EJN European Journal of Neuroscience

FENS Federation of European Neuroscience Societies

*European Journal of Neuroscience, Vol. 51, pp. 1364–1376, 2020* doi:10.1111/ejn.13992

# Electrocorticography reveals continuous auditory and visual speech tracking in temporal and occipital cortex

Cristiano Micheli,[1,2] Inga M. Schepers,[1,3] Müge Ozker,[4] Daniel Yoshor,[4,5] Michael S. Beauchamp[4] and Jochem W. Rieger[1,3]

[1]Department of Psychology, Carl von Ossietzky University, Oldenburg, Germany
[2]Donders Centre for Cognitive Neuroimaging, Radboud University, Nijmegen, The Netherlands
[3]Research Center Neurosensory Science, Carl von Ossietzky University, Oldenburg, Germany
[4]Department of Neurosurgery, Baylor College of Medicine, Houston, Texas
[5]Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas

*Keywords*: audiovisual speech, clear speech, continuous speech, multisensory, naturalistic stimuli

## Abstract

During natural speech perception, humans must parse temporally continuous auditory and visual speech signals into sequences of words. However, most studies of speech perception present only single words or syllables. We used electrocorticography (subdural electrodes implanted on the brains of epileptic patients) to investigate the neural mechanisms for processing continuous audiovisual speech signals consisting of individual sentences. Using partial correlation analysis, we found that posterior superior temporal gyrus (pSTG) and medial occipital cortex tracked both the auditory and the visual speech envelopes. These same regions, as well as inferior temporal cortex, responded more strongly to a dynamic video of a talking face compared to auditory speech paired with a static face. Occipital cortex and pSTG carry temporal information about both auditory and visual speech dynamics. Visual speech tracking in pSTG may be a mechanism for enhancing perception of degraded auditory speech.

## Introduction

While many studies of speech perception examine only single syllables or words, in natural situations, humans are typically presented with a parallel stream of visual (mouth movements) and auditory speech. Improved knowledge about the brain regions and neural responses that track concurrent features of the auditory and the visual speech signal will lead to a better understanding of how visual information improves speech perception (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954).

The posterior superior temporal gyrus and superior temporal sulcus (pSTG/pSTS) are likely candidates for auditory and visual speech tracking. These areas show multisensory responses to auditory and visual stimuli in fMRI (Beauchamp, Lee, Argall, & Martin, 2004; Lee & Noppeney, 2011; Noesselt et al., 2007; Stevenson & James, 2009), EEG/MEG (Arnal, Wyart, & Giraud, 2011; Schepers, Schneider, Hipp, Engel, & Senkowski, 2013) and with intracranial ECoG recordings (Ozker, Schepers, Magnotti, Yoshor, & Beauchamp, 2017; Rhone et al., 2016). However, the current understanding of audiovisual integration in the context of speech processing is rather incomplete. The evidence suggests that normal hearing listeners may use audiovisual integration for other purposes than speech comprehension (e.g., simultaneity detection, Noesselt et al., 2007). In fact, visual information can have a strong effect on speech perception, as demonstrated, for example, by the McGurk effect (McGurk & MacDonald, 1976; Nath and Beauchamp, 2012) and the enhancement of speech intelligibility by several dB SNR (Shannon et al., 1995). In the presence of noise, visual speech tracking in pSTG may be a mechanism for enhancing perception of degraded auditory speech (Ozker et al., 2017).

Most studies on audiovisual speech perception so far compared average neural response magnitude and latency differences between unimodal and multimodal experimental conditions (Besle et al., 2008; Rhone et al., 2016; Schepers, Yoshor, & Beauchamp, 2015; van Wassenhove, Grant, & Poeppel, 2005) but did not focus on the representation of dynamic speech characteristics of the different modalities (e.g., using stimuli with continuous sentence analysis or using different perceptual modalities such as audio and video streams). The effects of visual speech information on auditory speech tracking have been investigated with EEG in auditory noise-free (Crosse, Butler, & Lalor, 2015) and auditory noisy (Crosse, Di Liberto, & Lalor, 2016) conditions. The authors showed that representation of the auditory speech envelope is enhanced in noise-free conditions specifically with congruent visual speech and that in the presence of severe auditory noise, visual speech can improve early temporal tracking of the auditory speech stream. These EEG studies

*Correspondence*: Cristiano Micheli, as above. Email: michelic72@gmail.com

Received 14 August 2017, revised 19 May 2018, accepted 29 May 2018

Edited by John Foxe. Reviewed by Manuel Mercier, Université Toulouse, France; and Luc Arnal, University of Geneva, Switzerland.

All peer review communications can be found with the online version of the article.

did not differentially investigate which brain regions exhibit speech tracking and whether the parallel visual speech information is actually represented together with the auditory speech information in neural responses.

Recently, different models of the sensory speech input were related to the EEG signal during silent lip reading, and it was found that the unheard auditory speech envelope among other signals could predict the visual EEG signal (O'Sullivan, Crosse, Di Liberto, & Lalor, 2016). Additionally, visible lips movements accompanying speech have been related to low-frequency oscillations with MEG (Park, Kayser, Thut, & Gross, 2016). Park et al. observed speech lip movement tracking in visual cortex and left motor cortex for low-frequency responses (below 11 Hz). Yet, while these results are highly informative on the dynamic tracking of visual speech features in the brain, the authors focused in tracking of visual information only and did not simultaneously look at coherence between the auditory speech envelope and the brain responses, which could have informed about audiovisual effects.

In contrast to MEG and EEG, ECoG recordings offer spatial resolution in the millimetre range and temporal resolution that is sufficiently high to analyse ongoing speech coding in the human brain and to discriminate information encoded in slow and fast dynamic ranges. Moreover, ECoG recordings offer a higher signal-to-noise ratio in the fast dynamic signal range above 60 Hz than MEG and EEG recordings (Quandt et al., 2012). Combining ECoG recordings with statistical modelling techniques recent studies was able to reveal neural mechanisms specific of speech coding in the human brain with high spatial and temporal resolution (Chang et al., 2010; Holdgraf et al., 2016, 2017; Mesgarani & Chang, 2012; Mesgarani, Cheung, Johnson, & Chang, 2014). Zion Golumbic et al. (2013) showed with ECoG that high gamma responses in STG track the auditory speech envelope during audiovisual speech perception but did not investigate visual lip movement tracking. Brain regions tracking the auditory as well as the visual speech information are likely candidates for audiovisual integration of concurrent speech features as their neural activation contains information on both speech signals.

To simultaneously address the question of auditory as well as visual speech tracking in the human brain, clear auditory sentences were presented with either dynamic videos (AVdyn condition) or a static image (AVstatic condition) while ECoG signals were recorded. Partial correlation analyses were performed between slow (low frequency or LF, 6–30 Hz) and fast (high gamma or HG, 70–250 Hz) dynamic signal ranges and the auditory as well as the visual speech envelopes to investigate their representation in the brain responses separately. Previous research has suggested that the two different dynamic bands serve different computational functions in speech processing (Giraud & Poeppel, 2012). The HG band responses reflect local neural processing (Lachaux, Axmacher, Mormann, Halgren, & Crone, 2012; Ray & Maunsell, 2011) and auditory feature processing in primary and higher cortices (Holdgraf et al., 2016, 2017; Ozker et al., 2017). The role of LF band activity (from 6 Hz to 30 Hz) in sentence processing might be related to semantic processing (Wang et al., 2012), disruption of the perceptual status-quo (Engel & Fries, 2010) and sentence unification (Bastiaansen, Magyari, & Hagoort, 2010). Interestingly, noninvasive studies demonstrated that LF band brain responses can be entrained by speech (Ding and Simon 2014). Due to the putative differences in computational function, we analysed LF and HG bands separately.

Our results show tracking of both auditory and visual speech envelopes in pSTG and visual cortex, providing further support for the notion that these areas process multimodal speech information.

## Material and Methods

### *Subjects*

ECoG data of seven subjects (one male; six right-handed (self-reported); mean age = 41 years, standard deviation = 13 years) were collected and all of them had normal or corrected-to-normal vision. Subdural electrodes were implanted for the clinical purpose of monitoring the onset of seizures in patients affected by intractable epilepsy. The study was approved by the local ethics review board (Baylor College of Medicine Institutional Review Board and the University of Texas Committees for the Protection of Human Subjects), and subjects gave informed consent prior to the experimental session.

### *Stimuli*

All subjects were attending to videos or static pictures of a speaker uttering a sentence in English of the type "William bought eight white pencils" (either only with an audio track or with a video in which a speaker was uttering the sentences, that is the lip movements were visible on the monitor). A set of 210 different sentences were recorded with a female native English speaker and consisted of videos recorded with a JVC GC-PX100 camera and a Røde M2 cardioid microphone placed in front of the speaker (approximately 30 cm) and care was taken that the microphone did not appear in the image and that the values of the audio intensity were not clipped. The videos were acquired with a frame rate of 60 Hz and successively stored on disc in a standard format (AVI). The audio signal was recorded at 48 kHz with a fade-in/fade-out effect in the early and late 20 ms of the track. The sequence was low-pass filtered at 10 kHz with a 24th order forward–reverse Butterworth filter, 16 bits quantized, and normalized to the maximum root-mean-square value of all sentences, to avoid signal clipping.

