (permutations: 10,000; effect size = 0.187; p = 0.037). We performed two additional permutation tests on resultant vector length differences between hits and misses in the first and second tone windows separately to test whether visual entrainment was related to successful auditory processing. No significant difference of vector length was found in the first tone window (permutations: 10,000; effect size = 0.041; p = 0.405), whereas in the second tone window the resultant vector for hits tended strongly to be longer compared to misses (permutations: 10,000; effect size = 0.228; p = 0.052). These results confirmed the presence of a mean visual phase modulation across participants predicting tone detection performance in the second tone window, but not in the first tone window. Altogether, the two approaches (i.e. assuming either an individual phase locking, or a similar mean phase across participants) showed that the visual phase modulation induced by theta lip activity predicted better the detection of second tones and support the hypothesis of visual entrainment shaping auditory perception.

Finally, we investigated whether tone detection differed between the first and second tones windows across participants. Such a difference might reflect an auditory bias by visual inputs (Fig. 3B). We computed the hit and false alarm (FA) rates to calculate the sensitivity index ($d' = Z_{\text{Hit rate}} - Z_{\text{FA rate}}$). False alarms were calculated by sorting participant's responses in the absence of a tone, i.e. in the zero tone condition and according to their onsets (occurring either in the first window or in the second window). As FA rates were very low ($FA_{\text{first tone}} = 0.014 \pm 0.016$; $FA_{\text{second tone}} = 0.013 \pm 0.019$), extreme hit and FA rate values were adjusted using a standard d' correction by replacing rates of FA rate = 0 with FA rate = $0.5/n_{\text{noise trials}} = 0.005$ and hit rate = 1 with hit rate = $(n-0.5)/n_{\text{signal trials}} = 0.995$ (Macmillan and Kaplan, 1985). First, two independent one-sample t-tests established that participants detected the first and second tones in the two tones condition, as the d' scores were greater than zero (first tone: T(1,23) = 28.02; p < 0.001, two-tailed; second tone: T(1,23) = 28.699; p < 0.001, two-tailed). Second, a paired-samples *t*-test confirmed that the second tones were better detected than the first ones (T(1, 23) = 5.771 p < 0.001; two-tailed; Fig. 3B). Third, a paired-sample *t*-test applied on the hit reaction times showed that participants responded faster to second compared to first tones (T(1, 23) = 5.486; p < 0.001; two-tailed; Fig. 3C). Therefore, the tone position predicted performances in the TDT (d' and reaction times). Importantly, the improvement of the second tone detection could not be attributed to a simple attentional effect due to the presence of the preceding first one, as the single tone condition replicated the two tones condition performances (i.e. by sorting the single tones as first/second tones according to their onsets; see Figure S1B). Finally, a paired-samples *t*-test performed on the FA rates in the no tone condition confirmed that detection performance modulations did not reflect a change in response bias between the two windows (T(1, 23) = 0.627; p = 0.537; two-tailed). Additionally, the hit rates in both conditions confirmed that the calibration task worked efficiently with mean hit rate = 0.77 ± 0.10 in the two tones condition (first tone = 0.69 ± 0.13; second tone 0.85 ± 0.98) and mean hit rate = 0.74 ± 0.15 in the single tone condition (first tone = 0.67 ± 0.18; second tone = 0.81 ± 0.12). The effect of visual entrainment to theta lip activity on auditory perception was established by (1) the phasic modulation that predicted performance in the late window and, (2) the improvement of tone detection in the late window. In the next step, we addressed whether the perception of the same silent movies synchronised visual and auditory cortices through theta oscillations.

Visual cortex leads synchronization to left auditory cortex via theta oscillations during silent lips perception

The above results suggest that visual speech stimuli may recruit the auditory regions via entrainment to render some time-windows more sensitive to auditory detection than others. To test this hypothesis on a neural level, we recorded the EEG signal of 23 new participants during the perception of the same 60 silent movies used in the previous tone detection task. Addressing the neural correlates in a separate procedure was motivated to avoid signal contamination induced by the participants' motor responses to detected auditory tones. In doing so, we ensured that EEG signals only reflected the perception of silent moving lips. Participants were instructed to attend to each movie and rate its emotional content based on the speaker's face. The movies were presented in a single block and randomised across participants. First, the sources of interest responding to speakers' lip movements were identified applying a linearly constrained minimum variance beamforming method. Neural entrainment to lip movements was estimated by computing mutual information (MI) on the theta phase between the EEG epochs and corresponding lip signals in the equivalent first tone (0–2.5 s; early time-window) and second tone windows (2.5–5 s; late time-window). Just as in the behavioural data, we assessed whether entrainment increased over time by contrasting the MI between the early and late time window. Second, the EEG data at the identified visual and auditory sources were reconstructed to perform single-trial phase coupling analysis. The synchrony between visual and auditory sources was reflected by the distribution of theta phase angle differences $\phi_{A\text{-}V} = \phi_{\text{audio}} - \phi_{\text{visual}}$ at each time-point within the early and late time-windows, and the directionality of the coupling was evidenced with the sign of $\phi_{A\text{-}V}$ (i.e. a mean distribution of $\phi_{A\text{-}V} = 0$ would mean perfect phase alignment, while $\phi_{A\text{-}V} < 0$ would mean that the visual phase leads the auditory phase; see Material and Methods). To address potential issues of circularity, the directionality of information flow was also assessed using the Phase Slope Index (PSI).

Source localization analysis revealed that the maximum increases in MI_{late} as compared to MI_{early} were localised in the left visual and auditory cortices, as well as in the right visual cortex to a lesser extent (Fig. 4A). This result confirmed the expected recruitment of both visual and auditory sensory areas during the perception of speakers' lip movements even in the absence of speech sound based on previous studies (Crosse et al., 2019; Calvert et al., 1997). Two separate Rayleigh tests confirmed non-uniform distributions of $\phi_{A\text{-}V}$ in the early (n = 23; μ = −1.31 rad or −75°; r = 0.637; p < 0.001; Bonferroni-corrected) and late time-windows (n = 23; μ = −0.52 rad or - 30°; r = 0.919; p < 0.001; Bonferroni-corrected). An additional Kuiper two-sample test showed that the mean $\phi_{A\text{-}V}$ distributions between the early and late time-windows converged towards two different preferred angles ($k = 3.24 \times 10^5$; p < 0.001; Fig. 4C). The negative theta phase angle differences $\phi_{A\text{-}V}$ in both the early and late time-windows confirmed that the visual phase led the

