NEUROSCIENCE

The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech

Raphaël Thézé[1], Anne-Lise Giraud[1], Pierre Mégevand[1,2]*

When we see our interlocutor, our brain seamlessly extracts visual cues from their face and processes them along with the sound of their voice, making speech an intrinsically multimodal signal. Visual cues are especially important in noisy environments, when the auditory signal is less reliable. Neuronal oscillations might be involved in the cortical processing of audiovisual speech by selecting which sensory channel contributes more to perception. To test this, we designed computer-generated naturalistic audiovisual speech stimuli where one mismatched phoneme-viseme pair in a key word of sentences created bistable perception. Neurophysiological recordings (high-density scalp and intracranial electroencephalography) revealed that the precise phase angle of theta-band oscillations in posterior temporal and occipital cortex of the right hemisphere was crucial to select whether the auditory or the visual speech cue drove perception. We demonstrate that the phase of cortical oscillations acts as an instrument for sensory selection in audiovisual speech processing.

INTRODUCTION

Speaking entails moving, and watching the speaker brings the listener a wealth of information that complements the message conveyed by the voice. Visual speech cues become particularly valuable for intelligibility when the auditory speech signal is less reliable, for instance, in noisy environments (*1*). Another example of the importance of visual cues on speech processing is exemplified by the McGurk effect, where the mismatching of a phoneme-viseme pair creates notable perceptual illusions (*2*). The neuronal mechanisms responsible for the profound impact of visual cues on speech perception have not yet been elucidated.

The rhythmic nature of speech (*3*) affects neural activity in multiple cortical areas, which oscillate in phase with the auditory inputs (*4*–*7*). The magnitude of this oscillatory tracking correlates with intelligibility, suggesting that oscillations might be instrumental in the cortical processing of heard speech (*8*). Visual speech also entrains oscillations in many areas, especially occipital and posterior temporal cortex (*9*–*12*). Neuronal oscillations represent momentary fluctuations in neuronal excitability (*13*), which means that the phase angle of low-frequency oscillations in sensory cortex at a given moment in time determines the responsiveness of that region to incoming inputs (*14*). The phase of ongoing oscillations in sensory cortex can be reset by cross-modal stimuli from another sensory modality (*15*); visual cues could thus amplify or diminish the responsiveness of speech-processing cortex to incoming speech sounds via this mechanism (*16*). Alternatively, because the influence of cross-modal sensory input on cortical oscillations depends on attention (*17*, *18*), neuronal oscillations could act as a mechanism to select whether the auditory or the visual channel is selected for further cortical processing and dominates perception (*19*).

Here, we establish that cortical oscillations are a crucial mechanism in the processing of audiovisual speech. Building innovative, naturalistic audiovisual speech stimuli with speech synthesis and animated three-dimensional (3D) virtual characters, we inserted a mismatched phoneme-viseme pair into syntactically correct and semantically meaningful sentences. We thus created bistable stimuli where the perception of a key word was driven by either the visual or the auditory cue. Using high-density scalp electroencephalography (EEG) and intracranial EEG (iEEG), we show that the phase of prestimulus theta-band oscillations in the right posterior temporal and occipital cortex determines which sensory channel drives perception. Our findings strongly support cortical oscillations as an instrument of sensory selection in the processing of audiovisual speech.

RESULTS

We designed audiovisual speech stimuli by combining speech-synthesized sentences with virtual characters whose lip movements were animated synchronously to the speech sounds (*20*) (Fig. 1A). We created 10 such sentences, where we could voluntarily mismatch one phoneme-viseme pair in a key word to trigger McGurk effects. In a behavioral experiment, words that started with a /v/ viseme mismatched to a /b/ phoneme were perceived as /v/-leading words in about 60% of trials (Fig. 1B). Thus, this combination of auditory and visual speech cues created bistable perception and was used in further neurophysiological experiments to study the role of cortical oscillations in determining which sensory channel drove perception on a single-trial basis.

For that purpose, we recorded high-density EEG (hdEEG) in 15 healthy participants. Replicating our behavioral experiment, these participants perceived the crucial /b/-phoneme, /v/-viseme stimuli as /v/-leading words in approximately half the trials (median: 53%; interquartile range: 44 to 62%). We compared cortical responses, grouped as a function of subsequent perception, focusing on a 1-s period surrounding the presentation of the mismatched stimulus and on frequency bands from 1 to 13 Hz. First, we established that differences in perception were not associated with any strong differences in oscillatory power, either before or after the mismatched stimulus (movie S1).

[1]Department of Basic Neurosciences, Faculty of Medicine, University of Geneva, 1202 Geneva, Switzerland. [2]Division of Neurology, Department of Clinical Neurosciences, Geneva University Hospitals, 1205 Geneva, Switzerland.
*Corresponding author. Email: pierre.megevand@unige.ch