The speech data used in the experiment were selected from a speech corpus extensively used in speech audiometry (i.e., for the evaluation of speech performance of hearing aids, OLSA, English Oldenburger Satztest). Each sentence consisted of five consecutive words uttered at normal speed and representing five word categories (people names – verbs numbers – adjectives – objects). The vocabulary consisted respectively of 10 different words for each category recombined to form a corpus of 210 sentences of correct semantic meaning (e.g., "Rachel ordered four red flowers"). Half of the sentences (105) were presented with synchronized auditory and dynamic visual streams, denoted auditory with dynamic video (AVdyn) condition, the other half was presented with the originally recorded auditory track together with a static image of the speaker, denoted auditory with static video (AVstatic) condition (Figure 1a). The onset of the video was monitored with a photodiode placed in the lower left corner of the screen where a small rectangle of 60 × 60 pixels changed from black to white at video start.

### *Stimuli: Audio and video feature extraction*

The analyses reported here are based on the audio amplitude envelopes and the vertical lip distances (Figure 1b). We used PRAAT (Boersma & van Heuven, 2001) to calculate audio amplitude envelopes from audio spectrograms. The vertical lips distance was extracted from the visual stimuli using the IntraFace face-tracking software (Xiong & Torre, 2013) and interpolated to the sampling frequency of the brain recordings using a custom MatLab script.

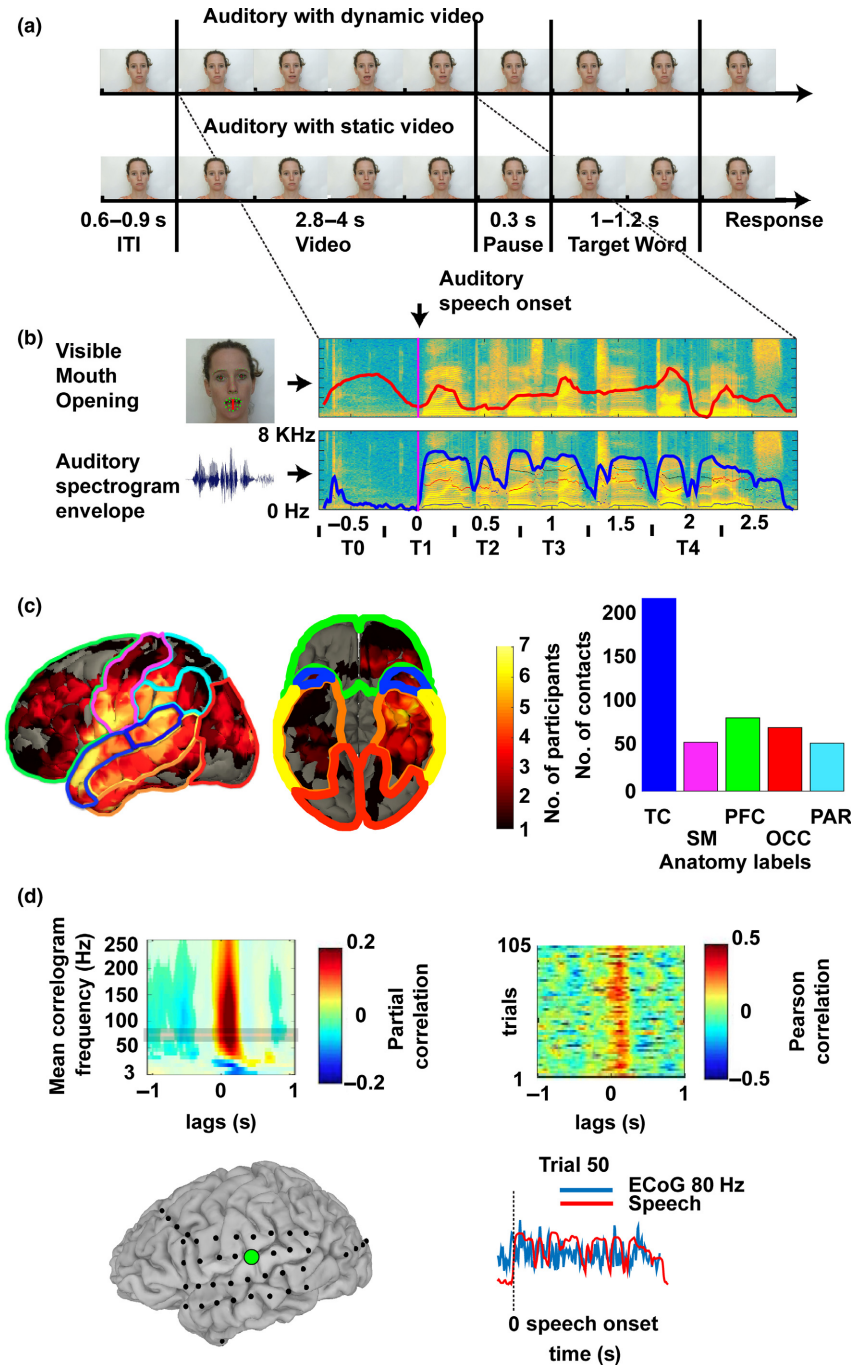


FIG. 1. (a) Trial structure: Subjects listened to sentences presented either with the speaker moving the lips (auditory with dynamic video, AVdyn) or with a static image of a speaker (auditory with static video, AVstatic). After a short pause, a target word was presented and the subjects had to answer whether the word was present in the previous sentence. (b) Two stimulus features were extracted from the sentence utterance interval: The time series of the vertical mouth opening (red) and the envelope of the spectrogram (blue). Example time series for one sentence are shown. T0-T4 denote intervals for the time-resolved analysis (T0: before audio onset ($-0.5 \pm 0.25$ s), T1: audio speech onset ($0 \pm 0.25$ s), T2: early sentence ($0.5 \pm 0.25$ s), T3: middle sentence ($1 \pm 0.25$ s), T4: late sentence ($2 \pm 0.25$ s)). Note that in the AVdyn condition, the speaker may already move her lips prior to auditory speech onset. (c) Location of electrodes for all subjects, projected on a template brain. Over subjects, the highest densities are along the STG (blue), inferior somatomotor (magenta), prefrontal (green), occipital (red), and parietal (cyan) cortices, indicated by the hotter colours. OCC: occipital; PAR: parietal; PFC: prefrontal; SM: sensory-motor; TC: temporal cortex. (d) Correlation between the neural response and the auditory envelope for a representative electrode over pSTG. Frequency-resolved correlogram (partial correlations) for a single electrode (top left). The frequency range highlighted in the plot ($80 \pm 10$ Hz) was used for the subsequent panels in d. Single trial correlation (Pearson's correlation) across AVdyn trials between the neural response and the auditory envelope for the different lags from $-1$ to 1 s (top right). Single sentence neural response and auditory envelope time courses ($r = 0.49$, lag of 160 ms of the auditory envelope compared to the neural response (bottom right).

### Experimental design

The experiment began with a grey background screen, instructing the subject to fixate the mouth of the speaker in the centre of the image. Each trial started with a still image of a speaker with a closed mouth maintained for a jittered time of 0.6–0.9 s (Figure 1a). Then, a video was delivered in AVdyn trials or the static image remained on the screen in AVstatic trials. Depending on the sentence length, this lasted between 2.8 and 4 s (mean duration = 3.2 s, *SD* = 0.33 s). Afterwards, a static image of the speaker corresponding to the last frame of the video was shown on AVdyn trials or the static image remained on the screen on AVstatic trials for a duration of 0.3 s. Subsequently, a word was presented (mean duration = 1.1 s, *SD* = 0.2 s) and the subject was asked to respond with the right mouse button if the word was included in the sentence or the left mouse button otherwise. This task was included to ensure that subjects were attending to the sentences during the presentation of the stimuli. The response mouse button triggered the start of the next trial.

All 210 trials were arranged in a pseudorandom succession with the limitation that a maximum of two sentences in a row from the same stimulus category (AVdyn or AVstatic) was allowed. Trials were presented in three blocks of 70 sentences, each of approximately 7 min' duration. Rest periods were included between blocks. If the subject got tired, the last block or blocks were not presented (Subject 1 and 7), in compliance with the IRB protocol.

### Data acquisition set-up

Each recording session took place in the subject's hospital room and subjects comfortably sat in their bed in front of a telescopic arm-held monitor and listened to sound stimuli delivered by two loudspeakers placed on the wall to the left and right of the subject's head. Stimuli were presented with Presentation software (Version 16-5; Neurobehavioral Systems, Inc., Albany, CA, USA) with a Dell 21,5'' monitor (resolution 1,920 × 1,080 pixels, aspect ratio 16-9) at a distance of approximately 60 cm (22'') from the subject. The presented images covered the entire screen. Sounds were delivered from loudspeakers close to the subject's back at a comfortable volume ensuring full intelligibility. Timestamps for the subjects' responses were generated by the presentation software after the participants answered the question "Did the target occur in the sentence?." A yes/no answer was recorded from, respectively, a left/right mouse button pressed with the right hand.

