MEG Activity in Visual and Auditory Cortices Represents Acoustic Speech-Related Information during Silent Lip Reading

Bröhl, Felix; Keitel, Anne; Kayser, Christoph

Published in:
eNeuro

DOI:
10.1523/ENEURO.0209-22.2022

Publication date:
2022

eNeuro

Research Article: New Research | Cognition and Behavior

MEG activity in visual and auditory cortices represents acoustic speech-related information during silent lip reading

https://doi.org/10.1523/ENEURO.0209-22.2022

Cite as: eNeuro 2022; 10.1523/ENEURO.0209-22.2022

Received: 30 May 2022
Accepted: 6 June 2022

MEG activity in visual and auditory cortices represents acoustic speech-related information during silent lip reading

Short title: Processing of visual speech in the brain

Felix Bröhl[1*], Anne Keitel[2], Christoph Kayser[1]

[1]*Department for Cognitive Neuroscience, Faculty of Biology, Bielefeld University, Universitätsstr. 25, 33615, Bielefeld, Germany*

[2]*Psychology, University of Dundee, Scrymgeour Building, Dundee DD1 4HN, UK*

[*] **Corresponding author:** Felix Bröhl (felix.broehl@uni-bielefeld.de)

Highlights

- Visual and auditory cortices represent unheard acoustic features during lip reading
- Auditory cortex emphasizes the acoustic envelope
- Visual cortex emphasizes a pitch signature
- Tracking of unheard features in auditory cortex is associated with behavior

Number of pages: 30

Number of figures: 5

Number of tables: 2

Number of words: Abstract 209, Introduction 749, Discussion 1843

Declaration of competing interest

We declare no conflict of interest.

Acknowledgement

This work was supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC, BB/L027534/1) and the European Research Council (ERC-2014-CoG; grant No 646657)

Abstract

Speech is an intrinsically multisensory signal and seeing the speaker's lips forms a cornerstone of communication in acoustically impoverished environments. Still, it remains unclear how the brain exploits visual speech for comprehension. Previous work debated whether lip signals are mainly processed along the auditory pathways or whether the visual system directly implements speech-related processes. To probe this, we systematically characterized dynamic representations of multiple acoustic and visual speech-derived features in source localized MEG recordings that were obtained while participants listened to speech or viewed silent speech. Using a mutual-information framework we provide a comprehensive assessment of how well temporal and occipital cortices reflect the physically presented signals and unique aspects of acoustic features that were physically absent but may be critical for comprehension. Our results demonstrate that both cortices feature a functionally specific form of multisensory restoration: during lip reading they reflect unheard acoustic features, independent of co-existing representations of the visible lip movements. This restoration emphasizes the unheard pitch signature in occipital cortex and the speech envelope in temporal cortex and is predictive of lip reading performance. These findings suggest that when seeing the speaker's lips, the brain engages both visual and auditory pathways to support comprehension by exploiting multisensory correspondences between lip movements and spectro-temporal acoustic cues.

Significance statement

Lip reading is central for speech comprehension in acoustically impoverished environments. Recent studies show that the auditory and visual cortex can represent acoustic speech features from purely visual speech. It is still unclear, however, what information is represented in these cortices and if this phenomenon is related to lip reading comprehension. Using a comprehensive conditional mutual information analysis applied to magnetoencephalographic data, we demonstrate that signatures of acoustic speech arise in both cortices in parallel, even when discounting for the physically presented stimulus. In addition, the auditory but not the visual cortex activity was related to successful lip reading across participants.

Keywords

Speech entrainment, lip reading, audio-visual, speech tracking, language, MEG

1. Introduction

Speech is an intrinsically multisensory stimulus that can be conveyed via acoustic and visual signals. It remains debated how the brain exploits the information derived from visual speech (Besle et al., 2008; Calvert et al., 1997; Calvert and Campbell, 2003; Grant and Seitz, 2000). One view is that the visual system directly contributes to establishing speech representations (Bernstein et al., 2011; O'Sullivan et al., 2017; Ozker et al., 2018), as oro-facial movements provide temporal information that can be predictive of concurrent acoustic signals and allow mapping visual cues onto phonological representations (Campbell, 2008; Lazard and Giraud, 2017). The visual cortex tracks dynamic lip signals (Park et al., 2016) and, as suggested recently, may also directly 'restore' the acoustic envelope of the visually presented speech (Hauswald et al., 2018; Suess et al., 2022). Another view is that visual speech is mainly represented in regions of the auditory pathways, possibly exploiting speech-specific processes of this system. Along this line, a recent study suggested that the early auditory cortex may also be capable of reflecting the unheard acoustic envelope of a spoken narrative (Bourguignon et al., 2020). Importantly, the evidence that visual speech is reflected along both auditory and the visual pathways may not be mutually exclusive, as both may contribute to a supramodal frame of reference for speech (Arnal et al., 2009; Rauschecker, 2012).

To probe the respective involvement of visual and auditory cortices in representing visual speech, many previous studies presented syllables or isolated words as stimuli (Calvert et al., 1997; Calvert and Campbell, 2003; Pekkola et al., 2005). However, these results come short of how they translate to continuous or natural speech. Furthermore, many studies did not probe a direct link to behavioral performance, leaving it unclear whether potential cerebral representations derived from visual speech are behaviorally relevant (Bourguignon et al., 2020; Ludman et al., 2000; Mégevand et al., 2020). The latter can be particularly challenging given that pure lip reading performance for every-day speech is often low (Altieri et al., 2011; Grant and Seitz, 2000).

The present study rests on the assumption that probing the roles of visual and auditory cortices in representing visual speech requires data from a paradigm based on continuous speech with carefully controlled levels of lip reading performance. In previous work we established such a paradigm and collected MEG data from participants during a word recognition task based on syntactically similar sentences that were presented either purely acoustically or purely visually. In the auditory condition participants were presented with the acoustic signal embedded in background noise, while in the visual condition they watched the muted speaker. The individual sentences were constructed from a closed-set of linguistic items with a common syntactic structure, similar to matrix-sentences used in standardized hearing assessment (Hagerman, 1982; Kollmeier et al., 2015). With this we achieved a comparable level of word recognition performance during auditory- and visual-only conditions and verified that this dataset allows linking neural representations of lexical information and speech dynamics to behavior (Keitel et al., 2020, 2018).

We leverage this paradigm to probe the roles of visual and auditory pathways in representing visual speech and facilitating lip reading performance. This led us to formulate the two following questions: First, whether representations of restored (e.g. unheard acoustic) features are independent of those physically present (e.g. lip movements). Second, we asked whether these representations of restored features are tied to word recognition performance. To be able to compare cerebral signatures of visual and acoustic speech-derived features, we rooted this analysis on the following systematic assessment: we quantified how well source-localized MEG signals track multiple acoustic and visual speech-derived features independently of each other, both when these features are physically present (e.g. the lip contour when watching the speaker) or absent (e.g. the pitch contour when watching the speaker). We focused this analysis on two regions of interest centered on early auditory and visual cortices, which previous studies have implied in supporting lip reading (Bourguignon et al., 2020; Hauswald et al., 2018; Park et al., 2016). The analysis was based on a mutual information approach that has been used in previous work to probe dynamic speech representations and which is well suited to address the statistical dependency between multiple variables (Daube et al., 2019; Keitel et al., 2018). Our results show that both occipital and temporal regions reflect unheard acoustic speech-derived features independently of the physically present lip movements. This 'restoration' of acoustic information in the temporal, but not the occipital, cortex is predictive of word recognition performance across participants.

2. Materials and Methods

The data analyzed in this study has been collected and analyzed in previous studies (Keitel et al., 2020, 2018). The analyses conducted here pose new questions and provide novel results beyond the previous work.

2.1 Participants and data acquisition

Data was collected from 20 native British-English speaking participants (9 female, age 23.6 ± 5.8 years mean ± SD). Due to prominent environmental artefacts in the MEG recordings, data from two participants were excluded from further analysis. Thus, the analyzed data was from 18 participants (7 female, age 24 ± 6.0 years mean ± SD). All participants were screened to exclude hearing impairment prior to data collection using the quick hearing check questionnaire (Koike et al., 1994), had normal or corrected-to-normal vision and were all right-handed (Oldfield, 1971). All participants provided written informed consent and received monetary compensation of 10 £/h. The experiment was approved by the College of Science and Engineering, University of Glasgow (approval number 300140078) and conducted in compliance with the Declaration of Helsinki.

MEG data was collected using a 248-magnetometer whole-head MEG system (MAGNES 3600 WH, 4-D Neuroimaging) with a sample rate of 1 kHz. Head positions were measured at the beginning and end of

each run, using five coils placed on the participants' heads. Coil positions were co-digitized with the participant's head-shape (FASTRAK®, Polhemus Inc., VT, USA). Participants were seated in an upright position in front of a screen. Visual stimuli were displayed with a DLP projector at 25 frames per second, a resolution of 1280 × 720 pixels, and covered a visual field of 25 × 19 degrees. Acoustic stimuli were transmitted binaurally through plastic earpieces and 370-cm long plastic tubes connected to a sound pressure transducer and were presented in stereo at a sampling rate of 22,050 Hz.

