Systems/Circuits

# Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech

Mathieu Bourguignon,[1,2,3] Martijn Baart,[1,4] Efthymia C. Kapnoula,[1] and Nicola Molinaro[1,5]

[1]Basque Center on Cognition, Brain and Language (BCLB), 20009 San Sebastian, Spain, [2]Laboratoire de Cartographie fonctionnelle du Cerveau and [3]Laboratoire Cognition Langage et Développement, UNI–ULB Neuroscience Institute, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium, [4]Department of Cognitive Neuropsychology, Tilburg University, 5037 AB Tilburg, The Netherlands, and [5]Ikerbasque, Basque Foundation for Science, 48013 Bilbao, Spain

唇读重要性

Lip-reading is crucial for understanding speech in challenging conditions. But how the brain extracts meaning from, silent, visual speech is still under debate. Lip-reading in silence activates the auditory cortices, but it is not known whether such activation reflects immediate synthesis of the corresponding auditory stimulus or imagery of unrelated sounds. To disentangle these possibilities, we used magnetoencephalography to evaluate how cortical activity in 28 healthy adult humans (17 females) entrained to the auditory speech envelope and lip movements (mouth opening) when listening to a spoken story without visual input (audio-only), and when seeing a silent video of a speaker articulating another story (video-only). In video-only, auditory cortical activity entrained to the absent auditory signal at frequencies <1 Hz more than to the seen lip movements. This entrainment process was characterized by an auditory-speech-to-brain delay of ~70 ms in the left hemisphere, compared with ~20 ms in audio-only. Entrainment to mouth opening was found in the right angular gyrus at <1 Hz, and in early visual cortices at 1–8 Hz. These findings demonstrate that the brain can use a silent lip-read signal to synthesize a coarse-grained auditory speech representation in early auditory cortices. Our data indicate the following underlying oscillatory mechanism: seeing lip movements first modulates neuronal activity in early visual cortices at frequencies that match articulatory lip movements; the right angular gyrus then extracts slower features of lip movements, mapping them onto the corresponding speech sound features; this information is fed to auditory cortices, most likely facilitating speech parsing.

*Key words:* audiovisual integration; lip-reading; magnetoencephalography; silent speech; speech entrainment

**Significance Statement**

Lip-reading consists in decoding speech based on visual information derived from observation of a speaker's articulatory facial gestures. Lip-reading is known to improve auditory speech understanding, especially when speech is degraded. Interestingly, lip-reading in silence still activates the auditory cortices, even when participants do not know what the absent auditory signal should be. However, it was uncertain what such activation reflected. Here, using magnetoencephalographic recordings, we demonstrate that it reflects fast synthesis of the auditory stimulus rather than mental imagery of unrelated, speech or non-speech, sounds. Our results also shed light on the oscillatory dynamics underlying lip-reading.

## Introduction

In everyday situations, seeing a speaker's articulatory mouth gestures, here referred to as lip-reading or visual speech, can help us decode the auditory speech signal (Sumby and Pollack, 1954). In fact, lip movements are intelligible even without an auditory signal, likely because there is a strong connection between auditory and visual speech (Munhall and Vatikiotis-Bateson, 2004; Chan-

Received May 14, 2019; revised Nov. 28, 2019; accepted Dec. 4, 2019.

Author contributions: M. Bourguignon, M. Baart, E.C.K., and N.M. designed research; M. Bourguignon, M. Baart, E.C.K., and N.M. performed research; M. Bourguignon and E.C.K. analyzed data; M. Bourguignon, M. Baart, E.C.K., and N.M. wrote the paper.

This work was supported by the Innoviris Attract program (Grant 2015-BB2B-10), by the Spanish Ministry of Economy and Competitiveness (Grant PSI2016-77175-P), and by the Marie Skłodowska-Curie Action of the European Commission (Grant 743562) to M. Bourguignon; by the Netherlands Organization for Scientific Research (VENI Grant 275-89-027) to M. Baart; by the Spanish Ministry of Economy and Competitiveness, through the Juan de la Cierva-Formación fellowship, and by the Spanish Ministry of Economy and Competitiveness (Grant PSI2017-82563-P) to E.C.K.; by the Spanish Ministry of Science, Innovation and Universities (Grant RTI2018-096311-B-I00), the Agencia Estatal de Investigación, the Fondo Europeo de Desarrollo Regional, and by the Basque government (Grant PI_2016_1_0014) to N.M.; and by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Programme for Centres/Units of Excellence in R&D" (SEV-2015-490) awarded to the BCBL. We thank Riitta Hari at Department of Art (Aalto University School of Arts, Design, and Architecture, Espoo, Finland) for helpful comments on the paper.

The authors declare no competing financial interests.

Correspondence should be addressed to Mathieu Bourguignon at mabourgu@ulb.ac.be.

https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Copyright © 2020 the authors

drasekaran et al., 2009). It is however not clear how the brain extracts meaning from visual speech.

Some evidence points to the possibility that visual speech is recoded into acoustic information. For example, seeing silent visual speech clips of simple speech sounds such as vowels or elementary words activates auditory cortical areas (Calvert et al., 1997; Pekkola et al., 2005), even when participants are not aware of what the absent auditory input should be (Calvert et al., 1997; Bernstein et al., 2002; Paulesu et al., 2003). However, recoding visual speech into an acoustic representation (here referred to as *synthesis*) is computationally demanding. It has therefore been suggested that meaning is directly extracted from visual speech within visual areas and heteromodal association cortices (Bernstein and Liebenthal, 2014; O'Sullivan et al., 2016; Lazard and Giraud, 2017; Hauswald et al., 2018). According to this view, activation in early auditory cortices driven by lip-reading might reflect imagery of unrelated (possibly speech) sounds (Bernstein and Liebenthal, 2014), but not a direct recoding of visual speech into its corresponding acoustic representation. As previous work has relied on time-insensitive neuroimaging techniques (Calvert et al., 1997; Bernstein et al., 2002; Paulesu et al., 2003; Pekkola et al., 2005), there was no empirical evidence to disentangle these two alternatives. Here, we took advantage of auditory cortical entrainment to look for decisive evidence to support the existence of a synthesis mechanism whereby visual speech is recoded into its corresponding auditory information.

When people listen to continuous natural speech, oscillatory cortical activity synchronizes with the auditory temporal speech envelope (Luo and Poeppel, 2007; Bourguignon et al., 2013; Gross et al., 2013; Peelle et al., 2013; Molinaro et al., 2016; Vander Ghinst et al., 2016; Meyer et al., 2017; Meyer and Gumbert, 2018). Such "speech-brain entrainment" originates mainly in auditory cortices at frequencies matching phrase (<1 Hz) and syllable rates (4–8 Hz), and is thought to be essential for speech comprehension (Ahissar et al., 2001; Luo and Poeppel, 2007; Peelle et al., 2013; Ding et al., 2016; Meyer et al., 2017). An electroencephalography study suggested that silent lip-read information entrains cortical activity at syllable rate when participants are highly familiar with speech content (Crosse et al., 2015). However, because participants knew what the absent speech sound should be in this study, it remains unclear whether entrainment is driven by the (1) lip-read information, (2) covert production or repetition of the speech segment, (3) top-down lexical and semantic processes, or (4) some combination of these factors.

Here, we address the following critical question: does the brain use lip-read input to bring auditory cortices to entrain to the audio speech signal even when there is no physical speech sound and participants do not know the content of the absent auditory signal? To do so, we evaluated entrainment to a spoken story without visual input (audio-only), and compared these data to a silent condition with a video of a speaker articulating another story (video-only). To determine the "lip-read specificity" of these entrainment patterns, we also included a condition in which the mouth configuration of the speaker telling another story was transduced into a dynamic luminance contrast (control-video-only). If the brain can synthesize properties of missing speech based on concurrent lip-reading in a timely manner, auditory cortical entrainment with the envelope of the audio signal should be similar in audio-only and video-only, even if the speech sound was not physically present in the latter condition.

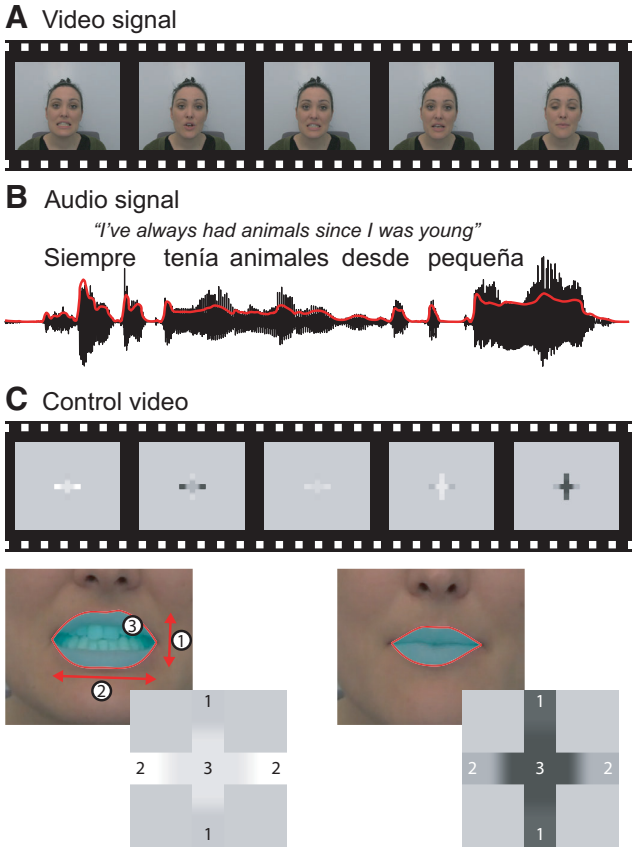


**Figure 1.** Experimental material. ***A***, ***B***, Two-second excerpt of video (***A***) and audio (***B***; auditory speech envelope in red) of the speaker telling a 5 min story about a given topic. There were eight different videos. Video without sound was presented in video-only, and sound without video was presented in audio-only. ***C***, Corresponding control video in which a flickering Greek cross encoded speaker's mouth configuration. Based on a segmentation of mouth contours, the cross encoded mouth opening (1), mouth width (2), and mouth surface (3). The resulting video was presented in control-video-only.

## Materials and Methods

*Participants.* Twenty-eight healthy human adults (17 females) aged 24.1 ± 4.0 years (mean ± SD) were included in the study. All reported being native speakers of Spanish and right-handed. They had normal or corrected-to-normal vision and normal hearing, had no prior history of neurological or psychiatric disorders, and were not taking any medication or substance that could influence the nervous system.

The experiment was approved by the BCBL Ethics Review Board and complied with the guidelines of the Helsinki Declaration. Written informed consent was obtained from all participants before testing.