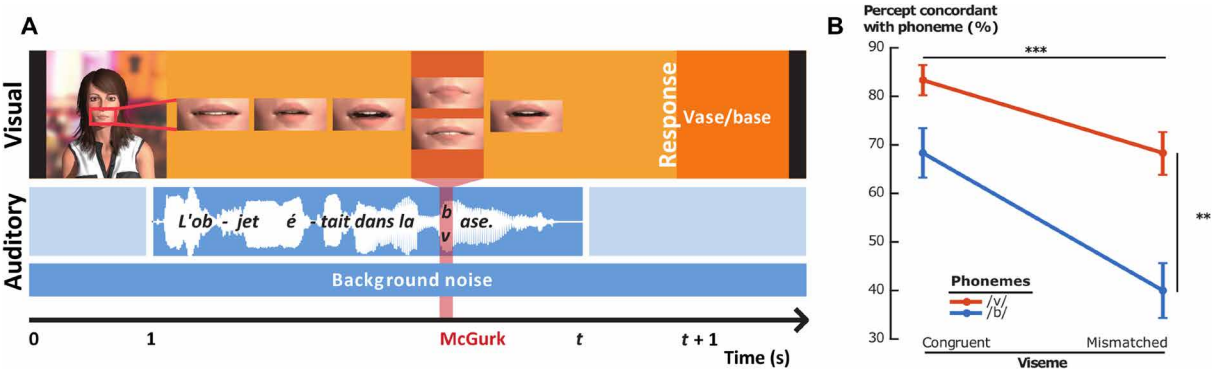

Fig. 1. Experimental design and behavioral results. (**A**) The time course of one trial is illustrated. At trial onset, an animated character appeared on the screen. Throughout the trial, background noise (audio recording of a café's interior) was playing. After 1 s, the character started speaking. One second after the sentence was over, participants had to report whether they perceived one word or another. The French sentence translates to "The object was in the base/mud"; note that the illusion would also be present in English with "base/vase," although the meaning differs. (**B**) When the phoneme-viseme pair of the key word was congruent, the participants' ($N = 24$) perception was concordant with the phoneme (and viseme) on most trials. When the phoneme-viseme pair was mismatched, perception was less often concordant with the phoneme [***repeated-measures analysis of variance (ANOVA), main effect of congruence: $F_{1,23} = 30.26$, $P = 1.4 \times 10^{-5}$]. When the phoneme was /v/, perception was more often concordant with the phoneme than when it was /b/ (**main effect of phoneme: $F_{1,23} = 10.33$, $P = 0.004$). Words that had a /b/ phoneme mismatched with a /v/ viseme were especially prone to be perceived as /v/-leading words (i.e., discordant with the phoneme; interaction: $F_{1,23} = 4.83$, $P = 0.038$), with about 60% of trials where perception was dominated by the visual cue.

To assess whether differences in oscillatory phase were associated with perception, we computed the phase opposition sum (POS) statistic (*21*). This statistic is maximal when instantaneous phase angle is consistent within trials of each group and differs by a half-cycle between groups. We found three periods where groups of electrodes showed a significantly high POS (Fig. 2 and movie S2). Of those, one occurred before the mismatched stimulus [largest z score: 4.41; corresponding P value: 0.011, corrected for multiple comparisons over all electrodes, frequencies, and time points using a false discovery rate (FDR) procedure; permutation testing; see Methods for details]. The electrodes involved were located in the right posterior quadrant (Fig. 3A); oscillatory phase difference at these electrodes was in the theta band (4 to 6 Hz) between 370 and 250 ms before the mismatched stimulus. Examination of the precise time course of EEG activity in a representative participant confirms that theta-band oscillations were concentrated at different phase angles according to the participant's subsequent perception (Fig. 3, B and C). On the other hand, prestimulus theta phase did not vary systematically as a function of subsequent perception in the homologous electrodes of the left cerebral hemisphere (fig. S1). These results indicate that the phase of theta-band cortical oscillations in the right posterior quadrant predicts, about 300 ms in advance, whether the auditory or the visual speech cue will eventually dominate perception.

We further examined the role of oscillatory phase in determining perception in a patient implanted with subdural iEEG electrodes over portions of the right frontal, temporal, and parietal lobes. The patient perceived /b/-phoneme, /v/-viseme stimuli as /v/-leading words in 48% of trials. In a group of electrodes on the posterior superior temporal gyrus, POS was significant in the theta band (3 to 6 Hz) between 460 and 120 ms before the mismatched stimulus (maximum z score: 6.22; corresponding P value: 3.20×10^{-7}, corrected over frequencies and time points; permutation testing; Fig. 4). These findings confirm the crucial role of prestimulus theta-band oscillatory phase and point to the right posterior superior temporal cortex as a key region in the processing of auditory and visual speech cues.

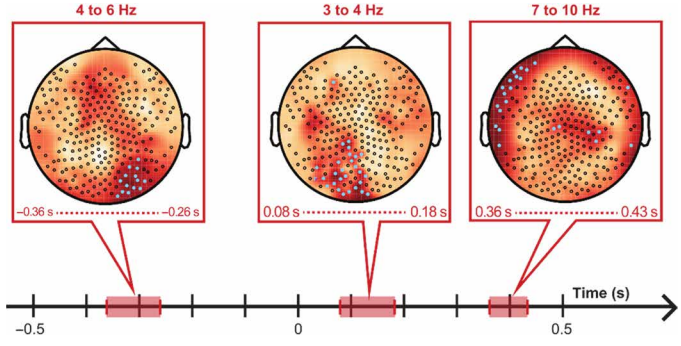

Fig. 2. Clusters of oscillatory phase separation as a function of perception. The POS, which quantifies differences in mean phase angle, is color-coded on topographic maps as a z score relative to its distribution according to the null hypothesis (permutation test). Note that the color scale is cropped at z values of [−2; +4], although some observed z values were larger. Electrodes with significant POS are highlighted in blue. Three groups of electrodes showed periods of significant differences in oscillatory phase as a function of perception. The first one, which occurred around 300 ms before the mismatched stimulus, involved right posterior electrodes in the theta band. The second one involved bilateral posterior electrodes around 100 ms after the mismatched stimulus. The third one involved mostly left temporal electrodes at the theta-alpha boundary around 400 ms after the mismatched stimulus.

To establish that cortical oscillations were present throughout the period preceding the mismatched stimulus and that they were not evoked as a phasic response to some external feature of the stimuli before the mismatched viseme-phoneme pair, we examined theta power in more detail. Note that we intentionally placed the mismatched viseme-phoneme pairs toward the end of the sentences (see table S1) to ensure that cortex was already processing the ongoing audiovisual speech input by the time the mismatch occurred. We did not observe any phasic increase in the power of theta oscillations in right posterior electrodes during the 1-s period preceding the mismatched stimulus (all $P > 0.05$, corrected for multiple comparisons