2.2 Stimulus material

The stimulus material comprised two structurally equivalent sets of 90 unique closed-set English sentences. Specifically, along the idea of matrix-style sentences using in standardized hearing assessment (Hagerman, 1982; Kollmeier et al., 2015), each sentence was constructed with the same sequence of linguistic elements, the order of which can be described with the following pattern [filler phrase, time phrase, name, verb, numeral, adjective, noun]. One such sentence for example was 'I forgot to mention (filler phrase), last Thursday morning (time phrase) Mary (name) obtained (verb) four (numeral) beautiful (adjective) journals (noun)'. For each element, a list of 18 different options was created and sentences were constructed so that each single element was repeated ten times. Sentence elements were randomly combined within each set of 90 sentences. This procedure yielded 180 structurally similar but distinct sentences. To measure word recognition performance for each sentence, a target word was defined in each sentence: either the adjective (first set of sentences) or the numeral (second set). Sentences lasted on average 5.4 ± 0.4 s (mean ± SD, ranging from 4.6 s to 6.5 s) and lasted a total of approximately 22 minutes. The speech material was spoken by a male British actor, who was tasked to speak clearly and naturally and to move as little as possible while speaking to assure that the lips center stayed at the same place in each video frame. Audiovisual recordings were gathered with a high-performance camcorder (Sony PMW-EX1) and an external microphone in a sound attenuating booth.

Participants were presented with audio-only (A-only), audiovisual or visual-only (V-only) speech material in three conditions (Keitel et al., 2020). However, for the present analysis we only focus on the A-only and V-only conditions, as in these one can best dissociate visual- and auditory-related speech representations given that only one physical stimulus was present. Furthermore, during the AV condition word recognition performance was near-ceiling (Keitel et al., 2020), making it difficult to link cerebral and behavioral data. Because performance would have been at ceiling with clear speech in the A-only condition, the acoustic speech was embedded in environmental noise. This noise for each trial was generated by randomly selecting 50 individual sounds from a set of sounds recorded from natural, everyday sources or scenes (e.g. car horns, talking people, traffic). These sounds were then added together to create a distracting noise scene for the duration of each trial. For each participant the individual noise level was further adjusted, as described previously (Keitel et al., 2020). This resulted in an average performance of approximately 70%

correct for both A-only and V-only conditions and allowed us to dissociate between correct and incorrect word recognition.

2.3 Experimental Design

Each participant was presented with each of the 180 sentences in three conditions (A-only, V-only and AV). The order of the conditions was fixed for all participants as A-only, AV and then V-only. This order exposed the participants to the stimuli twice before the lip reading task, which helped to increase performance and render it comparable to the A-only task. Each condition was divided into 4 blocks of 45 sentences each, with two blocks being 'adjective' and two 'number' blocks. For each participant, the order of sentences within each block was randomized. The first sentence of each block was a 'dummy' trial that was subsequently excluded from analysis. During each trial, participants either fixated a dot (in A condition) or a small cross overlaid onto the mouth of the speaker's face (in V condition). In the A condition, each sentence was presented as the respective audio recording, i.e. the spoken sentence, together with the background noise. In the V condition, only the video of the speaker's face was presented clearly and no sound was present. After each trial, four words were presented as response options (either four adjectives or four written numbers) on the screen and participants had to indicate using a button press which word they had perceived. Inter-trial intervals were set to last about two seconds.

2.4 Preprocessing of stimulus material

From the stimulus material we extracted the following auditory and visual features. Based on previous literature that demonstrated robust encoding of the amplitude envelope, it's temporal derivative and the fundamental frequency of speech, we derived these features from the acoustic speech recordings (Bröhl and Kayser, 2021; Oganian and Chang, 2019; Teoh et al., 2019). To derive the broadband envelope we filtered the acoustic waveform into twelve logarithmically spaced bands between 0.1 and 10 kHz (zero-phase 3rd order Butterworth filter with boundaries: 0.1, 0.22, 0.4, 0.68, 1.1, 1.7, 2.7, 4.2, 6.5, 10 kHz) and subsequently took the absolute value of the Hilbert transform for each band. The broadband amplitude envelope (hereon termed *aud env*) was then derived by taking the average across all twelve band-limited envelopes and was subsequently down-sampled to 50 Hz. We computed the slope of this broadband envelope (hereon termed *aud slope*) by taking its first derivative. To characterize the pitch contour we extracted the fundamental frequency (hereon termed *aud pitch*) over time using the Praat software ('to Pitch' method with predefined parameters) (Boersma and van Heuven, 2001). This was done using the original acoustic waveform at a sampling rate of 22,050 Hz. The resulting pitch contour was again down sampled to 50 Hz. All three acoustic features together are labelled *AudFeat* in the following.

In a similar fashion we derived the horizontal opening of the lips, the area covered by the lip opening, and its derivative from the video recordings. The lips were detected based on the color of the lips in the video material using a custom-made algorithm. From these we determined the contour of the lip opening based

on luminance values and deriving connected components from these (Giordano et al., 2017). The results were visually inspected to ensure accurate tracking of the lips. From this segmentation of the lip opening we derived the total opening (in pixels) (hereon termed *lip area*) and estimates of the respective diameters along the horizontal axes (hereon termed *lip width*): these were defined between the outermost points along the horizontal axis. These signals were initially sampled at the video rate of 25 fps. As for the auditory features, we computed the slope of the lip area (hereon termed *lip slope*). The time series of these visual features were then linearly interpolated to a sample rate of 50 Hz. Because the horizontal and vertical mouth openings are partially correlated with each other and with the total mouth opening, we selected the total area and the horizontal width as signals of interest, as the latter is specifically informative about the acoustic formant structure (Plass et al., 2020). We grouped the total lip area, it's temporal derivative and the lip-width as signatures of lip features (*LipFeat*), which are of the same dimensionality as the acoustic features (*AudFeat*) described above.

For comparison with previous studies (Chandrasekaran et al., 2009; Giordano et al., 2017; Hauswald et al., 2018; Park et al., 2016) we quantified the power spectra of these features and their cross-coherences using MATLAB's 'pwelch' and 'mscoher' functions using a window length of 1 s with 50% overlap and otherwise predefined parameters. The resulting spectra were log transformed and averaged across sentences. To visualize the cross-coherences we first obtained key frequency ranges of interest from our main results (c.f. Fig. 3) and averaged the coherences within two ranges of interest (0.5 - 1 Hz and 1 - 3 Hz).

2.5 MEG preprocessing

Preprocessing of MEG data was carried out using custom MATLAB scripts and the FieldTrip toolbox (Oostenveld et al., 2011). Each experimental block was processed separately. Individual trials were extracted from continuous data starting 2 s before sound onset and until 10 s after sound onset. The MEG data were denoised using a reference signal. Known faulty channels (N = 7) were removed. Trials with SQUID jumps (3.5% of trials) were detected and removed using FieldTrip procedures with a cut-off z-value of 30. Data were band-pass filtered between 0.2 and 150 Hz using a zero-phase 4th order Butterworth filter and subsequently down sampled to 300 Hz before further artefact rejection. Data were visually inspected to find noisy channels (4.37 ± 3.38 on average across blocks and participants) and trials (0.66 ± 1.03 on average across blocks and participants). Noise cleaning was performed using independent component analysis with 30 principal components (2.5 components removed on average). Data were further down sampled to 50 Hz and bandpass filtered between 0.8 and 30 Hz using a zero-phase 3rd order Butterworth filter for subsequent analysis.

2.6 MEG source reconstruction

Source reconstruction was performed using Fieldtrip, SPM8, and the Freesurfer toolbox based on T1-weighted structural magnetic resonance images (MRIs) for each participant. These were co-registered to

the MEG coordinate system using a semi-automatic procedure (Gross et al., 2013; Keitel et al., 2017). MRIs were then segmented and linearly normalized to a template brain (MNI space). We projected sensor-level time series into source space using a frequency-specific linear constraint minimum variance (LCMV) beamformer (Van Veen et al., 1997) with a regularization parameter of 7% and optimal dipole orientation (singular value decomposition method). The grid points had a spacing of 6 mm, thus resulting in 12,337 points. For whole-brain analyses, a subset of grid points corresponding to cortical gray matter regions only was selected (using the AAL atlas, Tzourio-Mazoyer et al., 2002), yielding 6,490 points in total. Within these we defined temporal and occipital regions of interest (ROI) based on the brainnetome atlas (Yu et al., 2011). The individual ROIs were chosen based on previous studies that demonstrate the encoding of acoustic and visual speech features in occipital and superior temporal regions (Di Liberto et al., 2018; Giordano et al., 2017; Keitel et al., 2020; Teng et al., 2018). As temporal ROI we included Brodmann area 41/42, caudal area 22 (A22c), rostral area 22 (A22r) and TE1.0 and TE1.2. As occipital ROI we defined the middle occipital gyrus (mOccG), occipital polar gyrus (OPC), inferior occipital gyrus (iOccG) and the medial superior occipital gyrus (msOccG).