*Experimental paradigm.* Figure 1 presents stimulus examples and excerpts. The stimuli were derived from 8 audio-visual recordings of a female native Spanish speaker talking for 5 min about a given topic (animals, books, food, holidays, movies, music, social media, and sports). Video and audio were simultaneously recorded using a digital camera (Canon, Legria HF G10) with an internal microphone. Video recordings were framed as head shots, and recorded at the PAL standard of 25 frames/s (videos were 1920 × 1080 pixels, 24 bits/pixel, with an auditory sampling rate of 44,100 Hz). The camera was placed ~70 cm away from the speaker, and the face spanned ~one-half of the vertical field-of-view. Final images were resized to a resolution of 1024 × 768 pixels.

For each video, a "control" video was created in which mouth movements were transduced into luminance changes (Fig. 1*C*). To achieve this we extracted lip contours from each individual frame of the video recordings with an in-house MATLAB code based on the approach of Eveno et al. (2004). In the control video, the luminance of a Greek cross changed according to mouth configuration (Fig. 1*C*). Its size (300 × 300 pixels)

was roughly matched with the extent of the eyes and mouth, which are the parts of the face people tend to look at when watching a speaker's face (Vatikiotis-Bateson et al., 1998). Mouth configuration variables (mouth opening, width, and surface) were rescaled so that their 1st and 99th percentiles corresponded to the minimum and maximum luminance levels. The center of the cross encoded the mouth surface area, its top and bottom portions encoded mouth opening, and its left- and right-most portions encoded mouth width. In this configuration, the three represented parameters were spatially and temporally congruent with the portion of the mouth they parametrized. All portions were smoothly connected by buffers along which the weight of the encoded parameters varied as a squared cosine. These control videos were designed to determine whether effects were specific to lip-reading. The transduced format was preferred to other classical controls such as meaningless lip movements or gum-chewing motions because it preserved the temporal relation between the visual input and underlying speech sounds.

For each sound recording, we derived a non-speech control audio consisting of white noise modulated by the auditory speech envelope. These control sounds were designed to determine whether uncovered effects were specific to speech. However, conditions that included these control sounds were not analyzed because they were uninformative about lip-reading driven oscillatory entrainment.

In total, participants completed 10 experimental conditions while sitting with their head in a MEG helmet. This included all nine possible combinations of three types of visual stimuli (original, control, no video) and three types of audio stimuli (original, control, no audio). The test condition with no audio and no video was trivially labeled as the *rest* condition and lasted 5 min. Each of the other 8 conditions was assigned to 1 of the 8 stories (condition–story assignment counterbalanced across participants). In this way, we ensured that each condition was presented continuously for 5 min, and that the same story was never presented twice. The tenth condition was a localizer condition in which participants attended 400 Hz pure tones and checkerboard pattern reversals lasting 10 min. This condition is not analyzed in this paper. All conditions were presented in random order, separated by short breaks. Videos were shown on a back-projection screen (videos were 41 × 35 cm) placed in front of the participants at a distance of ~1 m. Sounds were delivered at 60 dB (measured at ear-level) through a front-facing speaker (Panphonics) placed ~1 m behind the screen. Participants were instructed to watch the videos and listen to the sounds attentively.

To investigate our research hypotheses, we focused on the following conditions: (1) the original speech audio with no video, referred to as audio-only; (2) the original video with no audio, referred to as video-only; (3) the control video with no audio, referred to as the control-video-only; and (4) the rest.

*Data acquisition.* Neuromagnetic signals were acquired with a whole-scalp-covering neuromagnetometer (Vectorview, Elekta) in a magnetically shielded room. The recording pass-band was 0.1–330 Hz and the signals were sampled at 1 kHz. The head position inside the MEG helmet was continuously monitored by feeding current to four head-tracking coils located on the scalp. Head position indicator coils, three anatomical fiducials, and at least 150 head-surface points (covering the whole scalp and the nose surface) were localized in a common coordinate system using an electromagnetic tracker (Fastrak, Polhemus).

Eye movements were tracked with an MEG-compatible eye tracker (EyeLink 1000 Plus, SR Research). Participants were calibrated using the standard 9-point display and monocular eye movements were recorded at a sampling rate of 1 kHz. Eye-movements were recorded for the duration of all experimental conditions.

High-resolution 3D-T1 cerebral magnetic resonance images (MRIs) were acquired on a 3-tesla MRI scan (Siemens Medical System) facility available at the BCBL.

*MEG preprocessing.* Continuous MEG data were first preprocessed off-line using the temporal signal space separation method (correlation coefficient: 0.9, segment length: 10 s) to suppress external sources of interference and to correct for head movements (Taulu et al., 2005; Taulu and Simola, 2006). To further suppress heartbeat, eye-blink, and eye-movement artifacts, 30 independent components (Vigário et al., 2000; Hyvärinen et al., 2004) were evaluated from the MEG data low-pass filtered at 25 Hz using FastICA algorithm (dimension reduction: 30, nonlinearity: tanh). Independent components corresponding to such artifacts were identified based on their topography and time course and were removed from the full-rank MEG signals.

*Coherence analysis.* Coherence was estimated between MEG signals and (1) the auditory speech temporal envelope, (2) mouth opening, (3) mouth width, and (4) mouth surface. The auditory speech temporal envelope was obtained by summing the Hilbert envelope of the auditory speech signal filtered through a third octave filter bank (central frequency ranging linearly on a log-scale from 250 to 1600 Hz; 19 frequency bands), and was further resampled to 1000 Hz time-locked to the MEG signals (Fig. 1*B*). Continuous data from each condition were split into 2 s epochs with 1.6 s epoch overlaps, affording a spectral resolution of 0.5 Hz while decreasing noise on coherence estimates (Bortel and Sovka, 2014). MEG epochs exceeding 5 pT (magnetometers) or 1 pT/cm (gradiometers) were excluded from further analyses to avoid data contamination by artifact sources that had not been suppressed by the temporal signal space separation or removed with independent component analysis. These steps led to an average of 732 artifact-free epochs across participants and conditions (SD = 36). A one-way repeated-measures ANOVA revealed no differences between conditions ($F_{(2,54)} = 1.07$, $p = 0.35$). Next, we estimated sensor-level coherence (Halliday et al., 1995) and combined gradiometer pairs based on the direction of maximum coherence (Bourguignon et al., 2015). Only values from these gradiometer pairs are presented in the results.

In coherence analyses, we focused on four frequency ranges (0.5, 1–3, 2–5, and 4–8 Hz) by averaging coherence across the frequency bins they encompassed. The 2–5 and 4–8 Hz frequency ranges were well matched to the count rate of words (3.34 ± 0.12 Hz; mean ± SD. across the 8 videos) and syllables (5.91 ± 0.12 Hz), whereas the count rate of phrases (1.01 ± 0.20 Hz) fell in between the two lowest ranges. As in a previous study (Vander Ghinst et al., 2019), rates were assessed as the number of phrases, words, or syllables manually extracted from audio recordings divided by the corrected duration of the audio recording. For phrases, the corrected duration was trivially the total duration of the audio recording. For words and syllables, the corrected duration was the total time during which the talker was actually talking, that is the total duration of the audio recording (here 5 min) minus the sum of all silent periods when the auditory speech envelope was below one-tenth of its mean for at least 100 ms. Note that setting the threshold for the duration defining a silent period to a value obviously too low (10 ms) or too high (500 ms) changed the estimates of word and syllable count rates by only ~10%. These frequency ranges were selected also because auditory speech entrainment dominates at 0.5 and 4–8 Hz (Luo and Poeppel, 2007; Bourguignon et al., 2013; Gross et al., 2013; Peelle et al., 2013; Molinaro et al., 2016; Vander Ghinst et al., 2016; Meyer et al., 2017; Meyer and Gumbert, 2018) but is also present at intermediate frequencies (Keitel et al., 2018), and because lip entrainment has previously been identified at 2–5 Hz (Park et al., 2016; Giordano et al., 2017). Coherence maps were also averaged across participants for illustration purposes.

We only report coherence estimated between MEG signals and (1) the auditory speech envelope and (2) mouth opening. Although tightly related, the two latter signals displayed only a moderate degree of coupling, that peaked at 0.5 and 4–8 Hz (Fig. 2*Ai*), with a visual-to-auditory speech delay of ~120 ms (maximum cross-correlation between auditory speech envelope and mouth opening; Fig. 2*Bi*). Mouth opening and mouth surface were coherent at >0.7 across the 0–10 Hz range (Fig. 2*Aii*) and yielded similar results. Mouth width displayed a moderate level of coherence with mouth opening (Fig. 2*Aiii*) and an unclear visual-to-auditory speech delay (Fig. 2*Bii*). Mouth width was not included in the main analyses because it led to lower coherence values with MEG signals than mouth opening, but was retained as a nuisance factor in the partial coherence analyses (described later in this subsection).

It is worth noting that the magnitude of the coupling between the auditory speech envelope and mouth opening (as assessed by coherence) we report for our audiovisual stimuli was 2–3 times lower than that reported previously (Park et al., 2016; Hauswald et al., 2018). To ensure that this discrepancy was not due to the inadequacy of our lip-extraction procedure, we compared our time-series of mouth parameters to those

extracted using a deep-learning-based solution (Visage Technologies; face tracking and analysis). This revealed a good correspondence between the estimated time-series for mouth opening ($r = 0.95 \pm 0.01$; mean ± SD. across the 8 videos), mouth width ($r = 0.88 \pm 0.01$), and mouth surface ($r = 0.95 \pm 0.01$). The genuine difference between the level of audio-visual speech coupling found in our study compared with others might be due to the language used (Spanish here vs English elsewhere), or to the idiosyncrasies of our talker. Nevertheless, this relative decoupling between audio and visual speech signals provided an opportunity to separate their respective cortical representations more efficiently.

Coherence was also estimated at the source level. To do so, individual MRIs were first segmented using FreeSurfer software (Reuter et al., 2012; RRID:SCR_001847). Then, the MEG forward model was computed using the Boundary Element Method implemented in the MNE software suite (Gramfort et al., 2014; RRID: SCR_005972) for three orthogonal tangential current dipoles (corresponding to the 3 spatial dimensions) placed on a homogeneous 5 mm grid source space covering the whole brain. At each source, the forward model was further reduced to its two first principal components, which closely corresponded to sources tangential to the skull; the discarded component corresponded to the radial source which is close to magnetically silent. Coherence maps were produced within the computed source space at 0.5, 1–3, 2–5, and 4–8 Hz using a linearly constrained minimum variance beamformer built based on the *rest* data covariance matrix (Van Veen et al., 1997; Hillebrand and Barnes, 2005). Source maps were then interpolated to a 1 mm homogenous grid and smoothed with a Gaussian kernel of 5 mm full-width at half-maximum. Both planar gradiometers and magnetometers were used for inverse modeling after dividing each sensor signal (and the corresponding forward-model coefficients) by the SD of its noise. The noise variance was estimated from the continuous rest MEG data band-passed through 1–195 Hz, for each sensor separately.