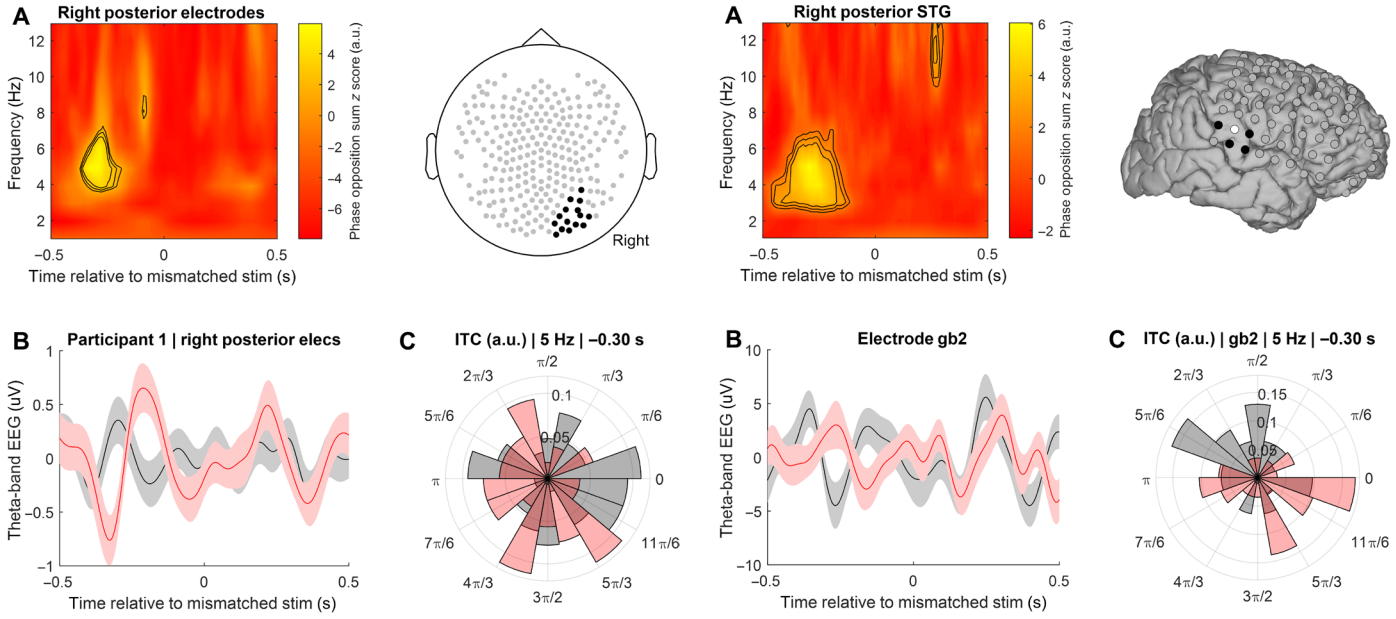

Fig. 3. The phase of theta-band oscillations in right posterior cortex predicts which sensory channel eventually drives perception. (**A**) The POS statistic quantifies the difference in mean phase angle according to eventual perception for the group of electrodes depicted in the inset at right and combined over participants ($N = 15$). POS is color-coded as a z score relative to its distribution according to the null hypothesis that there is no phase angle difference (permutation test). Black isocontour lines show the $P = 0.05$, 0.01, and 0.001 levels. Note that the POS is a one-sided statistical test, which explains why large negative z values are not deemed significant. a.u., arbitrary units. (**B**) Theta-band EEG responses in one representative participant, averaged over the electrodes depicted in (A). Black: trials perceived as /b/-leading words; red: trials perceived as /v/-leading words. The solid line shows mean voltage, and the shaded area SE over trials. Note that theta-band oscillatory phase diverges markedly before, but not after, the mismatched stimulus. (**C**) In the same representative participant, the phase angle of 5-Hz oscillations 300 ms before the mismatched stimulus is shown as a polar histogram, grouped according to eventual perception (black: /b/-leading words; red: /v/-leading words). The intertrial coherence (ITC) increases when many individual trials with similar phase are binned together. POS z score: 2.30, $P = 0.01$.

Fig. 4. The phase of theta-band oscillations in right posterior superior temporal cortex determines which sensory channel eventually drives perception. (**A**) POS, combined for the electrodes depicted in the inset at right, is plotted identically to Fig. 3A. Inset: The highlighted electrodes lie on the posterior superior temporal gyrus. Responses from electrode gb2, highlighted in white, are displayed in (B) and (C). (**B**) Theta-band EEG responses in one representative electrode are plotted identically to Fig. 3B. (**C**) In the same representative electrode, the phase angle of 5-Hz oscillations 300 ms before the mismatched stimulus is shown as a polar histogram, identically to Fig. 2C. POS z score: 3.44, $P = 2.9 \times 10^{-4}$.

over time points and perceptual outcomes; one-tailed paired t tests for power increase between successive 250-ms time points; Fig. 5A). Similarly, we did not find any phasic increase in theta power in the posterior superior temporal gyrus on iEEG (all $P > 0.05$, corrected; one-tailed paired t tests; Fig. 5B). This observation confirms that cortex is engaged in ongoing theta-band oscillations by the time the phase angle difference arises, and makes it unlikely that this phase angle difference is explained by a phasic cortical response to an external stimulus feature before the mismatch.

As a confirmation that the phase of prestimulus theta oscillations in right posterior cortex is determinant for the eventual perception of mismatched stimuli, we also examined responses to mismatched /b/-viseme, /v/-phoneme words. When presented with such a combination, participants perceived a /b/-leading word in 34% of trials (median; interquartile range: 19 to 50%). Using the POS to examine the link between oscillatory phase and perception, we did not find any electrode displaying significant POS (movie S3). Focusing on the theta band between 350 and 250 ms before the mismatched stimulus, the POS topography was somewhat similar to that shown in Fig. 2, with a group of right posterior electrodes approaching significance (fig. S2A). Similarly, the patient perceived a /b/-leading word in 33% of trials. iEEG electrodes on the right posterior superior temporal gyrus showed significant POS in the theta and alpha bands between 320 and 80 ms before the mismatched stimulus (maximum z score: 4.36; corresponding P value: 3.09×10^{-4}, corrected over frequencies and time points; permutation testing; fig. S2B). Even though these complementary results are not as robust as our main findings, they also point toward prestimulus theta-band oscillatory phase as a factor in determining which sensory channel eventually drives perception.

Oscillations beyond the theta band are also involved in the cortical processing of speech (*8*). We thus explored the relationship between oscillatory power or phase and perception between 14 and 30 Hz. At those frequencies, there was no significant power difference as a function of perception (movie S4). Phase angle tended to differ as a function of perception in the high-beta/low-gamma band for a brief period (24 to 26 Hz, 70 to 50 ms before the mismatched stimulus) in a small group of left-sided frontal electrodes (fig. S3A and movie S5). Beta oscillations (*22*) and neuronal activity in left frontal cortex (*23*) have been implicated in predicting sensory perception. This exploratory finding suggests that the phase, and not just the power, of beta-gamma oscillations in left frontal cortex might play an important role in the processing of audiovisual speech.

iEEG affords the possibility to investigate high-frequency activity (HFA), which correlates with local neuronal firing (*24*, *25*). In the right posterior superior temporal cortex, HFA tended to differ as a function of eventual perception for a brief period that coincided with that of significant theta phase differences (fig. S3B). More precisely,