The audio output from the presentation laptop was split into two signals, one was sent to the external loudspeaker and the other was sent to the ECoG amplifier and recorded as additional ECoG channel to document the actual audio presentation. Finally, the photodiode signal was recorded in an additional channel of the ECoG amplifier.

### Electrodes and data acquisition

Standard subdural recording electrodes were used (Ad-Tech Medical Instrument Corporation, Racine WI, USA). ECoG was recorded with a 128-channel amplifier (Cerebrus; Blackrock Microsystem, Salt Lake City, UT, USA) composed of four banks of 32 channels for each connector. The electrodes consisted of grids or strips of platinum alloy discs embedded in a silicon rubber support with 2.2 mm diameter exposed conductive surface and 1 cm centre-to-centre spacing. The ECoG signals were digitized at 2 kHz, low-pass filtered at 500 Hz (Butterworth filter, filter order of 4), high-pass filtered at 0.3 Hz (Butterworth filter, filter order of 1) and saved on disc. During the recordings, the electrodes were referenced to an intracranial electrode turned towards the skull.

The number of electrodes recorded was, respectively, 68 (S1), 51 (S2), 78 (S3), 70 (S4), 66 (S5), 51 (S6) and 108 (S7), for a total of 492 electrodes across seven subjects.

### Anatomical localization

We coregistered the individual postsurgical CT-scans to the individual presurgical T1-weighted MR-scans using SPM8 (www.fil.ion.ucl.ac.uk/spm) to determine the location of the implanted electrodes with respect to the individual MRI anatomy. The cortical surface of each individual T1-weighted scan was extracted with FreeSurfer (Dale, Fischl, & Sereno, 1999; Fischl, Sereno, & Dale, 1999). Following the methods for electrode localization suggested by Hermes, Miller, Noordmans, Vansteensel, and Ramsey (2010), electrode positions were first manually marked in a postoperative CT-scan to capture potential brain shifts due to the craniotomy and then assigned to the nearest node on the cortical surface using the CTMR software (Hermes et al., 2010). Finally, the electrode coordinates from the individual anatomies were transformed into MNI-space using SPM8 (Ashburner & Friston, 2009). Anatomical labelling of the electrodes was performed with a custom script that finds the correspondence between the normalized electrode positions and the AAL atlas volumetric parcels (Micheli, Kaping, Westendorff, Valiante, & Womelsdorf, 2015; Tzourio-Mazoyer et al., 2002) for all subjects. Each electrode was mapped to a particular atlas parcel if its position was within 0.5 cm distance from that area.

### Electrode visualization and grouping

We grouped electrodes into areas defined in the AAL atlas to conduct region-based analyses. The grouping resulted in nine areas (Figure 1c): superior temporal gyrus (STG); middle temporal gyrus (MTG), inferior temporal cortex, parahippocampal, and fusiform gyrus (IT-FG-PHC), inferior and superior parietal cortex (PAR), sensory-motor cortex (SM), prefrontal cortex (PFC) and occipital cortex (OCC). In addition, we subdivided STG at Heschl's gyrus into an anterior STG (aSTG), which is less multimodal, and a posterior STG (pSTG) region, which is more multimodal (Ozker et al., 2017). Electrodes were visualized on the surface by means of FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) and custom MatLab scripts (Figure 1c). The brain overlays in Figure 1 and all other figures were generated with a smoothing function with 0.75 cm radius around each electrode to be able to plot the cumulative results of the statistical analysis across subjects.

### Audio and video onset extraction

The concurrent recording of the photodiode signal and the audio waveform in the electrophysiology amplifier allowed us to extract precise timestamps from the experiment, which were successively employed to epoch the data sets. None of the subjects reported delays between the audio and video streams. In detail, we cross-correlated the envelope of the audio track in the video with the actual audio envelope recorded in the amplifier to detect and correct potential delays between audio and video onsets with high precision. This cross-correlation showed a peak at the latency with the best match between audio track in the video and the actual audio envelope and allowed for correction of potential latency jitters in the onset of the audio stimulus presentation. The precision of this method is in the

order of 1 ms as the audio track was recorded at 2 kHz in the ECoG amplifier. We determined the auditory speech onset in each trial as the time when the sound wave was trespassing an amplitude threshold. Note that the video started before the auditory speech onset and that the speaker sometimes moved her mouth prior to auditory speech onset (e.g., Figure 1b upper panel).

### *ECOG data processing and neural activation spectrograms*

Only data from electrodes that did not exhibit epileptic activity and that were not found to be sites of seizure onset were analysed. Bad channels were excluded after visual inspection and we used Infomax-independent component analysis (ICA) implemented in EEGlab to identify and project out line noise, large amplitude drifts, and high-amplitude transients that may originate from head, body or eye movements. Only a few ICA components (<2%) were discarded per session. Subsequently, the data sets were common average rereferenced using FieldTrip (Oostenveld et al., 2011).

In a next step, each channel's amplitude variation was *z*-scored. The *z*-scoring was performed on the entire session, prior to epoching. However, the set of parameters for *z*-scoring, mean and standard deviation, were calculated from the intertrials segments (i.e., having no stimulation). Note that such intervals were not represented by the entire session, but by disjoint intervals. Subsequently, the continuous data were epoched considering data centred around speech onset (time before onset: 2 s, duration after onset: sentence length). Finally, spectrograms of neural activation were calculated from the preprocessed data using FieldTrip toolbox (Oostenveld et al., 2011; Thomson, 1982). Low frequencies (LF: 6–30 Hz) and high frequencies in the high gamma band (HG: 70–250 Hz) were analysed separately via harmonic decomposition of 0.5 s time windows advanced in steps of 50 ms. Data were tapered using a Hanning window for the low frequencies, and four dpss tapers for the high frequencies. The frequency resolution was 2 Hz in the LF band and 10 Hz in the HG band. From the complex coefficients, we calculated the amplitude for each frequency. We chose the lower bound of the LF band at 6 Hz as this includes the temporal speech and mouth envelope modulations (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009) and 30 Hz as the upper bound of the beta band, which is modulated by AV speech stimulation (Schepers et al., 2015). The functional segregation into high- and low-frequency bands is confirmed by the effect of the time-frequency analysis as shown in Supporting Information Figure S3.

### *Stimulus-brain activation correlations*

We correlated the neural activation frequency bins of each electrode's spectrogram separately with the auditory envelopes and the vertical mouth opening time courses (Figure 1b). We used partial correlation analysis to calculate the correlation between brain activation time series and auditory or visual stimulus time series. Partial correlation removes the potential effects of a common third variable from the two variables to be correlated (Kunihiro, Shibata, & Sibuya, 2004). If two time series are correlated, as is likely the case for the auditory and the visual speech envelope, partial correlation can remove the common variance of the visual envelope from the correlation between the auditory and the brain activation time series as well as the common variance of the auditory envelope from the correlation of the visual and the brain activation time series.

In detail, the partial correlation (or PC) used in the study computes a Pearson's (linear) correlation between two time series (in this case the ECoG time series and the audio envelope), controlling for a third variable named *z* (in this case a time series of the video envelope) using MatLab function *rho = partialcorr(x,z),* where x contains the two time series to correlate. In our case, *z* is the other modality's stimulus envelope, also called external variable. The mathematical formulation of partial correlation is equivalent to solving a two associate linear regression problem, then to calculate their residuals and to finally correlate the residuals with Pearson's index. In formulas:

$$\mathbf{w}^*_{ecog} = \underset{w}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (ecog_i - \langle \mathbf{w}, \mathbf{vid} \rangle_i)^2 \right\}$$

$$\mathbf{w}^*_{aud} = \underset{w}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (aud_i - \langle \mathbf{w}, \mathbf{vid} \rangle_i)^2 \right\}$$

where *ecog* and *aud* represent, respectively, the preprocessed intracranial data and the *z*-scored intensity envelope of the audio track, *i* is the trial index, *N* is the number of trials, **w** are the weights of the two linear regressions and **vid** is the external variable to regress (the *z*-scored vertical lip distance). We then calculate the residuals:

$$res_{ecog,i} = ecog_i - \mathbf{w}^*_{ecog}, \mathbf{vid}_i$$

$$res_{aud,i} = aud_i - \mathbf{w}^*_{aud}, \mathbf{vid}_i$$

The partial correlation between audio envelope and ECoG signal is the sample correlation of the two above-mentioned residuals.

### *Statistics to determine whether an electrode is tracking speech envelopes*

To test if the 2D correlogram indicates tracking of speech envelopes by neural responses we employed a cluster based nonparametric randomization test suggested by Maris and Oostenveld (2007) and implemented in the Fieldtrip toolbox. This procedure consists of three steps: (a) generate a reference distribution of correlograms with random brain data, (b) derive a cluster level statistical threshold correcting for family wise error and (c) electrode-wise statistical testing of time-frequency correlation clusters.

