Title: Cortical tracking of speech reveals top-down reconstructive processes

Authors: Sankar Mukherjee\textsuperscript{1,\textsuperscript{x}}, Alice Tomassini\textsuperscript{1,\textsuperscript{x}}, Leonardo Badino\textsuperscript{1}, Aldo Pastore\textsuperscript{1}, Luciano Fadiga\textsuperscript{1,2}, Alessandro D’Ausilio\textsuperscript{1,2}

Affiliations:

\textsuperscript{1}Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy;
\textsuperscript{2}Section of Human Physiology, University of Ferrara, Ferrara, Italy;
\textsuperscript{x}Equal contribution.

Corresponding Author
Sankar Mukherjee
IIT - Istituto Italiano di Tecnologia
Via Fossato di Mortara, 17/19; 44121 – Ferrara
Email: sankar1535@gmail.com

Alessandro D’Ausilio
Human Physiology @ Università di Ferrara
IIT - Istituto Italiano di Tecnologia
Via Fossato di Mortara, 17/19; 44121 – Ferrara
Email: Alessandro.dausilio@iit.it
Abstract

Cortical entrainment to the (quasi-) rhythmic components of speech seems to play an important role in speech comprehension. It has been suggested that neural entrainment may reflect top-down temporal predictions of sensory signals. Key properties of a predictive model are its anticipatory nature and its ability to reconstruct missing information. Here we put both these two properties to experimental test. We acoustically presented sentences and measured cortical entrainment to both acoustic speech envelope and lips kinematics acquired from the speaker but not visible to the participants. We then analyzed speech-brain and lips-brain coherence at multiple negative and positive lags. Besides the well-known cortical entrainment to the acoustic speech envelope, we found significant entrainment in the delta range to the (latent) lips kinematics. Most interestingly, the two entrainment phenomena were temporally dissociated. While entrainment to the acoustic speech peaked around +0.3 s lag (i.e., when EEG followed speech by 0.3 s), entrainment to the lips was significantly anticipated and peaked around 0-0.1 s lag (i.e., when EEG was virtually synchronous to the putative lips movement). Our results demonstrate that neural entrainment during speech listening involves the anticipatory reconstruction of missing information related to lips movement production, indicating its fundamentally predictive nature and thus supporting analysis by synthesis models.
Introduction

Verbal interaction is a remarkable part of human behaviour and our brains are tuned to decode speech from the senses. Neural oscillations are thought to play an essential role in shaping speech perception over time (Giraud, Poeppel, 2012; Kösem et al., 2018; Meyer, 2018; Zoefel et al., 2018). It has been shown that neural oscillatory activities in auditory areas are coupled to the rhythmic properties of speech (Ahissar et al., 2001; Luo and Poeppel, 2007; Gross et al., 2013; Ding et al., 2014). The strength of this oscillatory coupling positively scales with speech intelligibility (Ghitza, 2012; Peelle et. al. 2013; Ding, Simon 2013; Kayser et al., 2015; Riecke et al., 2018) and is associated with speech comprehension performance (Ahissar et al., 2001; Luo, Poeppel, 2007; Peelle et. al. 2013). Investigations on the cocktail party effect (Cherry, 1953; Ding, Simon, 2012) show that selective attention translates into increased neural entrainment to the attended acoustic stream (Golumbic et al., 2012; Kerlin et al., 2010; Golumbic et al., 2013a; OSullivan et al., 2014; Vander Ghinst et al., 2016), and this is further boosted in the presence of congruent visual cues (Crosse et al., 2015). Indeed, the oscillatory dynamics within the visual cortex is also coupled to the periodicity of visual information related to lips movements (Park et al., 2016). As for speech-brain coupling, cortical visual coupling is affected by top-down modulation (Park et al., 2015; 2016; 2018a) and enhanced when the acoustic signal is degraded (Giordano et al., 2017). Not surprisingly, speech comprehension mostly benefits from visual cues in suboptimal listening conditions (Sumby, Pollack, 1954; Schroeder et al., 2008; Golumbic et al., 2013b).