Coherence maps were also produced at the group level. A nonlinear transformation from individual MRIs to the MNI brain was first computed using the spatial normalization algorithm implemented in Statistical Parametric Mapping (SPM8; Ashburner et al., 1997; Ashburner and Friston, 1999; RRID:SCR_007037) and then applied to individual MRIs and coherence maps. This procedure generated a normalized coherence map in the MNI space for each subject and frequency range. Coherence maps were then averaged across participants.

Individual and group-level coherence maps for the auditory speech envelope (mouth opening, respectively) were also estimated after controlling for mouth opening and mouth width (the auditory speech envelope, respectively) using partial coherence (Halliday et al., 1995). Partial coherence is the direct generalization of partial correlation (Kendall and Stuart, 1968) to the frequency domain (Halliday et al., 1995).

The same approach was used to estimate coherence between MEG (in the sensor and source space) and global changes (or edges) in the visual stimulus, and to partial out such "global visual change" from coherence maps for the auditory speech envelope. The global visual change signal was computed at every video frame as the sum of squares of the difference between that frame and the previous frame, divided by the sum of squares of the previous frame. This signal predominantly identified edges corresponding to periods when the speaker moved her head, eyebrows and jaw (Fig. 3). The rationale being that these periods may tend to co-occur with the onset of phrases and sentences (Munhall et al., 2004) and could modulate oscillatory activity in auditory cortices (Schroeder et al., 2008).

Finally, individual and group-level coherence maps for the auditory speech envelope in video-only were estimated after shifting the auditory speech envelope by ~30, ~60, … ~240, and ~270 s. For each subject and time-shift, the exact time-shift applied was selected within a ±10 s window around the target time-shift, at the silent period for which the auditory speech envelope smoothed with a 1-s square kernel was at the minimum. Ensuing values of coherence were used to rule out the possibility that coherence with the genuine auditory speech envelope results from general temporal characteristics of auditory speech.

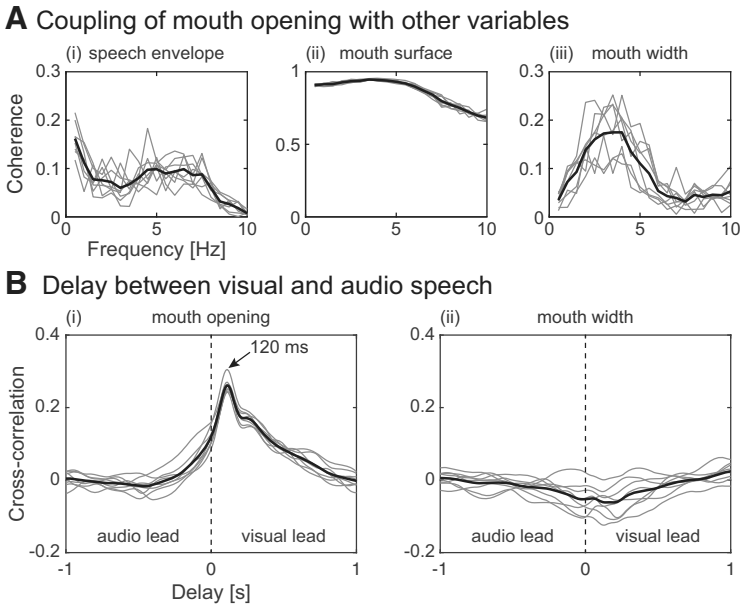


**Figure 2.** Relation between audio and visual speech signals. ***A***, Frequency-dependent coupling (coherence) of mouth opening with auditory speech envelope (***Ai***), mouth surface (***Aii***), and mouth width (***Aiii***). There is one gray trace per video (8 in total), and thick black traces are the average across them all. ***B***, Delay between visual and audio speech assessed with cross-correlation of auditory speech envelope with mouth opening (***Bi***) and mouth width (***Bii***).

*Estimation of temporal response functions.* We used temporal response functions (TRFs) to model how the auditory speech envelope affected the temporal dynamics of auditory cortical activity. Based on our results, TRFs were estimated only for the 0.2–1.5 Hz frequency range, in the audio-only and video-only conditions. A similar approach has been used to model brain responses to speech at 1–8 Hz (Lalor and Foxe, 2010; Zion Golumbic et al., 2013), and to model brain responses to natural force fluctuations occurring during maintenance of constant hand grip contraction (Bourguignon et al., 2017). TRFs are the direct analog of evoked responses in the context of continuous stimulation.

We used the mTRF toolbox (Crosse et al., 2016) to estimate the TRF of auditory cortical activity associated with the auditory speech envelope. In all conditions, source signals were reconstructed at individual coordinates of maximum 0.5 Hz coherence with the auditory speech envelope in audio-only. These two-dimensional source signals were projected onto the orientation that maximized the coherence with the auditory speech envelope at 0.5 Hz. Then, the source signal was filtered at 0.2–1.5 Hz, the auditory speech envelope was convolved with a 50-ms square smoothing kernel and both were down-sampled to 20 Hz (note that for auditory speech envelope, this procedure is equivalent to taking the mean over samples 25 ms around sampling points). For each subject, the TRFs were modeled from −1.5 to +2.5 s, for a fixed set of ridge values ($\lambda = 2^0$, $2^1$, $2^2$… $2^{20}$). We adopted the following tenfold cross-validation procedure to determine the optimal ridge value: for each subject, TRFs were estimated based on 90% of the data, and used to predict the 10% of data left out and the Pearson correlation was then estimated between predicted and measured signals. The square of the mean correlation value across the 10 runs provided an estimate of the proportion of variance explained by entrainment to the auditory speech envelope. TRFs were recomputed based on all the available data for the ridge value maximizing the mean explained variance. To deal with sign ambiguity, the polarity of each TRF was adapted so that correlation with the first singular vector of all subjects' TRF in the range −0.5 to 1.0 s is positive.

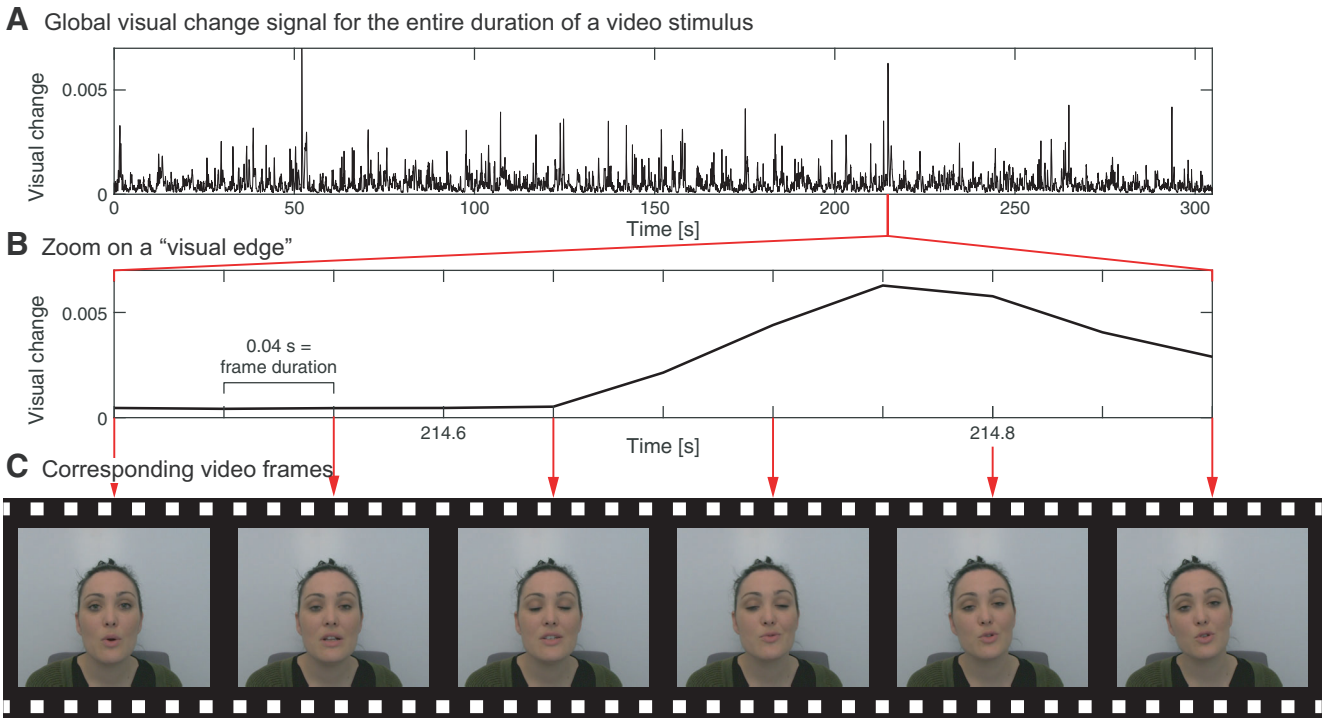


**Figure 3.** Global visual changes in the visual stimuli. ***A***, The global visual change signal as a function of time for the entire duration of a video stimulus. ***B***, Zoom on one of the most prominent edges (peaks) of the global visual change signal. ***C***, Video frames corresponding to this visual edge, showing that it was due to head movements.

Based on our results, the TRF framework was also used to model brain responses to mouth opening and the global visual change signal at 0.2–1.5 Hz and mouth opening at 2–5 Hz, and to model the evolution of the auditory speech envelope at 0.2–1.5 Hz associated with the time course of (1) mouth opening, (2) global visual change, and (3) the Hilbert envelope of mouth opening in the 2–5 Hz band. Note that the last TRF seeks phase–amplitude coupling between auditory speech envelope at 0.2–1.5 Hz (phase) and mouth opening at 2–5 Hz (amplitude), with the, perhaps not that common, perspective that the amplitude signal drives the phase signal. We used exactly the same parameters as reported in the previous paragraph, except the data for the brain response to mouth opening at 2–5 Hz were downsampled to 50 Hz and modeled from −0.7 to 1.2 s.

*Eye-tracking data.* As in previous studies using eye-tracking (McMurray et al., 2002; Kapnoula et al., 2015), eye-movements were automatically parsed into saccades and fixations using default psychophysical parameters. Adjacent saccades and fixations were combined into a single "look" that started at the onset of the saccade and ended at the offset of the fixation.

A region of interest was identified for each of the three critical objects: mouth and eyes in video-only and flickering cross in control-video-only (Fig. 4). In converting the coordinates of each look to the object being fixated, the boundaries of the regions of interest were extended by 50 pixels to account for noise and/or head-drift in the eye-tracking recording. This did not result in any overlap between the eye and mouth regions.