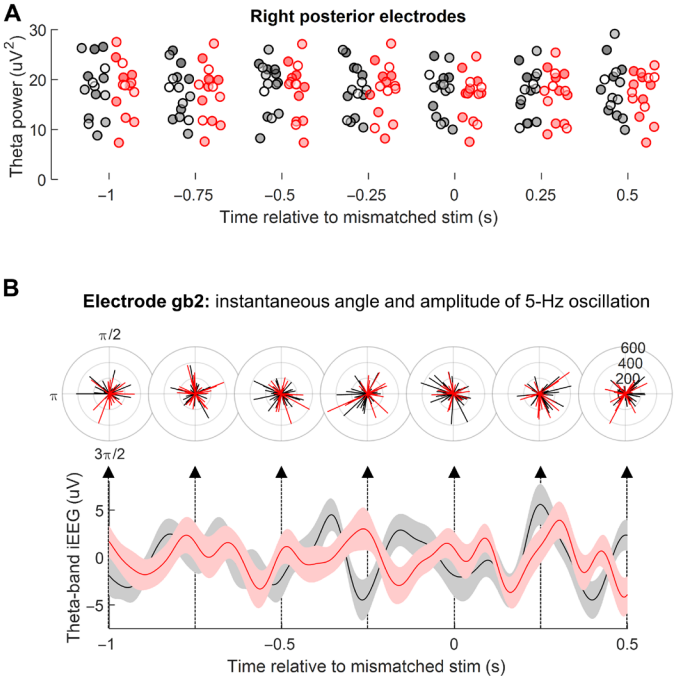

Fig. 5. No phasic increase in theta-band oscillatory power preceding the mismatched stimulus. (**A**) For each participant, power between 4 and 7 Hz at posterior electrodes is averaged over right posterior electrodes and over trials at selected time points, grouped as a function of perception (black: /b/-leading words; red: /v/-leading words; shades of colors correspond to individual participants). There is no phasic increase in theta power in the 1-s period preceding the mismatched stimulus. (**B**) Single-trial phase angle and power of 5-Hz oscillations at a posterior superior temporal gyrus electrode are shown on top of an extended version of Fig. 4B. The radial scale (quantifying power in uV^2) is the same for all polar plots. There is no phasic power increase in the 1-s period preceding the mismatched stimulus.

HFA when the patient's perception was driven by the viseme dipped below HFA associated with phoneme-driven perception (five successive time points reached significance from 180 to 130 ms before the mismatched stimulus; maximum *z* score: 2.39; corresponding two-tailed *P* value: 0.017, not corrected for multiple comparisons over time points; permutation testing). Given the relationship between theta phase and HFA (*26*), we speculate that differences in phase angles as a function of perception might lead to different states of auditory cortical excitability and therefore responsiveness to incoming speech sounds (*16*). We will need to examine this question in more detail with a larger iEEG dataset spanning more cortical areas, including core auditory cortex (*27*).

DISCUSSION

Past work established that perceptual awareness of difficult-to-detect stimuli in the visual and somatosensory modalities varies as a function of prestimulus oscillatory phase, suggesting that phase could be a determinant of perception (*28*–*30*). Recently, the integration or segregation of near-simultaneous, basic auditory and visual stimuli was shown to vary as a function of the phase of prestimulus oscillations as well (*31*). The artificial manipulation of cortical oscillations has perceptual consequences that depend on the imposed period of the oscillatory cycle (*32*). These previous findings strongly implicate prestimulus oscillatory phase as a determinant of cortical responsiveness to sensory stimulation. The mechanistic role played by oscillations in the perception and processing of audiovisual speech, however, remained to clarify.

In this study, we show that the phase angle of theta-band cortical oscillations in the posterior quadrant of the right cerebral hemisphere predicts whether it is the auditory or the visual speech cue that will dominate perception; hence, prestimulus phase determines subsequent perception. We thus provide evidence that cortical oscillations function as an instrument of sensory selection in the processing of audiovisual speech (*19*). The unique features of our experimental design reinforce the validity of our observations. First, we seamlessly embedded mismatched auditory and visual speech cues in syntactically and semantically correct sentences, which markedly improve the ecological relevance of our stimuli. Second, the fact that the mismatched cues occurred toward the end of sentences means that cortical oscillatory activity around that time was not contaminated by nonspecific stimulus-onset evoked responses and was already being entrained by the speech signals. Third, the bistable perception induced by our mismatched stimuli means that we could contrast cortical responses to stimuli that were, in essence, physically identical: Only perception varied on a trial-by-trial basis. The remarkable convergence of our neurophysiological results at two levels of granularity (hdEEG and iEEG), especially in terms of frequency and latency, further highlight their robustness.

It is expected that oscillations in the theta band are involved in the processing of audiovisual speech. Auditory cortex spontaneously generates oscillations in the theta range (*14*, *33*). Furthermore, the syllabic rate, which most strongly determines the rhythmicity of both auditory and visual speech cues, is centered at 4 to 6 Hz (*3*). Given that oscillatory phase correlates with neuronal excitability (*13*), the theta cycle represents an alternation of relatively higher and lower cortical responsiveness states, which is regularly realigned to the precise timing of the continuous but quasi-rhythmic speech input. In our situation, this alignment favors the processing of either the visual or the auditory speech cue so that it dominates perception on a trial-by-trial basis. Data from nonhuman primates indicate that low-frequency oscillations in sensory cortex are under the influence of attention (*17*, *18*). Hence, our results suggest that theta-band oscillations also reflect the attentional selection of a sensory channel in the processing of audiovisual speech.

Our hdEEG results point toward the right posterior quadrant as a key player in sensory selection during the processing of audiovisual speech. While the iEEG data, provided by a single patient, cannot fully resolve which exact cortical areas are involved, the posterior superior temporal cortex is clearly included. The region is sensitive to both auditory and visual speech cues (*10*, *12*, *34*), and there is considerable evidence that it is a major hub for multisensory integration (*35*–*38*). Right-sided temporal and parietal regions track the temporal dynamics of speech as much as, if not more than, their left-sided counterparts (*39*, *40*), suggesting that the right hemisphere plays a hitherto little recognized role in the processing of naturalistic auditory speech. Our data extend those findings to audiovisual speech. Of note, although our results highlight right hemispheric regions, it has been amply demonstrated that the left hemisphere, and especially the left superior temporal cortex, also participates in audiovisual speech integration (*10*, *41*–*43*).