2.7 MEG analysis

The questions outlined in the introduction require quantifying how well the source reconstructed MEG data reflect the visual and or acoustic features. For this we relied on a previously established and validated mutual information (MI) framework (Ince et al., 2017). The analysis relies on the notion that a significant temporal relation between a cerebral signal and sensory features is indicating the cerebral encoding (or tracking) of the respective features in temporally entrained brain activity (Bröhl and Kayser, 2021; Keitel et al., 2018; Park et al., 2016). In the following we use the term 'tracking' when referring to such putative cerebral representations characterized using MI (Obleser and Kayser, 2019). To quantify the tracking of a given stimulus feature, or of a feature group, we concatenated the trial-wise MEG data and features along the time dimension and filtered these (using 3rd order Butterworth IIR filters) into typical frequency bands used to study dynamic speech encoding: 0.5 - 1 Hz, 1 - 3 Hz, 2 - 4 Hz, 3 - 6 Hz and 4 - 8 Hz (and 0.5 - 8 Hz). These were chosen to cover the typical modulation spectra of these features (Fig. 1C,D) and similar to previous work (Bröhl and Kayser, 2021; Etard and Reichenbach, 2019; van Bree et al., 2020; Zuk et al., 2021). The first 500 ms of each sentence were discarded to remove the influence of the transient sound-onset response. To compute the MI between filtered MEG and stimulus features, we relied on a complex-valued representation of each signal, which allowed us to include both the amplitude and phase information in the analysis: we first derived the analytic signal of both the MEG and stimulus feature(s) using the Hilbert transform and then calculated the MI using the Gaussian copula approach including the real and imaginary part of the Hilbert signals (Daube et al., 2019; Ince et al., 2017).

In a first step, we used this framework to visualize the tracking of AudFeat and LipFeat within the entire source space (Fig. 2A,B). This was mainly done to assert that the predefined ROIs used for the subsequent

analysis indeed covered the relevant tracking of these features. This analysis relied on a frequency range from 0.5 to 8 Hz and a range of stimulus-to-brain lags from 60 to 140 ms after stimulus onset. As a second step, we then quantified the tracking of auditory or visual features and their dependencies specifically within these ROIs and individual frequency bands (Fig. 3,4,5). To facilitate these analyses, we first determined the optimal lags for each feature, ROI and frequency band, given that the encoding latencies may differ between features and regions (Giordano et al., 2017). For this we determined at the group-level and for each set of features (i.e. AudFeat and LipFeat) and for each ROI and frequency band the respective lag yielding the largest group-level MI value (across participants and both A-only and V-only trials). This was done by computing the MI between each set of features and the MEG in a range of lags between 0 and 500 ms in 20 ms steps. For the subsequent analyses, we used these optimal lags and computed averaged MI values in a time window of -60 to 60 ms around these lags (computed in 20 ms steps).

The first question of this study as outlined in the introduction concerns the tracking (MI) of individual feature groups in temporal and occipital ROIs and the two experimental conditions, respectively, and more so if a given ROI reflects a given feature (e.g. the unheard acoustic envelope) independently of the physically present other feature (e.g. the visible lip movements in the visual-only condition). To quantify whether the tracking of each feature group (in a given ROI and frequency band) is statistically redundant with (or possibly complementary to) the other group, we calculated the conditional mutual information (CMI) between MEG and one feature group, partialling out the respective other group (Fig. 3, CMI values) (Giordano et al., 2017; Ince et al., 2017). Specifically, the CMI measure allows us to quantify the unique information shared between a variable and the MEG while controlling for the information provided by the conditional variable. Mathematically, it can be described as

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$$

where I denotes the conditional mutual information and H the joint entropies between combinations of variables X, Y and the conditional variable Z. Similarly, we also determined the CMI between the MEG and each individual feature, obtained by partialling out all other visual and auditory features (Fig. 4). To be able to compare the MI and CMI estimates directly, we ensured that both estimates had comparable statistical biases. To achieve this, we effectively derived the MI as a conditional estimate, in which we partialled out a statistically-unrelated variable. That is, we defined

$$MI(feature\ ;\ MEG) - MI(feature\ ;\ MEG|\ time\_shifted\_feature)$$

Here, time_shifted_feature is a representation of the respective feature(s) with a random time lag and hence no expected causal relation to the MEG. Each MI estimate was obtained by averaging this estimate over 2,000 repetitions of a randomly generated time-shifted feature vector. To render the (conditional) MI estimates meaningful relative to the expectation of zero MI between MEG and stimulus features, we furthermore subtracted an estimate of the null-baseline of no systematic relation between signals. This was obtained by computing (conditional) MI values after randomly time-shifting the stimulus feature(s) and averaging the resulting surrogate MI estimates over 100 randomizations.

2.8 Relating MI to word recognition performance

The behavioral performance for each participant and condition was obtained as the percent correctly (PC) reported target words (obtained in a 4-choice task). To probe the second question of whether the tracking of restored features relates to word recognition performance we relied on partial regression. Specifically, we probed the linear relation of word recognition performance and feature tracking across participants while accounting for potential spurious correlations between these due to variations in the individual signal-to-noise ratio in each participants' MEG data. We predicted the PC in the visual-only trials based on i) the individual MI for *aud env* in the temporal ROI and the MI for *aud pitch* in the and occipital ROI as the primary variables of interest, and ii) the tracking of LipFeat (MI) in the occipital ROI in visual trials and iii) the tracking of AudFeat in the temporal ROI in auditory trials. The last two serve as potentially confounding variables, as they provide a proxy to the overall SNR of the speech and lip tracking in the respective dataset. By focusing on *aud env* / *aud pitch* in the temporal/occipital ROIs respectively, we predicted task performance based on the individual features that were most associated with the tracking of AudFeat (c.f. Fig. 4C,D). To establish these regression models, we z-scored the MI values of interest (variables i - iii) and the PC across participants. For the confounding variables, we applied the z-scoring for each frequency band and subsequently averaged the z-scored values across bands. For each frequency band, we created a single model containing all target and confounding variables. From the respective models we obtained the significance of each predictor of interest. Furthermore, we compared the predictive power of this full model with that of a reduced model not featuring the predictors of interest (variable i). From the likelihoods of each model we derived the relative Bayes factor (BF) between these based on the respective BIC values obtained from each model. For visualization we used partial residual plots using the procedure described by Velleman and Welsch (Velleman and Welsch, 1981). This procedure was applied to each individual feature of interest (i.e. *aud env* and *aud pitch*).

2.9 Statistical analysis

Statistical testing of mutual information data was based on a non-parametric randomization approach incorporating corrections for multiple comparisons (Nichols and Holmes, 2003). To test whether the group-level median MI (or CMI) values were significantly higher than expected based on the null hypothesis of no systematic temporal relation between sensory features and MEG, we proceeded in a similar fashion as in previous work (Bröhl and Kayser, 2021; Giordano et al., 2017): we obtained a distribution of 2,000 MI values between randomly time-shifted MEG and the stimulus vectors, while keeping the temporal relation of individual features to each other constant. This distribution was obtained for each participant, frequency band, feature group (AudFeat and LipFeat), ROI (temporal, occipital) and condition (A-only, V-only) separately. To correct for multiple comparisons, we generated a single random distribution by pooling the randomly generated MI values across all dimensions except frequency bands, given that the MI values decreased considerably across bands (c.f. Fig. 3), and selecting the maximum 2000 values, thereby creating

a random maximum null distribution (Nichols and Holmes, 2003). We then tested the group-level median against the 99th percentile of this maximum distribution as a significance threshold, which effectively implements a one-sided randomization test at $p < 0.01$ corrected for all dimensions except frequency bands. To test for differences between MI and CMI values for a given condition, band and ROI, we also used a permutation approach combined with a Wilcoxon signed-rank test: first, we established the respective true Wilcoxon z-statistic between MI and CMI values; then we created a distribution of surrogate z-statistics under the null hypothesis of no systematic group-level effect, obtained by randomly permuting the labels of MI and CMI values 5,000 times. From this we obtained the maximum across features, bands, ROIs and conditions to correct for multiple comparisons and used the 99th percentile of this randomization distribution to determine the significance of individual tests.

The CMI values for individual features in Figure 4 were compared using a one-way repeated measure Kruskal-Wallis rank test, followed by a post-hoc Tukey Kramer multiple comparison. We used the same procedure to test for differences between CMI values in the sub-areas composing each ROI (Table 1). To test CMI values between hemispheres, we used a Wilcoxon signed rank test (Table 2). The resulting p-values were corrected for false discovery rate using the Benjamini-Hochberg procedure within each set of comparisons (Benjamini and Hochberg, 1995). In all tests an alpha level of $\alpha < 0.01$ was deemed significant. For all statistical tests we provide exact p-values, except for randomization tests where the approximate p-values were smaller than the inverse of the number of randomizations.