Based on these regions of interest, we estimated the proportion of eye fixation to the combined regions of interest encompassing eyes and mouth in video-only and flickering cross in control-video-only. Eyes and mouth regions were combined because these are the parts of the face people tend to look at when watching a talking face (Vatikiotis-Bateson et al., 1998). Importantly, even when people are looking at the eyes, lip movements, in the periphery of the field-of-view, still benefit speech perception (Paré et al., 2003; Kaplan and Jesse, 2019). The two resulting areas were of comparable size: 100,800 pixels for the flickering cross versus 77,300 pixels for the eyes and mouth. Data from one participant were excluded because of technical issues during acquisition, and eye fixation analyses were thus based on data from 27 participants.

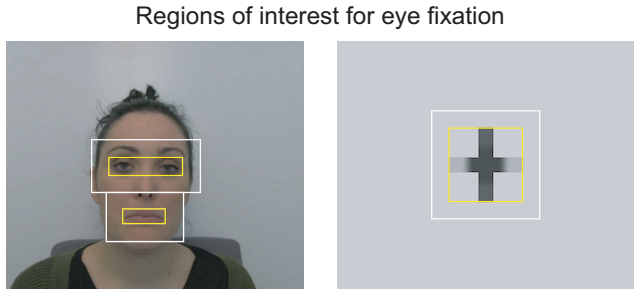


**Figure 4.** Regions of interest for eye fixation. The initial regions of interest are delineated in yellow, and the extended ones in white. Eye fixation analyses were based on extended regions. In video-only (left), the final region of interest comprised the mouth and the eyes. In control-video-only (right), it encompassed the flickering cross.

*Experimental design and statistical analyses.* Sample size was based on previous studies reporting entrainment to lip movements, which included 46 (Park et al., 2016) and 19 (Giordano et al., 2017) healthy adults.

The statistical significance of the local coherence maxima observed in group-level maps was assessed with a nonparametric permutation test that intrinsically corrects for multiple spatial comparisons (Nichols and Holmes, 2002). Subject- and group-level rest coherence maps were computed in a similar way to the genuine maps; MEG signals were replaced by rest MEG signals while auditory/visual speech signals were identical. Group-level difference maps were obtained by subtracting genuine and rest group-level coherence maps. Under the null hypothesis that coherence maps are the same regardless of the experimental condition, genuine and rest labels should be exchangeable at the subject-level before computing the group-level difference map (Nichols and Holmes, 2002). To reject this hypothesis and to compute a threshold of statistical significance for the correctly labeled difference map, the permutation distribution of the maximum of the difference map's absolute value was computed for a subset of 1000 permutations. The threshold at $p < 0.05$ was computed as the 95th percentile of the permutation distribution

**Table 1. Significant peak of speech and lip entrainment: peak MNI coordinates, significance level, confidence volume, and anatomical location**

| | Peak coordinates, mm | *p* | Mean ± SD | Confidence volume, $cm^3$ | Anatomical location |
|---|---|---|---|---|---|
| Speech entrainment at 0.5 Hz | | | | | |
| Audio-only | −64, −19, 8 | $<10^{-3}$ | | | Left auditory cortex |
| | 64, −21, 6 | $<10^{-3}$ | 0.075 ± 0.046 | 5.5 | Right auditory cortex |
| Video-only | −46, −30, 11 | 0.003 | 0.025 ± 0.017 | 35.5 | Left auditory cortex |
| | 68, −14, −2** | 0.029 (0.085) | 0.024 ± 0.015 | 5.6 | Right auditory cortex |
| | −57, 25, 15 | 0.005 | 0.021 ± 0.013 | 9.6 | Left inferior frontal gyrus |
| | −58, −15, 41* | 0.018 (0.063) | 0.023 ± 0.012 | 21.3 | Left inferior precentral sulcus |
| Lip entrainment at 0.5 Hz | | | | | |
| Video-only | 49, −46, 10 | 0.002 | 0.022 ± 0.014 | 30.9 | Right angular gyrus |
| Control-video-only | 10, −89, −21 | $<10^{-3}$ | 0.028 ± 0.023 | 6.3 | Inferior occipital area |
| | 25, −96, −1 | 0.008 | 0.027 ± 0.023 | 11.7 | Right lateral occipital cortex |
| | −23, −97, −4 | 0.046 | 0.023 ± 0.014 | 39.1 | Left lateral occipital cortex |
| Speech entrainment at 1–3 Hz | | | | | |
| Audio-only | −62, −15, 11 | $<10^{-3}$ | 0.031 ± 0.017 | 0.17 | Lett auditory cortex |
| | 66, −10, 9 | $<10^{-3}$ | 0.036 ± 0.022 | 0.22 | Right auditory cortex |
| Video-only | −51, −65, −16 | 0.020 | 0.012 ± 0.004 | 58.8 | Left inferior temporal gyrus |
| | −67, −20, −12* | 0.005 (0.22) | 0.012 ± 0.004 | 2.8 | Left middle temporal gyrus |
| Lip entrainment at 1–3 Hz | | | | | |
| Video-only | 5, −92, −13 | $<10^{-3}$ | 0.015 ± 0.007 | 22.6 | Calcarine cortex |
| | 33, −92, 6 | 0.001 | 0.014 ± 0.007 | 5.2 | Right lateral occipital sulcus |
| | −15, −96, 12 | $<10^{-3}$ | 0.015 ± 0.005 | 18.3 | Left calcarine cortex |
| Control-video-only | 1, −98, 10 | $<10^{-3}$ | 0.029 ± 0.016 | 0.3 | Calcarine cortex |
| | 34, −92, −3 | $<10^{-3}$ | 0.028 ± 0.016 | 0.9 | Right lateral occipital cortex |
| | −28, −94, −10 | $<10^{-3}$ | 0.023 ± 0.012 | 3.4 | Left lateral occipital cortex |
| Speech entrainment at 2–5 Hz | | | | | |
| Audio-only | 67, −11, 10 | $<10^{-3}$ | 0.020 ± 0.008 | 0.3 | Left auditory cortex |
| | −62, −14, 13 | $<10^{-3}$ | 0.016 ± 0.007 | 0.4 | Right auditory cortex |
| Lip entrainment at 2–5 Hz | | | | | |
| Video-only | −14, −97, 11 | $<10^{-3}$ | 0.018 ± 0.007 | 8.3 | Left calcarine cortex |
| | 2, −93, −2 | $<10^{-3}$ | 0.018 ± 0.008 | 15.1 | Calcarine cortex |
| Control-video-only | −1, −98, 11 | $<10^{-3}$ | 0.026 ± 0.016 | 1.8 | Calcarine cortex |
| | 25, −97, −8 | $<10^{-3}$ | 0.025 ± 0.012 | 1.9 | Right lateral occipital cortex |
| | −28, −94, −10 | $<10^{-3}$ | 0.024 ± 0.012 | 0.3 | Left lateral occipital cortex |
| Speech entrainment at 4 – 8 Hz | | | | | |
| Audio-only | −64, −18, 7 | $<10^{-3}$ | 0.013 ± 0.005 | 1.4 | Left auditory cortex |
| | 67, −13, 5 | $<10^{-3}$ | 0.020 ± 0.009 | 0.3 | Right auditory cortex |
| Lip entrainment at 4 – 8 Hz | | | | | |
| Video-only | 10, −94, −4 | $<10^{-3}$ | 0.013 ± 0.005 | 19.3 | Right calcarine cortex |
| | −11, −95, 9 | 0.001 | 0.013 ± 0.004 | 18.8 | Left calcarine cortex |
| Control-video-only | −4, −88, −18 | $<10^{-3}$ | 0.016 ± 0.008 | 5.3 | Inferior occipital cortex |
| | 27, −94, −5 | $<10^{-3}$ | 0.016 ± 0.007 | 22.5 | Right lateral occipital cortex |
| | −5, −97, 15 | 0.011 | 0.015 ± 0.006 | 30.0 | Calcarine cortex |

Only significant peaks of speech entrainment that survived partialing out lip movements (exceptions marked with *) and global visual changes (exceptions marked with **) are presented here. Likewise, only peaks of significant lip entrainment that survived partialing out the auditory speech envelope are presented here. For the exceptions, *p* values are displayed in parentheses.

(Nichols and Holmes, 2002). Permutation tests can be too conservative for voxels other than the one with the maximum observed statistic (Nichols and Holmes, 2002). For example, dominant coherence values in the right auditory cortex could bias the permutation distribution and overshadow weaker coherence values in the left auditory cortex, even if these were highly consistent across subjects. Therefore, the permutation test described above was conducted separately for left- and right-hemisphere voxels. All suprathreshold local coherence maxima were interpreted as indicative of brain regions showing statistically significant coupling with the auditory or visual signal.

A confidence volume was estimated for all significant local maxima, using the bootstrap-based method described by Bourguignon et al. (2018). The location of the maxima was also compared between conditions using the same bootstrap framework (Bourguignon et al., 2018).

For each local maximum, individual maximum coherence values were extracted within a 10 mm sphere centered on the group level coordinates, or on the coordinates of maxima for audio-only. Coherence values were compared between conditions or signals of reference with two-sided paired *t* tests.

The bootstrap method was used to assess the timing of peak TRFs (Efron and Tibshirani, 1993). As a preliminary step, TRFs were upsampled by spline interpolation to 1000 Hz. A bootstrap distribution based on 10,000 random drawings of subjects (or videos) was then built for the timing of peak TFR, from which we extracted the mean and SD. Also the bias-corrected and accelerated bootstrap (Efron and Tibshirani, 1993) was used to compare the timing of peak TRF between conditions.

For the eye-tracking data, individual proportions of fixations were transformed using the empirical-logit transformation (Collins et al., 1992). Fixations to eyes and mouth in video-only were compared with fixations to the flickering cross in control-video-only using a two-sided paired *t* test across participants.

*Data and software availability.* MEG and eye-tracking data as well as video stimuli are available on request from the corresponding author.

## Results

Table 1 provides the coordinates and significance level of the loci of statistically significant coherence with the auditory speech envelope (henceforth, speech entrainment) and mouth opening (henceforth, lip entrainment) in all conditions (audio-only, video-only, and control-video-only) at all the selected frequency ranges (0.5, 1–3, 2–5, and 4–8 Hz).