Beyond the three clusters depicted in Fig. 2, close examination of our movie S3 suggests other periods and regions (including the left hemisphere) where oscillatory phase could be associated with

sensory selection. Our relatively small sample size might have prevented us from detecting more subtle contrasts. It will be interesting to expand our study of sensory selection in mismatched audiovisual speech by acquiring a larger dataset of iEEG participants, including broader coverage of visual, auditory, and language-related cortex in both cerebral hemispheres.

Our data do not explain what intrinsic factors would cause the phase of theta oscillations to change on a trial-by-trial basis. One general explanation is that attention to ongoing multisensory stimulus streams could spontaneously fluctuate between sensory channels, similar to the spontaneous fluctuations observed in individual sensory systems (*44*). Another hypothesis (the two are not mutually exclusive) is that top-down semantic predictions from language-related cortex could preactivate the sensory representation of distinct words or phonemes, similar to the role played by the inferior frontal cortex in phonemic restoration effects (*23*). Such top-down effects could well be implemented mechanistically by controlling oscillatory phase in downstream areas. This could be tested by building richer sentences for our characters to utter, where the meaning of the sentence influences expectations of what the mismatched word would be.

Our observations feed into a larger body of work that implicates the phase of oscillatory activity as a major mechanism for representing information among large neuronal populations (*45*, *46*). Further work will establish how phase coding in the processing of audiovisual speech enables bottom-up, sensory-driven predictions and top-down predictions based on memory and semantics.

METHODS

Experimental design

Stimuli consisted of 3D-animated virtual characters uttering short sentences in French. Ten sentence pairs, where members of a pair differed from each other by a single phoneme, were generated (table S1). The auditory stimuli were synthesized with the ReadSpeaker text-to-speech software (www.readspeaker.com/). Each sentence was synthesized once with a female voice and once with a male voice. The pitch of the soundtracks was then raised or lowered by 10% in Unity (see below) to create three different-sounding voices per gender, for a total of six different voices per sentence.

The visual stimuli were synthesized with the 3D computer graphics software Adobe Fuse CC (www.adobe.com/products/fuse.html), from which six characters (three female) were selected. Nonspeech animations (respirations and small movements of the eyes, face, and shoulders) were generated with Adobe Mixamo (www.mixamo.com/#/). The 3D models were imported into the 3D development platform Unity (https://unity.com/). The LipSync Pro plugin (https://lipsync.rogodigital.com/) was used to generate articulatory movements for a variety of visemes (AI, E, U, O, CDGKNRSTHYZ, FV, L, MBP, WQ, or rest). For each sentence, key frames for each viseme were manually aligned onto the corresponding phoneme of the soundtrack.

Using fully synthetic auditory and visual speech allowed us to voluntarily mismatch one phoneme-viseme pair, pairing a /v/ viseme with a /b/ phoneme or vice versa, while maintaining complete synchrony of all audiovisual events. The combination of a /v/ viseme with a /b/ phoneme tends to be perceived as "v," whereas the reverse combination tends to be perceived as "b" (*47*). Thus, our stimuli were designed to manipulate the perception of a single phoneme-viseme pair, which results in the corresponding word being perceived as either a /b/-leading or a /v/-leading word.

Generating our stimuli using 3D-animated virtual characters and computerized speech synthesis allowed us to ensure that there were no differences in low-level visual and auditory properties (phonetics, intonation, and stress) between pairs of stimuli. We also made sure that the mismatched words in each pair were the same parts of speech and were inflected identically where applicable. To assess the potential effect of semantical expectations on the rate of illusory perception, we conducted an online survey where 220 French speakers rated which one of the two words would more likely complete each sentence (on a 1 to 5 scale, 3 indicating equal likelihood). We then correlated these ratings with the rate of illusory perception in our pilot behavioral study (N = 24 French speakers) and found essentially no correlation (r^2 = 0.026, P = 0.91, Pearson's correlation). Removing the sentences with the most unbalanced likelihood scores did not notably alter the rate of illusory perception either. This led us to conclude that semantic expectations did not play an important role in determining the rate of illusory perception.

The experiment was implemented as a collection of C# scripts in Unity. It was presented using a Dell Precision 5530 laptop computer with an Intel Core i7-8850H processor, an Nvidia Quadro P1000 graphic card, 16-GB RAM, and a 1920 × 1080 light-emitting diode screen, running Windows 10. All sounds were played on the computer's native sound card and speakers at 44.1 kHz. Facial animations were generated and executed on a fixed time basis set to the screen's refresh rate, 60 Hz. With this setup, we obtained a maximum delay of ±25 ms between the visual and auditory stimuli, as controlled by recording the experiment's time course with a photodiode and a microphone.

Each trial started with the fade-in on-screen appearance of a character, in front of an out-of-focus picture of a café's interior, together with a soundtrack of background conversations as one could hear in a café. After 1 s, the character started uttering a sentence. Toward the end of each sentence, a key word was presented with either a congruent or mismatched leading phoneme-viseme pair. One second after the character had finished speaking, participants were presented with a forced two-choice task where they had to indicate by button press whether they had heard a key word as a /v/-leading or a /b/-leading word. The corresponding words were written on-screen, below the character, and remained on-screen until participants gave a response. Reaction speed was not recorded. The participant's response triggered a fade-out and the onset of the next trial.

The full experiment consisted of 40 audiovisual sentences (10 pairs of auditory sentences, each presented once with the congruent visual sentence and once with the mismatched visual sentence), each uttered once by each of the 6 characters, for a total of 240 unique trials. Healthy participants sat in a darkened, soundproofed Faraday cage and went through all trials twice, with a break in the middle. The patient was recorded in his hospital bed and went through all trials once.

Compliance with ethical regulations

This study was approved by the Commission cantonale d'éthique de la recherche sur l'être humain de la République et canton de Genève (project no. 2018-00911) and was conducted in accordance with the relevant Swiss laws and regulations and international guidelines on research on human subjects. Informed consent was obtained in writing from all participants.

Participants

All participants were either native French speakers or completely fluent in French. All had normal hearing and normal or corrected-to-normal vision, according to their own report. Twenty-four healthy participants accomplished the behavioral experiment (13 women; mean age 28 years, range 22 to 44). Twenty-five healthy participants were recruited to participate in the hdEEG experiment. Two patients were excluded after the behavioral screening test (described below) because they did not have a sufficient rate of illusory perception. One further participant's data had to be discarded because of an insufficient rate of illusory perception during the EEG recording, despite passing the screening test. Last, data from seven participants had to be discarded because a technical failure of the recording amplifier strongly contaminated their EEG signals with 50-Hz line noise. Therefore, 15 participants were included in the EEG analyses (seven women; mean age 27 years, range 19 to 42). Participants in the behavioral and EEG experiments were paid proportionally to the duration of their participation.