(a) First, we generated reference distributions for the expected mean partial correlation in each entry of the correlograms under the assumption that the brain data contain no information about the stimulus envelopes. Therefore, we correlated the stimulus envelopes of each sentence with spectrograms of gaussian random noise with zero mean and standard deviation one, to mimic the *z*-scored brain data, and averaged over the single sentence correlograms to obtain a mean correlogram. We repeated this procedure 500 times to obtain a Monte Carlo distribution of the time-frequency correlogram entries that would be expected if the brain data contain no information about the stimulus envelopes. Note that we obtained reference distributions for the correlograms in the AVstatic and the AVdyn conditions separately, as in the experimental condition from 105 sentences for each condition.

(b) In the second step, the clustering step, we thresholded the correlogram reference distributions at the two-sided 1% percentile and found the time-frequency clusters of correlations that survived the criterion for each of the Monte Carlo correlograms. We then summed the correlations in the clusters and picked the cluster with the highest sum (maxsum criterion). The distribution of the maxsum cluster values of the Monte Carlo correlograms served as the test statistics to control the false alarm rate of the statistical test (Maris & Oostenveld, 2007).

(c) As the third step, we performed statistical testing for the LF and HG bands of each electrode separately. Therefore, we derived clusters in the mean correlograms of the actually measured data using the correlation thresholds derived in the second step and calculated the sum of the correlations in each cluster. We considered an electrode significant if the sum of correlations exceeded in two clusters, the maxsum distribution at an alpha level of $p < 0.05$ (two sided) and the cluster size was at least twice the number of frequencies. The second criterion aims to avoid that short artifactual spikes may create spurious significant clusters.

### *Differences between neural activation spectrograms*

To test for differences between the spectrograms derived from the neural activations in the AVdyn and AVstatic conditions, we first epoched the spectrograms to an interval of −2 to 2.5 s around audio onset and baseline corrected the epochs with the formula (X-BL)/BL (Oostenveld et al., 2011), where BL is the mean of the baseline from −2 to −1.75 s before audio speech onset. Testing for differences between neural activation in the AVdyn and AVstatic was performed for the 2-s interval starting from auditory speech onset and we used a similar procedure as described in the previous section. However, here, we created the reference distributions for the expected differences between spectrograms by permuting labels (AVdyn or AVstatic) across trials and calculated *t*-values for each time-frequency entry in the spectrogram. We thresholded these time-frequency matrices of *t*-values at *t*-values corresponding to $p < 0.01$ (two sided), found time-frequency clusters of t-values crossing this threshold, calculated the sum of *t*-values for each cluster and selected the maximum of the clusters sums. This process was repeated 500 times to obtain the maxsum test statistics to control the false alarm rate of the statistical test (Maris & Oostenveld, 2007). Finally, we performed statistical testing for the LF and HG bands of each electrode separately. Therefore, we derived *t*-value clusters for the differences between spectrograms with the actually measured AVdyn and AVstatic labels and calculated the sum of *t*-values for each cluster. We considered an electrode significant if the sum of *t*-values in a cluster exceeded the alpha level of $p < 0.05$ (two-tailed).

### *Spectrogram of the neural activation for each single condition*

We carried out this analysis with the aim to visualize the time courses of the different frequency bands separately for each condition. To do so, we baseline corrected the magnitude of the time-frequency Fourier coefficients with a relative change approach, according to the formula (X – BL)/BL. The variable X is the time-frequency magnitude (before correction) and BL is the mean of the time-frequency coefficients between −2 and −1.75 s before audio onset. Each baseline correction was applied to each frequency bin separately.

## Results

The population activations recorded with intracranial electrodes can provide information about the auditory and visual stimulus features encoded in the neural activation with high spatial and temporal specificity. The wide dynamic range of the intracranially recorded signals enabled us to report information encoding in different dynamic ranges of the brain signals, namely slow (low frequency, LF: 6–30 Hz) and fast (high gamma, HG: 70–250 Hz) oscillations, which are thought to reflect different functions (Lachaux et al., 2012; Varela, Lachaux, Rodriguez, & Martinerie, 2001).

Using partial correlation analysis, we found that neuronal responses in electrodes over posterior superior temporal gyrus (pSTG) and medial occipital cortex tracked, respectively, the auditory and visual speech envelopes. In addition, we found a cross-modal effect of pSTG tracking visual speech while occipital cortex tracks auditory speech envelope. Using magnitude difference between conditions with and without dynamic video we determined that pSTG is an important location of significant magnitude modulation.

### *Behavioural results*

All subjects showed high performance in detecting the target words. We calculated percentage correct as the number of correct responses over the total number of sentences in the word-to-sentence match task. The percentage of correct answers varied between 88% and 99% (S1: 97.1%, S2 94.7%, S3: 95.7%, S4: 98.6%, S5: 87.6%, S6: 97.1%, and S7: 99.3%) indicating that all subjects attended to the presented sentences. As we found no statistically significant difference between the AVdyn and the AVstatic condition, we report only the overall performance.

### *Electrode localizations*

Figure 1c shows the density of electrodes over the patients' sample (seven volunteers) analysed in the current study. In total, we retained 423 electrodes after rejection of electrodes with artifacts or epileptic activity. There were 180 electrodes over the temporal lobe with 26 of them over posterior superior temporal gyrus (pSTG, 100% in left hemisphere), 17 electrodes over anterior superior temporal gyrus (aSTG, 100% left), 56 over the sensorimotor cortex (95% left), 73 electrodes over occipital lobe (71% left), 84 electrodes over frontal cortex (95% left) and 55 electrodes over parietal cortex (89% left). The anatomical labels are based on the AAL atlas (Tzourio-Mazoyer et al., 2002). In total, about 171 electrodes showed brain activation modulations by auditory or visual speech information in at least one of the analyses presented here (tracking of envelopes or difference across AVstatic and AVdyn conditions).

### *Tracking of auditory and visual speech information*

Our first aim was to identify electrodes that track the amplitude envelope of the auditory speech signal and/or the time course of the visual vertical mouth opening. Electrodes that track both auditory and visual speech information can be considered candidates for audiovisual integration of speech signals as their neural activation contains information about both sensory modalities. However, vertical mouth movements can be correlated with some extent with the amplitude envelope of the auditory speech signal as they open and close the vocal tract. To disambiguate the correlations between brain activation and the auditory or visual stimulus features, we applied partial correlation analysis, which orthogonalizes the variables of interest with respect to the variable controlled for before calculating their correlation. The neural response profile at most electrodes consisted of a LF response that was usually decreased compared to baseline and a broadband HG response that was usually enhanced compared to baseline. Therefore, several electrodes that showed a correlation with the auditory speech envelope exhibited negative correlations for LF responses and positive correlations for HG responses (Figure 2b). This is shown for an example pSTG electrode in Figure 1d (top left panel). For the same electrode, auditory envelope tracking was consistent across trials and HG envelope