As outlined above, the functional relevance of speech-brain entrainment has been associated with the attentional-driven top-down selection of multimodal sensory cues over time (Park et al., 2015) which eventually aid speech comprehension. Importantly, selective attention exploits knowledge about the periodicity of speech (Poeppel, 2003) and the statistical associations between audio-visual speech signals (Chandrasekaran et al., 2009; Park et al., 2016; But see Schwartz, Savariaux, 2014). It is not clear, however, whether this top-down influence relies on predictive modeling and is therefore capable of reconstructing missing speaker-related information.

Here we designed an EEG experiment with participants engaged in a sentence listening task followed by a delayed word recognition task. After the listening phase, participants had to indicate by a button press which...
word between two options (displayed only when the sentence was completed) was spoken in the previous sentence. The sentences (200 for each subject) were derived from a dataset containing both acoustic data and lips kinematics recorded via electromagnetic articulography (EMA). We quantified brain entrainment to speech envelope and lips kinematics as commonly done in unimodal (Destoky et al., 2019; Luo and Poeppel, 2007; Peelle et al., 2013; Bourguignon et al., 2013) and audio-visual speech studies (Park et al., 2016; Park et al., 2015). However, differently from these previous studies, no lips-related information was ever visually shown to our participants nor subjects were asked to recall speech-related visual cues. One key prediction of our experiment was that if neural entrainment describes a fundamentally top-down synthetic process, we should find significant entrainment to lips kinematics beyond that to acoustic speech. Furthermore, to be considered predictive (Park et al., 2018b), we expect a temporal dissociation between neural entrainment to attended (speech envelope) and reconstructed (lips kinematics) cues, with the latter preceding the former.

Results

Participants performed the task with high accuracy, selecting the correct word in nearly 100% of the trials (98.4 ± 1 %; MEAN±SD). This demonstrates that they were paying attention during the listening phase and that the sentences were perfectly intelligible. Trials (n=7.44± 4.29; MEAN±SD) in which participants gave an incorrect response were discarded from the analysis.

In line with previous work (Ahissar et al., 2001; Peelle et al., 2013; Golumbic et al., 2012; Park et al., 2016), both speech envelope (hereinafter referred to as speech) and corresponding lips kinematics (hereinafter referred to as lips) are dominated by low-frequency components, as indicated by their power spectra showing distinct peaks in the delta (1-4 Hz) and theta (4-8 Hz) range (Figure 1c).

Based on the prominent low-frequency content of the speech and lips signals and on previous evidence (Ding and Simon, 2013; Gross et al., 2013; Kayser et al., 2015), we restricted our analyses to three low-frequency bands: delta (1-4 Hz), theta (4-8 Hz) and alpha (8-12 Hz).
First, we computed speech-brain coherence for temporally aligned signals (zero delay) as well as for EEG signals that were time-shifted (relative to speech signals) by different temporal delays: from -0.5 s (EEG preceding speech by 0.5 s) up to +0.5 s (EEG following speech by 0.5 s) in steps of 50 ms (see Figure 1b). Brain activities are coherent with the speech envelope at all tested frequency bands (cluster-based statistics against...
trial-shuffled surrogate data; p<0.025; Figure 2a & 2b), although to a different degree (delta=0.038±0.002, theta=0.035±0.001, alpha=0.032±0.0008, mean ± s.e.; F2,48=6.758, p=0.003; one-way ANOVA for repeated measures with frequency-band [delta, theta, alpha] as within-subject factor; data collapsed across channels and delays), with coherence for the delta-band being significantly stronger than that for the alpha-band (p<0.001; paired-sample t-test, Bonferroni corrected). Coherence shows comparable spatial distribution for all frequency bands, with maximal concentration over bilateral fronto-central sites (see Figure 2a; Peelle et al., 2013; Vander Ghinst et al., 2016; Molinaro, Lizarazu, 2018). Most interestingly, irrespective of the frequency band, speech-brain coherence is weak and non-significant at 0-s delay and steadily increases at positive delays, reaching maximal values when the EEG followed the acoustic speech by ~0.25 s (i.e. at +0.2/0.3-s delays; Crosse et al., 2015; O'Sullivan et al., 2014; Figure 2a).