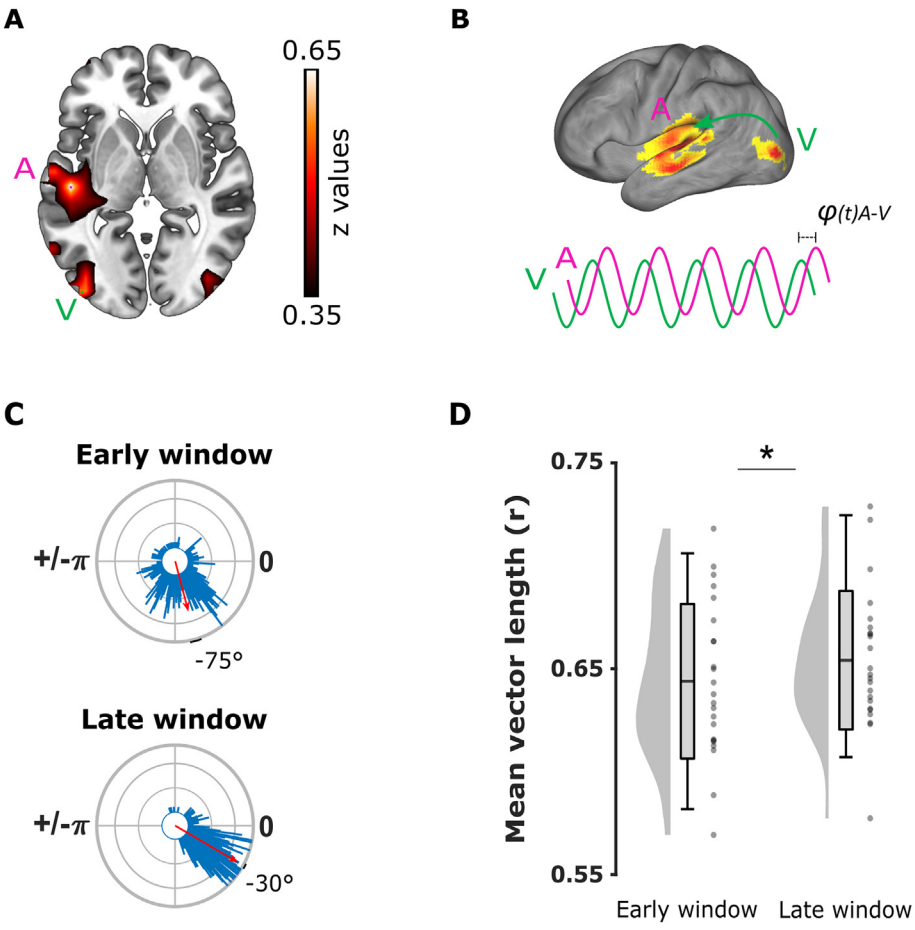


Fig. 4. Theta phase coupling analysis between visual and auditory areas during lips perception. (A) Difference of mutual information between the late and early time-windows ($MI_{late} > MI_{early}$ contrast; z values; coordinate of the slice: $z = 0$). Auditory (Pink dot; MNI coordinates of maximum voxel: [-50 -21 0]; Left Middle Temporal cortex) and visual (Green dot; MNI coordinates of maximum voxel: [-40 -89 0]; Left Middle Occipital cortex) sources were localized in the left hemisphere. (B) $MI_{early} > MI_{late}$ contrast projected on brain's surface for illustrative purpose: Synchronization was estimated through $\phi_{A\text{-}V}$ theta phase offset between theta oscillations at identified auditory (pink line) and visual sources (green line) by mean of phase coupling analysis. (C) Audio-visual phase coupling in the early and late time-windows corresponding to the time-windows containing the first and second tones in the TD Task. The mean $\phi_{A\text{-}V}$ offset between auditory and visual theta phases (red arrows) confirmed that oscillations entrained by lip movements in the visual cortex preceded oscillations in the auditory cortex by 75° (~37 ms) and 30° (~15 ms), respectively in the early and late time-windows. (D) Theta synchronization between visual and auditory areas improves with entrainment. The resultant vector length r of the distance between the observed $\phi_{A\text{-}V}$ and the theoretical $\phi_{A\text{-}V} = 0$ was greater in the late than the early window, suggesting a dynamic communication reflected by a decrease of time lag between visual and auditory activities. The graphs depict the density, the grand average (mean ± standard deviation; errors bars indicate 5th and 95th percentiles), and individual resultant vector length r (grey dots). Significance evidenced with a star ($p < 0.05$).

auditory phase, in line with the idea of visual oscillations responding first to the lips inputs and then directing theta oscillations in the auditory cortex. Additional PSI analyses revealed negative values in both time-windows (respectively $PSI_{early} = -0.029 \pm 0.068$ and $PSI_{late} = -0.037 \pm 0.06$; seed region: auditory source) and confirmed that the left visual source led the left auditory source during silent moving lips perception. Together, these results support our hypothesis that visual cortex led synchronization to left auditory cortex via theta oscillations during silent lips perception. Further, we compared the resultant vector length r of the distance between the $\phi_{A\text{-}V}$ phase difference observed in the data and a theoretical $\phi_{A\text{-}V} = 0°$ in the early and late windows separately. We hypothesized that the phase difference $\phi_{A\text{-}V}$ indexes the lag of information flow to travel from the visual to the auditory cortex. A change of $\phi_{A\text{-}V}$ between the early and late window suggests that visuo-auditory communication adjusts over time and does not reflect a simple passive transfer of information between the two sensory areas (in which case the phase difference would be constant during the entire visual stimulation). To address the latter, we compared the resultant vector length r of the distance between the $\phi_{A\text{-}V}$ phase difference observed in the data and a theoretical $\phi_{A\text{-}V} = 0°$ in the early and late time-windows separately. Two one-sample t-tests showed that the resultant vector length r of the distance between the observed $\phi_{A\text{-}V}$ and the zero $\phi_{A\text{-}V}$ was significantly greater than zero in both the early ($\phi_{A\text{-}V\ early} = 0.644 \pm 0.038$; $T(1, 22) = 82.26$; $p < 0.001$; one-tailed; Bonferroni-corrected) and late time-windows ($\phi_{A\text{-}V\ late} = 0.654 \pm 0.033$; $T(1, 22) = 93.28$; $p < 0.001$; one-tailed; Bonferroni-corrected), confirming that visual activity always preceded auditory activity during the trials. Finally, a paired-samples *t*-test showed that the resultant vector length r of the distance between the observed $\phi_{A\text{-}V}$ and the zero $\phi_{A\text{-}V}$ was significantly greater in the late time-window ($T(1, 22) = -1.937$; $p = 0.033$; one-tailed; Cohen's $d = 0.285$), confirming that coupling between auditory and visual sources dynamically adjusted with time (Fig. 4B, C and D).

Discussion

In two complementary experiments, we first established that visual entrainment to theta lip phase modulated auditory detection, even if information from silent movies was irrelevant to perform the task. Second, the perception of silent moving lips entrained theta oscillations in the visual cortex followed by the auditory cortex. Together, these results suggest that the brain's natural reaction to visual speech stimuli might be to align the excitability of the auditory cortex with sharp mouth-openings because that is when one expects to hear corresponding acoustic syllable edges (Park et al., 2016; Chandrasekaran et al., 2009; Giraud and Poeppel, 2012; Hickok and Poeppel, 2007; Peelle and Sommers, 2015). Such a neural process could be a very effective filtering method to increase the sensitivity of the auditory cortex in these relevant time windows for speech comprehension.

Our EEG results suggest that theta oscillations in the left visual cortex encoded the lips' activity first. Then information travelled to the left auditory cortex via phase coupling to shape its activity. Previous findings reported that the auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation during audiovisual movie perception (Luo et al., 2010). Other studies reported that the perception of silent lips also recruited the auditory regions (Bourguignon et al., 2020; Calvert et al., 1997; Crosse et al., 2015). Our findings go beyond by establishing how theta oscillations orchestrate visual and auditory cortices through phase coupling to ensure cross-region communication even in a unimodal condition. Furthermore, it is commonly agreed that entrainment takes several cycles from

rhythmic inputs to build up (Thut et al., 2011; Doelling et al., 2014; Lakatos et al., 2008; Zoefel et al., 2018). Behavioural and neural indicators of entrainment reported here consistently increased from the first (i.e. early) to the second (i.e. late) time-window of the trial in both experiments. This supports the idea that we indeed observed neural entrainment to lip movements and sheds light on the functional relevance of visual inputs modulating auditory theta rhythms. As visual onsets naturally lead corresponding auditory onsets by 100-to-300 ms in audiovisual speech (Chandrasekaran et al., 2009; van Wassenhove et al., 2005; Pilling, 2009), visual entrainment to lips may act as a filter by increasing excitability in the auditory cortex to windows containing relevant acoustic features. This hypothesis is corroborated by our phase coupling results where visual theta phase systematically led auditory theta phase during silent movie presentation. This result aligns well with previous findings reported in an intracranial EEG study by (Besle et al., 2008) investigating evoked potentials between visual and auditory cortices during lip movement perception. The authors showed that lip movements activate the visual motion area (~140 ms after stimulus onset), followed by similar responses in the auditory cortex ~10 ms later. They hypothesized that direct feedforward processes support audiovisual interactions during moving lip perception via direct projections from the visual cortex to the auditory cortex (Hishida et al., 2003; Cappe and Barone, 2005). Our study goes beyond by showing that both sensory areas communicate with each other via a phase coupling mechanism.

Whether such filtering reflected a direct feedforward modulation from the visual cortex to the auditory cortex or involved top-down modulations remains unclear. Firstly, our results show that the theta phase lag between auditory and visual cortices decreases across time, which speaks against a process that would just passively relay activity from visual to the auditory cortex (as such a passive process should be constant over time). Secondly, additional top-down controls may adjust theta phase synchronization between the sensory cortices. Indeed, higher-level sensorimotor areas also activate during speech perception (Park et al., 2016, 2018; Cogan and Poeppel, 2011; Arnal et al., 2009; Pulvermüller et al., 2006; Wilson et al., 2004). Assaneo and Poeppel (2018) demonstrated recently that activity in the motor and auditory cortices couple at theta rate during syllable perception, correlating with the strength of coupling between speech signal and EEG in the auditory cortex. On the other hand, motor areas play a role in temporal analysis of rhythmic sensory stimulation (Biau and Kotz, 2018; Arnal et al., 2015; Fujioka et al., 2015; Morillon et al., 2019). Entrainment to lip movements may provide the temporal theta structure of speech signal to motor cortex, which in turn adjusts downstream auditory excitability at critical windows containing the corresponding acoustic features in a top-down fashion (in line with (Park et al., 2015)). Thorne et al. (Thorne and Debener, 2014) associated cross-modal phase resetting with neural temporal predictions occurring when rhythmic visual input precedes auditory input within a stable time-window (~30–100 ms). Our phase coupling results are in line with this finding (Thorne and Debener, 2014), and suggest that the visual and auditory systems adjust over time to maintain this optimal delay in the cortex as well, potentially to allow for neural temporal predictions.