A 34-year-old man with drug-resistant epilepsy also participated in the experiment. He was implanted with subdural iEEG electrodes covering portions of his right frontal, temporal, and parietal lobes. After iEEG monitoring, resection of a focal cortical dysplasia in the right middle frontal gyrus led to freedom from further seizures at 3-month follow-up and did not cause any new neurological deficit.

Neurophysiological recordings

High-density scalp EEG was recorded using a 256-electrode cap, amplified and digitized at 1000 Hz for offline processing (Electrical Geodesics Inc., Eugene, OR). iEEG was recorded using 84 subdural electrodes (Ad-Tech Medical, Oak Creek, WI), amplified and digitized at 2048 Hz for offline processing (Brain Quick LTM, Micromed, S.p.A., Mogliano Veneto, Italy). A photodiode detected a brief white flash on the screen's upper left corner, whose onset was synchronous with that of the key phoneme-viseme pair.

hdEEG preprocessing

EEG analysis was carried out using the toolboxes FieldTrip (www.fieldtriptoolbox.org/; RRID:SCR_004849) (*48*), EEGLAB (https://sccn.ucsd.edu/eeglab/index.php; RRID:SCR_016333) (*49*), and custom-made MATLAB functions and scripts (The MathWorks Inc., Natick, MA; RRID:SCR_001622). First, the timeline was reconstructed from the photodiode channel. EEG epochs lasted the entire duration of speech stimuli, plus 2 s of data padding at each end. Because the hdEEG recording system caused EEG signals to be contaminated with 50-Hz line noise, an infinite impulse response notch filter was applied to the data.

Data were then cleaned from bad electrodes and trials with artifacts. Electrodes whose amplitude reached ±100 μV for more than 5% of the duration of all trials were flagged as noisy. In addition, a *z* score calculated over trials for each electrode identified electrodes with recurrent periods of abnormally low amplitude over prolonged periods. Channels with a *z* score larger than 3 for more than 10% of trials were flagged as noisy. Flagged electrodes were then reviewed manually to ensure that no more than 5% of all electrodes were flagged and to add electrodes that were noted to be broken at the time of EEG recording. The signal from flagged electrodes was replaced by an interpolation from neighboring electrodes using FieldTrip.

To identify artifactual trials, the covariance between all electrode pairs was computed for each trial. The single-trial covariance matrices were then compared to their average over trials using a distance metric, which was converted into a *z* score over trials. Individual trials with a *z* score higher than 2.3 (indicating higher variance between individual electrodes during those trials) were removed from further analysis. Additional artifacts from eye blinks, muscle activity, and other sources of noise were further identified and removed using an independent component analysis. Components were computed with the runica algorithm (*49*). Visualization and semiautomatic identification of noncerebral components were performed using the SASICA (Semi-Automated Selection of Independent Components of the electroencephalogram for Artifact correction) algorithm (*50*). Artifact-free epochs were rereferenced to average reference, filtered between 1 and 30 Hz (Butterworth filters, low-pass order 6, high-pass order 4), and corrected with respect to a baseline from 250 to 50 ms before stimulus onset.

iEEG electrode localization

iEEG electrodes were localized and displayed using the iELVis (intracranial ELectrode Visualization) toolbox (http://ielvis.pbworks.com/w/page/116347253/FrontPage; RRID:SCR_016109) (*51*). Briefly, the patient's pre-implant 3D T1 magnetic resonance imaging (MRI) was processed using FreeSurfer (http://surfer.nmr.mgh.harvard.edu/; RRID:SCR_001847) (*52*) for skull stripping, extraction of the pial surface, and automatic parcellation of gyri and sulci. The post-implant 3D computed tomography was coregistered to the pre-implant MRI using FSL (www.fmrib.ox.ac.uk/fsl/; RRID:SCR_002823) (*53*). Electrodes were manually localized using BioImage Suite 3 (https://medicine.yale.edu/bioimaging/suite/; RRID:SCR_002986) (*54*). Brain shift was corrected using an inverse gnomonic map projection (*55*).

Time-frequency analysis

The power and phase of low-frequency oscillations were computed using a wavelet transform in FieldTrip. Wavelets (three-cycle width) were centered every hertz from 1 to 13 Hz and every 10 ms from 500 ms before to 500 ms after the key stimulus. Differences in mean phase angle between perceptual outcomes were quantified by the *POS* (*21*): $POS = ITC_b + ITC_v - 2 \times ITC_{all}$, where ITC_b and ITC_v are the mean resultant vectors of single-trial phase angles for trials where a /b/-leading or a /v/-leading word was perceived, respectively, and ITC_{all} is the mean resultant vector of single-trial phase angles for all trials together, irrespective of perception. Power and POS were expressed as *z* scores relative to their estimated distribution under the null hypothesis that they do not differ according to perceptual outcome, as described below.

Representative waveforms of theta-band activity (single-participant hdEEG and single-electrode iEEG) were obtained by filtering EEG signals (fourth-order Butterworth filters, low-pass 2 Hz, high-pass 8 Hz), averaging over neighboring electrodes (hdEEG only) and then over trials. These examples appear in Figs. 3B (hdEEG) and 4B and 5B (iEEG). Representative phase angles at selected frequencies and latencies, obtained from the wavelet transform described above, are shown in Figs. 3C (hdEEG) and 4C (iEEG).

HFA, indexing local neuronal firing (*24*), was obtained by filtering the iEEG signal in 10-Hz bands between 75 and 175 Hz (fourth-order Butterworth filters), computing band-limited power through a Hilbert transform, dividing band-limited power by its own mean over time to compensate for the 1/*f* power drop, and averaging over 10-Hz bands (*56*). Single-trial HFA was baseline

corrected to the 250-ms period immediately preceding the onset of each sentence. HFA plots appear in fig. S3B.

Statistical analysis

Participants' responses in the behavioral experiment were analyzed with a repeated-measures analysis of variance (ANOVA) with phoneme-viseme congruence (congruent versus mismatched) and phoneme identity (/b/ versus /v/) as within-subject factors. The result of this analysis is shown in Fig. 1B.