Next, we examined whether brain activity was also coherent with the articulatory signal – the lips – and found statistically significant coherence in the delta band (p=0.003 at 0-s delay; Figure 2a & 2b). Although sharing a similar fronto-central topography, lips-brain coherence is temporally dissociated from speech-brain coherence, being maximal at earlier delays (between -0.05 s and +0.15 s; see Figure 2a). This suggests that neural entrainment to the (reconstructed) lip movements preceded in time the entrainment to the speech sounds envelope.
Figure 2. (a) Topographies of speech- (top) and lips- (bottom) brain coherence in the delta (1-4 Hz), theta (4-8 Hz) and alpha (8-12 Hz) bands are shown for -0.5-s delay (leftmost graph) and from 0-s to +0.5-s delays, in steps of 0.1 s. Black dots represent significant channels (p<0.025) according to (group-level) cluster-based permutation tests against surrogate data (generated by trial-shuffling). (b) Speech- (black) and lips- (grey) brain coherence averaged across channels is shown at 0-s (left) and at +0.3-s delay (right) as a function of
frequency band. Asterisks indicate significant (p<0.025) frequency-bands for speech- (black) and lips- (grey) brain coherence, according to cluster-based permutation statistics. Error bars represent standard error of the mean.

We statistically evaluated the temporal dissociation by comparing speech-brain and lips-brain coherence at the different delays (from -0.5 to +0.5 s). Given that lips-brain coherence was found to be statistically significant only in the delta range, this analysis was confined to this frequency band. A group-level cluster-based permutation test yielded two significant clusters, showing comparable fronto-central topography: one negative, indexing lips-brain coherence larger than speech-brain coherence, spanning from -0.05 s to +0.15 s delay (p=0.017 at 0-s delay), and one positive, indexing speech-brain coherence larger than lips-brain coherence, extending from +0.2 to +0.45 s delay (p<0.0001 at +0.3-s delay; see Figure 3a & 3b). This significant temporal dissociation further confirms that delta-band brain activities are anticipatorily coupled to (reconstructed) lips oscillatory dynamics while this coupling is later replaced by entrainment to the ensuing acoustic speech signal.

One possible confound to exclude is that lips-brain coherence emerges as a by-product of the coherence between lips and speech envelope. Two considerations make this unlikely. First, coherence between the speech envelope and the lips is maximal (at all delays) in the theta, not in the delta, range (see Figure 1d). Indeed, theta-band speech-lips coherence might reflect the rate of syllables production and it has been previously reported (Hauswald et al., 2018). Secondly, and most importantly, lips-brain and speech-brain coherence show a different dependence from the temporal alignment between the two signals: the former is stronger when the EEG signal is virtually synchronous to the lips signal associated with the sentence production, while the latter is established at later times, when the EEG follows the acoustic speech signal by a few hundreds of milliseconds.

However, to statistically ascertain the relation between lips- and speech-brain coherence we run a partial coherence analysis, i.e., we re-computed coherence at two critical delays (0 s and +0.3 s) by partialling out the contribution of speech and lips, respectively. As expected, this analysis reveals a certain degree of inter-dependency between the two phenomena, as both speech- and lips-brain coherence values are diminished (at
both delays). However, this analysis crucially confirms the previous pattern of results: at 0 s delay, lips-brain (partial) coherence is significantly stronger than speech-brain (partial) coherence (p=0.019; cluster-based statistics), and the reverse is true at +0.3-s delay, where speech-brain (partial) coherence exceeds lips-brain (partial) coherence (p<0.0001; cluster-based statistics; see Figure 3c).