Alternatively, mouth-opening perception may target internal articulatory representations and help to identify the corresponding sounds in the auditory signal. A recent MEG study investigated the visuo-phonological mapping mechanisms mediated by top-down motor areas during silent moving lip presentation (Hauswald et al., 2018). The authors found a stronger coherence between theta activity evoked in the visual cortex by the perception of lip movements and the absent corresponding auditory signal when the moving lips were presented forward as compared to backward. Further, they reported a stronger connectivity between visual cortex and precentral areas in the forward condition, supporting a role of top-down motor controls to map phonological representations with intelligible mouth openings. Here, theta activity found in the auditory cortex may reflect the contribution of inferences generated from such a visuo-phonological mapping. Although speculative, this could partially explain why the increase of entrainment in the auditory cortex was left lateralized, i.e. by recruiting language-related representations classically associated with the left hemisphere. The silent moving lips were taken from speakers speaking in English and presented to native English speakers, which might facilitate such lip-sound mapping. This hypothesis fits with recent debates on whether neural tracking during speech processing reflects online cooperation between pure entrainment to external features and endogenous rhythms providing abstract representations (Meyer et al., 2019; Obleser and Kayser, 2019; Haegens and Zion Golumbic, 2018; Rimmele et al., 2018). However, this would not explain why visual speech information improved the detection of unrelated pure tones here, which will be addressed in future experiments.

The present study focuses on the theta band that reflected the syllabic structure and was dominant in our visual stimuli. Our specific interest towards theta-syllable activity in moving lips was motivated by the fact that the mouth constitutes preponderant direct access to visual speech information mapping sounds (Thompson and Malloy, 2004; Thompson, 1995). However, other parts from the speaker's face bear quasi-rhythmic speech features at distinct time-scales (Kösem and van Wassenhove, 2017). For instance, eyebrow and head movements temporally align with auditory prosody occurring at 0.5–3 Hz delta rate during continuous speech (Ghitza, 2017; Biau et al., 2016; Krahmer and Swerts, 2007; Munhall et al., 2004). Therefore, the proposed filtering method increasing the sensitivity of the auditory cortex may extend to other frequency bands. Future experiments presenting silent videos focusing on speaker's eyebrows will need to test whether visual entrainment to prosodic features recruits auditory cortex and supports cross-region communication as well; or whether long-range synchronization relies specifically on theta even when the dominant activity in stimuli peaks in other frequency bands.

Finally, additional post-hoc analysis suggested that the phasic modulation across participants predicted auditory performance differently between two populations (Figure S3). Splitting our participants into two subgroups based on their preferred mean phase revealed that the increase of performance between the first and second tone windows was greater in the group 2 as compared to the group 1 (respectively pink and blue colours in Figure S3). One could speculate that participants from the group 2 were fine-tuned to a preferred visual theta phase that represents an optimal time-window. This optimal window allowed information to travel to the auditory cortex either directly or via top-down modulations, and reset auditory activity at “perfect” moments when a tone occurred. Interestingly, a recent study also reported subpopulations of “synchronizers” exhibiting differences of behaviours in a spontaneous auditory synchronization task, which related to morphological differences in frontal-to-auditory white matter pathways (Assaneo et al., 2019). Although speculative, one could assume that our results may reflect similar anatomical differences that influence how participants synchronize to visual lip activity and how information travels from visual to auditory cortex. Nevertheless, this data-driven interpretation contrasts somehow with our initial assumption of a similar mean phase modulation induced in the brain by the same stimuli across participants, which will need to be unravelled in the future.

Back to our filtering hypothesis, visual theta entrainment would increase auditory excitability coinciding with more time windows containing a tone in this subpopulation regardless of the nature of sounds. During audiovisual speech perception instead, mouth openings may predict critical windows containing envelope peaks relevant for signal processing (i.e., syllable onsets and nuclei). Inversely, mouth closings may loosen up the filtering action when auditory envelope becomes less informative (e.g., silences or noise). Future experiments will determine whether the visual filter tunes auditory activity independently from its relevance, or whether the influence of visual information on auditory activity is restricted to useful time-windows (i.e., only when lip movements predict sound features). We acknowledge that we did not test for effects of visual entrainment on speech perception per se and therefore cannot conclude that such visual filtering does influence speech

perception. For instance, it would be relevant to investigate whether the visual filter tunes auditory activity independently, or the influence of visual information on auditory activity is restricted to useful time-windows in speech signal, i.e. only when lip movements predict sound features. Further, if a visual filtering builds up over time as hypothesized here, one may argue that listeners would struggle more to understand the beginning of sentences until the auditory system synchronises and tunes to improve speech processing. Nevertheless, this is not the case in most of speech perception conditions as (1) auditory signal often provides enough information to achieve comprehension without visual information. (2) Sentences are rarely processed in isolation in real life, and previous semantic context allows generating predictions that facilitate next sentence comprehension. These are interesting questions for future research.

Conclusion

Although the auditory signal alone often provides enough structural information for the early analytic steps of continuous speech, e.g. telephone conversations, a visual filter may be especially helpful to sharpen auditory perception when hearing is impaired or in elders (Grant et al., 1998). Our results provide an important step toward understanding how visual information functionally drives auditory speech perception, and suggest future directions to investigate hearing loss compensation, i.e. to improve lip-reading along with hearing correction.

Material and methods

Tone detection experiment

Experimental model and subject details

Twenty-eight healthy English native speakers (mean age = 19 years ±0.69; 21 females) took part in the first behavioural experiment. Five participants were left-handed. All of them reported normal or corrected-to-normal vision and hearing. All participants were granted experimental participation credit. The data from four participants were excluded because of extreme overall performances and the final analysis were applied on twenty-four data sets (for details see Tone detection performances section). All the participants signed informed consent and ethical approval was granted by the University of Birmingham Research Ethics Committee, complying with the Declaration of Helsinki.

Method details

Apparatus

The task was programmed with Matlab (R2018a; The MathWorks, Natick, MA, USA) and presented with Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). The silent videos were presented on a 21-inch CRT display with a screen refresh rate of 75 Hz (Nvidia Quadro K600 graphics card: 875 MHz graphics clock, 1024 MB dedicated graphics memory; Nvidia, Santa Clara, CA, USA). The auditory stimuli were presented through EEG-compatible insert earphones (ER-3C; Etymotic Research, Elk Grove Village, IL). The accuracy of movie and sound presentation timing was optimised by detecting a small white square displayed on the left of the first frame of each visual stimulus with a photodiode (ThorLabs DET36A, thorlabs.de), and Psychophysics Toolbox (PsychPort Audio and ASIO4ALL extensions for Matlab). Additionally, a parallel audio port was used to record the online audio signal of each trial during presentation. Continuous photodiode and audio data during trials were recorded through a BioSemi Analog Input Box (AIB) adding two separate channel inputs into BioSemi ActiveTwo system: the BioSemi AD-box was connected with the AIB through optical fibres. The input from the photodiode was connected through a BNC connector and the input from the microphone was connected through the 3.5 mm audio. Those two inputs were connected to the AIB though a 37 pin Sub-D connector. Data were digitized using the BioSemi ActiView software, with a sampling rate of 2048 Hz. Offline analysis were performed to calculate the real delay between visual and audio stimuli offset using in-house Matlab codes. Any lag between real video and sound onsets (i.e. detected first frame and auditory tone onsets) was later compensated in the data analyses when computing the corresponding visual theta phase to the tone onsets as follows: the detected lag was added or removed from the theoretical tone onset (depending on whether the sound was presented sooner or later than theoretically expected respect to the video onset) to obtain the real time-point at which the tone onset occurred in the theta phase of the lip signal. This correction was applied for each individual trial across participants (average video-sound lag across trial and participants = 5.30 ± 0.98 ms). The experiment was run from a solid-state hard drive on a Windows 7-based PC (3.40 GHz processor, 16 Gb RAM). Participants used a standard computer keyboard to respond to the task.

Stimuli

Movies

Sixty 5-s movies were extracted from natural face-to-face interviews published on YouTube (www.youtube.com) by various universities channels and downloaded via free online application. Satisfying movies containing meaningful content (i.e. one complete sentence, speaker facing toward the camera) were edited using Shotcut (Meltytech, LLC). For each movie, the video and the sound were exported separately (Video:.mp4 format, 1280 × 720 resolution, 25 frame per second, 200 ms linear ramp fade in/out; Audio:.wav format, 44,100 Hz sampling rate, mono).