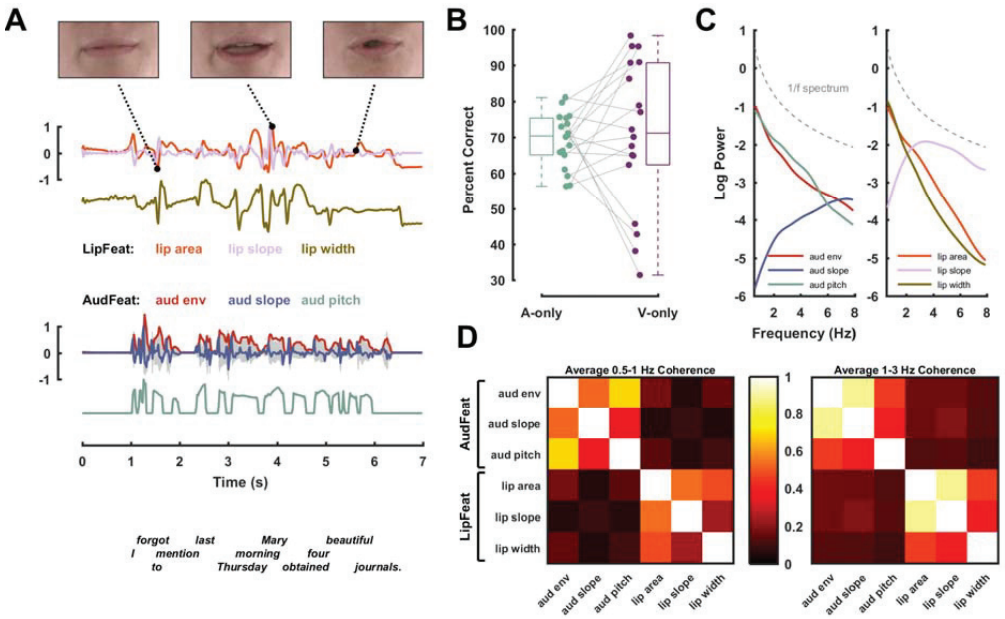

Fig. 1. Stimulus material and experimental methodology. *Acoustic and visual features were extracted from audiovisual speech material and were used to quantify their cerebral tracking during audio-only and visual-only presentations. (**A**) The stimulus material consisted of 180 audiovisual recordings of a trained actor speaking individual*

*English sentences. For visualization here only the mouth is shown, but participants were presented with the entire face. From the video recordings we extracted three features describing the dynamics of the lip aperture: the area of lip opening (lip area), its slope (lip slope), and the width of lip opening (lip width); collectively termed 'LipFeat'. From the audio waveform we extracted three acoustic features: the broadband envelope (aud env), its slope (aud slope), and a measure of dominant pitch (aud pitch); collectively termed 'AudFeat'. (**B**) Trial-averaged percent correctly (PC) reported target words in auditory (A-only) and visual-only (V-only) conditions, with dots representing individual participants. (**C**) Logarithmic power spectra for individual stimulus features. For reference, a 1/f spectrum is shown as a dashed grey line. (**D**) Coherence between pairs of features averaged within two predefined frequency bands (0.5 - 1 Hz left; 1 - 3 Hz right, see Methods for details).*

3. Results

3.1 Acoustic and visual features are tracked in temporal and occipital cortices

Participants were presented with either spoken speech (in A-only trials) or a silent video of the speaking face (in V-only trials) and were asked to report a target word for each sentence in a 4-choice word recognition task. The behavioral data show that participants were well able to detect the correct word both during acoustic speech embedded in noise and during lip reading and achieved overall similar levels of performance in both conditions (Fig. 1B, median fraction correct responses for A-only = 0.7, V-only = 0.71; n = 18). To quantify the tracking of relevant features, we defined three auditory (AudFeat) and three visual (LipFeat) features respectively based on the acoustic waveform and the lip trajectory (Fig. 1A). An analysis of their temporal coherences revealed that they were coherent in the frequency bands of interest (e.g. 1 - 3 Hz, envelope-lip area coherence of ~0.2) (Fig. 1D). The overall pattern of coherence and the degree of temporal relation between acoustic features and lip movements in the present material is comparable with those in other datasets (Chandrasekaran et al., 2009; Giordano et al., 2017; Hauswald et al., 2018; Park et al., 2016).

Previous work has shown that in the dataset analyzed here temporal and occipital brain regions reflect auditory and visual speech signals respectively (Keitel et al., 2020). We extend this observation to the entire group of acoustic (AudFeat; Fig. 2A) or lip features (LipFeat; Fig. 2B) using a mutual information (MI) approach. The whole-brain maps show the expected prevalence of acoustic (visual) tracking in temporal (occipital) regions. Given that our main questions concerned the tracking of features specifically in occipital and temporal brain regions, we focused the subsequent work on atlas-based regions of interest (Fig. 2C; the temporal ROI shaded in mint and the occipital ROI shaded in purple, see methods for details).

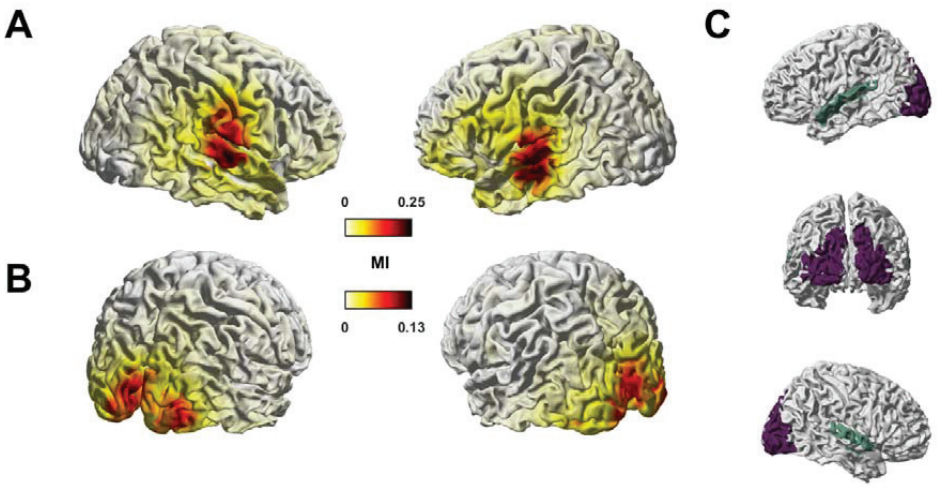

Fig. 2. Tracking of auditory and visual features in MEG source space. *The figure shows group-level median MI values for auditory (AudFeat; panel **A**) and lip features (LipFeat; panel **B**) in the frequency range from 0.5 - 8 Hz (n = 18 participants). (**C**) Colored shading indicates regions of interest: temporal region in mint includes Brodmann area 41/42, caudal area 22 (A22c), rostral area 22 (A22r) and TE1.0 and TE1.2; occipital region in purple includes middle occipital gyrus (mOccG), occipital polar gyrus (OPC), inferior occipital gyrus (iOccG) and medial superior occipital gyrus (msOccG).*

3.2 Temporal and occipital cortex represent acoustic speech features during silent lip reading

To address the question of whether temporal and occipital cortices represent auditory and visual speech features during lip reading, we performed a comprehensive analysis of the tracking of both sets of features across a range of frequency bands during auditory (A-only) and visual (V-only) conditions (MI values; Figure 3). To further quantify whether the tracking of each feature group is possibly redundant with the tracking of the respective other feature group, we derived CMI values for each feature group, obtained by partialling out the respective other group (CMI values). By comparing MI and CMI values we can test, for example, whether the temporal ROI tracks the unheard speech envelope during silent lip reading also when discounting for the actually presented lip trajectory. In the following we discuss the results per sensory modality and region of interest.

As expected, when listening to speech (A-only), the temporal ROI significantly tracks auditory features (AudFeat) in all frequency bands tested (Fig. 3, top row, red MI data; non-parametric randomization test, all bands: $p < 5 \times 10^{-5}$). This tracking persists when discounting potential contributions of the not-seen visual features (red CMI data all individually significant: $p < 5 \times 10^{-5}$), though in some bands the CMI values were significantly lower than the unconditional MI (Wilcoxon signed rank test comparing MI vs. CMI, 2 - 4 Hz: z =

3.59, 3 - 6 Hz: z = 3.68, 4 - 8 Hz: z = 3.42, all comparisons: $p < 2 \times 10^{-5}$). During the same auditory trials, lip features are only marginally reflected in the temporal ROI, as shown by low but significant MI and CMI values above 1Hz (Fig. 3, top row, cyan MI and CMI data; all bands above 1 Hz: $p < 5 \times 10^{-5}$). This tracking of visual features was significantly reduced when partialling out the physically presented auditory features (2 - 4 Hz: z = 3.59, 3 - 6 Hz: z = 3.68, 4 - 8 Hz: z = 3.42, all comparisons: $p < 2 \times 10^{-5}$).

During lip reading (V-only), the temporal ROI tracks the unheard auditory features, particularly below 1 Hz (Fig. 3, 2nd row, red MI data; all bands: $p < 5 \times 10^{-5}$). Except in the 2 - 4 Hz range, the temporal ROI tracks the unheard AudFeat to a similar degree as when discounting the actually presented visual signal (significant red CMI values, all bands: $p < 5 \times 10^{-5}$) as there were no significant differences between MI and CMI values except one band (2 - 4 Hz: z = 3.42, $p < 1 \times 10^{-4}$, see asterisks). The physically presented lip movements during these V-only trials were also tracked significantly in the temporal ROI (Fig. 3, 2nd row; cyan MI and CMI data, 1 - 6 Hz: $p < 5 \times 10^{-5}$) but the CMI values were only marginally above chance level, suggesting that genuine visual representations in the temporal region is weak.