Figure 3. (a) Topographies of the difference between speech- and lips-brain coherence are shown for delays from -0.15 to +0.45 s (in steps of 0.15 s). Black dots represent significant channels (p<0.025) as resulted by (group-level) cluster-based permutation test. (b) Speech- (black) and lips- (grey) brain coherence averaged across channels is shown as a function of delay (from -0.5 to +0.5 s). Horizontal lines indicate significant (p<0.025) delays according to cluster-based permutation test (black: speech>lips; grey: lips>speech). (c) Bars show speech- (black) and lips- (grey) brain partial coherence averaged across channels at 0-s (top) and +0.3-s (bottom)
delays. Asterisks indicate statistical significance (p<0.025) as resulting from cluster-based permutation statistics. **Right.** Topographies of the difference between speech- and lips-brain partial coherence at 0-s (top) and +0.3-s (bottom) delays. Black dots represent significant channels (p<0.025) according to cluster-based permutation test.

**Discussion**

A large body of research has investigated the coupling of brain activities to speech-related information (Giraud, Poeppel, 2012; Kosem et al., 2018; Meyer, 2018; Zoefel et al., 2018). Recently, it has been reported that frontal brain activities could modulate the strength of neural entrainment to speech in sensory cortices, with a relatively short delay of 50-60 ms (Park et al., 2015). This has been taken as evidence of a top-down attentional modulation exerted upon a fundamentally bottom-up, sensory-driven process (Park et al., 2018). In fact, one important tenet of a top-down processing resides in its anticipatory nature.

However, these recent studies do not clarify whether top-down neural processes can synthetize missing information related to others’ speech production. Demonstrating a reconstructive capability is the prerequisite to consider neural entrainment as instantiating an analysis-by-synthesis computation (Bever, Poeppel, 2010; Crosse et al., 2015). Here we tested at the same time the two defining properties of a top-down predictive process, that is its anticipatory and its reconstructive nature. To this end, we specifically evaluated whether neural activities are coupled to latent information (lips kinematics) and how this coupling evolves in time.

In agreement with previous investigations, we showed significant speech-brain coherence (delta, theta and alpha frequencies) with a positive lag of 100-300 ms (Crosse et al., 2015; O’Sullivan et al., 2014) and a fronto central topography (Peelle et al., 2013; Vander Ghinst et al., 2016; Bourguignon et al., BioRxiv, 2018). Crucially, lips-brain coherence was also significant and had a similar topography, but peaked around 0-100 ms. More importantly, the temporal dynamics of neural entrainment significantly differed between lips and speech, even after partialling out the contribution of the other signal.
Notably, brain entrainment to absent cues was shown without introducing any explicit or implicit request to focus on the visual or the motor description of the acoustically presented sentences. Moreover, behavioral performance was at ceiling, suggesting that only a limited advantage could have been derived from exploiting lips-related (real or reconstructed) information. These two aspects seem to suggest that reconstruction of missing cues is independent from a specific attentional effort and, more importantly, automatically put in place in the absence of a specific behavioral need. Hence, to grasp the functional meaning of this reconstructive process, we should consider the temporal relationship between acoustic and lips-related speech signals as well as their different informational content.

Mouth opening and speech envelope are temporally modulated in the 2–7 Hz frequency range (Chandrasekaran et al. 2009), which overlaps with the timescale of syllable production as well as with the rhythmic fingerprint of auditory cortices (Giraud, Poeppel, 2012). These commonalities are suggestive of a potential functional link between rhythmicity in speech articulatory movements, in their acoustic outcome and in the intrinsic activity of sensory cortices. However, in natural conditions mouth opening is the visible part of a far more complex pattern of phono-articulatory configurations needed to produce speech sounds (Catford 1977). Still, notwithstanding the fact that visual signals are far less informative than acoustic ones (Fisher, 1968), they entrain visual cortex oscillations during silent lip-reading (O’Sullivan et al., 2017; Hauswald et al., 2018; Ozker et al., 2018).

Yet, there seems to be an asymmetric relationship between activities in primary visual and auditory cortices during audio-visual speech perception. The auditory cortex tracks the temporal dynamics of visual speech cues, while the tracking of auditory speech cues by the visual cortex is far less pronounced (Megevand et al., In press). This asymmetry could be explained by the fact that the acoustic signal is in principle sufficient for understanding the message, while the visual one plays a supportive role, which may gain more relevance only in the presence of a degraded acoustic signal (Soto-Faraco et al., 2007).