Lip movements' detection

Lips contour signal was extracted for each video using in-house Matlab codes. We computed the area information (area contained within the lips contour), the major axis information (horizontal axis within lip contour) and minor axis information (vertical axis within lip contour) as described in (Park et al., 2016). In the present study, we used vertical aperture information of the lips contour to establish the theta correspondence between lips and auditory speech (i.e. aperture between the superior and inferior lips) but using area information gave very similar results, as also reported in (Park et al., 2016). The lips time-series was resampled at 250 Hz for further analyses with corresponding auditory speech envelope.

Auditory speech signal

The amplitude envelope of each movie sound was computed using in-house Matlab codes (Park et al., 2016, 2018; Chandrasekaran et al., 2009). First, eight equidistant frequency bands spanning on the cochlear map in the range 100–10,000 Hz were constructed (Smith et al., 2002). Then, narrow band sound signals were then band-pass filtered with a fourth-order Butterworth filter (forward and reverse). Absolute Hilbert transform was applied to obtain amplitude envelopes for each narrow band. These signals were then averaged across bands and resulted in a unique wideband amplitude envelope per sound signal. Each final signal was resampled to 250 Hz for further theta correspondence analyses.

Mutual information between lip movements and corresponding auditory speech signal

To identify the main oscillatory activity conveyed by the lip movements in each visual stimulus, we determined at which theta frequency the auditory and visual speech signals showed significant dependencies. To do so, we examined the audiovisual speech frequency spectrum (1–20 Hz) and computed the mutual information (MI) between the minor axis

information and speech envelope signals sampled at 250 Hz. MI measures the statistical dependence between two variables with no prior hypothesis, and with a meaningful effect size measured in bits (Ince et al., 2017; Shannon, 1948). We applied the Gaussian Copula Mutual Information (GCMI) approached described in (Ince et al., 2017) in which the MI between two signals corresponds to the negative entropy of their joint copula transformed distribution. This method provides a robust, semi-parametric lower bound estimator of MI by combining the statistical theory of copulas together with the closed-form solution for the entropy of Gaussian variables, allowing good estimation over circular variables, like phase as well as power. For each movie, the complex spectrum is normalized by its amplitude to obtain a 2D representation of the phase as points lying on the unit circle for both the lip movements and auditory envelope time-series. The real and imaginary parts of the normalized spectrums are rank-normalized separately and the phase dependence for each frequency between the two 2D signals is estimated using the multivariate GCMI estimator giving a lower bound estimate of the MI between the phases of the two signals. Here, we applied the GCMI analyses in two conditions to determine the frequency of interest in each movie: first, we computed MI between corresponding lips and envelope signals as well as non-matching signals (i.e. lips time-series paired with random auditory envelope signals). For the matching signals, the averaged MI spectrum revealed a greater peak in the expected 4–8 Hz theta frequencies, reflected by a bump in the band of interest. In contrast, there was no relationship between random auditory and visual signal pairs, which depicts a flat line profile along the whole spectrum (see Supplementary Information Figure S4A). These results are well in line with previous studies using coherence or MI measures, and confirm the temporal coupling between lips and auditory speech streams at the expected syllable rate in our videos (Park et al., 2016, 2018; Chandrasekaran et al., 2009). Second, for each movie, we performed a peak detection on the MI spectrum to determine which specific frequency carried most theta information to maximize entrainment in the tone detection and silent movie perception tasks (4 Hz frequency peak: 16 videos; 5 Hz frequency peak: 15 videos; 6 Hz frequency peak: 9 videos; 7 Hz frequency peak: 13 videos; 8 Hz frequency peak: 7 videos. See Supplementary Information Figure S4B).

Audio tones and white noise

Auditory pure tones and white noise stimuli were generated using in-house Matlab codes. The target tone consisted in a sinusoidal signal of 100 ms at one kHz (sampling rate: 44,100 Hz). The same noise consisted in a Gaussian white noise lasting 2 s for the calibration task and 5 s in the tone detection task (the white noise has been generated only once and loaded during each procedure to ensure that all the participants were tested with the same noise; sampling rate: 44,100 Hz). Both the tone and the white noise signals were normalized between - 1 and 1 (arbitrary units). During the entire procedure, the auditory stimuli (white noise with embedded tone) were displayed at constant ~72 dB SPL and across participants.

Tones onsets

For each trial, the target tones were embedded in the white noise at predetermined pseudo-random onsets counterbalanced across conditions (zero, one or two tones per trial, 100 trials per condition). In the calibration task serving to determine the individual threshold of target tone detection (see below for the general procedure), there could be only zero or one tone maximum per trial. For the one tone condition, the onset of the target tone always randomly occurred between 300 and 1400 ms after the trial onset to allow participants to detect it properly and have time to respond before the end of the trial. In the zero tone condition, the auditory track consisted of 2 s of white noise only. In the tone detection task, there could be zero, one or two tones per trial. In the one tone condition, the onset of the target always occurred randomly between 300 and 4500 ms after the trial onset. In the two tones condition, the first tone randomly occurred in a time-window centred on the first half of the trial length, between 300 and 3000 ms (mean first tone onsets = 1.68 ± 0.78 s). The second tone occurred in a time-window centred on the second half of the trial length, between 1000 ms after the first tone onset and 4500 ms after the trial onset (mean second tone onsets = 3.45 ± 0.62 s). This design provided participants with enough time to detect and respond to both tones, and kept the two tones temporally unrelated from each other. In the zero tone condition, the auditory track consisted in 5-s of white noise only. The signal-to-noise ratio between target tones and white noise was determined for each participant individually with the calibration task performances and adjusted consequently in the following tone detection task (see below). As tones onsets were randomly generated, we controlled that the random generation satisfactorily sampled the visual phase across participants (i.e., with no particular preferred phase angle across participants). The distribution of probability of tone onsets p(onsets) along the visual phase was computed for the first and second tone windows, in the two tones and single tone conditions. The visual phase ($-\pi$ to $+\pi$, $\pi/8$ step) was binned in $n = 16$ bins, and the proportion of tones p(onsets) in every bin was computed for each participant. P(onsets) of each phase bin was assessed statistically against its expected proportion from a random distribution by applying one-sample t-tests against $1/n = 0.0625$ (In the case of a random distribution of tone onsets, each phase bin should contain 6.25% of the total tones). A Benjamini-Hochberg correction for multiple comparisons was applied on the p-value of each bin in the four distributions ($\alpha = 0.05$ (Benjamini and Hochberg, 1995);). Results revealed no preferred bin significantly different from chance level (i.e. 0.0625) and confirmed that the tone onsets satisfactorily sampled the visual phase for the first and second tone windows, in the two tones and single tone conditions.

Procedure of the calibration task and tone detection task (TDT)

The experiment began after the completion of a safety-screening questionnaire and the provision of informed consent. Participants sat in a well-lit testing room at approximatively 60 cm from the centre of the screen and wore the insert earphones for sound presentation. Participants performed first a short pure tone detection task with no visual stimuli (i.e. calibration task). This task served to determine the individual threshold at which each participant detected ~70–80% of the target tones in auditory modality only, and the signal-to-noise ratio (SNR) to be implemented between the amplitude of the target tones and the white noise in the following tone detection task (TDT). The calibration task was composed of a four-trial practice to identify the target tone itself, followed by five blocks containing 20 trials each. Each trial began with a black fixation cross (500–1000 ms duration, jittered) followed by the presentation of a red cross over a grey background during 2 s to indicate the period of possible target tones occurrence. A continuous white noise was displayed during the red cross presentation. In 50% of the trials, a unique audio tone was embedded in the white noise at unpredictable onset, and participants had to press "1" key as fast and accurately as possible only when they perceived a target tone. The pseudo-random sequence of the procedure ensured that there were never more than two consecutive trials of the same condition. The participants received no feedback and the procedure continued to the next trial after the end of the 2-s white noise. The signal-to-noise ratio was adjusted following an adapted "two-down one-up" staircase procedure, while keeping the overall loudness of the stimulus fixed (Leek, 2001): For the first five trials, the SNR was fixed (mean white noise power of 0.981) and served as a starting point across participants. After each trial, the keypress response of the participant was stored to adjust the SNR for the next trial as following: for two successive hits, the SNR was decreased by 2% of the starting signal energy in the next trial. For two successive correct rejections (i.e. no response when no tone occurred) or one correct rejection following a hit, the SNR was kept identical for the next trial. After a miss or a false alarm, the SNR was always increased by 2% of the starting

signal energy. At the end of the calibration task, the individual SNR was averaged over the last 30 trials and stored for the following real tone detection task (mean calibration accuracy rate: 0.75 ± 0.05). The participants took a short break and were recalled the instructions before starting the proper tone detection task. The calibration task lasted approximatively 7 min.