Only EEG data from trials where the key word contained a /b/ phoneme mismatched with a /v/ viseme were analyzed here. For hdEEG, oscillatory power and phase were compared between trials grouped according to perceptual outcome. To compare oscillatory power, power was averaged over trials for each participant, electrode, frequency, and time point, separately for each perceptual outcome. The differences in mean power between perceptual outcomes were computed (one value per participant, electrode, frequency, and time point) and then compared to their distribution under the null hypothesis that there was no significant difference as a function of perception. The distribution was estimated using a permutation test with 1000 repetitions, from which the mean and SD were calculated. The observed values were expressed as a z score relative to that distribution, and a two-sided P value was computed from the z score. At each electrode, frequency, and time point, P values were then combined over participants using Stouffer's method (*20*). The resulting P values express the probability that the observed values be observed if all null hypotheses are true. The time course of combined P values was then subjected to correction for multiple comparisons using the FDR procedure (*57*). This stringent correction ensured that only strongly significant differences in power would be detected. The result of this analysis is shown in movie S1. No power difference survived FDR correction. Movie S4 was generated in identical fashion.

To compare phase angle, POS was calculated as described above. The observed POS statistic (one value per participant, electrode, frequency, and time point) was then expressed as a z score of its estimated distribution under the null hypothesis that there was no phase angle difference as a function of perception. A one-sided P value was computed from the z score (one-sided because POS is a signless quantity). At each electrode, frequency, and time point, P values were combined over participants using Stouffer's method. The time course of combined P values was then subjected to FDR correction for multiple comparisons. This stringent correction ensured that only strongly significant differences in mean phase angle would be detected. The result of this analysis is shown in movie S2. There were three periods where groups of electrodes showed a POS statistic that was large enough to survive the stringent FDR correction; of those, one occurred before the mismatched stimulus. The pre-stimulus period of POS significance concerned 16 electrodes, of which 15 were neighbors in the right posterior quadrant. For illustrative purposes, those electrodes' POS z scores were again combined using Stouffer's method and plotted as a contour plot. This appears in Fig. 3A (and fig. S1 for homologous electrodes in the left posterior quadrant). In that figure, data from two electrodes that showed a significant POS z score were omitted from the contour plot: one electrode that was isolated in the right frontal region and one left occipital electrode that was also part of a lower-frequency post-stimulus group of significant electrodes. Both electrodes can be seen on movie S2. Movies S3 and S5 were generated in identical fashion.

For iEEG, phase angle differences as a function of perception were quantified with the POS statistic as described above, which was expressed as a z score of its estimated distribution under the null hypothesis. Visual inspection of results for each electrode showed a group of five neighboring electrodes on the posterior superior temporal gyrus that showed significant POS in the theta band and during the pre-mismatched stimulus period. Those electrodes' z scores were combined using Stouffer's method and plotted as a contour plot, which appears in Fig. 4A and fig. S2B.

Data availability

The EEG and iEEG data from this study are freely available on the University of Geneva's institutional repository, Yareta (https://doi.org/10.26037/yareta:cripcwu4nbh5vprpqzhorbycry).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/45/eabc6348/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. W. H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).
2. H. McGurk, J. Macdonald, Hearing lips and seeing voices. *Nature* **264**, 746–748 (1976).
3. S. Greenberg, H. Carvey, L. Hitchcock, S. Chang, Temporal properties of spontaneous speech—A syllable-centric perspective. *J. Phon.* **31**, 465–485 (2003).
4. E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, M. M. Merzenich, Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13367–13372 (2001).
5. J. E. Peelle, J. Gross, M. H. Davis, Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* **23**, 1378–1387 (2012).
6. N. Ding, J. Z. Simon, Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 11854–11859 (2012).
7. J. Gross, N. Hoogenboom, G. Thut, P. Schyns, S. Panzeri, P. Belin, S. Garrod, Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biol.* **11**, e1001752 (2013).
8. A.-L. Giraud, D. Poeppel, Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
9. H. Park, C. Kayser, G. Thut, J. Gross, Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife* **5**, e14521 (2016).
10. C. Micheli, I. M. Schepers, M. Ozker, D. Yoshor, M. S. Beauchamp, J. W. Rieger, Electrocorticography reveals continuous auditory and visual speech tracking in temporal and occipital cortex. *Eur. J. Neurosci.* **51**, 1364–1376 (2020).
11. A. E. O'Sullivan, M. J. Crosse, G. M. Di Liberto, E. C. Lalor, Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Front. Hum. Neurosci.* **10**, 679 (2016).
12. P. Mégevand, M. R. Mercier, D. M. Groppe, E. Z. Golumbic, N. Mesgarani, M. S. Beauchamp, C. E. Schroeder, A. D. Mehta, Crossmodal phase reset and evoked responses provide complementary mechanisms for the influence of visual speech in auditory cortex. *J. Neurosci.* 10.1523/JNEUROSCI.0555-20.2020 (2020).
13. M. Volgushev, M. Chistiakova, W. Singer, Modification of discharge patterns of neocortical neurons by induced oscillations of the membrane potential. *Neuroscience* **83**, 15–25 (1998).
14. P. Lakatos, A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, C. E. Schroeder, An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.* **94**, 1904–1911 (2005).
15. P. Lakatos, C.-M. Chen, M. N. O'Connell, A. Mills, C. E. Schroeder, Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* **53**, 279–292 (2007).
16. C. E. Schroeder, P. Lakatos, Y. Kajikawa, S. Partan, A. Puce, Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* **12**, 106–113 (2008).
17. P. Lakatos, G. Karmos, A. D. Mehta, I. Ulbert, C. E. Schroeder, Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* **320**, 110–113 (2008).
18. P. Lakatos, M. N. O'Connell, A. Barczak, A. Mills, D. C. Javitt, C. E. Schroeder, The leading sense: Supramodal control of neurophysiological context by attention. *Neuron* **64**, 419–430 (2009).
19. C. E. Schroeder, P. Lakatos, Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* **32**, 9–18 (2009).