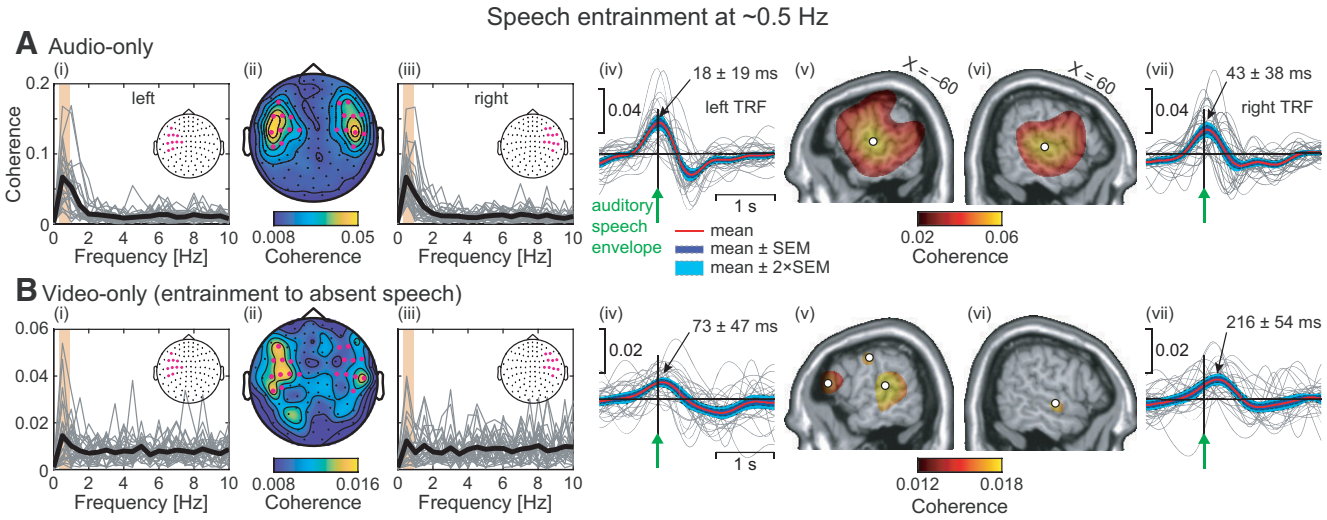


**Figure 5.** Speech entrainment at 0.5 Hz. ***A***, Speech entrainment in audio-only. ***Ai–Aiii***, Sensor distribution of speech entrainment at 0.5 Hz quantified with coherence (***Aii***) and its spectral distribution at a selection of 10 sensors in the left (***Ai***) and right hemisphere (***Aiii***) of maximal 0.5 Hz coherence (magenta). Gray traces represent individual subject's spectra at the sensor of maximum 0.5 Hz coherence within the preselection, and the thick black trace is their group average. ***Aiv–Avii***, Brain distribution of significant speech entrainment quantified with coherence in the left (***Av***) and right hemispheres (***Avi***) and the TRF associated with auditory speech envelope at coordinates of peak coherence (white discs) in the left (***Aiv***) and right hemispheres (***Avii***). In brain images, significant coherence values at MNI coordinates $|X| > 40$ mm were projected orthogonally onto the parasagittal slice of coordinates $|X| = 60$ mm. ***B***, Same as in ***A*** for video-only, illustrating that seeing speaker's face was enough to elicit significant speech entrainment at auditory cortices. Note that coherence spectra were estimated at the subject-specific sensor selected based on coherence in audio-only.

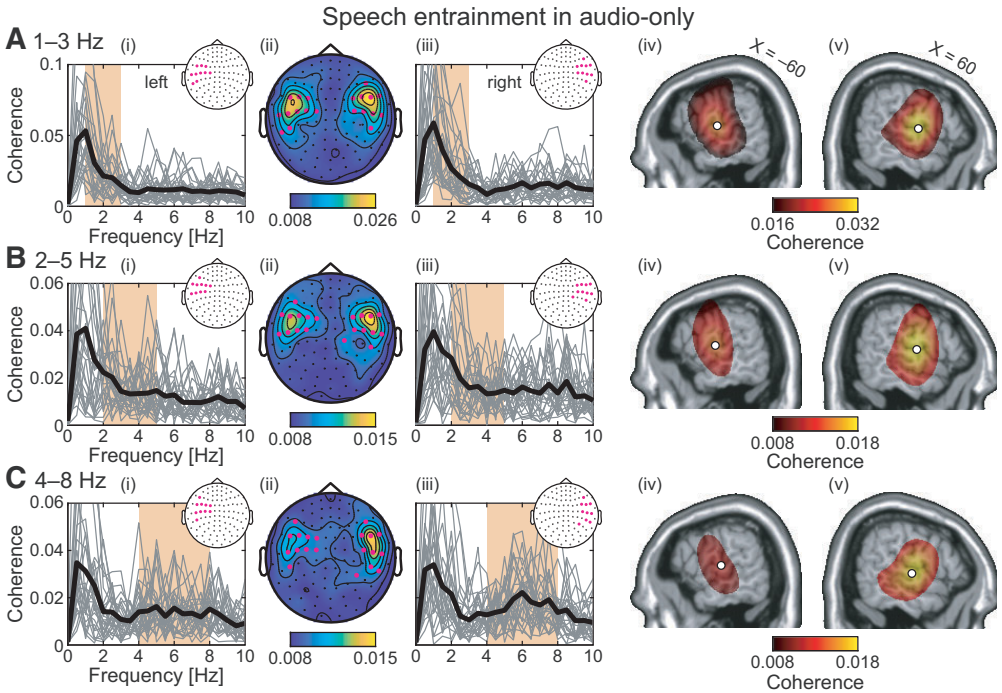


**Figure 6.** Speech entrainment quantified with coherence in audio-only at 1–3 Hz (***A***), 2–5 Hz (***B***), and 4–8 Hz (***C***). ***Ai–Ciii***, Sensor distribution of speech entrainment (***ii***) and its spectral distribution at a selection of 10 left-hemisphere (***i***) and right-hemisphere (***iii***) sensors of maximum coherence (magenta). Gray traces represent individual subject's spectra at the sensor of maximum coherence across the considered frequency range and within the preselection, and the thick black trace is their group average. (***Aiv–Cv***) Brain distribution of significant speech entrainment in the left (***iv***) and right hemispheres (***v***) produced as described in Figure 5.

### Entrainment to heard speech

In audio-only, significant speech entrainment peaked at sensors covering bilateral auditory regions in all the explored frequency ranges: 0.5 Hz (Fig. 5*A*), 1–3 Hz (Fig. 6*A*), 2–5 Hz (6*B*), and 4–8 Hz (6*C*). Underlying sources were located in bilateral auditory cortices (Figs. 5*A*, 6; Table 1).

### Auditory cortices entrain to absent speech at frequencies <1 Hz

In visual-only, there was significant 0.5 Hz entrainment to the speech sound that was actually produced by the speaker, but not heard by participants (Fig. 5*B*; Table 1). The significant loci for speech entrainment were the bilateral auditory cortices, the left

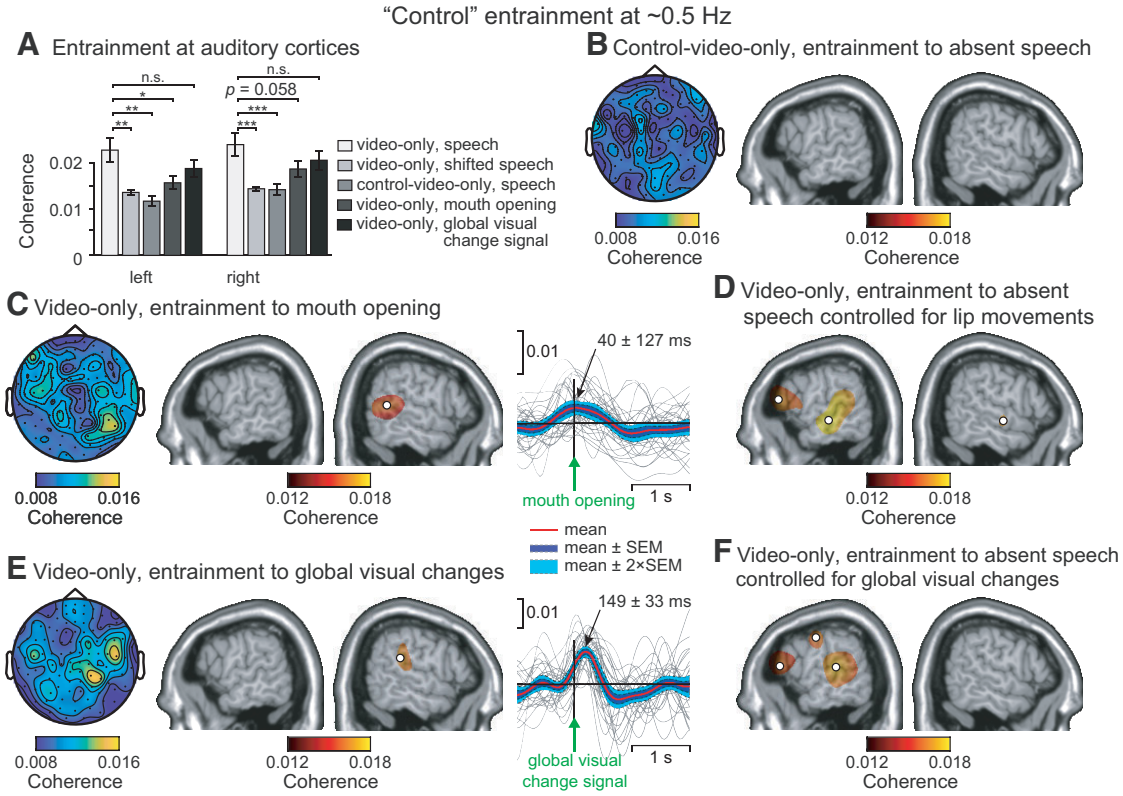


**Figure 7.** Control for the entrainment to absent speech at 0.5 Hz. ***A***, Entrainment values quantified with coherence at coordinates identified in audio-only (mean ± SD. across participants). ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, n.s. non significant. ***B***, Sensor and brain distribution of auditory speech entrainment in control-video-only wherein speech entrainment was not significant. ***C***, Sensor and brain distribution of lip entrainment in video-only and associated temporal evolution. Lip entrainment was significant only in the right angular gyrus. ***D***, Brain distribution of significant speech entrainment at 0.5 Hz after partialing out lip movements (mouth opening and mouth width). ***E***, ***F***, As in ***C*** and ***D*** but for the global visual change signal instead of mouth opening. Brain images were produced as described in Figure 5.

inferior frontal gyrus, and the inferior part of the left precentral sulcus (Fig. 5*B*; Table 1). Critically, the location of the auditory sources where we observed maximum 0.5 Hz entrainment did not differ significantly between audio-only and video-only (left: $F_{(3,998)} = 1.62$, $p = 0.18$; right: $F_{(3,998)} = 0.85$, $p = 0.47$). Not surprisingly, the magnitude of 0.5 Hz speech entrainment was higher in audio-only than in video-only (left: $t_{(27)} = 6.36$, $p < 0.0001$; right: $t_{(27)} = 6.07$, $p < 0.0001$). Nevertheless, brain responses associated with speech entrainment at ~0.5 Hz displayed a similar time course in audio-only and video-only (Fig. 5*A*,*B*). In the left hemisphere, brain response peaked after the auditory speech envelope with a delay that did not differ significantly between the two conditions (audio-only: 18 ± 19 ms, video-only: 73 ± 47 ms; $p = 0.27$); in the right hemisphere this delay was significantly shorter for audio-only (43 ± 38 ms) than video-only (216 ± 54 ms; $p = 0.019$). These results demonstrate that within the auditory cortices, neuronal activity at ~0.5 Hz is modulated similarly by heard speech sounds and absent speech when lip-read information is available, but incurs an additional delay in the right hemisphere. Next, we address four critical questions related to this effect: (1) Can it be explained by the general temporal characteristics of auditory speech? (2) Is it unspecific to seeing the speaker's face? (3) Is it a direct result of lip-reading-induced visual activity simply being fed to auditory areas? (4) Is it mediated by edges in the visual stimuli (predominantly reflecting head, eyebrow, and jaw movements) that would prime phrase/sentence onset and modulate auditory cortical activity. A negative answer to these four questions would support the view that auditory speech envelope is "synthesized" through internal models that map visual speech onto sound features.