The main structure of the TDT was the same as in the preceding calibration task. The TDT was composed of a short four-trial practice followed by 300 trials divided in 6 blocks of 50 trials each and separated by breaks (the sixty silent movies were repeated five times each to generate the total 300 trials). Each trial began with a red fixation cross presentation (500–1250 ms duration, jittered). Then, a random 5-s silent movie was presented with a black fixation cross in the centre of the screen to give the participants a point to gaze at and reduce saccades. The continuous white noise was displayed together with the silent movie according to the three random conditions: no tone (100 trials), one single tone (100 trials) or two tones (100 trials) hidden in the white noise. Participants were instructed to press "1" key as fast and accurately as possible only when they perceived a target tone. The participants received no feedback on their responses and the procedure continued with the next trial after the end of the silent movie. The SNR between the tones and the white noise was determined in the previous calibration task as explained above. The TDT lasted approximatively 50 min.

TDT conditions

The condition of interest containing the two tones (i.e. first and second tone) served to assess our main hypothesis that entrainment increases in time with the perception of visual information conveyed by the speakers' lip movements. According to this, the second tones should be better detected and associated to a greater theta entrainment as compared to the first tones to reflect the modulation of the auditory system by the entrained visual system to lip movements. The zero and single tone conditions were additional control conditions: the zero tone condition served to determine the false alarm rates (i.e. participants' keypresses in the absence of tone) and controlled whether participants tended to press more together with the tone onset delays (i.e. time-dependent response bias). The single tone condition served to counterbalance the number of trials containing two tones and control for the predictability of the second tone. The replication of the performances observed in the two tones condition by sorting the single tones according to their onsets equivalent to either first or second tone onsets would confirm that the detection of the second tone is not due to its predictability from a preceding tone but its position in time only. The pseudo-random sequence of the procedure ensured that there were never more than three consecutive trials of the same condition.

Quantification and statistical analysis

The Tone Detection task was within-subject design.

Tone detection performances

The hits (i.e. correctly detected tones) and false alarms (i.e. keypress responses during the zero tone condition allocated to the first or second tone windows depending on their onsets) rates were computed to calculate the individual mean sensitivity index (i.e. d') in the two conditions for each participant (i.e. single tone and two tones conditions). The reaction times of the hits were computed to calculate the individual mean reaction times in the two conditions for each participant (i.e. single tone and two tones conditions). Additionally, we calculated the mean correct response rates and reaction times of the two conditions concatenated together of each individual to exclude blindly potential outliers without favouring the results towards our hypothesis and performing as following: below chance level (correct response rate < 0.5) or perfectly (correct response rate $= 1$), or with mean reaction times outside the grand averaged reaction times ± two standard deviations range. Accordingly, four participants were excluded from analyses (two participants performed below chance level, one participant performed perfectly and one participant's reaction times were slower than the grand average mean + two standard deviations). A paired-samples *t*-test was conducted on the averaged d' scores and hit reaction times between the first and second tones in the two tones condition and single tone condition separately. Additionally, a paired-samples *t*-test was performed on false alarm rates from the first and second windows in the zero tone condition to control for any response bias with time.

Visual entrainment to theta activity conveyed by lip movements

To bridge visual entrainment to auditory processing together, we related the tone target onsets to the theta activity conveyed by the lip movements during silent movies perception: First, for each movie the theta phase of the lip movements' time-series was computed by applying a Hilbert transform with a bandpass filter centred on the frequency bin of MI peak ±2 Hz, accordingly to the mean theta frequency determined in MI stimuli analyses. Second, we computed the instantaneous theta phase of the lips signal corresponding to the onset of the tones occurring during each trial. All further circular statistics on angular scale were performed using the CircStat toolbox on Matlab (Berens, 2009).

(1) In a first approach, we assumed that participants entrained at different preferred phases and tested for a phase locking within participants after realigning individual phases on their preferred phase (for further details on this approach, see (Zoefel et al., 2019)). First, the visual phase was binned in 20 bins (-π to π, π/10 step) and the accuracy in the binned phase was computed across separate hit and miss trials at first and second tones for each participant. Second, after obtaining the distribution of accuracy across trials along the visual phase, the "preferred phase" bin yielding the best performance was determined by means of an automatic peak detection (i.e., visual phase bin with the highest accuracy). The preferred phase bin position was shifted to the centre bin ($bin_{preferred\ phase} = 20/2 + 1$) and the remaining bins were phase-wrapped accordingly. This operation was computed for hit and miss trials separately. Third, visual entrainment in hit and miss trials was estimated by computing the distance between the preferred bin and a baseline averaged across bins as follows: the average response of the two bins adjacent to the bin opposite to the preferred phase was subtracted from the average response of the two bins adjacent to the preferred phase (Zoefel et al., 2019). Fourth, the phase modulation i.e. the statistical difference of entrainment between hit and miss trials, was estimated by the t-value from a paired-sample T-test (one-tailed) at first and second tones (effect size: $t\text{-value}_{hit\text{-}miss}$). To circumvent the unbalanced numbers of hit and miss trials, we tested the phase modulation in the first and second tone windows with two separate permutation tests as follows: First, for each participant we generated 10,000 iterations for which the hit/miss trial labels were shuffled for the first and second tones in the two tones condition. Second, two balanced subsamples of shuffled hit/miss trials were selected, with a number matching the smallest number available between the hit and miss trials. Third, phase realignment and visual entrainment were computed across hit and miss trials of each permuted data, as for the original data. Fourth, the $t\text{-value}_{hit\text{-}miss}$ was computed between the 10,000 pairs of hit and miss permuted data sets of the first and second tone, and the resultant 10,000 $t\text{-value}_{hit\text{-}miss}$ were sorted in descending order. To estimate the final p-value and test the null hypothesis, the original $t\text{-value}_{hit\text{-}miss}$ was ranked in the sorted permuted $t\text{-value}_{hit\text{-}miss}$ +1 and divided by the total number of permutations+1. If the p-value was smaller than $\alpha = 0.05$, we rejected the null hypothesis H_0 = there is no difference of phase modulation between hit and miss trials

(i.e. visual entrainment is significantly greater in the hit trials). The interaction between phasic modulation and tone position was assessed by means of a third permutation testing the statistical difference of t-value$_{\text{hit-miss}}$ between the first and second tones [t-value$_{\text{hit-miss}}$]$_{\text{second tone}}$ - [t-value$_{\text{hit-miss}}$]$_{\text{first tone}}$, applying the exact same method as describe above.

(2) As participants viewed the same videos, which theoretically imposed the same phase modulations onto the brain activity, we hypothesized that a similar phase should emerge across participants. In the second analytic approach, we tested visual entrainment across participants. We compared the mean theta phase distributions between first and second tone onsets across participants. The circular uniformity in the first and second tones windows within and across participants were estimated separately by applying Rayleigh tests to calculate the mean direction and resultant vector length from hits/miss trials. To assess statistically the difference of visual entrainment at tones detected in the first and second windows in the two tones condition (i.e. hits only), we performed a permutation test on the resultant vector length difference (z-value) second tone minus first tone reflecting the effect size (effect size: z-value$_{\text{second tone}}$ - z-value$_{\text{first tone}}$). For each participant, we generated 10,000 iterations as following: first, the hit trial labels were shuffled between the first and second tones in the two tones condition. Second, two balanced subsamples of shuffled trials were selected, with a number matching the smallest number of trials between the first and second tone hits. Third, the mean phase of the first and second tone shuffled trials were computed for each iteration and per participant. Fourth, a Rayleigh's test of uniformity was applied on the mean phases across participants to determine a resultant vector length r at the first and second tones per iteration (i.e. z-value). For each iteration, we computed the difference of z-value$_{\text{second tone}}$ - z-value$_{\text{first tone}}$ to quantify its effect size, and the resultant 10,000 z-value differences were sorted in descending order. To estimate the final p-value and test the null hypothesis, the difference of z-value between the original first and second tone data was ranked in the sorted permuted z-value differences +1 and divided by the total number of permutations+1. If the p-value was smaller than $\alpha = 0.05$, we rejected the null hypothesis H_0 = there is no difference of resultant vector length between the first and second tones (i.e. the visual entrainment is significantly greater in the second tone window). To assess the phasic modulation in the first and second windows across participants, two separate permutation tests were performed on the resultant vector length difference between hit and miss trials (effect size: z-value$_{\text{hit-miss}}$ = z-value$_{\text{hit}}$ - z-value$_{\text{miss}}$). The existence of an interaction between visual phasic modulation and tone position across participants was assessed with a permutation test on the difference of vector length r ([z-value$_{\text{hit-miss}}$]$_{\text{second tone}}$ - [z-value$_{\text{hit-miss}}$]$_{\text{first tone}}$). Permutations and p-values were computed as described above after randomly shuffling the hit/miss labels and tone position information from the original data accordingly.