As expected, during lip reading (V-only) the occipital ROI tracks lip features (LipFeat) across frequency bands (Fig. 3, bottom row, cyan MI values; all bands: $p < 5 \times 10^{-5}$). Again, this tracking persists after partialling out the non-presented acoustic features (cyan CMI values; all bands: $p < 5 \times 10^{-5}$), although the CMI values were significantly lower than the MI (all bands above 1 Hz: $z \geq 3.72$, $p < 2 \times 10^{-5}$). This indicates some redundancy between the tracking of the physically present lip trajectory and that of the unheard auditory features. Confirming this, occipital tracking of the physically presented lip signals emerges in parallel with that of the non-presented auditory features (Fig. 3, bottom panel, red MI data; all bands: $p < 5 \times 10^{-5}$). This occipital tracking of unheard auditory features was significantly reduced when partialling out the lip signal (MI vs. CMI data; all bands above 1 Hz: $z \geq 3.72$, $p < 2 \times 10^{-5}$) but remained statistically significant (red CMI data; all bands: $p < 5 \times 10^{-5}$).

Finally, when listening to speech (A-only), the occipital ROI shows significant but weak tracking of auditory (Fig. 3, 3rd row, red MI data; 1 - 6 Hz: $p < 5 \times 10^{-5}$) and visual features (cyan MI data; only 3 - 6 Hz: $p < 5 \times 10^{-5}$), suggesting that purely acoustic signals have a weak influence on the occipital brain region.

Collectively, these results show the expected representations of auditory features in temporal cortex during listening to speech and of lip features in occipital cortex during lip reading. In addition, they reveal that during lip reading, both temporal and occipital regions represent unheard auditory features and do so independently of co-existing representations of the physically presented lip movements. In the auditory cortex this 'restoration' of auditory signals prevails in the low delta band (0.5 - 1 Hz), in the visual cortex this emerges in multiple bands.

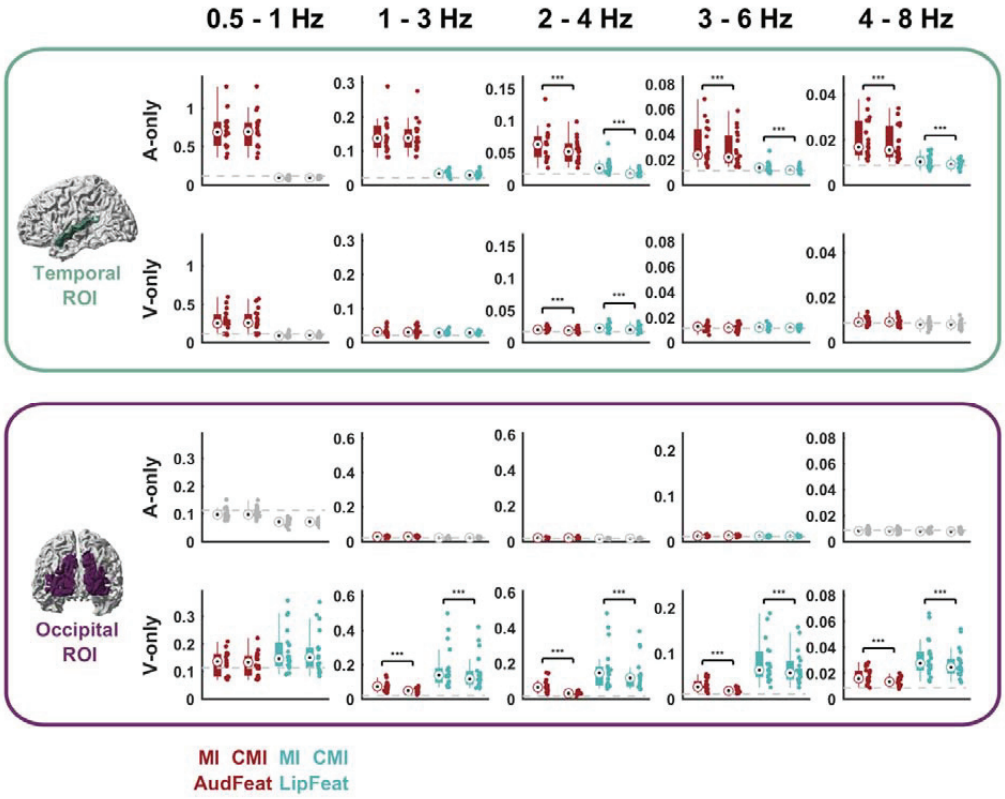

Fig. 3. Feature tracking across regions of interest and conditions. *For both conditions (A-only and V-only) and ROIs (temporal and occipital) the figure illustrates the strength of feature tracking for presented and physically not-present features (MI values) and the respective strength of tracking after partialling out the respective other feature group (CMI values). Each panel depicts (from left to right) the MI for AudFeat, the CMI for AudFeat partialling out LipFeat, the MI for LipFeat, and the CMI for LipFeat partialling out AudFeat. Dots represent individual participants (n = 18). Bars indicate the median, 25th and 75th percentile. The grey dashed line indicates the 99th percentile of the frequency-specific randomized maximum distribution correcting for all other dimensions. Conditions below a group-level significance threshold of 0.01 are greyed out. Brackets with asterisks indicate significant differences between MI and CMI values, based on a Wilcoxon signed-rank test (* p < 0.01, ** p < 0.005, *** p < 0.001). Units for MI and CMI are in bits.*

To obtain an estimate of the effect size of the restoration of the unheard AudFeat during lip reading we expressed the respective CMI values relative to those of the tracking of the respectively modality-preferred inputs of each ROI (Fig. 4A,B): for the temporal region the tracking of AudFeat during A-only trials and for the occipital region the tracking of LipFeat during V-only trials. In the temporal ROI, the restoration effect size, i.e. the tracking of AudFeat during lip reading, was about a third as strong as this feature's tracking while directly listening to speech (Fig. 4A; $AudFeat_{V\text{-}only}$ /$AudFeat_{A\text{-}only}$; 0.5 - 1 Hz: median = 0.37, 1 - 3 Hz:

447 median = 0.24). In the occipital ROI, the tracking of AudFeat was about half as strong or stronger compared
448 to the tracking of lip features when seeing the speaker (Fig. 4B; $AudFeat_{V\text{-}only}$ /$LipFeat_{V\text{-}only}$; 0.5 - 1 Hz:
449 median = 0.84, 1 - 3 Hz: median = 0.4). Albeit smaller than the tracking of the respective modality-preferred
450 sensory inputs, the restoration of unheard auditory features still results in a prominent signature in
451 temporally aligned brain activity in both cortices.

452 **3.3 Feature tracking is bilateral and prevails across anatomical brain areas**

453 Having established the tracking of auditory and lip features in both temporal and occipital ROIs, we probed
454 whether this tracking is possibly lateralized in a statistical sense and whether it potentially differs among
455 the individual anatomical areas grouped into temporal and occipital ROIs respectively. While these analyses
456 do not directly concern our main hypotheses outlined in the introduction, the issue of lateralization is
457 pervasive in the literature on speech, and hence is addressed here for the sake of completeness. For this
458 analysis we focused on the conditional tracking of each feature group. Comparing CMI values among
459 anatomical areas (averaged across hemispheres) for each ROI (occipital, temporal), frequency band (0.5 - 1
460 and 1 - 3 Hz), condition and feature group revealed a significant effect of area for AudFeat tracking in the
461 temporal ROI during A-only trials (Table 1; 0.5 - 1 Hz: $\chi^2(3) = 27.02$, $p = 4.7 \times 10^{-6}$, $\varepsilon^2 = 0.35$; 1 - 3 Hz: $\chi^2(3) =$
462 29.62, $p = 2.7 \times 10^{-6}$, $\varepsilon^2 = 0.39$; p-values FDR-corrected). Post hoc comparisons revealed that in both bands,
463 tracking of AudFeat was higher in A41/42 and A22c compared to TE1.0/1.2 and A22r (Tukey-Kramer test, all
464 tests $p < 10^{-5}$). The effect of Area was close to but not significant for LipFeat tracking in the occipital ROI
465 during V-only trials (0.5 - 1 Hz: $\chi^2(3) = 12.3$, $p = 0.026$, $\varepsilon^2 = 0.14$; 1 - 3 Hz: $\chi^2(3) = 14.57$, $p = 0.012$, $\varepsilon^2 = 0.17$).
466 Importantly, these results suggest that while the tracking of auditory features was stronger in the the early
467 auditory region during A-only trials, the restoration of unheard auditory features during lip reading
468 emerges to a similar degree among the individual temporal and occipital areas.

469 We performed a similar analysis comparing the CMI values within temporal or occipital ROIs between
470 hemispheres. This revealed no significant effects of hemispheres (Table 2), hence offering no evidence for a
471 statistical lateralization of feature tracking in the present data.