Although carrying relatively low information, visual cues can still bias acoustic sensory processing if they could be used anticipatorily. Predictive models, in general, envision that prior knowledge biases sensory processing by means of top-down projections (Friston, Frith, 2015). As such, they could potentially convey critical predictive
value for the decoding of the true objects of perception which reside in the acoustic signal (Schroeder et al., 2008; Golumbic et al., 2013b). This putative predictive power of visual cues in audio-visual speech perception has already been described as a reduction in the latency of Event-Related Potentials (ERPs) evoked by speech listening (Van Wassenhove et al., 2005). Manipulation of the amount of information carried by the visual signal was also linked to the magnitude of the anticipatory effects (Van Wassenhove et al., 2005; Stekelenburg, Vroomen, 2007; Arnal et al., 2009).

In the present study, we found evidence that speech listening does not only entail neural tracking of speech envelope but also includes the anticipatory synthesis of missing lips-related cues. The neuro-functional origin of these reconstructive processes may reside in the computations run in the frontal cortex. Indeed, unimodal neural entrainment to speech envelope or mouth movements is driven by neural sources located in the left motor and premotor cortex (Park et al., 2015; Park et al., 2018). A precentral origin of these top-down modulations is also suggested by the fact that corticobulbar excitability is modulated by passive listening to speech (Fadiga et al., 2002; Watkins et al., 2003; D’Ausilio et al., 2014; Schmitz et al., 2018) and that transient perturbation of the activity in premotor and motor areas produces somatotopically organized modulation of speech discrimination performance (D’Ausilio et al., 2009; Mottonen et al., 2009; Sato et al., 2009; D’Ausilio et al., 2012; Bartoli et al., 2015; Murakami et al., 2018).

In conclusion, our data provide evidence that neural entrainment during speech listening is reconstructive and predictive. We speculate that the brain’s ability to anticipate and reconstruct lips movements might rely on the sensorimotor mapping (including acoustic, somatosensory and motor signals) built during the acquisition of speech production (Bruderer et al., 2015; Kuhl et al., 2014).
Materials and Methods

Participants

Twenty-five healthy native Italian speakers participated in the study (16 females; mean age=23.5; age range=20-28 years). All participants were right-handed and had normal self-reported hearing. Participants were all naïve with respect to the aims of the study and were all paid (€10/h) for their participation. The study and experimental procedures were approved by the local ethics committee (Comitato Etico della Provincia di Ferrara). Participants provided written, informed consent after explanation of the task and experimental procedures in accordance with the guidelines of the local ethics committee and the Declaration of Helsinki.

Stimuli

The stimuli were chosen from the Multi-SPeaking-style Articulatory corpus (MSPKA; Canevari et. al. 2015), an Italian dataset of simultaneous recordings of audio and articulatory data (lips, jaw and tongue). Audio was recorded at a sampling rate of 22.5 kHz. Articulators were tracked at a sampling frequency of 400 Hz by means of an electromagnetic articulography system (EMA; NDI Wave, Northern Digital Instruments, Canada; Berry, 2011). In the present study, we used data from two 5-degrees-of-freedom (x, y, z positions, pitch and roll) sensor coils glued on the upper lip (UL) and lower lip (LL). For head movement correction, a 6-degrees-of-freedom sensor coil was fixed on the bridge of a pair of glasses worn by the participants. The EMA data provides a very accurate characterization of lips kinematics and it is commonly used in speech technology research (Savariaux et al., 2017). Here we used articulatory and acoustic data were recorded from one speaker reading news-related sentences and tongue twisters. For this study, we selected 200 sentences from the MSPKA corpus (female speaker). The (acoustic) stimuli were manually checked to ensure there were no missing or noisy data. All stimuli were then normalized to the same average intensity (71 dB). It should be noted that the sentences had different duration (4.36±1.73 s; MEAN±SD), with an average number of phonemes equal to 13.32 ± 3.5 phonemes/s (MEAN±SD) and an average phoneme duration of 0.08±0.02 s (MEAN±SD).
**Experimental Procedure**