The same approaches (1) and (2) were applied on the single tone condition as well, after sorting the tones as first or second according to their onset (see Supplementary Information).

Subpopulations of group 1 and group 2

Participants were sorted in two subpopulations according to their preferred theta phase in the second tone window, where visual entrainment supposedly took place after enough lip movements inputs in the condition of interest (i.e. two tones condition). The Rayleigh tests revealed non-uniform distributions of preferred phase at the second tones for group 1 ($n = 11$; $\mu = 348.83°$; $p < 0.001$) and group 2 ($n = 10$; $\mu = 249.63°$; $p < 0.001$). A Kuiper two-sample test confirmed that the mean preferred phases were different between group 1 and 2 ($k = 121$; $p < 0.01$). Results are presented in the Supplementary Information.

Silent movie perception-EEG experiment

Experimental model and subject details

Twenty-five healthy English native speakers (mean age = 21.52 years ±3.86; 17 females) took part in the first behavioural experiment. All of them reported normal or corrected-to-normal vision and hearing, and were right-handed. Twenty participants were granted credits and five participants received financial compensation for their participation (£20). The data from two participants were excluded from the final analyses due to too noisy EEG data. All the participants signed informed consent and ethical approval was granted by the University of Birmingham Research Ethics Committee, complying with the Declaration of Helsinki.

Method details

Apparatus

The task was programmed with Matlab (R2018a; The MathWorks, Natick, MA, USA) and presented with Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). The silent videos were presented in the same manner as described in the previous TDT section.

Stimuli

Movies

The movies presented during the silent lips perception task were the exact same 60 movies used in the previous tone detection task. The order of movies was randomised across participants.

Procedure

Participants sat in a well-lit testing room at approximatively 60 cm from the centre of the screen to complete a safety-screening questionnaire and the provision of informed consent first. After the correct preparation of the EEG cap, the participants were instructed to attend to all the movies quietly and to avoid movements during the presentation. Each trial was preceded by a central fixation cross (500–1250 ms duration, jittered) followed by the presentation of a random 5-s movie. A central fixation cross was displayed during the movie presentation to give participants a point to gaze at and reduce excessive saccades. Participants were instructed to attend to each movie carefully and rate its emotional content based on speaker's facial gestures by using the number keys on the keyboard after the presentation (i.e. 1 for neutral through 5 for very emotional; results not reported). The total presentation of the sixty movies lasted approximatively 10 min.

Online EEG recordings: Continuous EEG signal was recorded using a 128 channel BioSemi ActiveTwo system (BioSemi, Amsterdam, Netherlands). Vertical and horizontal eye movements were recorded from additional electrodes placed approximatively one cm to the left of the left eye, one cm to the right of the right eye, and one cm below the left eye. Online EEG signals were digitalized using BioSemi ActiView software at a sampling rate of 2048 Hz. For each participant, the position of the electrodes on the scalp were tracked using a Polhemus FASTRAK device (Colchester) and recorded with Brainstorm (Tadel et al., 2011) implemented in MATLAB.

Offline EEG preprocessing: EEG data were preprocessed offline using Fieldtrip (Oostenveld et al., 2011) and SPM 8 toolboxes (Wellcome Trust Centre for Neuroimaging). Continuous EEG signals were bandpass filtered between one and 100 Hz and bandstop filtered (48–52 Hz and 98–102 Hz) to remove line noise at 50 and 100 Hz. Data were epoched from 2000 ms before stimulus onset to 7000 ms after stimulus onset, and

downsampled to 512 Hz. Bad trials and channels with artefacts were excluded by visual inspection and numerical criteria (e.g., variance as well as kurtosis) before applying an independent component analysis (ICA) to remove components related to ocular artefacts. Bad channels were then interpolated using the method of triangulation of nearest. After re-referencing the data to average reference, trials with artefacts were manually rejected by a last visual inspection. On average, 4.48 ± 2.48 trials were removed and 4.04 ± 1.82 channels were interpolated per participants.

Head models

For the 22 participants without individual MRI scans, the MNI-MRI and the volume conduction templates provided by Fieldtrip were used to construct the head models. Electrode positions of each participant were aligned to the template head model. Source models were prepared with the template volume conduction model and the aligned individuals' electrode positions following standard procedures. One participant provided his own MRI scans and his head model was built using his structural scans (Michelmann et al., 2016): the MRI scans were segmented into four layers (i.e. brain, CSF, skull and scalp) using the Statistical Parametric Mapping 8 (SPM8; http://www.fil.ion.ucl.ac.uk/spm) and Huang toolboxes (Huang et al., 2013). The volume conduction model was constructed using the dipoli method implemented in Fieldtrip. Participant's electrode positions were aligned to his individual head model. Finally, his MRI was warped into the same MNI template MRI of Fieldtrip and the inverse of the warp was applied to a template dipole grid to have each grid point position in the same normalized MNI space as the other participants for further group analyses.

Source localization during silent movie perception

Source analyses on EEG data recorded during silent movies presentation were run using individual electrode positions, grid positions and template volume conduction model. For the participant who had his MRI scans, source analyses were calculated using normalized grid positions instead. Source activity was reconstructed using a linearly constrained minimum variance beamforming method implemented in Fieldtrip (LCMV; see (Van Veen et al., 1997)). The neural entrainment to lip movements at source level was determined by computing mutual information between EEG epochs and the lip movements during silent movie presentation (i.e. lips time-series from the silent movie presented during the trial). To test our hypothesis that entrainment builds up in time with perceived theta lips activity, we contrasted the MI between the equivalent time-window to the second tone window (MI_{late}), and the equivalent time-window to the first tone window (MI_{early}) in the previous TDT. Accordingly, we expected first to observe an increase of theta activity in the visual cortex reflecting entrainment to lip movements. Second, we expected an equivalent pattern in the auditory correlates reflecting a tuning from visual activity. For each single trial, MI was first computed separately in the time-windows equivalent to the first (0.5–2.5 s after trial onset; MI_{early}) and second tone time-windows of the TDT (2.5–4.5 s after trial onset; MI_{late}) at the 2020 virtual electrodes by using the same approach described in the stimuli analysis section (i.e. where we established which frequency carried most correspondence between lips and envelope signals for each video; using a wavelet transform to compute the phase). The first and last 500 ms at the edges of the epoch were not included into MI analyses to avoid the trial onset and offset responses. Second, for each single trial, the MI spectrum was realigned with respect to the frequency bin (±2 Hz) corresponding to the peak of MI between lips and envelope signal established in the movie analyses. This step was done to be able to average all the trials together taking into account the main theta activity carried in each individual movie. For instance, if the peak of MI between lip movements and auditory envelope was found at 4 Hz in the video number 1, the realigned MI spectrum between EEG and lips signals from the trials presenting video number 1 was now 4 ± 2 Hz (2–6 Hz; 1 Hz bin) to insure that the central bin of each single trial corresponds to the objectively determined frequency peak of theta activity. Third, the realigned MIs of single trials were averaged across trials within each participant for further group analyses. For each participant, we calculated the normalized difference of MI at the frequency bin of interest in the late minus early time-window at all the 2020 virtual electrodes (MI normalization: $(MI_{late} - MI_{early})/MI_{early}$; third bin in the realigned MI spectrum). Finally, the normalized difference of MI between second and first tone time-windows was grand averaged across participants and the grand average was interpolated to the MNI MRI template. The coordinates for auditory and visual sources of interest were determined by finding the maximum of $MI_{late} - MI_{early}$ differences in regions corresponding to the auditory (left Middle Temporal cortex) and visual areas (left Middle Occipital cortex), and defined using the automated anatomical labelling atlas (AAL (Tzourio-Mazoyer et al., 2002),).

Source reconstruction

We performed time-series reconstruction analysis to investigate the synchronization at theta activity between the two sources of interest during silent movie presentation. The time series data were reconstructed and extracted at the visual and auditory coordinates determined by source localization analysis. LCMV beamformer reconstruction can cause random direction of source dipoles and eventually affect phase analysis results. To get around this issue, the event-related potentials (ERP) time-locked to movie onsets at visual and auditory sources were plotted to identify the visual component, i.e. N1–P2–N2 waveform (Wang et al., 2018). After visual inspection, the sign of the reconstructed data were flipped in direction by multiplying the time-series by −1 if any visual or auditory source ERPs showed the opposite of the expected direction of a visual component (i.e. negative-positive-negative polarity). This "flipping" correction was applied consistently across all trials before sorting data between early and late time-windows, thus it did not bias results towards our hypothesis. The same phase coupling analyses were computed with unflipped source data as a control. Phase angle differences between visual and auditory theta activities in the early and late windows were also non-uniformly distributed according to Rayleigh tests with significantly different mean angles according to a Kuiper's test, confirming that the flipping procedure only better reflected phase coupling modulation with entrainment.