### Below-1-Hz entrainment to absent speech is not explained by the general temporal characteristics of auditory speech

In video-only, auditory sources (coordinates identified in audio-only) entrained significantly more to the corresponding, though absent, auditory speech than to unrelated auditory speech, here taken as the corresponding speech shifted in time (left: $t_{(27)} = 3.08$, $p = 0.0047$; right: $t_{(27)} = 3.78$, $p = 0.0008$; Fig. 7*A*). In this analysis, individual subject values were computed as the mean value across all considered time shifts. In addition, inspection of the maps of entrainment to unrelated speech did not reveal any special tendency to peak in auditory regions. This demonstrates that entrainment to absent speech in auditory cortices is not a consequence of the general temporal characteristics of auditory speech.

### Below-1-Hz entrainment to absent speech is specific to seeing speaker's face

Analysis of a control-visual-only condition revealed that entrainment to unheard speech at auditory cortices was specific to seeing the speaker's face. In the control condition, participants were looking at a silent video of a flickering Greek cross whose luminance pattern dynamically encoded the speaker's mouth configuration. We observed luminance-driven entrainment at 0.5 Hz at occipital cortices (Table 1), but no significant entrainment with unheard speech ($p > 0.1$; Fig. 7*B*). Importantly, speech entrain-

ment at auditory sources (coordinates identified in audio-only) was significantly higher in video-only than in control-video-only (left: $t_{(27)} = 3.44$, $p = 0.0019$; right: $t_{(27)} = 4.44$, $p = 0.00014$; Fig. 7*A*). These differences in auditory speech entrainment cannot be explained by differences in attention as participants attended the flickering cross in control-video-only approximately as much as speaker's eyes and mouth in video-only (81.0 ± 20.9 vs 87.5 ± 17.1%; $t_{(26)} = 1.30$, $p = 0.20$: fixation data derived from eye-tracking recordings). This demonstrates that auditory cortical entrainment to unheard speech is specific to seeing the speaker's face.

### Below-1-Hz entrainment to absent speech does not result from a direct feeding of lip movements to auditory cortices

Although driven by lip-read information, auditory cortical activity at ~0.5 Hz in visual-only entrained more to unheard speech than to seen lip movements. Indeed, speech entrainment was stronger than lip entrainment at the left auditory source coordinates identified in audio-only ($t_{(27)} = 2.52$, $p = 0.018$; Fig. 7*A*). The same trend was observed at the right auditory source ($t_{(27)} = 1.98$, $p = 0.058$; Fig. 7*A*). However, at 0.5 Hz, lip movements entrained brain activity in the right angular gyrus (Fig. 7*C*; Table 1), a visual integration hub implicated in biological motion perception (Allison et al., 2000; Puce and Perrett, 2003). Such entrainment entailed a visual-speech-to-brain delay of 40 ± 127 ms. Note that the dominant source of lip and speech entrainment were ~4 cm apart ($F_{(3,998)} = 4.68$, $p = 0.0030$). Still, despite being distinct, their relative proximity might be the reason why speech entrainment was only marginally higher than lip entrainment in the right auditory cortex. Indeed, because of issues inherent to reconstructing brain signals based on extracranial signals (known as source leakage), lip entrainment estimated at the auditory cortex was artificially enhanced by the source in the angular gyrus. This leads us to conclude that entrainment in bilateral auditory cortices occurred with unheard speech rather than with seen lip movements. As further support for this claim, speech entrainment was still significant bilaterally in auditory cortices after partialing out lip movements (mouth opening and width; Fig. 7*D*). In the right hemisphere, it peaked 2.2 mm away from sources observed without partialing out lip movements. In the left hemisphere, the peak in the partial coherence map was displaced toward the middle temporal gyrus (MNI coordinates: −64, −21, −9). Although it did not peak in the left auditory cortex, the source distribution of the partial coherence was clearly pulled toward that brain region.

### Below-1-Hz entrainment to absent speech is not explained by modulation of auditory activity by edges in the visual stimulus

Speech entrainment did not differ significantly from entrainment to the global visual change signal at the coordinates of bilateral auditory sources identified in audio-only (left: $t_{(27)} = 1.17$, $p = 0.25$; right: $t_{(27)} = 1.10$, $p = 0.28$; Fig. 7*A*). However, entrainment to the global visual change signal at ~0.5 Hz was significant only in the posterior part of the right superior temporal gyrus (MNI coordinates: 62, −32, 21), with a visual-change-to-brain delay of 149 ± 33 ms (Fig. 7*E*). Most importantly, speech entrainment corrected for the global visual change signal still peaked and was significant in three left hemisphere sources that were <2.5 mm away from those of uncorrected speech entrainment (Fig. 7*F*). Corrected speech entrainment in the right hemisphere peaked 1 mm away from the right auditory source of uncorrected speech entrainment and was only marginally significant ($p = 0.085$). In sum, global changes in the visual stimulus modulated oscillatory brain activity at ~0.5 Hz in the right posterior superior temporal gyrus, but such modulation did not mediate the entrainment to absent speech.

Altogether, our results support the view that auditory speech envelope is synthesized through lip-reading.

### Entrainment to absent speech at other frequencies

At 1–3 Hz, there was significant entrainment to the absent speech in visual-only but not in control-visual-only (Table 1). Significant entrainment to absent speech in visual-only peaked in the posterior part of the left inferior temporal gyrus, and in the central part of the middle temporal gyrus (Table 1).

Entrainment in the posterior part of the left inferior temporal gyrus was specific to seeing the speaker's face (comparison visual-only vs control-visual-only: $t_{(27)} = 2.72$, $p = 0.011$) but did not entail a synthesis process because speech entrainment at this location was not significantly different from lip entrainment ($t_{(27)} = 1.30$, $p = 0.20$) and it did not reach significance after partialing out mouth movements (Table 1).

Entrainment in the central part of the middle temporal gyrus was not specific to seeing the speaker's face (comparison visual-only vs control-visual-only: $t_{(27)} = 1.48$, $p = 0.15$) and did not entail a synthesis process because speech entrainment at this location was not significantly different from lip entrainment ($t_{(27)} = -0.10$, $p = 0.92$) despite surviving partialing out of mouth movements.

At 2–5 Hz, there was no significant entrainment to the absent speech in visual-only nor in control-visual-only.

At 4–8 Hz, there was significant entrainment to the absent speech in video-only and control-video-only, but only in occipital areas, and it vanished after partialing out the contribution of lip movements.

### Entrainment to lip movements

Lip entrainment at 1–3 Hz, 2–5 Hz, and 4–8 Hz trivially occurred in occipital cortices in video-only and control-video-only (Table 1). Figure 8 illustrates entrainment at 2–5 Hz, which we planned to focus on based on previous reports (Park et al., 2016; Giordano et al., 2017). Brain responses associated with lip entrainment at 2–5 Hz peaked with a delay of 115 ± 8 ms (first source) and 159 ± 8 ms (second source).

Our data do not suggest the presence of entrainment to unseen lip movements in visual cortices in audio-only. Indeed, in that condition, significant lip entrainment at 0.5 Hz occurred only in auditory cortices, and disappeared when we partialed out entrainment to the auditory speech envelope. No significant lip entrainment in this condition was found at any of the other tested frequency ranges: 1–3, 2–5, and 4–8 Hz.

### Delays between auditory and visual speech

Time-efficient synthesis of the auditory speech envelope might rely on the visual-to-auditory lag inherent to natural speech. Indeed, in our audio-visual stimuli, the ~0.5 Hz auditory speech envelope peaked 87 ± 9 ms after the ~0.5 Hz mouth-opening time course (Fig. 9, left). But our results indicate that in visual-only, visual activity entrains to 2–5 Hz mouth movements, whereas auditory activity entrains to an ~0.5 Hz absent auditory speech envelope. The simplest way to connect these oscillations is through phase–amplitude coupling, whereby the amplitude of 2–5 Hz visual activity modulates the phase of ~0.5 Hz auditory activity. Accordingly, we also estimated the delay from the envelope of 2–5 Hz mouth opening time course to ~0.5 Hz auditory speech envelope, and found it was 170 ± 7 ms (Fig. 9, middle).