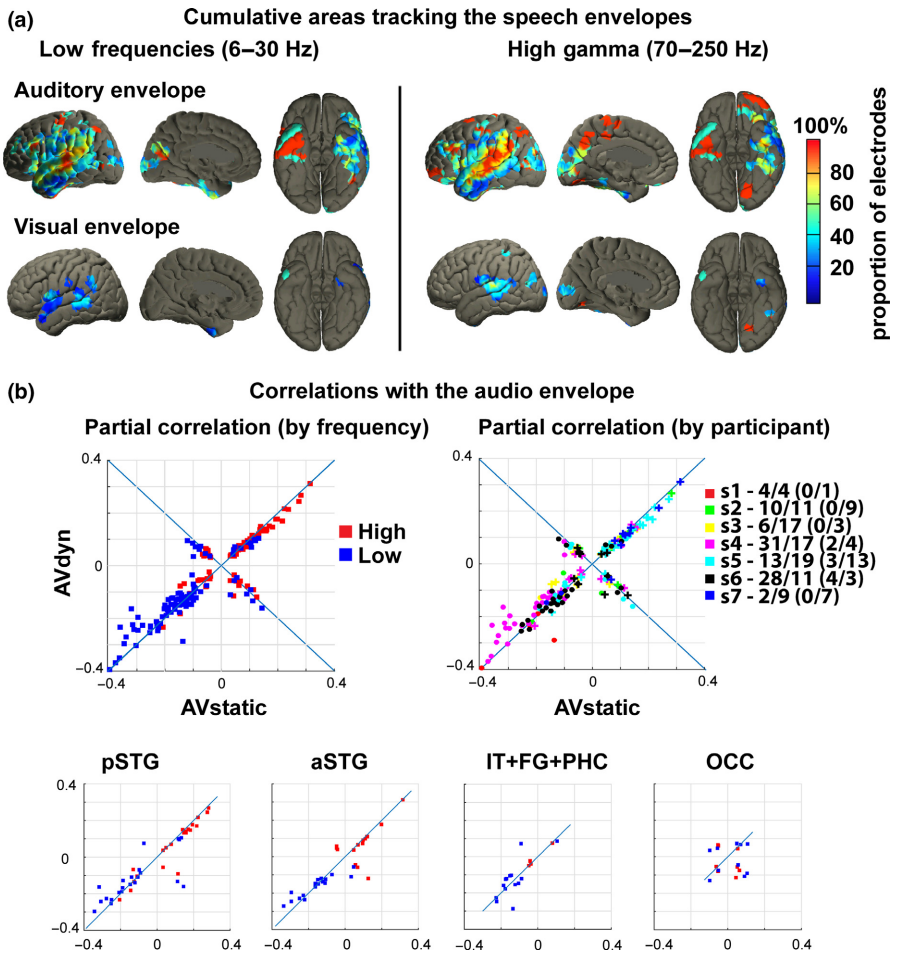


FIG. 2. (a) MNI template localization of areas that showed tracking of the auditory speech envelope (upper row) or the envelope of the visible mouth opening (lower row) as revealed by partial correlation. The colour coded overlay represents the proportion of electrodes per area with significant tracking, calculated over electrodes from all subjects. Note the high proportions over pSTG and medial occipital cortex, indicating tracking of the auditory and the visual envelope of the speech signal. The effect is more pronounced in the HG- than in the LF band. Supporting Information Figure S1 reports the absolute number of participants with significant electrodes and Figure 1C shows the proportion of participants which had electrode coverage in an area. (b) Upper row: The panels show the maxima of partial correlations for all significant electrodes. Each point corresponds to one electrode (electrodes can appear twice in the low and high frequencies). Left: positive correlations are distributed mainly in the HG band (red), negative correlations in the LF band (blue). Right: correlations distinguished by subject. The colours on the right indicate the different subjects (s1 to s7). Crosses indicate HG band and dots LF band. The numbers indicate significant electrodes per subject in the LF/HG bands. The numbers in parentheses indicate the electrodes in the positive quadrant. Lower row: significant electrodes in different cortical areas differentiated per band. Acronyms: aSTG: anterior STG; IT+FG+PHC: inferior temporal, fusiform gyrus, parahippocampal cortex; OCC: occipital cortex; pSTG: posterior STG.

tracking occurred at a consistent lag between the brain response and the auditory stimulus (Figure 1d, top right panel). Figure 1d also shows a single trial HG response time course and the respective auditory envelope with a lag that showed maximal correlation (lag = 160 ms, $r = 0.49$). We investigated the distribution of electrodes in which the neural activation exhibited statistically significant tracking in our population of subjects. In Figure 2a, the auditory envelope tracking (upper row) and visual envelope tracking (bottom row) are separately depicted for LF and HG responses. We found auditory envelope tracking most consistently across subjects in the HG activity in electrodes over STG. In most subjects ($n = 6$), the HG response as well as the LF amplitude variations in electrodes over pSTG tracked the auditory envelope (Figure 2a, Table 1b). This finding is in concordance with previous studies showing that posterior STG/STS is involved in auditory speech processing (Fedorenko & Thompson-Schill, 2014; Hickok & Poeppel, 2007; Skeide & Friederici, 2016). Importantly, several electrodes over the pSTG tracked the vertical mouth movements in addition to the auditory envelope and simultaneous tracking for both modalities was found more consistently in the HG compared to the LF response (HG: 8/26, 31% of electrodes; LF: 3/26, 12% of electrodes; Table 1). These findings indicate that HG neural activity in pSTG carries information about the auditory speech envelope as well as the corresponding mouth movements. Only one electrode in aSTG tracked both the audio envelope and the visible mouth movements in LF and HG neural activation variations, while a large proportion of electrodes in aSTG tracked the auditory speech envelope (HG: 12/17, 71% of electrodes; LF: 12/17, 71% of electrodes; Table 1). Electrodes over anterior inferior temporal cortex tracked mainly the auditory envelope both in LF (auditory: 20, 19%) and fast HG (auditory 13, 13%; mouth movements: 2, 2%) amplitude variations.

In Figure 2b, we note that AVdyn and AVstatic correlations do not differ substantially from each other, and the effect has a comparable trend for both hemispheres (see Supporting Information Figure S4). Note that the different signs of partial correlation reflect

TABLE 1. Number of aSTG, pSTG, occipital and inferior temporal electrodes that showed auditory, video or auditory and video envelope tracking in the LF and HG responses

| | Audio (A) tracking | Video (V) tracking | A+V tracking |
|---|---|---|---|
| (a) Low-frequency (LF) envelope tracking (%) | | | |
| pSTG | 21/26 (81) | 3 (12) | 3 (12) |
| aSTG | 12/17 (71) | 1 (6) | 1 (6) |
| Occipital | 7/22 (32) | – | – |
| Inferior temporal cortex | 20/104 (19) | – | – |
| (b) High gamma (HG) envelope tracking (%) | | | |
| pSTG | 23 (88) | 8 (31) | 8 (31) |
| aSTG | 12 (71) | 1 (6) | 1 (6) |
| Occipital | 7 (32) | 2 (9) | 2 (9) |
| Inferior temporal cortex | 13 (13) | 2 (2) | – |

*Note*. Total number of electrodes recorded in each region is provided in Table 1a and the percentage of electrodes that showed tracking is given in brackets.

solely the different signs of brain oscillatory magnitudes (stimulus envelope always positive), represented mainly by increases in the HG band and decreases in the LF band after speech onset (see Supporting Information Figure S3). Off-diagonal elements in Figure 2b represent correlations with discrepant sign for AVdyn and AVstatic, but also showing similar magnitude. The change in sign is due to an increase in brain magnitude (in one band) for one condition at speech onset, but a decrease for the other condition.

In addition, two electrodes over medial occipital cortex, in the vicinity of the calcarine sulcus, tracked the auditory envelope and the vertical mouth movements in the HG response (HG: 2/22, 9%; LF: none) and five additional occipital electrodes showed only auditory envelope tracking (Table 1b). This result, seen in the AVstatic condition only, is particularly interesting as we observe audio tracking independent of the visual input in a region typically sensitive to visual stimuli. However, as the occipital electrode coverage was low, it should be taken with caution (see also Supporting Information Table S1).

### *Modulation of auditory-evoked neural activation levels by visual information*

The observed effect of neural tracking of the audio envelope in both AVdyn and AVstatic conditions (Figure 2b) does not necessarily imply magnitude differences in neural HG and LF brain responses between the two conditions. A previous study has shown differences in HG responses between audiovisual clear and audiovisual noisy speech in pSTG (Ozker et al., 2017). Therefore, we tested for which electrodes the HG and LF responses differed significantly in magnitude between the AVstatic and the AVdyn stimuli. Figure 3a depicts the distribution of electrodes in our population of subjects with statistically significant differences between the two conditions in the 2-s interval after speech onset overlaid on a standard MNI brain ($p < 0.05$, two-tailed and corrected for multiple comparisons). Similar to the tracking results, the neural activation amplitude in electrodes over pSTG, occipital cortex, aSTG, and anterior inferior temporal cortex is altered in both HG and LF responses when visual information about mouth movements is added to the audio speech.

Figure 3b shows the effects of adding visual speaker information on neural activation amplitude at greater temporal detail. The bars show the proportion of significant electrodes separately for five different 500-ms-long intervals, with the first interval starting 500 ms before auditory speech onset (T0-T4, Figure 1b), four dynamic bands (4–15 Hz, 15–30 Hz, 30–70 Hz, and 70–250 Hz), for different anatomical brain areas, and accumulated over subjects. In addition to that, Figure 3b shows the direction of the effects (increase: AVdyn > AVstatic; decrease: AVdyn<AVstatic). The reason for the subdivision in five time epochs was to show the difference in activation during the AVstatic and AVdyn condition right after the speech stimulus onset and at later time intervals in the sentence (Figure 1b). The plots indicate that a higher number of electrodes show differences during the earlier intervals in the sentence. We found the strongest effects in pSTG and occipital cortex where the highest proportions of electrodes were modulated. Figure 3c depicts example time courses of the HG (top) and LF (bottom) responses at four pSTG electrodes from a single subject for the AVdyn and AVstatic condition. The amplitude differences in the HG band between the AVstatic and AVdyn conditions (four electrodes for subject 4, and four pSTG electrodes for subject 5, results shown for subject four only in Figure 3c) appear to be accompanied by a speech onset latency difference. In addition, we found smaller proportions of significant electrodes in anterior inferior temporal cortex (IT+FG+PHC), and in aSTG. Importantly, of the areas that showed activation amplitude differences mainly pSTG and the occipital cortex showed auditory in concomitance with visual envelope tracking. This effect was also observed to a lesser extent for inferior temporal cortex. The presence of both envelope tracking and amplitude changes in pSTG is necessary in an audiovisual speech integration area and further supports the notion that auditory and visual speaker information is represented in this area. Note, however, that in contrast to the partial correlation analysis, activation amplitude modulations by adding dynamic visual input are harder to interpret. It was also notable that additional visual speech information increased HG and reduced LF amplitudes in all three regions and all electrodes with significant signal changes. This suggests different coding strategies in different dynamic bands, in concordance with previous work (van Kerkoerle et al., 2014; Lachaux et al., 2012; Varela et al., 2001).

## Discussion

Extending the results of previous studies on audiovisual speech processing, our study demonstrates the simultaneous representation of speech-specific auditory and visual envelope information in pSTG and to a lesser extent in occipital cortex. These findings add further support for the notion that pSTG/STS is an important multimodal brain area in speech processing and that visual areas also represent auditory speech dynamics.