20. R. Thézé, M. A. Garidi, L. Albert, A. Provost, A.-L. Giraud, P. Mégevand, Animated virtual characters to explore audio-visual speech in controlled and naturalistic environments. *Sci. Rep.* **10**, 15540 (2020).
21. R. VanRullen, How to evaluate phase differences between trial groups in ongoing electrophysiological signals. *Front. Neurosci.* **10**, 426 (2016).
22. L. H. Arnal, A.-L. Giraud, Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* **16**, 390–398 (2012).
23. M. K. Leonard, M. O. Baud, M. J. Sjerps, E. F. Chang, Perceptual restoration of masked speech in human cortex. *Nat. Commun.* **7**, 13619 (2016).
24. N. E. Crone, D. L. Miglioretti, B. Gordon, R. P. Lesser, Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain* **121**, 2301–2315 (1998).
25. S. Ray, N. E. Crone, E. Niebur, P. J. Franaszczuk, S. S. Hsiao, Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).
26. R. T. Canolty, E. Edwards, S. S. Dalal, M. Soltani, S. S. Nagarajan, H. E. Kirsch, M. S. Berger, N. M. Barbaro, R. T. Knight, High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**, 1626–1628 (2006).
27. P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, N. Mesgarani, Joint representation of spatial and phonetic features in the human core auditory cortex. *Cell Rep.* **24**, 2051–2062.e2 (2018).
28. N. A. Busch, J. Dubois, R. VanRullen, The phase of ongoing EEG oscillations predicts visual perception. *J. Neurosci.* **29**, 7869–7876 (2009).
29. K. E. Mathewson, G. Gratton, M. Fabiani, D. M. Beck, T. Ro, To see or not to see: Prestimulus α phase predicts visual awareness. *J. Neurosci.* **29**, 2725–2732 (2009).
30. T. J. Baumgarten, A. Schnitzler, J. Lange, Beta oscillations define discrete perceptual cycles in the somatosensory domain. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12187–12192 (2015).
31. N. Ikumi, M. Torralba, M. Ruzzoli, S. Soto-Faraco, The phase of pre-stimulus brain oscillations correlates with crossmodal synchrony perception. *Eur. J. Neurosci.* **49**, 150–164 (2019).
32. R. Cecere, G. Rees, V. Romei, Individual differences in alpha frequency drive cross-modal illusory perception. *Curr. Biol.* **25**, 231–235 (2015).
33. A.-L. Giraud, A. Kleinschmidt, D. Poeppel, T. E. Lund, R. S. J. Frackowiak, H. Laufs, Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* **56**, 1127–1134 (2007).
34. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
35. A. A. Ghazanfar, C. E. Schroeder, Is neocortex essentially multisensory? *Trends Cogn. Sci.* **10**, 278–285 (2006).
36. L. E. Bernstein, E. Liebenthal, Neural pathways for visual speech perception. *Front. Neurosci.* **8**, 386 (2014).
37. J. E. Peelle, M. S. Sommers, Prediction and constraint in audiovisual speech perception. *Cortex* **68**, 169–181 (2015).
38. P. Riedel, P. Ragert, S. Schelinski, S. J. Kiebel, K. von Kriegstein, Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. *Cortex* **68**, 86–99 (2015).
39. D. A. Abrams, T. Nicol, S. Zecker, N. Kraus, Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *J. Neurosci.* **28**, 3958–3965 (2008).
40. A. M. Alexandrou, T. Saarinen, S. Mäkelä, J. Kujala, R. Salmelin, The right hemisphere is highlighted in connected natural speech production and perception. *Neuroimage* **152**, 628–638 (2017).
41. M. S. Beauchamp, A. R. Nath, S. Pasalar, fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *J. Neurosci.* **30**, 2414–2417 (2010).
42. J. Keil, N. Müller, N. Ihssen, N. Weisz, On the variability of the McGurk effect: Audiovisual integration depends on prestimulus brain states. *Cereb. Cortex* **22**, 221–231 (2012).
43. H. Park, R. A. A. Ince, P. G. Schyns, G. Thut, J. Gross, Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLOS Biol.* **16**, e2006558 (2018).
44. R. VanRullen, Perceptual cycles. *Trends Cogn. Sci.* **20**, 723–735 (2016).
45. A. Sirota, S. Montgomery, S. Fujisawa, Y. Isomura, M. Zugaro, G. Buzsáki, Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. *Neuron* **60**, 683–697 (2008).
46. A. J. Watrous, L. Deuker, J. Fell, N. Axmacher, Correction: Phase-amplitude coupling supports phase coding in human ECoG. *eLife* **4**, e12810 (2015).
47. J. Jiang, L. E. Bernstein, Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 1193–1209 (2011).
48. R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156859 (2011).
49. A. Delorme, S. Makeig, EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
50. M. Chaumon, D. V. M. Bishop, N. A. Busch, A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *J. Neurosci. Methods* **250**, 47–63 (2015).
51. D. M. Groppe, S. Bickel, A. R. Dykstra, X. Wang, P. Mégevand, M. R. Mercier, F. A. Lado, A. D. Mehta, C. J. Honey, iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *J. Neurosci. Methods* **281**, 40–48 (2017).
52. B. Fischl, FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
53. M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, S. M. Smith, FSL. *Neuroimage* **62**, 782–790 (2012).
54. A. Joshi, D. Scheinost, H. Okuda, D. Belhachemi, I. Murphy, L. H. Staib, X. Papademetris, Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics* **9**, 69–84 (2011).
55. A. I. Yang, X. Wang, W. K. Doyle, E. Halgren, C. Carlson, T. L. Belcher, S. S. Cash, O. Devinsky, T. Thesen, Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage* **63**, 157–165 (2012).
56. T. Golan, I. Davidesco, M. Meshulam, D. M. Groppe, P. Mégevand, E. M. Yeagle, M. S. Goldfinger, M. Harel, L. Melloni, C. E. Schroeder, L. Y. Deouell, A. D. Mehta, R. Malach, Human intracranial recordings link suppressed transients rather than 'filling-in' to perceptual continuity across blinks. *eLife* **5**, e17243 (2016).
57. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).

Acknowledgments: We thank the patient for participating in the experiment, the health care professionals of the Geneva University Hospitals' epileptology unit for accommodating us, M. Gadiri for assistance with data collection and analysis, and L. Albert and A. Provost for assistance with preparing the task and stimuli. **Funding:** This work was supported by the Swiss National Science Foundation (grant no. 167836 to P.M.). **Author contributions:** P.M. developed the concept. A.-L.G. and P.M. obtained funding. R.T. and P.M. designed the experiment, analyzed the data, and prepared the figures. R.T. collected the data. All authors interpreted the data and wrote the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Data are publicly available on the University of Geneva's institutional repository (https://yareta.unige.ch). Additional data related to this paper may be requested from the authors.

Submitted 5 May 2020
Accepted 17 September 2020
Published 4 November 2020
10.1126/sciadv.abc6348

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