Theta phase coupling between auditory and visual sources

First, auditory signal was projected orthogonally onto the visual signal (i.e. the reconstructed source level time-series activities measured at voxels exhibiting maximal MI with the lip movement signal in auditory and visual areas respectively) applying a Gram-Schmidt process (GSP (Hipp et al., 2012);) for single trials before computing phase information. This was done to reduce the noise correlation patterns reflecting activity from a common source (i.e. volume conduction) estimate captured at different electrodes (in that case, the phase alignment reflects the same source activity and not the phase coupling between two distinct source activities). The GSP increases the signal-to-noise ratio by leaving intact the proper activities conveyed at the two distinct electrodes while reducing noise correlation weight (see (Hipp et al., 2012)). Second, for each trial the instantaneous theta phase of the auditory and visual orthogonalized time-series were computed by applying a Hilbert transform with a bandpass filter centred on the frequency bin of MI peak ±2 Hz, accordingly to the mean theta frequency of the video presented during the trial. Third, the difference of unwrapped instantaneous phase between auditory and visual sources was computed for each single trial at each time-point in two windows corresponding to the time-windows containing the first (0.5–2 s after trial onset; early window) and second tones in the TDT task (3–4.5 s after trial onset; late window). The first and last 500 ms at the edges of the epoch were not included into phase coupling analyses to avoid the trial onset and offset responses. To further

control for any potential issues of circularity, the phase-slope index (PSI) was calculated in the early and late time-windows between the left auditory and the left visual sources to assess the directionality of information flow between our two sources of interest (Nolte et al., 2008) but see also (Bastos and Schoffelen, 2016). The PSI was computed separately in the early and late windows with a bandpass filter centred on the frequency bin of MI peak ±2 Hz of each single trial, using the left auditory source as the seed region. For each frequency of the band, the change in the phase difference between neighbouring frequency bins, weighted with the coherence, was estimated to calculate the resultant PSI across trials. A positive PSI indicates that the seed region leads the visual source, while a negative PSI indicates that the visual source leads the auditory source during the visual perception of silent moving lips.

Quantification and statistical analysis

The silent lips perception task were within-subject design

Audio-visual theta synchrony

The phase coupling between auditory and visual sources was estimated through their theta phase angle difference in time-windows equivalent to the first and second tone windows from the TDT (i.e. MI_{early} and MI_{late} time-windows). To assess that the $\phi_{A\text{-}V}$ theta phase coupling between visual and auditory cortices was modulated in time (i.e. between early and late time-windows), we computed and compared the resultant vector length r of the distance between the observed $\phi_{A\text{-}V}$ in the data and a theoretical angle $\phi_{A\text{-}V} = 0^{\circ}$ in the early and late time-windows separately. We hypothesized that the $\phi_{A\text{-}V}$ reflects the speed at which information from the visual cortex is conveyed to the auditory cortex. If the $\phi_{A\text{-}V}$ is constant over time, this means that it reflects the time lag of a passive transfer of information between the two sensory areas. In contrast, a decrease of theta phase distance over time (i.e. the $\phi_{A\text{-}V}$ in the late time-window gets closer to zero degree than the $\phi_{A\text{-}V}$ in the early time-window) suggests that visuo-auditory communication gets more efficient. For each trial, we calculated the resultant vector length of the distance between the real auditory-visual phase offset and an arbitrary fixed phase offset at 0° at each time point in the two time-windows. The resultant vector length was collapsed across time in the first and second windows separately, resulting in two values per trial. Single-trial values in the first and second windows were then averaged across trials for each participant, and the difference of phase entrainment values was assessed with a paired samples *t*-test.

Additional information

Reagent or resource	Source	Identifier
Software and Algorithms		
MATLAB	The MathWorks	R2018a
Psychophysics Toolbox	http://psychtoolbox.org	3
FieldTrip	http://www.fieldtriptoolbox.org	v.20161231
SPM8	Wellcome Trust Centre for Neuroimaging	8
ASIO4All	Steinberg Media Technologies	2.12
ActiView	BioSemi B.V. Amsterdam, Netherlands	7
Shotcut	Meltytech, LLC	v.18.06.02
Brainstorm Toolbox	https://neuroimage.usc.edu/brainstorm/	
CARET	Washington University School of Medicine	5.65
Circular Statistics Toolbox	https://uk.mathworks.com/matlabcentral/fileexchange/10676-circular-statistics-toolbox-directional-statistics	v.1.21.0.0
Other		
BioSemi ActiveTwo system	BioSemi B.V. Amsterdam, Netherlands	EEG system
ER-3C system	Etymotic Research, Elk Grove Village, IL	EEG compatible earphones
Fastrak	Polhemus, Colchester, VT, USA	Electromagnetic digitiser
ThorLabs DET36A	https://thorlabs.de	Photodetector

Contact for reagent and resource sharing

Data for individual participants and associated scripts will be made available after publication of the manuscript, as consent for sharing data at the level of the individual participant was received. Further information or requests should be directed to the corresponding author.

Author contribution

E.B, H.P and S.H designed the experiments and paradigms. E.B and D.W collected and analysed the data. E.B, D.W, H.P, O.J and S.H wrote the paper. All the authors discussed the results and commented on the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by a Sir Henry Wellcome Postdoctoral Fellowship awarded to E.B. (Grant reference number: 210924/Z/18/Z), as well as grants from the ERC (Consolidator Grant 647954) and ESRC (ES/R010072/1) awarded to S.H., who is further supported by the Wolfson Foundation and Royal Society. The authors would like to thank people from the Cognition and Oscillations Lab and David Poeppel for their valuable comments and inputs during the preparation of this manuscript.

Appendix A. Peer Review Overview and Supplementary data

A Peer Review Overview and (sometimes) Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.crneur.2021.100014.

References

Arnal, L.H., Morillon, B., Kell, C.A., Giraud, A.-L., 2009. Dual neural routing of visual facilitation in speech processing. J. Neurosci. 29, 13445–13453. https://doi.org/10.1523/jneurosci.3194-09.2009.

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

Assaneo, F.M., Poeppel, D., 2018. The coupling between auditory and motor cortices is rate-restricted: evidence for an intrinsic speech-motor rhythm. Sci. Adv. 4 (2) https://doi.org/10.1126/sciadv.aao3842.

Assaneo, M.F., Ripollés, P., Orpella, J., Lin, W.M., de Diego-Balaguer, R., Poeppel, D., 2019. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. Nat. Neurosci. 22 (4), 627–632. https://doi.org/10.1038/s41593-019-0353-z.

Bastos, A.M., Schoffelen, J.M., 2016. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Front. Syst. Neurosci. 9, 175. https://doi.org/10.3389/fnsys.2015.00175.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57 (1), 289–300.

Berens, P., 2009. CircStat: a MATLAB toolbox for circular statistics. J. Stat. Software 31 (10), 1–21. https://doi.org/10.18637/jss.v031.i10.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., Giard, M.-H., 2008. Visual activation and audiovisual interactions in the auditory cortex during speech