In natural audiovisual speech, the auditory signal and the visual signal vary dynamically and are correlated (Chandrasekaran et al., 2009; Park et al., 2016). Therefore, it is important to partial out the visual signal when the relationship between the auditory speech signal and the neuronal response is investigated and to partial out the auditory signal when the relationship with the visual signal is investigated. With this approach, we could show which brain regions track only the auditory speech signal, only the visual speech signal, or both.

Several studies have now shown that different auditory speech features of the dynamic speech input are reflected in brain signals recorded noninvasively (EEG, MEG) and in neural responses directly recorded from human cortex (ECoG), particularly pSTG (Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Peelle, Gross, & Davis, 2013; Zion Golumbic et al., 2013). Our findings of auditory

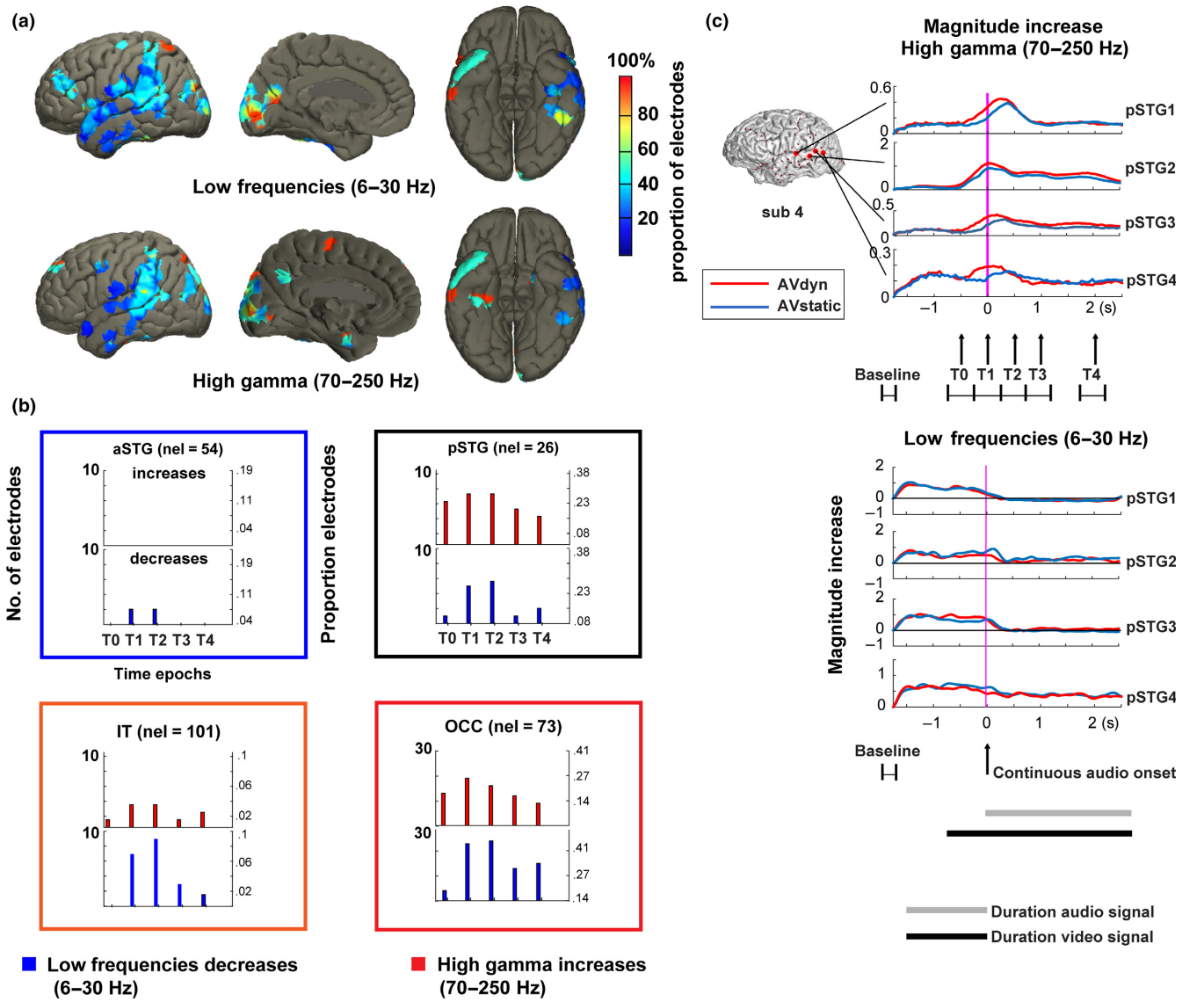


FIG. 3. Neural activation differences between AVdyn and AVstatic stimuli. (a) Cortical distribution of electrodes that show significant effects (corrected for multiple comparisons) of adding speaker mouth movements (AVdyn) to auditory speech (AVstatic) over total number of electrodes from all subjects. The effect is calculated from spectrograms of neural activation in two dynamic bands and aggregated over subjects. Activation in pSTG and in medial visual cortex was modulated by additional visual speech information in both dynamic bands. Supporting Information Figure S2 reports the absolute number of participants with significant electrodes and Figure 1c shows the density of participants per area. (b) Time resolved effects in four anatomical areas defined in Figure 1c. The bars indicate the proportion of electrodes in each area that show a significant difference in the neural activation spectrograms of the respective dynamic ranges (6–30 Hz, 70–250 Hz) in one of five (T0-T4) 500 ms time intervals starting before speech onset (T0) and ending 2 s after auditory speech onset (T4). Nel = total number of electrodes in the respective brain area. The strongest effect, in terms of proportion of electrodes with significant effects, is in pSTG and occipital cortex (OCC). The inferior anterior temporal cortices show smaller effects. Note that adding visual speaker information decreases amplitudes in the LF bands, below 30 Hz and increase neural activation in HG bands above 70 Hz. (c) Examples of average time courses across trials for the two experimental conditions (AVdyn = red, AVstatic = blue) in four electrodes of one subject (S4). The time courses are shown for the HG (top) and LF (bottom) responses to the continuous speech input. The baseline time interval is −2 to −1.75 s with respect to speech onset, and it is chosen to make sure that no audio input is presented within its time frame. Note that the AVdyn response onsets precede the AVstatic latency in most electrodes.

speech envelope tracking in a large proportion of pSTG electrodes are in line with these previous studies. In addition to auditory envelope tracking in pSTG, we observed auditory envelope tracking in aSTG and occipital cortex. This is consistent with a recently proposed division in STG between auditory responses more anteriorly and multisensory responses more posteriorly (Ozker et al., 2017).

As we partialed out the visual envelope information in the correlation analysis with the auditory speech envelope, the auditory envelope tracking in occipital cortex is particularly interesting. It shows that early visual cortex already contains information specifically on the auditory speech stream and extends previous findings of phase reset and audiovisual interaction effects in visual cortex (Mercier et al., 2013; Noesselt et al., 2007) by providing evidence that also auditory speech information is represented in the neural responses in medial occipital cortex. This is likely moderated by top-down influences, presumably from frontal cortex, which modulate visual cortex during speech perception according to top-down task demands (Schepers et al., 2015).

Tracking of the auditory envelope was not only most pronounced in pSTG, but this brain region also most consistently showed robust lip movement tracking. All pSTG electrodes showing auditory envelope tracking also showed visual envelope tracking. Thus, both dynamic auditory and visual speech features seem to be represented in pSTG. Our results demonstrate that both HG and LF neural responses show auditory and visual speech tracking and suggest that tracking is more pronounced in the high-frequency range, which has been directly linked to local neuronal activity (Lachaux et al., 2012; Ray & Maunsell, 2011). Auditory envelope, but not visual envelope tracking has previously been shown in the HG band response of ECoG recordings in pSTG during audiovisual speech perception (Zion Golumbic et al., 2013). For auditory-only speech, enhanced phase locking to the auditory speech signal has been shown with increased speech comprehension in temporal cortex with MEG (Peelle et al., 2013). We report simultaneous tracking of both auditory and visual speech signals in pSTG. In line with our results of simultaneous tracking of both auditory and visual speech in pSTG, a recent fMRI study demonstrated that regions in pSTS that preferentially responded to the mouth region of dynamic visual stimuli also showed strong responses to voices and preferred vocal to non-vocal sounds (Zhu & Beauchamp, 2017). The dynamic integration of the two speech signals may therefore occur in pSTG/STS, which should result in better speech perception performance when the auditory signal is degraded. This question could be addressed in a future study relating auditory and visual speech envelope tracking in pSTG to speech comprehension performance of degraded audiovisual speech. However, as correlation patters do not differ drastically in the AVstatic and AVdyn conditions (Figure 2b), brain activity related to audio intensity probably does not get modulated by the video input. Given the significant difference of magnitude across conditions, we speculate that brain responses to video input could interact with audio responses for higher level features, such as formants, phonemes or words/sentences, as already previously documented (Chang et al., 2010; McGurk & MacDonald, 1976).