Subjects sat in a dimly lit, sound and electrically shielded room in front of a screen and two loudspeakers (distance: ~1 m) and with their left and right hands resting on a response box (RB-840 Response Pad, Cedrus Corporation). The experiment was composed of two blocks with a rest period of 2 min in between. Each block consisted of 100 trials that were structured as follows. Each trial started with the appearance of a white fixation cross at the center of the screen. Subjects were required to maintain fixation throughout the duration of the trial. After a variable time (2-2.5 s), the fixation cross changed color (from white to green) and a randomly selected sentence was acoustically presented from the two loudspeakers. When the sentence was completed, the fixation cross turned back white. After a variable pause (1-1.5 s) the fixation cross was replaced by two words displayed above and below the center of the screen. Participants were asked to indicate with a button press which of the two words was spoken in the previously presented sentence. The above/below word was chosen by pressing the up/down key of the response box with the left and right index finger, respectively. Participants were not prompted to respond as quickly as possible, but they were encouraged to select the correct word. However, they were aware that if they did not respond within 5 s, the next trial was automatically run. Stimuli presentation and response acquisition were controlled with Matlab Psychtoolbox-3.

**EEG recording**

EEG data were recorded continuously during the experiment with a 64-channel system using Ag/AgCl impedance-optimized active electrodes (Brain Products GmbH, Gilching, Germany). Electrodes were placed according to the international 10–20 system and on-line referenced to the right mastoid. Electrooculograms (EOGs) were recorded with four electrodes removed from their original scalp sites (FT9, FT10, PO9, PO10) and placed at the bilateral outer canthi (horizontal eye movements) and below and above the right eye (vertical eye movements, blinks). Electrode impedance was kept <5 kΩ. Data were acquired at a sampling rate of 1000 Hz.

**Preprocessing of EEG, speech and articulatory data**
All the analyses were performed with the MNE-Python software (Gramfort et al., 2013). Continuous EEG data were first bandpass-filtered between 0.5 and 40 Hz (two-pass Butterworth filter, third-order). Data were then epoched in shorter segments corresponding to the duration of each trial and visually inspected for bad channels and/or artifacts in the time domain. To identify and remove artifacts related to participants’ eye movements and muscle activity, we used Independent Component Analysis (ICA). Noisy channels (not included in the ICA) were then interpolated using a distance-weighted nearest-neighbor approach. Finally, data were re-referenced using a common average reference over all electrodes.

The amplitude envelope of the acoustic speech signal was calculated by adapting a previously described method (Smith et. al., 2002). We used the Chimera toolbox and defined eight frequency bands in a range 100–10000 Hz that are equally spaced on the cochlear map. The speech signal was first filtered within those eight frequency bands (two-pass Butterworth filter, fourth-order). Then, we computed the absolute value of the Hilbert transform for each bandpass-filtered signal. The result of this computation was then averaged across frequency bands, yielding the wideband speech envelope.

Lips kinematics was derived from the EMA data (x-y midsagittal coil positions) as the absolute distance between UL and LL in the y-plane, providing a measure of the amount of opening of the lips during speech production. The lips signal was subsequently smoothed by applying an adaptive median filter with a window of range of 10-50 ms and an elliptic low-pass filter (20 Hz cutoff frequency). Both the EEG and speech signals were downsampled to match the lower sampling rate of the lips signal (i.e., 400 Hz). Example traces of the raw audio signal, speech envelope and lips kinematics are shown in Figure 1a.

*Spectral analysis*

We analyzed EEG data recorded during stimulus presentation (i.e., during sentence listening) and corresponding speech and lips data. To exclude the stimulus-evoked components associated with speech onset (in the EEG signal), we discarded the initial 0.5 s of each sentence (see Figure 1b). All data were then cut into 3-s segments.
As a consequence, sentences shorter than 3 s were discarded from the analysis (42 out of 200 sentences) while sentences longer than 6 s contributed with more than one segment to the analysis.