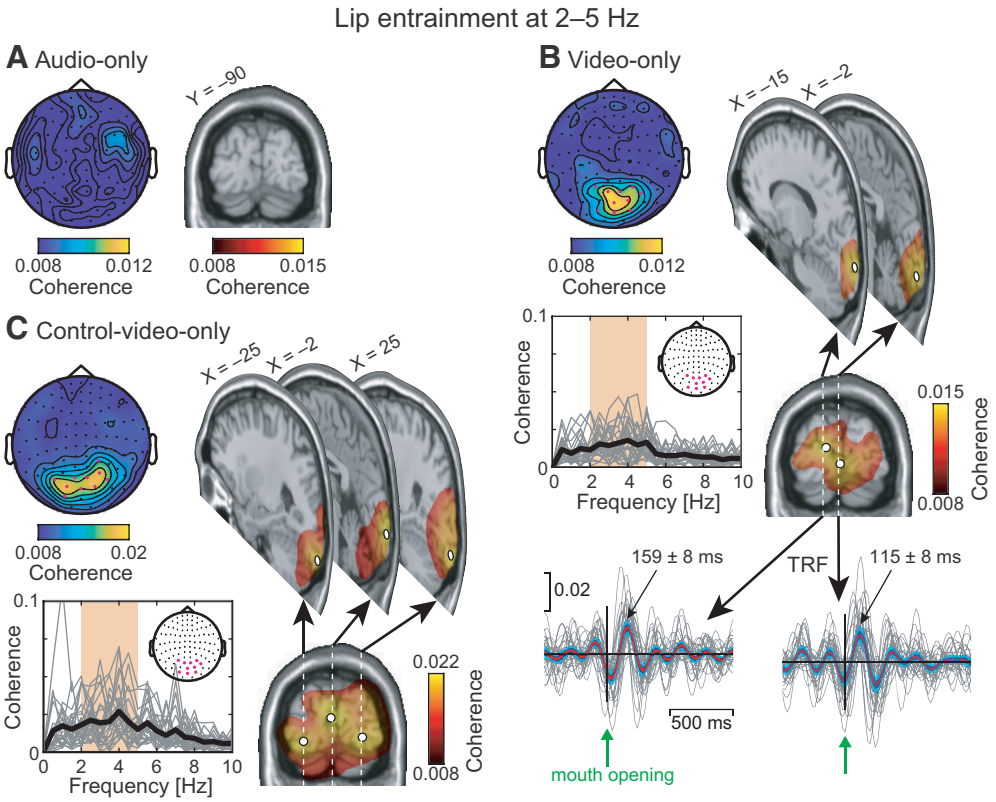


**Figure 8.** Lip entrainment at 2–5 Hz in audio-only (***A***), video-only (***B***), and control-video-only (***C***). Lip entrainment is presented both in the sensor space and on the brain in all conditions (audio-only, video-only, control-video-only). In brain maps, significant coherence values at MNI coordinates $Y < -70$ mm were projected orthogonally onto the coronal slice of coordinates $|Y| = 90$ mm. Locations of peak coherence are marked with white discs. Note that coherence was not significant in audio-only. Additional parasagittal maps are presented for all significant peaks of coherence. In these maps, the orthogonal projection was performed for significant coherence values at *Y*-coordinates <5 mm away from the selected slice *Y*-coordinate. The figure also presents a spectral distribution of coherence at a selection of 10 sensors of maximum 2–5 Hz coherence (magenta) in video-only and control-video-only. Gray traces represent individual subject's spectra at the sensor of maximum 2–5 Hz coherence within the preselection, and the thick black trace is their group average. Finally, TRFs to mouth opening are presented for the two significant sources of peak entrainment to mouth opening in video-only.

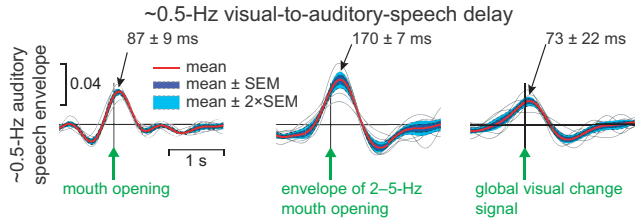


**Figure 9.** Visual-to-auditory-speech delays at ~0.5 Hz. TRF of auditory speech envelope filtered through 0.2–1.5 Hz associated with the time course of mouth opening (left), 2–5 Hz envelope of mouth opening (middle), and global visual changes in video stimuli. There is one gray trace per video (8 in total), and thick red traces are the average across them all.

Also important is the interplay between global changes in the visual stimulus (mainly driven by head, eyebrow, and jaw movements) and auditory speech envelope. This is because global visual changes could in principle modulate auditory cortical activity and hence mediate entrainment to absent speech. And indeed, in our audio-visual stimuli, the ~0.5 Hz auditory speech envelope peaked 73 ± 22 ms after the ~0.5 Hz global visual change signal (Fig. 9, right), meaning that low-level visual changes can cue slow changes in speech envelope (indicating phrase/sentence boundaries). However, the global visual change signal and the auditory speech envelope were only weakly coupled at ~0.5 Hz (mean ± SD. coherence across the 8 video stimuli: 0.051 ± 0.022) and in the other frequency ranges we explored. For a comparison, this degree of coupling was significantly lower than that between mouth opening and the auditory speech envelope ($t_{(7)} = 5.63$, $p = 0.0008$; paired *t* test on the coherence values for the 8 videos). In other words, lip movements provide more information about speech envelope than global changes in the visual stimulus, and similar temporal lead on auditory speech envelope (Fig. 9). This further supports the view that auditory cortical entrainment to silent speech results from a fast synthesis process driven by lip-reading rather than from modulation of auditory activity driven by the identification of low-level cross-sensory correspondences.

## Discussion

We have demonstrated that the brain synthesizes the slow (<1 Hz) temporal dynamics of unheard speech from lip-reading. Specifically, watching silent lip-read videos without prior knowledge of what the speaker is saying leaves a trace of the auditory speech envelope in auditory cortices that closely resembles that left by the actual speech sound.

### Entrainment to unheard speech in auditory cortices

Our most striking finding was that lip-reading induced entrainment in auditory cortices to the absent auditory speech at frequencies <1 Hz. This entrainment (1) was specific to lip-reading, (2) was not a consequence of the general temporal characteristics of auditory speech, (3) was not a mere byproduct of entrainment to lip movements, and (4) was not mediated by low-level changes in the visual stimulus (at least in the left hemisphere). Instead, this genuine entrainment is similar to the entrainment induced

by actual auditory speech: both are rooted in bilateral auditory cortices and are characterized by similar time courses, though with an additional delay of ~200 ms in the right hemisphere. This suggests the existence of a time-efficient synthesis mechanism that maps facial articulatory mouth gestures onto corresponding speech sound features. Such a mechanism would likely leverage the natural visual-to-auditory speech delay (90–170 ms) and could be explained by visually-driven predictive coding (Friston and Kiebel, 2009). Likewise, auditory-driven predictive coding could account for the short (<50 ms) latencies observed here in audio-only (Park et al., 2015).

Importantly, such auditory entrainment is unlikely to be driven by auditory imagery. Auditory imagery reflects perceptual auditory processing not triggered by external auditory stimulation (Nanay, 2018). In principle, observation of lip movements could lead to auditory imagery of related or unrelated speech or non-speech sounds. Clearly, auditory imagery of the actual speech sounds was never an option because participants were not professional lip-readers and were not cued about speech content. Furthermore, our results demonstrate that the auditory entrainment we observed cannot be linked to auditory imagery of unrelated sounds either because it was stronger for the corresponding but absent sound than for either seen lip movements or unrelated speech. Accordingly, the fast synthesis hypothesis we have suggested seems to be the most likely interpretation of the observed entrainment.

The synthesis mechanism we have uncovered is likely grounded in the fact that lip-read information is coupled to the auditory signal in space and time (Munhall and Vatikiotis-Bateson, 2004; Chandrasekaran et al., 2009). In addition, the phonetic identity of each phoneme is supported by sound as well by the configuration of the lips. Even young infants are sensitive to this type of correspondence (Kuhl and Meltzoff, 1982), and phonetic integration continues to develop into adulthood, where the first traces of speech-specific phonetic integration are observed within ~250 ms after sound onset (Stekelenburg and Vroomen, 2012; Baart et al., 2014). Presumably, the tight audiovisual coupling in speech lies at the foundation of lip-read-induced entrainment to absent auditory speech in the brain, and there is indeed much evidence for entrainment to auditory speech at phrase and syllable rates (Luo and Poeppel, 2007; Bourguignon et al., 2013; Gross et al., 2013; Peelle et al., 2013; Molinaro et al., 2016; Vander Ghinst et al., 2016; Meyer et al., 2017; Meyer and Gumbert, 2018).

Frequencies <1 Hz match with phrasal, stress, and sentential rhythmicity. Accordingly, corresponding entrainment to heard speech sounds has been hypothesized to subserve parsing or chunking of phrases and sentences (Ding et al., 2016; Meyer et al., 2017), or to help align neural excitability with syntactic information to optimize language comprehension (Meyer and Gumbert, 2018). Hence, our data suggest that such entrainment/alignment can be obtained through lip-reading, thereby facilitating speech chunking, parsing, and extraction of syntactic information.

As 4–8-Hz frequencies match with syllable rate, corresponding entrainment has been hypothesized to reflect parsing or chunking of syllables. Supporting this view, 4–8 Hz entrainment is enhanced when listening to intelligible speech compared with non-intelligible speech (Ahissar et al., 2001; Luo and Poeppel, 2007; Peelle et al., 2013). However, we did not observe such entrainment during silent lip-reading, which may suggest that the brain does not synthesize the detailed phonology of unfamiliar silent syllabic structures based on lip-read information only. After all, lip-reading is a very difficult task, even for professional lip-readers (Chung et al., 2017). This is because different phonemes correspond to very similar lip configurations (e.g., /ba/, /pa/, and /ma/). However, when the auditory signal is known, this ambiguity in the mapping between lip-reading and the corresponding phonemes disappears. Indeed, it has been suggested that lip-reading can induce entrainment in auditory cortices at frequencies >1 Hz when participants are aware of the content of the visual-only speech stimuli (Crosse et al., 2015).

### Entrainment to lip movements

During silent lip-reading, activity in early visual cortices entrained to lip movements mainly at frequencies >1 Hz, in line with previous studies (Park et al., 2016; Giordano et al., 2017). Such occipital lip entrainment was reported to be modulated by audio-visual congruence (Park et al., 2016). This is probably the first necessary step for the brain to synthesize features of the absent auditory speech. Our results suggest that corresponding signals are forwarded to the right angular gyrus (Hauswald et al., 2018).

The right angular gyrus was the dominant source of lip entrainment at frequencies <1 Hz. It is the convergence area for the dorsal and ventral visual streams and is specialized for processing visual biological motion (Perrett et al., 1989; Allison et al., 2000; Puce and Perrett, 2003; Marty et al., 2015). The right angular gyrus, or more precisely an area close to it termed the temporal visual speech area (Bernstein et al., 2011; Bernstein and Liebenthal, 2014), activates during lip-reading (Calvert et al., 1997; Allison et al., 2000; Campbell et al., 2001) and observation of mouth movements (Puce et al., 1998). It has also been suggested that it maps visual input onto linguistic representation during reading (Démonet et al., 1992), and lip-reading (Hauswald et al., 2018). Our results shed light on the oscillatory dynamics underpinning such mapping during lip-reading: based on visual input at dominant lip movement frequencies (>1 Hz), the angular gyrus presumably extracts features of lip movements <1 Hz, which can then serve as an intermediate step to synthesize speech sound features. Given the short lip-to-brain delay observed in this brain area (~40 ms), such extraction might rely on the prediction of mouth movements.