perception: intracranial recordings in humans. J. Neurosci. 28 (52), 14301–14310. https://doi.org/10.1523/JNEUROSCI.2875-08.2008.
Biau, E., Kotz, S.A., 2018. Lower beta: a central coordinator of temporal prediction in multimodal speech. Front. Hum. Neurosci. 12 https://doi.org/10.3389/fnhum.2018.00434.
Biau, E., Fernandez, L.M., Holle, H., Avila, C., Soto-Faraco, S., 2016. Hand gestures as visual prosody: BOLD responses to audio-visual alignment are modulated by the communicative nature of the stimuli. Neuroimage 132, 129–137.
Bourguignon, M., Baart, M., Kapnoula, E.C., Molinaro, N., 2020. Lip-reading enables the brain to synthesize auditory features of unknown silent speech. J. Neurosci. 40 (5), 1053–1065. https://doi.org/10.1523/jneurosci.1101-19.2019.
Brainard, D.H., 1997. The psychophysics toolbox. Spatial Vis. 10, 433–436.
Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276 (5312), 593–596. https://doi.org/10.1126/science.276.5312.593.
Cappe, C., Barone, P., 2005. Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. Eur. J. Neurosci. 22, 2886–2902.
Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. PLoS Comput. Biol. 5 (7), e1000436 https://doi.org/10.1371/journal.pcbi.1000436.
Cogan, G.B., Poeppel, D., 2011. A mutual information analysis of neural coding of speech by low-frequency MEG phase information. J. Neurophysiol. 106 (2), 554–563. https://doi.org/10.1152/jn.00075.2011.
Crosse, M.J., ElShafei, H.A., Foxe, J.J., Lalor, E.C., 2015. Investigating the temporal dynamics of auditory cortical activation to silent lipreading. In: 7th Annual International IEEE EMBS Conference on Neural Engineering.
Crosse, M.J., Butler, J.S., Lalor, E.C., 2019. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci.: Off. J. Soc. Neurosci. 35 (42), 14195–14204. https://doi.org/10.1523/jneurosci.1829-15.2015.
Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85 (Pt 2), 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035.
Fujioka, T., Ross, B., Trainor, L.J., 2015. Beta-band oscillations represent auditory beat and its metrical hierarchy in perception and imagery. J. Neurosci. 35 (45), 15187–15198. https://doi.org/10.1523/jneurosci.2397-15.2015.
Ghitza, O., 2017. Acoustic-driven delta rhythms as prosodic markers. Lang. Cogn. Neurosci. 32 (5), 545–561. https://doi.org/10.1080/23273798.2016.1232419.
Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15 (4), 511–517. https://doi.org/10.1038/nn.3063.
Grant, K.W., Walden, B.E., Seitz, P.F., 1998. Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. J. Acoust. Soc. Am. 103 (5 Pt 1), 2677–2690. https://doi.org/10.1121/1.422788.
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 Biol. 11 (12), e1001752 https://doi.org/10.1371/journal.pbio.1001752.
Haegens, S., Zion Golumbic, E., 2018. Rhythmic facilitation of sensory processing: a critical review. Neurosci. Biobehav. Rev. 86, 150–165. https://doi.org/10.1016/j.neubiorev.2017.12.002.
Hanslmayr, S., Axmacher, N., Inman, C.S., 2019, July 1. Modulating human memory via entrainment of brain oscillations. Trends Neurosci. 42, 485–499. https://doi.org/10.1016/j.tins.2019.04.004.
Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., Weisz, N., 2018. A visual cortical network for deriving phonological information from intelligible lip movements. Curr. Biol.: CB 28 (9), 1453–1459. https://doi.org/10.1016/j.cub.2018.03.044 e3.
Hickok, G., Poeppel, D., 2007, May. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402. https://doi.org/10.1038/nrn2113.
Hipp, J.F., Hawellek, D.J., Corbetta, M., Siegel, M., Engel, A.K., 2012. Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat. Neurosci. 15 (6), 884–890. https://doi.org/10.1038/nn.3101.
Hishida, R., Hoshino, K., Kudoh, M., Norita, M., Shibuki, K., 2003. Anisotropic functional connections between the auditory cortex and area 18a in rat cerebral slices. Neurosci. Res. 46, 171–182.
Huang, Y., Dmochowski, J.P., Su, Y., Datta, A., Rorden, C., Parra, L.C., 2013. Automated MRI segmentation for individualized modeling of current flow in the human head. J. Neural. Eng. 10, 066004.
Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., Schyns, P.G., 2017. A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. Hum. Brain Mapp. 38 (3), 1541–1573. https://doi.org/10.1002/hbm.23471.
Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol. 16 (3), e2004473 https://doi.org/10.1371/journal.pbio.2004473.
Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., 2007. What's new in psychtoolbox-3. Perception 36, 1–16.
Kösem, A., van Wassenhove, V., 2017. Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. Lang. Cogn. Neurosci. 32 (5), 536–544.
Krahmer, E., Swerts, M., 2007. The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. J. Mem. Lang. 57 (3), 396–414.
Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320 (5872), 110–113. https://doi.org/10.1126/science.1154735.
Leek, M.R., 2001. Adaptive procedures in psychophysical research. Percept. Psychophys. 63 (8), 1279–1292. https://doi.org/10.3758/BF03194543.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54 (6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004.
Luo, H., Liu, Z., Poeppel, D., 2010. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. 8 (8), 25–26. https://doi.org/10.1371/journal.pbio.1000445.
Macmillan, N.A., Kaplan, H.L., 1985. Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. Psychol. Bull. 98, 185–199.
Meyer, L., Sun, Y., Martin, A.E., 2019. Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. Lang. Cogn. Neurosci. 1–11. https://doi.org/10.1080/23273798.2019.1693050.
Michelmann, S., Bowman, H., Hanslmayr, S., 2016. The temporal signature of memories: identification of a general mechanism for dynamic memory replay in humans. PLoS Biol. 14, e1002528.
Morillon, B., Arnal, L.H., Schroeder, C.E., Keitel, A., 2019, December 1. Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. Neurosci. Biobehav. Rev. 107, 136–142. https://doi.org/10.1016/j.neubiorev.2019.09.012.
Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci. 15 (2), 133–137.
Nolte, G., Ziehe, A., Nikulin, V.V., Schlögl, A., Krämer, N., Brismar, T., Müller, K.R., 2008. Robustly estimating the flow direction of information in complex physical systems. Phys. Rev. Lett. 100 (23), 234101 https://doi.org/10.1103/PhysRevLett.100.234101.
Obleser, J., Kayser, C., 2019. Neural entrainment and attentional selection in the listening brain. Trends Cognit. Sci. 23, 913–926. https://doi.org/10.1016/j.tics.2019.08.004.
Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci., 156869
Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., Gross, J., 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol. 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. https://doi.org/10.7554/eLife.14521.
Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., Gross, J., 2018. Representational interactions during audiovisual speech entrainment: redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. PLoS Biol. 16 (8).
Peelle, J.E., Davis, M.H., 2012. Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. 3, 320. https://doi.org/10.3389/fpsyg.2012.00320.
Peelle, J.E., Sommers, M.S., 2015. Prediction and constraint in audiovisual speech perception. Cortex; J. Devoted Stud. Nerv. Syst. Behav. 68, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006.
Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spatial Vis. 10, 437–442.
Pilling, M., 2009. Auditory event-related potentials (ERPs) in audiovisual speech perception. J. Speech Lang. Hear. Res.: JSLHR (J. Speech Lang. Hear. Res.) 52 (4), 1073–1081. https://doi.org/10.1044/1092-4388 (2009/07-0276.
Pulvermüller, F., Huss, M., Kherif, F., Del Prado Martin, F.M., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U. S. A 103 (20), 7865–7870. https://doi.org/10.1073/pnas.0509989103.
Rimmele, J.M., Morillon, B., Poeppel, D., Arnal, L.H., 2018. Proactive sensing of periodic and aperiodic auditory patterns. Trends Cognit. Sci. 22, 870–882. https://doi.org/10.1016/j.tics.2018.08.003.
Shannon, C.E., 1948. The mathematical theory of communication. Bell Syst. Tech. J. 27, 379.
Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416 (6876), 87–90. https://doi.org/10.1038/416087a.
Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M., 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci., 879716
Thompson, L.A., 1995. Encoding and memory for visible speech and gestures: a comparison between young and older adults. Psychol. Aging 10, 215–228.
Thompson, L.A., Malloy, D., 2004. Attention resources and visible speech encoding in older and younger adults. Exp. Aging Res. 30, 241–252.
Thorne, J.D., Debener, S., 2014. Look now and hear what's coming: on the functional role of cross-modal phase reset. Hear. Res. 307, 144–152.
Thut, G., Veniero, D., Romei, V., Miniussi, C., Schyns, P., Gross, J., 2011. Rhythmic TMS causes local entrainment of natural oscillatory signatures. Curr. Biol. 21 (14), 1176–1185. https://doi.org/10.1016/j.cub.2011.05.049.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289.
Van Veen, B.D., van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44, 867–880.
van Wassenhove, V., Grant, K.W., Poeppel, D., 2005. Visual speech speeds up the neural processing of auditory speech. Proc. Natl. Acad. Sci. U. S. A 102 (4), 1181–1186. https://doi.org/10.1073/pnas.0408949102.

Wang, D., Clouter, A., Chen, Q., Shapiro, K.L., Hanslmayr, S., 2018. Single-trial phase entrainment of theta oscillations in sensory regions predicts human associative memory performance. J. Neurosci.: Off. J. Soc. Neurosci. 38 (28), 6299–6309. https://doi.org/10.1523/JNEUROSCI.0349-18.2018.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7 (7), 701–702. https://doi.org/10.1038/nn1263.

Zoefel, B., Archer-Boyd, A., Davis, M.H., 2018. Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. Curr. Biol. 28 (3), 401–408. https://doi.org/10.1016/j.cub.2017.11.071 e5.

Zoefel, B., Davis, M.H., Valente, G., Riecke, L., 2019. How to test for phasic modulation of neural and behavioural responses. Neuroimage 202, 116175. https://doi.org/10.1016/j.neuroimage.2019.116175.