All the analyses were then performed on 2-s segments (the first and last 0.5 s of the original 0.3-s segments were used for the analysis based on time-shifting of the EEG signal, see below). The power spectral density (PSD) of both the speech and lips signals was estimated for frequencies between 0.5 and 12 Hz using the multitaper method (0.5 Hz steps, 2 Hz smoothing; Percival, Walden, 1993).

The cross-spectral density between EEG and speech, EEG and lips as well as speech and lips, was computed on single trials with multitaper frequency transformation. The multitaper method was applied at the following frequency ranges: delta (1-4 Hz, centered at 2.5 Hz; 1.5 Hz smoothing), theta (4-8 Hz, centered at 6 Hz; 2 Hz smoothing) and alpha (8-12 Hz, centered at 10 Hz; 2 Hz smoothing).

Coherence was then computed for each subject, frequency-band (delta, theta and alpha) and EEG channel for temporally aligned signals (0-s delay) as well as by shifting the EEG signal in time by different amounts of delays: from -0.5 s (EEG precedes speech and lips by 0.5 s) up to +0.5 s (EEG follows speech and lips by 0.5 s) in steps of 50 ms (see Figure 1b). Speech-lips coherence was also computed at the same delays by time-shifting the speech signal accordingly (see Figure 1d).

**Statistical analysis**

Statistical analysis at the group level was performed using non-parametric cluster-based permutation tests (Maris, Oostenveld, 2007). First, we evaluated statistically both speech-brain and lips-brain coherence at different delays (from -0.5 to +0.5 s in steps of 50 ms) and for the three frequency bands (delta, theta and alpha). To this aim, we generated surrogate data by breaking the original association between the EEG and the speech/lips signals. In practice, for each subject and channel, we randomly reassigned the EEG segments (used to calculate coherence in the original analysis) to speech/lips signals that corresponded to different stimulus segments compared to the original ones. Note that the randomly reassigned speech/lips signals could still be part...
of the same sentence as the original one (but a different and non-overlapping 2-s segment). This surrogate data does not contain anymore the original temporal dependencies between the EEG and the speech/lips signals and can thus serve as control data. Coherence maps were then calculated for each subject-specific surrogate data in the same way as described above for the original data. Then, for every sample (here defined as [channel, delay]), a dependent-sample two-tailed t value was computed (between the original and surrogate datasets). All samples for which this t value exceeded an a priori decided threshold (uncorrected p < 0.05) were selected and subsequently clustered on the basis of spatial and temporal contiguity. Cluster-level statistics was computed by taking the sum of t-values in each cluster. The cluster yielding the maximum sum was subsequently used for evaluating the difference between the two datasets (with the maximum sum used as test statistic). We permuted the data across the two datasets (swapping the coherence values for the two datasets in a random subset of subjects), and for each random permutation (10000 iterations), we calculated again the test statistics in the same way as previously described for the original data. This procedure generates a surrogate distribution of maximum cluster t-values against which we can evaluate the actual data. The p-value of this test is given by the proportion of random permutations that yields a larger test statistic compared to that computed for the original data.

Besides evaluating both speech-brain and lips-brain coherence against the coherence that can be expected by chance for each pair of signals (as indexed by the surrogate data), we also evaluated whether neural activity is more strongly related to the acoustic speech or to the corresponding lip kinematics at different temporal delays. This analysis was limited to the delta-band (1-4 Hz), as this was the only band where we found significant coherence between EEG and lips (previously described analysis). Group-level statistical analysis was performed again by applying non-parametric cluster-based permutation statistics (described above).

Finally, partial (speech-brain and lips-brain) coherence was computed at two different delays, at 0-s and at +0.3-s delay and for both delays we compared the obtained coherence values for the two signals by means of cluster-based statistics (as described above).
References


Schwartz, J.L. and Savariaux, C., 2014. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLoS Computational Biology, 10(7), pp.e1003743.