Several electrodes in visual cortex showed only auditory envelope tracking or both lip movement and auditory envelope tracking. The visual lip movement tracking in visual cortex we observed is in line with the MEG study by Park et al. (2016) showing visual lip movement tracking in visual cortex during audiovisual speech presentation, albeit only in the lower frequencies (<11 Hz). Park et al. (2016) partialed out the auditory speech envelope, here we also partialed out the visual lip movement information and demonstrate that the auditory envelope information is represented in visual cortex as well. Our finding of speech tracking in visual cortex also relates to a previous ECoG study, which showed that specifically concurrent auditory speech but not auditory noise reduced visual cortex responses to visual speech (Schepers et al., 2015). Based on this previous finding and our finding of auditory and visual envelope tracking in medial occipital cortex, we would expect more pronounced visual speech envelope tracking in visual cortex when the auditory speech signal is severely degraded or missing because the listener then needs to rely on the visual speech signal for comprehension. Supporting Information Table S1 shows audio entrainment of both occipital and interior temporal electrodes in both AVdyn and AVstatic condition, which argues for a multimodal function of visual cortex (Murray et al., 2016).

In addition to the speech envelope tracking, we show that the HG response magnitude is increased in pSTG for AVdyn compared to AVstatic speech and that the LF response in pSTG is decreased. In general, HG band responses were enhanced compared to baseline and LF responses were decreased compared to baseline and the dynamic visual speech information increased the HG response and reduced the LF response (enhanced decrease). Enhanced HG band responses and decreased LF responses compared to baseline have been previously reported to audiovisual speech with ECoG (Rhone et al., 2016; Schepers et al., 2015; Uno et al., 2015) and seem to be a general response profile in population electrophysiological responses to sensory stimulation (e.g., Scheeringa, Koopmans, van Mourik, Jensen, & Norris, 2016; Siegel, Donner, Oostenveld, Fries, & Engel, 2008). Magnitude changes in neural brain responses reported here, particularly in the broad HG range, which has been linked to neuronal spiking activity (Ray & Maunsell, 2011), might underlie BOLD response changes seen in previous fMRI studies (Mukamel et al., 2005). The increased magnitude of HG responses in pSTG for AVdyn vs. AVstatic speech is in line with fMRI studies showing enhanced pSTG/STS BOLD responses to audiovisual compared to auditory-only speech (Stevenson & James, 2009; Ye, Rüsseler, Gerth, & Münte, 2017). Conversely, the role of LF band activity in sentence processing might be related to disruption of the perceptual status-quo (Engel & Fries, 2010). If we consider the current task set to be unrelated to motor activity and follow this interpretation, we could speculate that LF suppression is related to disruption of the current perceptual set and predicts the probability of new processing demands. Alternatively, low-frequency inhibition might be due to endogenous working memory processing (Spitzer & Haegens, 2017), or termination of inhibition (Jensen & Mazaheri, 2010), in contrast to high gamma excitation due to population spiking (Whittingstall & Logothetis, 2009).

As opposed to pSTG, we observed no magnitude differences in HG responses (AVdyn vs. AVstatic) in aSTG and only a single electrode in aSTG showed visual lip movement tracking in the LF and HG frequency range, pointing to a dissociation in audiovisual speech processing between pSTG and aSTG. For LF and HG responses, we found auditory envelope tracking in aSTG, albeit in fewer electrodes than in pSTG. This dissociation is in line with the results by Ozker et al. (2017), who reported different response profiles with respect to the magnitude of HG band responses in aSTG and pSTG to audiovisual clear and noisy speech with greater responses to auditory clear speech in aSTG. Moreover, they observed greater HG response variability to audiovisual speech with an auditory noisy component in aSTG than pSTG suggesting that the additional visual speech information improves the auditory speech representation of the noisy auditory speech input in pSTG, but not aSTG.

One could ask whether the difference in magnitude between the two conditions could be caused by differences in attentional load due to the audiovisually incongruent stimuli (Arnal, Morillon, Kell, & Giraud, 2009) in the AVstatic condition. One would expect cross-modal attentional effects if auditory signal was not informative. However, in the AVstatic condition, the auditory signal is informative. Even if we expected a mismatch response in pSTG, for example greater response in AVstatic (when subjects hear the voice but don't see the accompanying mouth movements), Figure 3b shows the opposite of this expectation. We see that AVdyn > AVstatic responses in both visual cortex and in pSTG, hence we infer that magnitude differences might be due to an influence of the video input on the audio speech perception.

The mechanisms of audiovisual interaction in speech processing in the human brain are currently not very clear. With depth electrodes in human auditory cortex, Besle et al. (2008) measured event-related potentials to audiovisual syllables and observed audiovisual interactions in secondary auditory cortex from 30 ms after auditory speech onset reflected in changes in magnitude of the ERPs. A study in nonhuman primates observed a latency reduction

in spiking activity to audiovisual compared to unisensory speech but did not find consistent effects on the magnitude of the response in early auditory cortex (Chandrasekaran, Lemus, & Ghazanfar, 2013). One possibility is that visual speech information, which precedes auditory information by 100–200 ms in natural speech (Chandrasekaran et al., 2009) induces changes in the ongoing neural oscillations in auditory cortex so that the auditory speech signal is more robustly represented (Peelle & Sommers, 2015; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). Specific visual speech features may, however, not be represented in early auditory cortex. Further knowledge about audiovisual integration could be gained in future studies by adding a visual only condition, which would allow for tests of enhancement of multisensory compared to unisensory response magnitudes (Stein & Stanford, 2008).

Our findings suggest that visual speech information about the most visible articulatory movements is represented only at later processing stages in temporal cortex. This timing information on visual lip movements available in pSTG may support auditory speech processing, for example, through enhanced temporal prediction of the auditory speech input (Peelle & Sommers, 2015).

In sum, our study demonstrates auditory envelope tracking and visual mouth movement tracking in electrodes over pSTG suggesting that multimodal pSTG has temporal information about visible mouth movements during speech utterances that may help to improve speech processing under adverse conditions. We also found, although to weaker extent, tracking of auditory envelopes in visual cortex suggesting enhanced communication between posterior temporal cortex and visual cortex during multimodal audiovisual integration (Bernstein & Liebenthal, 2014; Nath & Beauchamp, 2011; Noesselt et al., 2007). However, further research is needed to determine how visual brain areas and pSTG/pSTS communicate during audiovisual speech comprehension and how the information on the concurrent visual and auditory speech signal is integrated within and across regions.

## Supporting Information

Additional supporting information can be found in the online version of this article:

Fig. S1. (Upper panel) MNI template localization of areas that showed tracking of the auditory speech envelope (upper row) or the envelope of the visible mouth opening (lower row) as revealed by partial correlation. The color coded overlay represents the absolute number of participants per area with significant electrodes over all local electrodes. Note the high density over pSTG and medial occipital cortex, indicating tracking of the auditory and the visual envelope of the speech signal. The effect is more pronounced in the HG- than in the LF-band. See also Fig. 2 in the main text for a comparison with the proportion of significant electrodes per area. (Lower panel) Proportion of significant electrodes in different cortical areas differentiated per band (left right) and per modality of tracking (Auditory and Visual envelopes). Note that these results are equivalent to the results reported in Table 1 in the main text. Acronyms: pSTG, posterior STG; aSTG, anterior STG; IT, inferior temporal + fusiform gyrus + parahippocampal cortex; OCC, occipital cortex.

Fig. S2. Anatomical distribution of participants that show significant effects (corrected for multiple comparisons) of adding speaker mouth movements (AVdyn) to auditory speech (AVstatic) calculated from spectrograms of neural activation in two dynamic bands and aggregated over subjects.

Fig. S3. Left: time course of a pSTG electrode (subject 4) in different frequency bands. Center: two time-frequency plots for different pSTG electrodes (subject 4 and 5). The 0 on the *x*-axis denotes speech onset. Right: zoom in of the LF band (same plots as Center). Note that the decrease in magnitude after speech onset shows a consistent pattern in the LF and HG bands. The effect of the LF band is clearly band-limited and distinguishable from a (non-functional) 1/f effect magnitude decay along the frequency axis.

Fig. S4. The panels show the maxima of partial correlations for all significant electrodes in the left hemisphere (left) and in the right hemisphere (right). Each point corresponds to one electrode.

Fig. S5. Cumulative plot representing the proportion of subjects with audio-tracking areas, with respect to the total number of electrodes in that area.

Fig. S6. Individual subject's brains (S1 to S7) representing the electrodes tracking the audio envelope.

Fig. S7. Tracking correlograms (partial correlation) for the AUDIO envelope.

Table S1. Number of visual cortex electrodes (and subject number in parentheses) that present audio envelope tracking in both the AVstatic and the AVdyn conditions.

## Acknowledgements

We would like to thank Virginie van Wassenhove for helpful scientific discussions and Susann Bräuer for support with the video recordings. This research was supported by grant number SFB/TRR 31 A16 from the German Research Foundation (DFG) to JWR and Veterans Administration Clinical Science Research and Development Merit Award Number 1I01CX000325-01A1 to DY and NIH R01NS065395 to MSB.

## Conflict of Interest

The authors declare that they have no conflict of interest to declare.

## Data Accessibility

Anonymized data can be made available on request.

## Authors' Contributions

Drafting and revision of the manuscript: CM, IMS, MO, MSB, JWR. Data analysis: CM, JWR, MSB. Stimulus construction and acquisition of data: IMS, MO, DY, MSB, JWR.