		0.5 - 1 Hz				1 - 3 Hz			
ROI	**Anatomical area**	**AudCMI**	**Chisq; pval**	**LipCMI**	**Chisq; pval**	**AudCMI**	**Chisq; pval**	**LipCMI**	**Chisq; pval**
A-only trials									
temporal	A41/42	0.97	***27.02 ; 4.7e-05***	0.096	2.47 ; 0.59	0.19	***29.62 ; 2.7e-05***	0.032	5.14 ; 0.32
	TE1.0/1.2	0.56		0.099		0.11		0.028	
	A22c	0.86		0.098		0.18		0.033	
	A22r	0.5		0.093		0.095		0.029	
occipital	mOccG	0.1	3.50 ; 0.47	0.073	0.66 ; 0.88	0.025	2.71 ; 0.58	0.02	1.97 ; 0.66

	OPC	0.09		0.069		0.025		0.02	
	iOccG	0.11		0.068		0.027		0.021	
	msOccG	0.11		0.074		0.029		0.021	
V-only trials									
temporal	A41/42	0.33	4.59 ; 0.36	0.1	5.14 ; 0.32	0.034	5.24 ; 0.32	0.027	1.00 ; 0.85
	TE1.0/1.2	0.25		0.088		0.031		0.028	
	A22c	0.34		0.1		0.035		0.027	
	A22r	0.22		0.085		0.028		0.029	
occipital	mOccG	0.13	3.90 ; 0.44	0.17	12.30 ; 0.026	0.045	8.20 ; 0.13	0.15	14.57 ; 0.012
	OPC	0.14		0.19		0.06		0.2	
	iOccG	0.14		0.2		0.048		0.17	
	msOccG	0.11		0.11		0.039		0.082	

Table 1. Feature tracking in individual anatomical areas within temporal and occipital ROIs. *The table lists CMI values of either set of features (AudCMI, LipCMI) and a statistical comparison between the individual atlas-defined areas of the temporal and occipital ROIs (Kruskal-Wallis tests, reporting chi-squares (Chisq) and p-values (pval)). Bold numbers indicate statistically significant results. P-values are FDR-corrected within this table.*

		0.5 - 1 Hz				1 - 3 Hz			
ROI	**Hemisphere**	**AudCMI**	**z; pval**	**LipCMI**	**z; pval**	**AudCMI**	**z; pval**	**LipCMI**	**z; pval**
A-only trials									
temporal	left	0.8	1.20 ; 0.59	0.094	-0.33 ; 0.74	0.13	-0.81 ; 0.59	0.028	-1.11 ; 0.59
	right	0.64		0.099		0.15		0.031	
occipital	left	0.1	-0.37 ; 0.74	0.07	-0.33 ; 0.74	0.027	0.81 ; 0.59	0.021	0.63 ; 0.65
	right	0.1		0.072		0.025		0.02	
V-only trials									
temporal	left	0.31	0.89 ; 0.59	0.098	0.76 ; 0.59	0.035	0.85 ; 0.59	0.025	-1.85 ; 0.26
	right	0.26		0.091		0.03		0.03	
occipital	left	0.11	-2.24 ; 0.2	0.14	-2.98 ; 0.046	0.043	-1.68 ; 0.3	0.13	-2.07 ; 0.21
	right	0.15		0.2		0.054		0.18	

Tab. 2. Feature tracking in each hemisphere. *The table lists CMI values of either set of features (AudCMI, LipCMI) and a statistical comparison between hemispheres of each ROI (Wilcoxon signed rank tests, reporting z values (z) and p-values (pval)). P-values are FDR-corrected within this table.*

3.4 Occipital cortex reflects pitch more than other acoustic features during lip reading

Having established that occipital and temporal regions track unheard auditory features, we then asked how individual features contribute to these representations. For this we focused on the following condition: the tracking of AudFeat in the delta range in V-only trials (Fig. 4C,D). We quantified the CMI for each individual feature, while discounting the evidence about all other left-out visual and auditory features, hence focusing on the unique tracking of each individual acoustic feature.

For the temporal ROI this revealed the prominent tracking of aud env (Fig. 4C). In the in the 0.5 - 1 Hz band only the CMI for aud env was above chance ($p < 5 \times 10^{-5}$) and there was a significant effect of feature (Kruskal-Wallis rank test $\chi^2(2) = 9.27$, $p = 9.1 \times 10^{-4}$, $\varepsilon^2 = 0.14$). Post-hoc tests revealed that the CMI for aud env differed significantly from that of aud slope (Tukey-Kramer test, $p = 6.2 \times 10^{-4}$; the other comparisons were not significant; $p = 0.35$ for env vs. slope and $p = 0.22$ for slope vs. pitch). In the 1 - 3 Hz band, the tracking of all auditory features was significant (all features: $p < 5 \times 10^{-5}$) and there was no significant effect of features ($\chi^2(2) = 4.14$, $p = 0.13$, $\varepsilon^2 = 0.04$).

For the occipital ROI, this revealed a dominance of aud pitch (Fig. 4D). In the 0.5 - 1 Hz band, only the CMI of aud pitch was above chance ($p < 5 \times 10^{-5}$), a direct comparison revealed a significant effect of features (0.5 - 1 Hz: $\chi^2(2) = 18.28$, $p = 1.07 \times 10^{-4}$, $\varepsilon^2 = 0.32$) and post-hoc tests revealed a significant difference between aud pitch and aud slope ($p = 7.03 \times 10^{-5}$), while the other comparisons were not significant ($p = 0.26$ for pitch vs. env and $p = 0.02$ for env vs. slope). In the 1 - 3 Hz range, the tracking of all features was significant (all features: $p < 5 \times 10^{-5}$), there was a significant effect of features $\chi^2(2) = 19.2$, $p = 6.77 \times 10^{-5}$, $\varepsilon^2 = 0.34$), and post-hoc tests revealed a significant difference between pitch and slope ($p = 3.61 \times 10^{-5}$), while the other comparisons were not significant ($p = 0.05$ for pitch vs. env and $p = 0.12$ for env vs. slope). Collectively these results suggest that the restoration of acoustic signals in the occipital region emphasizes spectral pitch, while in the temporal region this emphasizes the temporal speech envelope.

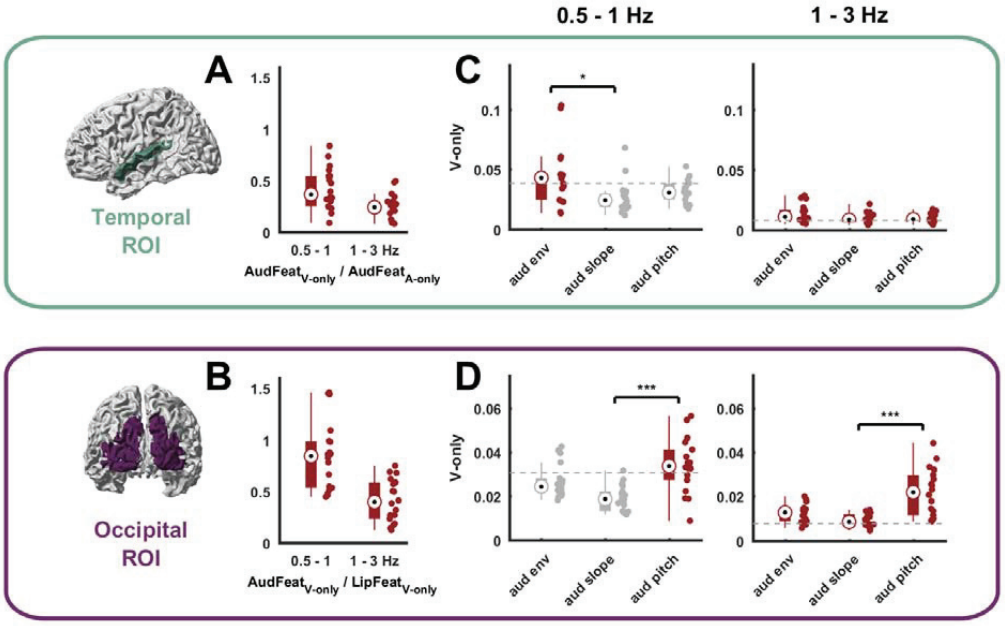

Fig. 4. Modality dominance and tracking of individual auditory features during lip reading. *(**A**,**B**) Comparison of the tracking of unheard AudFeat over the tracking of the modality-preferred sensory input in each ROI (i.e. AudFeat during A-only trials in the temporal ROI; LipFeat during V-only trials in the occipital ROI). (**C**,**D**) Tracking of individual auditory features during V-only trials conditioned on all other auditory and lip features in temporal (**C**) and occipital (**D**) ROIs. Brackets with asterisks indicate levels of significance from one-way Kruskal-Wallis rank test with post-hoc Tukey-Kramer testing (* $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$). Dots represent individual data points. Bars indicate the median, 25th and 75th percentile. The grey dashed line indicates the 99th percentile of the frequency-specific randomized maximum distribution correction for all other features. Units in (**A**) and (**B**) are a ratio, in panels (**C**) and (**D**) units are in bits.*

3.5 Tracking of auditory features is associated with lip reading performance

Finally, we probed the second main question of whether the restoration of unheard auditory features during silent lip reading relates to word recognition performance. For this we probed the predictive power of the MI about specific auditory features in either ROI for word recognition performance during V-only trials (Fig. 5). We specifically focused on the tracking of aud env in the temporal ROI and of aud pitch in the occipital ROI as the dominant feature-specific representations (c.f. Fig. 4C,D). Using linear models we predicted word recognition scores across participants based on the tracking indices of interest and while discounting for potential confounds from differences in signal-to-noise ratio in the MEG data.

The results show that variations in word recognition scores are well predicted by the collective measures of feature tracking (0.5 - 1 Hz: R^2 = 0.74, 1 - 3 Hz: R^2 = 0.8). Importantly, the tracking of aud env in the temporal ROI was significantly predictive of lip reading performance (Fig. 5, 0.5 - 1 Hz: β = 0.6, p = 0.037; 1 - 3 Hz: aud env β = 0.6, p = 2.8 x 10^{-4}), while tracking of pitch in the occipital ROI was not (0.5 - 1 Hz: β = -0.13, p = 0.56; 1 - 3 Hz: β = -0.026, p = 0.91). This conclusion is also supported by Bayes factors for the added predictive power of aud env and aud pitch to these models (aud env in the temporal ROI; 0.5 - 1 Hz: BF = 3.12; 1 - 3 Hz: BF = 26.34; aud pitch in the occipital ROI; 0.5 - 1 Hz BF = 0.3; 1 - 3 Hz BF = 0.24).