### Entrainment to unheard speech in visual cortices

Previous studies that have examined the brain dynamics underlying lip-reading of silent connected visual speech have essentially focused on visuo-phonological mapping in occipital cortices (O'Sullivan et al., 2016; Lazard and Giraud, 2017; Hauswald et al., 2018). For example, it was shown that occipital 0.3–15-Hz EEG signals are better predicted by a combination of motion changes, visual speech features and the unheard auditory speech envelope than by motion changes alone (O'Sullivan et al., 2016). Also, visual activity has been reported to entrain more to absent speech at 4–7 Hz when a video is played forward rather than backward (Hauswald et al., 2018). Importantly, this effect was not driven by entrainment to lip movements because lip entrainment was similar for videos played forwards and backwards. Instead, it came with increased top-down drive from left sensorimotor cortices to visual cortices, indicating that visuo-phonological mapping had already taken place in early visual cortices through top-down mechanisms (O'Sullivan et al., 2016; Hauswald et al., 2018). Our study complements these results by showing that auditory cortices also entrain to unheard speech, but at frequencies <1 Hz, probably based on earlier processes taking place in the occipital regions and the right angular gyrus.

### Limitations and future perspectives

We did not collect behavioral data from our participants. Further studies should clarify how the synthesis mechanism we have uncovered relates to individual lip-reading abilities, or susceptibility to the McGurk effect.

It also remains to be clarified what features of speech are synthesized, and under which circumstances auditory cortices can entrain to absent speech at higher frequencies (especially 4–8 Hz).

Finally, it will be important to specify which features of the articulatory mouth gestures lead to <1 Hz auditory entrainment to absent speech. This would require visual control conditions in which, for example, lip movements are shown in isolation, or replaced by point-light stimuli.

### Conclusion

Our results demonstrate that the brain can quickly synthesize a representation of coarse-grained auditory speech features in early auditory cortices and shed light on the underlying oscillatory dynamics. Seeing lip movements first modulates neuronal activity in early visual cortices at frequencies that match articulatory lip movements (>1 Hz). Based on this activity, the right angular gyrus, putatively the temporal visual speech area, extracts and possibly predicts the slower features of lip movements. Finally, these slower lip movement dynamics are mapped onto their corresponding speech sound features and this information is fed to auditory cortices. Receiving this information likely facilitates speech parsing, in line with the hypothesized role of entrainment to heard speech at frequencies <1 Hz.

## References


Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A 98:13367–13372.

Allison T, Puce A, McCarthy G (2000) Social perception from visual cues: role of the STS region. Trends Cogn Sci 4:267–278.

Ashburner J, Friston KJ (1999) Nonlinear spatial normalization using basis functions. Hum Brain Mapp 7:254–266.

Ashburner J, Neelin P, Collins DL, Evans A, Friston K (1997) Incorporating prior knowledge into image registration. Neuroimage 6:344–352.

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

Bernstein LE, Auer ET Jr, Moore JK, Ponton CW, Don M, Singh M (2002) Visual speech perception without primary auditory cortex activation. Neuroreport 13:311–315.

Bernstein LE, Jiang J, Pantazis D, Lu ZL, Joshi A (2011) Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. Hum Brain Mapp 32:1660–1676.

Bernstein LE, Liebenthal E (2014) Neural pathways for visual speech perception. Front Neurosci 8:386.

Bortel R, Sovka P (2014) Approximation of the null distribution of the multiple coherence estimated with segment overlapping. Signal Process 96:310–314.

Bourguignon M, De Tiège X, de Beeck MO, Ligot N, Paquier P, Van Bogaert P, Goldman S, Hari R, Jousmäki V (2013) The pace of prosodic phrasing couples the listener's cortex to the reader's voice. Hum Brain Mapp 34:314–326.

Bourguignon M, Piitulainen H, De Tiège X, Jousmäki V, Hari R (2015) Corticokinematic coherence mainly reflects movement-induced proprioceptive feedback. Neuroimage 106:382–390.

Bourguignon M, Piitulainen H, Smeds E, Zhou G, Jousmäki V, Hari R (2017) MEG insight into the spectral dynamics underlying steady isometric muscle contraction. J Neurosci 37:10421–10437.

Bourguignon M, Molinaro N, Wens V (2018) Contrasting functional imaging parametric maps: the mislocation problem and alternative solutions. Neuroimage 169:200–211.

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596.

Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS (2001) Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). Brain Res Cogn Brain Res 12:233–243.

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

Chung JS, Senior A, Vinyals O, Zisserman A (2017) Lip reading sentences in the wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, July.

Collins JJ, Fanciulli M, Hohlfeld RG, Finch DC, Sandri G v. H, Shtatland ES (1992) A random number generator based on the logit transform of the logistic variable. Comput Phys 6:630.

Crosse MJ, ElShafei HA, Foxe JJ, Lalor EC (2015) Investigating the temporal dynamics of auditory cortical activation to silent lipreading. 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER). Montpellier, France, April.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604.

Démonet JF, Chollet F, Ramsay S, Cardebat D, Nespoulous JL, Wise R, Rascol A, Frackowiak R (1992) The anatomy of phonological and semantic processing in normal subjects. Brain 115:1753–1768.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164.

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. New York: Chapman and Hall.

Eveno N, Caplier A, Coulon PY (2004) Accurate and quasi-automatic lip tracking. IEEE Trans Circuits Syst Video Technol 14:706–715.

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364:1211–1221.

Giordano BL, Ince RAA, Gross J, Schyns PG, Panzeri S, Kayser C (2017) Contributions of local speech encoding and functional connectivity to audio-visual speech perception. eLife 6:e24763.

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

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:e1001752.

Halliday D, Rosenberg JR, Amjad AM, Breeze P, Conway BA, Farmer SF (1995) A framework for the analysis of mixed time series/point process data: theory and application to the study of physiological tremor, single motor unit discharges and electromyograms. Prog Biophys Mol Biol 64:237–278.

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 28:1453–1459.e3.

Hillebrand A, Barnes GR (2005) Beamformer analysis of MEG data. Int Rev Neurobiol 68:149–171.

Hyvärinen A, Karhunen J, Oja E (2004) Independent component analysis. Hoboken NJ: Wiley.

Kaplan E, Jesse A (2019) Fixating the eyes of a speaker provides sufficient visual information to modulate early auditory processing. Biol Psychol 146:107724.

Kapnoula EC, Packard S, Gupta P, McMurray B (2015) Immediate lexical integration of novel word forms. Cognition 134:85–99.

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

Kendall MG, Stuart A (1968) The advanced theory of statistics. Statistician 18:163.

Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. Science 218:1138–1141.

Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31:189–193.

Lazard DS, Giraud AL (2017) Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome. Nat Commun 8:14872.

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54:1001–1010.

Marty B, Bourguignon M, Jousmäki V, Wens V, Op de Beeck M, Van Bogaert P, Goldman S, Hari R, De Tiège X (2015) Cortical kinematic processing of executed and observed goal-directed hand actions. Neuroimage 119:221–228.

McMurray B, Tanenhaus MK, Aslin RN (2002) Gradient effects of within-category phonetic variation on lexical access. Cognition 86:B33–B42.

Meyer L, Gumbert M (2018) Synchronization of electrophysiological responses with speech benefits syntactic information processing. J Cogn Neurosci 30:1066–1074.

Meyer L, Henry MJ, Gaston P, Schmuck N, Friederici AD (2017) Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cereb Cortex 27:4293–4302.

Molinaro N, Lizarazu M, Lallier M, Bourguignon M, Carreiras M (2016) Out-of-synchrony speech entrainment in developmental dyslexia. Hum Brain Mapp 37:2767–2783.

Munhall KG, Vatikiotis-Bateson E (2004) Spatial and temporal constraints on audiovisual speech perception. In: The handbook of multisensory processes (Calvert GA, Spence C, Stein BE, eds), pp 177–188. Cambridge, MA: MIT.

Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E (2004) Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol Sci 15:133–137.

Nanay B (2018) Multimodal mental imagery. Cortex 105:125–134.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

O'Sullivan AE, Crosse MJ, Di Liberto GM, Lalor EC (2016) Visual cortical entrainment to motion and categorical speech features during silent lipreading. Front Hum Neurosci 10:679.

Paré M, Richler RC, ten Hove M, Munhall KG (2003) Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. Percept Psychophys 65:553–567.

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

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

Paulesu E, Perani D, Blasi V, Silani G, Borghese NA, De Giovanni U, Sensolo S, Fazio F (2003) A functional-anatomical model for lipreading. J Neurophysiol 90:2005–2013.

Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23:1378–1387.

Pekkola J, Ojanen V, Autti T, Jääskeläinen IP, Möttönen R, Tarkiainen A, Sams M (2005) Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport 16:125–128.

Perrett DI, Harries MH, Bevan R, Thomas S, Benson PJ, Mistlin AJ, Chitty AJ, Hietanen JK, Ortega JE (1989) Frameworks of analysis for the neural representation of animate objects and actions. J Exp Biol 146:87–113.

Puce A, Perrett D (2003) Electrophysiology and brain imaging of biological motion. Philos Trans R Soc Lond B Biol Sci 358:435–445.

Puce A, Allison T, Bentin S, Gore JC, McCarthy G (1998) Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 18:2188–2199.

Reuter M, Schmansky NJ, Rosas HD, Fischl B (2012) Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61:1402–1418.

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neuronal oscillations and visual amplification of speech. Trends Cogn Sci 12:106–113.

Stekelenburg JJ, Vroomen J (2012) Electrophysiological evidence for a multisensory speech-specific mode of perception. Neuropsychologia 50:1425–1431.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215.

Taulu S, Simola J (2006) Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys Med Biol 51:1759–1768.

Taulu S, Simola J, Kajola M (2005) Applications of the signal space separation method. IEEE Trans Signal Process 53:3359–3372.

Van Veen BD, 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.

Vander Ghinst M, Bourguignon M, Op de Beeck M, Wens V, Marty B, Hassid S, Choufani G, Jousmäki V, Hari R, Van Bogaert P, Goldman S, De Tiège X (2016) Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. J Neurosci 36:1596–1606.

Vander Ghinst M, Bourguignon M, Niesen M, Wens V, Hassid S, Choufani G, Jousmäki V, Hari R, Goldman S, De Tiège X (2019) Cortical tracking of speech-in-noise develops from childhood to adulthood. J Neurosci 39:2938–2950.

Vatikiotis-Bateson E, Eigsti IM, Yano S, Munhall KG (1998) Eye movement of perceivers during audiovisualspeech perception. Percept Psychophys 60:926–940.

Vigário R, Särelä J, Jousmäki V, Hämäläinen M, Oja E (2000) Independent component approach to the analysis of EEG and MEG recordings. IEEE Trans Biomed Eng 47:589–593.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." Neuron 77:980–991.