## References

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

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, **14**, 797–801. https://doi.org/10.1038/nn.2810

Ashburner, J., & Friston, K. J. (2009). Computing average shaped tissue probability templates. *NeuroImage*, **45**(2), 333–341. https://doi.org/10.1016/j.neuroimage.2008.12.008

Bastiaansen, M., Magyari, L., & Hagoort, P. (2010). Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *Journal of Cognitive Neuroscience*, **22**(7), 1333–1347. https://doi.org/10.1162/jocn.2009.21283

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, **41**, 809–823. https://doi.org/10.1016/S0896-6273(04)00070-4

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, **8**, 386.

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

humans. *Journal of Neuroscience*, **28**, 14301–14310. https://doi.org/10.1523/JNEUROSCI.2875-08.2008
Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, **5**, 341–347.
Chandrasekaran, C., Lemus, L., & Ghazanfar, A. A. (2013). Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E4668–E4677. https://doi.org/10.1073/pnas.1312518110
Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, **5**, e1000436.
Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, **13**, 1428–1432. https://doi.org/10.1038/nn.2641
Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, **35**, 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015
Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, **36**, 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016
Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, **9**, 179–194. https://doi.org/10.1006/nimg.1998.0395
Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*, **28**, 8–311.
Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, **19**, 158–164. https://doi.org/10.1038/nn.4186
Engel, A. K., & Fries, P. (2010). Beta-band oscillations–signalling the status quo? *Current Opinion in Neurobiology*, **20**(2), 156–165. https://doi.org/10.1016/j.conb.2010.02.015
Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, **18**, 120–126. https://doi.org/10.1016/j.tics.2013.12.006
Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, **9**, 195–207. https://doi.org/10.1006/nimg.1998.0396
Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, **15**, 511–517.
Hermes, D., Miller, K. J., Noordmans, H. J., Vansteensel, M. J., & Ramsey, N. F. (2010). Automated electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of Neuroscience Methods*, **185**, 293–298. https://doi.org/10.1016/j.jneumeth.2009.10.005
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, **8**, 393–402. https://doi.org/10.1038/nrn2113
Holdgraf, C. R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., … Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, **7**, 13654. https://doi.org/10.1038/ncomms13654
Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, **11**, 61. https://doi.org/10.3389/fnsys.2017.00061
Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Front Human Neuroscience*, **4** (4), 186.
Kunihiro, B., Shibata, R., & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, **46**(4), 657–664. https://doi.org/10.1111/j.1467-842X.2004.00360.x
Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E., & Crone, N. E. (2012). High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, **98**, 279–301. https://doi.org/10.1016/j.pneurobio.2012.06.008
Lee, H., & Noppeney, U. (2011). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *Journal of Neuroscience*, **31**, 11338–11350. https://doi.org/10.1523/JNEUROSCI.6510-10.2011
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, **164**, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**(5588), 746–748. https://doi.org/10.1038/264746a0. PMID 1012311
Mercier, M. R., Foxe, J. J., Fiebelkorn, I. C., Butler, J. S., Schwartz, T. H., & Molholm, S. (2013). Auditory-driven phase reset in visual cortex: Human electrocorticography reveals mechanisms of early multisensory integration. *NeuroImage*, **79**, 19–29. https://doi.org/10.1016/j.neuroimage.2013.04.060
Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, **485**, 233–236.
Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, **343**, 1006–1010. https://doi.org/10.1126/science.1245994
Micheli, C., Kaping, D., Westendorff, S., Valiante, T. A., & Womelsdorf, T. (2015). Inferior-frontal cortex phase synchronizes with the temporal-parietal junction prior to successful change detection. *NeuroImage*, **119**, 417–431. https://doi.org/10.1016/j.neuroimage.2015.06.043
Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., & Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science*, **309**, 951–954. https://doi.org/10.1126/science.1110913
Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2016). The multisensory function of the human primary visual cortex. *Neuropsychologia*, **83**, 161–169. https://doi.org/10.1016/j.neuropsychologia.2015.08.011
Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, **31**, 1704–1714. https://doi.org/10.1523/JNEUROSCI.4853-10.2011
Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, **59**(1), 781–787. https://doi.org/10.1016/j.neuroimage.2011.07.024
Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Journal of Neuroscience*, **27**, 11431–11441. https://doi.org/10.1523/JNEUROSCI.2252-07.2007
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci*, **2011**, 156869.
O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Front Human Neuroscience*, **10**, 679.
Ozker, M., Schepers, I. M., Magnotti, J. F., Yoshor, D., & Beauchamp, M. S. (2017). A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. *Journal of Cognitive Neuroscience*, **29**, 1044–1060. https://doi.org/10.1162/jocn_a_01110
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.
Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, **23**, 1378–1387. https://doi.org/10.1093/cercor/bhs118
Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, **68**, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006
Quandt, F., Reichert, C., Hinrichs, H., Heinze, H.-J., Knight, R. T., & Rieger, J. W. (2012). Single trial discrimination of individual finger movements on one hand: a combined MEG and EEG study. *Neuroimage*, **59**, 3316–3324.
Ray, S., & Maunsell, J. H. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, **9**, e1000610.
Rhone, A. E., Nourski, K. V., Oya, H., Kawasaki, H., Howard, M. A., & McMurray, B. (2016). Can you hear me yet? An intracranial investigation of speech and non-speech audiovisual interactions in human cortex. *Language, Cognition and Neuroscience*, **31**, 284–302. https://doi.org/10.1080/23273798.2015.1101145
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of

speech comprehension in noisy environments. *Cerebral Cortex*, **17**(5), 1147–1153.

Scheeringa, R., Koopmans, P. J., van Mourik, T., Jensen, O., & Norris, D. G. (2016). The relationship between oscillatory EEG activity and the laminar-specific BOLD signal. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 6761–6766. https://doi.org/10.1073/pnas.1522577113

Schepers, I. M., Schneider, T. R., Hipp, J. F., Engel, A. K., & Senkowski, D. (2013). Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. *NeuroImage*, **70**, 101–112. https://doi.org/10.1016/j.neuroimage.2012.11.066

Schepers, I. M., Yoshor, D., & Beauchamp, M. S. (2015). Electrocorticography reveals enhanced visual cortex responses to visual speech. *Cerebral Cortex*, **25**, 4103–4110. https://doi.org/10.1093/cercor/bhu127

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, **12**, 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270** (5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Siegel, M., Donner, T. H., Oostenveld, R., Fries, P., & Engel, A. K. (2008). Neuronal synchronization along the dorsal visual pathway reflects the focus of spatial attention. *Neuron*, **60**, 709–719. https://doi.org/10.1016/j.neuron.2008.09.010

Skeide, M. A., & Friederici, A. D. (2016). The ontogeny of the cortical language network. *Nature Reviews Neuroscience*, **17**, 323–332. https://doi.org/10.1038/nrn.2016.23

Spitzer, B., & Haegens, S. (2017). Beyond the status quo: A role for beta oscillations in endogenous content (Re)activation. *eNeuro*, **4**(4), https://doi.org/10.1523/ENEURO.0170-17.2017.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, **9**(4), 255–266. https://doi.org/10.1038/nrn2331

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage*, **44**, 1210–1223. https://doi.org/10.1016/j.neuroimage.2008.09.034

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212–215. https://doi.org/10.1121/1.1907309

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, **70**, 1055–1096. https://doi.org/10.1109/PROC.1982.12433

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, **15**, 273–289. https://doi.org/10.1006/nimg.2001.0978

Uno, T., Kawai, K., Sakai, K., Wakebe, T., Ibaraki, T., Kunii, N., … Saito, N. (2015). Dissociated roles of the inferior frontal gyrus and superior temporal sulcus in audiovisual processing: Top-down and bottom-up mismatch detection. *PLoS ONE*, **10**, e0122580. https://doi.org/10.1371/journal.pone.0122580

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 14332–14341. https://doi.org/10.1073/pnas.1402773111

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 1181–1186. https://doi.org/10.1073/pnas.0408949102

Varela, F., Lachaux, J. P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, **2**, 229–239. https://doi.org/10.1038/35067550

Wang, L., Jensen, O., van den Brink, D., Weder, N., Schoffelen, J. M., Magyari, L., … Bastiaansen, M. (2012). Beta oscillations relate to the N400 m during language comprehension. *Human Brain Mapping*, **33**(12), 2898–2912. https://doi.org/10.1002/hbm.21410

Whittingstall, K., & Logothetis, N. K. (2009). Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. *Neuron*, **64** (2), 281–289. https://doi.org/10.1016/j.neuron.2009.08.016

Xiong, X., & Torre, F. D. la (2013). *Supervised descent method and its applications to face alignment*. In 2013 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539.

Ye, Z., Rüsseler, J., Gerth, I., & Münte, T. F. (2017). Audiovisual speech integration in the superior temporal region is dysfunctional in dyslexia. *Neuroscience*, **356**, 1–10. https://doi.org/10.1016/j.neuroscience.2017.05.017

Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *Journal of Neuroscience*, **37**, 2697–2708. https://doi.org/10.1523/JNEUROSCI.2914-16.2017

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, **77**, 980–991. https://doi.org/10.1016/j.neuron.2012.12.037