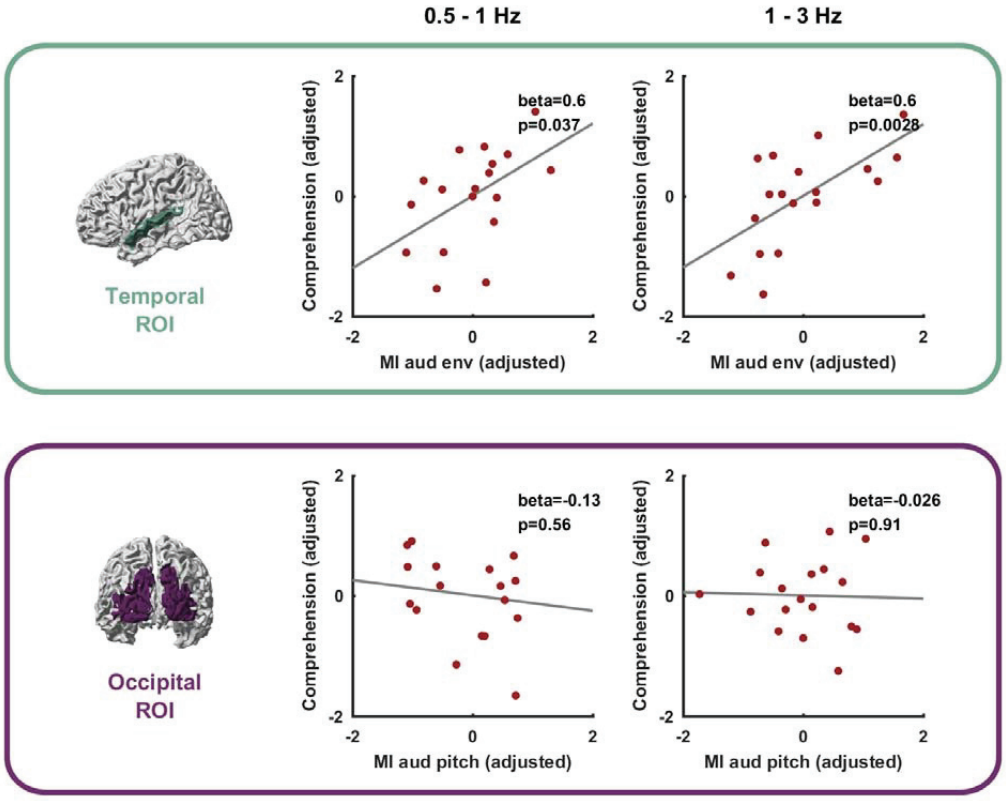

Fig. 5. Association between lip reading performance and tracking of auditory features. *Across participants the tracking of aud env during V-only trials in the temporal ROI but not the tracking of aud pitch in the occipital ROI was significantly associated with* word recognition *performance (PC) across participants in visual trials. Graphs show partial residual plots, dots represent individual data points and the line indicates the linear fit to the target variable from the full regression model.*

4. Discussion

Natural face-to-face speech is intrinsically multidimensional and provides the auditory and visual pathways with partly distinct acoustic and visual information. These pathways could in principle focus mainly on the processing of their modality-specific signals, effectively keeping the two input modalities largely separated. Yet, many studies highlight the intricate multisensory nature of speech-related representations in the brain, including multisensory convergence at early stages of the hierarchy (Bernstein and Liebenthal, 2014; Crosse et al., 2015; Schroeder et al., 2008; Schroeder and Lakatos, 2009) as well as in classically amodal speech regions (Keitel et al., 2020; Mégevand et al., 2020; Scott, 2019). However, as the present results suggest, the auditory and visual pathways are also capable of apparent 'restoring' information about an absent modality-specific speech component: while seeing a silent speaker, both auditory and visual cortices track the temporal dynamics of the speech envelope and the pitch contour respectively, in a manner that is independent on the physically visible lip movements. These 'restored' representations of acoustic features relate to participants' word recognition, suggesting that they may form a central component of silent lip reading.

4.1 Auditory and visual cortex reflect acoustic speech features during lip reading

We systematically quantified the tracking of auditory and visual speech features during unisensory auditory and visual (lip reading) conditions in dynamically entrained brain activity. As expected, this confirmed that early auditory and visual regions reflect acoustic and visual features respectively at the time scales of delta (< 4 Hz) and theta (4 - 8 Hz) band activity, in line with previous work (Aiken and Picton, 2008; Bauer et al., 2020; Doelling et al., 2014; Giraud and Poeppel, 2012; Haegens and Zion Golumbic, 2018; Obleser and Kayser, 2019). In addition, we found that during lip reading both regions contained significant information about unheard auditory features, also when discounting for the physically presented lip movements. This representation of acoustic features prevailed in low delta in auditory and delta and theta bands in visual cortex. Interestingly, this representation emphasized the temporal speech envelope in auditory cortex and spectral pitch in visual cortex. These results not only support that both regions are active during lip reading (Besle et al., 2008; Calvert et al., 1997; Calvert and Campbell, 2003; Ludman et al., 2000; Luo et al., 2010), but directly show that they contain temporally and feature-specific representations derived from lip movements that are relevant for comprehension.

These results advance our understanding of how the brain exploits lip movements in a number of ways. The restoration of auditory features during silent lip reading has been suggested in previous studies, one quantifying the coherence of temporal brain activity with the non-presented speech envelope (Bourguignon et al., 2020) and others quantifying the coherence between occipital activity and the envelope (Hauswald et al., 2018; Suess et al., 2022). Yet, these studies differed in their precise experimental designs, their statistical procedures revealing the 'restoration' effect, and did not probe a direct link to behavioral performance. The present data demonstrate that such tracking of auditory speech-derived features indeed emerges in parallel and in the same participants. Our data reveal the restoration of

572 unheard acoustic features also when discounting the physically present lip signals (i.e. when using
573 conditional mutual information). This finding is important, as the mere coherence of brain activity with the
574 acoustic speech envelope may otherwise simply reflect amodal information contained in the physically-
575 present visual speech that is directly redundant with the acoustic domain (Daube et al., 2019).

576 Furthermore, they show that this effect is largely bilateral and emerges across a number of anatomically-
577 identified areas, suggesting that it forms a generic property of the respective pathways. Interestingly, the
578 unheard auditory features during V-only trials were restored dominantly in the lower frequencies (0.5 - 3
579 Hz), similarly to recent results (Bourguignon et al., 2020). In principle, activity at these slow timescales may
580 possibly reflect oro-facial cues such as head, eye or eyebrow movements (Munhall et al., 2004; Schroeder
581 et al., 2008). We aimed to mitigate such confounds by instructing the speaker to move their head as little as
582 possible and to avoid gestures, and by instructing participants to focus their gaze on the speaker's lips.
583 Moreover, our results align with recent work showing the restoration of the unheard acoustic envelope
584 even when controlling for the speaker's movement during visual presentation (Bourguignon et al., 2020).
585 One may speculate whether this restoration reflects the synthesis of speech-specific elements. However,
586 linguistic elements at this time scale mostly encompass phrasal structures, prosody or speech rhythm
587 (Gross et al., 2013; Keitel et al., 2018; Meyer et al., 2017) and few of these are probably restored during lip
588 reading in detail. Possibly, the restoration of the unheard envelope based on lip movements reflects
589 processes for the temporal segmentation of speech-related information based on low-frequency activity
590 (Doelling et al., 2014; Ghitza, 2017; Nidiffer et al., 2021).

591 These results come with an important caveat: the capability to read from lips alone is generally low in naïve
592 listeners (Altieri et al., 2011; Grant and Seitz, 2000), which poses an intricate problem when studying the
593 cerebral basis of lip reading. To solicit a sufficient number of trials with successful lip reading and to balance
594 word recognition performance between visual-only and auditory-only trials, we relied on a specifically
595 designed experimental paradigm with two critical features. First, this paradigm relied on sentences
596 constructed based on a repeating set of linguistic elements and a forced-choice task with a closed set of
597 options. This limits the generalizability of the results towards naturally-produced every day speech, as
598 participants could in principle learn the mapping of only target words onto lip movements and choose the
599 most likely one during the course of the experiment. Although we did not strictly control for this, both the
600 chosen elements in each sentence as well as the target and distractor words were chosen randomly.
601 Second, to familiarize participants with the material, the A-only condition preceded the V-only condition
602 during the experiment. This may allow for memory-related processes to contribute to the observed
603 restoration effects. However, the use of 180 syntactically similar but unique sentences makes it in our view
604 highly unlikely that participants solely relied on the stimulus repetition and memory to solve the word
605 recognition task. Rather, we believe that the restoration in the occipital cortex reflects the active parsing of
606 the lip movement signal and engages specific visuo-phonetic transformations, as speculated previously
607 (Hauswald et al., 2018; Nidiffer et al., 2021). This poses a possible solution to how the brain finds the best

match between visually perceived speech and a word from a limited set of options. Nevertheless, the visual system might be primed to perceive visual speech after being familiarized with the underlying acoustic stimulus in a previous condition. Even in naturalistic listening situations one is likely to do so when observing a moving face. Therefore, we do not expect any priming of the visual system to confound the nature of lipreading in this paradigm compared to real-life situations. More so, this alludes to the origin of speech-related information during lip reading in general, as bottom-up processes may be aided by sentence-level predictions or expectations that contribute in a top-down manner and partially predict acoustic and lexical information based on the immediately preceding material (Cope et al., 2017). Given that lip reading performance was higher than in other studies or real-life circumstances (Altieri et al., 2011; Grant and Seitz, 2000) it is possible that top-down processes exerted a stronger influence on early visual and auditory cortices in this data compared to real-life circumstances.

4.2 Lip reading activates a network of occipital and temporal regions

Previous work has shown that lip movements activate a network of temporal, parietal and frontal regions (Bourguignon et al., 2020; Calvert et al., 1997; Capek et al., 2008; O'Sullivan et al., 2017; Ozker et al., 2018; Paulesu et al., 2003; Pekkola et al., 2005) and that both occipital and motor regions can align their neural activity to the dynamics of lip movements (Park et al., 2018, 2016). The present data substantiate this, but also show that the representation of the physically visible lip trajectory along visual pathways is accompanied by the representation of spectral pitch, a type of selectivity not directly revealed previously (Suess et al., 2022). Spectral features are vital for a variety of listening tasks (Albouy et al., 2020; Bröhl and Kayser, 2021; Ding and Simon, 2013; Tivadar et al., 2020, 2018), and oro-facial movements provide concise information about the spectral domain. Importantly, seeing the speaker's mouth allows discriminating formant frequencies and provides a comprehension benefit particularly when spectral features are degraded in the underlying acoustic speech (Plass et al., 2020). This suggests a direct and comprehension-relevant link between the dynamics of the lip contour and spectral speech features (Campbell, 2008). Hence, a representation of acoustic features during silent lip reading may underlie the mapping of lip movements onto phonological units such as visemes, a form of language-specific representation emerging along visual pathways (Nidiffer et al., 2021; O'Sullivan et al., 2017). This emphasize the role of the visual system as an active agent during audio-visual speech processing.

Our results corroborate the notion that multisensory speech reception is not contingent only on high-level and amodal representations. Rather, they suggest that the brain likely exploits cross-modal correspondences between auditory and visual speech along a number of dimensions, including basic temporal properties (Bizley et al., 2016; Chandrasekaran et al., 2009) as well as mid-level features, such as pitch or visual object features, whose representation is traditionally considered to be modality specific (Crosse et al., 2015; Plass et al., 2020; Schroeder et al., 2008; Zion Golumbic et al., 2013). Previous work has debated whether visual speech is mainly encoded along the auditory pathways or whether occipital regions

643 contribute genuine speech-specific representations (O'Sullivan et al., 2017; Ozker et al., 2018). Our results
644 speak in favor of occipital regions supporting speech reception by establishing multiple forms of speech-
645 related information, including those aligned with the acoustic domain revealed here, and those establishing
646 visemic categories based on complementary visual signals (Nidiffer et al., 2021; Suess et al., 2022). Which
647 precise occipital areas and by which patterns of connectivity they contribute to comprehension remains to
648 be investigated, but both kinds of representations may well emerge from distinct temporal-occipital
649 networks (Bernstein and Liebenthal, 2014). While visemic information may be driven by object-related
650 lateral occipital regions, the more auditory-aligned representations such as the restoration of spectral
651 signatures may be directly driven by the connectivity between occipital areas and superior temporal
652 regions, which play a key role for audio-visual speech integration (Arnal et al., 2009; Lazard and Giraud,
653 2017). In the auditory cortex, the alignment of neural activity to the unheard speech envelope may reflect
654 the predictive influence of visual signals on guiding the excitability of auditory pathways via low frequency
655 oscillations (Schroeder et al., 2008). This alignment of auditory cortical activity to attended or expected
656 sounds is well documented and has been proposed as a cornerstone of multisensory speech integration in
657 general (Lakatos et al., 2008; Schroeder and Lakatos, 2009; Stefanics et al., 2010), and as shown here,
658 directly relates to participants comprehension performance.

659 ## Credit author statement

660 Conceptualization: A.K., C.K., Project administration: A.K., C.K., Funding acquisition: C.K., Methodology: F.B.,
661 A.K., C.K., Software: F.B., A.K., Formal Analysis: F.B., Investigation: F.B., Data Curation: F.B., Supervision:
662 C.K., Writing Original Draft: F.B., C.K., Writing Review & Editing: F.B., A.K., C.K.

663 ## Data and code availability

664 Data and code used in this study are publicly available on the Data Server of the University of Bielefeld
665 (https://gitlab.ub.uni-bielefeld.de/felix.broehl/fb02).

References

Aiken, S.J., Picton, T.W., 2008. Human cortical responses to the speech envelope. Ear Hear. 29, 139–157. https://doi.org/10.1097/AUD.0b013e31816453dc

Albouy, P., Benjamin, L., Morillon, B., Zatorre, R.J., 2020. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. Science (80-.). 367, 1043–1047. https://doi.org/10.1126/science.aaz3468

Altieri, N.A., Pisoni, D.B., Townsend, J.T., 2011. Some normative data on lip-reading skills (L). J. Acoust. Soc. Am. 130, 1–4. https://doi.org/10.1121/1.3593376

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

Bauer, A.K.R., Debener, S., Nobre, A.C., 2020. Synchronisation of Neural Oscillations and Cross-modal Influences. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2020.03.003

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B.

Bernstein, L.E., Jiang, J., Pantazis, D., Lu, Z.-L., Joshi, A., 2011. Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. Hum. Brain Mapp. 32, 1660–1676. https://doi.org/10.1002/hbm.21139

Bernstein, L.E., Liebenthal, E., 2014. Neural pathways for visual speech perception. Front. Neurosci. https://doi.org/10.3389/fnins.2014.00386

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., Giard, M.H., 2008. Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. J. Neurosci. 28, 14301–14310. https://doi.org/10.1523/JNEUROSCI.2875-08.2008

Bizley, J.K., Maddox, R.K., Lee, A.K.C., 2016. Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. Trends Neurosci. 39, 74–85. https://doi.org/10.1016/j.tins.2015.12.007

Boersma, P., van Heuven, V., 2001. PRAAT, a system for doing phonetics by computer. Glot Int. 5, 341–347.

Bourguignon, M., Baart, M., Kapnoula, E.C., Molinaro, N., 2020. Lip-reading enables the brain to synthesize auditory features of unknown silent speech. J. Neurosci. 40, 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Bröhl, F., Kayser, C., 2021. Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. Neuroimage 233, 117958. https://doi.org/10.1016/j.neuroimage.2021.117958

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science (80-.). 276, 593–596. https://doi.org/10.1126/science.276.5312.593

Calvert, G.A., Campbell, R., 2003. Reading speech from still and moving faces: The neural substrates of visible speech. J. Cogn. Neurosci. 15, 57–70. https://doi.org/10.1162/089892903321107828

Campbell, R., 2008. The processing of audio-visual speech: Empirical and neural bases. Philos. Trans. R. Soc.

B Biol. Sci. 363, 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Capek, C.M., MacSweeney, M., Woll, B., Waters, D., McGuire, P.K., David, A.S., Brammer, M.J., Campbell, R., 2008. Cortical circuits for silent speechreading in deaf and hearing people. Neuropsychologia 46, 1233–1241. https://doi.org/10.1016/j.neuropsychologia.2007.11.026

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. PLoS Comput. Biol. 5, e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Cope, T.E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P.S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R.P., Griffiths, T.D., Davis, M.H., Rowe, J.B., 2017. Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat. Commun. 8, 2154. https://doi.org/10.1038/s41467-017-01958-7

Crosse, M.J., Butler, J.S., Lalor, E.C., 2015. Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. J. Neurosci. 35, 14195–204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Daube, C., Ince, R.A.A., Gross, J., 2019. Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. Curr. Biol. 29, 1924-1937.e9. https://doi.org/10.1016/J.CUB.2019.04.067

Di Liberto, G.M., Lalor, E.C., Millman, R.E., 2018. Causal cortical dynamics of a predictive enhancement of speech intelligibility. Neuroimage 166, 247–258. https://doi.org/10.1016/j.neuroimage.2017.10.066

Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33, 5728–35. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Etard, O., Reichenbach, T., 2019. Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. J. Neurosci. 39, 5750–5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019

Ghitza, O., 2017. Acoustic-driven delta rhythms as prosodic markers. Lang. Cogn. Neurosci. 32, 545–561. https://doi.org/10.1080/23273798.2016.1232419

Giordano, B.L., Ince, R.A.A., Gross, J., Schyns, P.G., Panzeri, S., Kayser, C., 2017. Contributions of local speech encoding and functional connectivity to audio-visual speech perception. Elife 6. https://doi.org/10.7554/eLife.24763

Giraud, A.L., Poeppel, D., 2012. Cortical oscillations and speech processing: Emerging computational principles and operations. Nat. Neurosci. 15, 511–517. https://doi.org/10.1038/nn.3063

Grant, K.W., Seitz, P.-F., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108, 1197. https://doi.org/10.1121/1.1288668

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech Rhythms and

738 Multiplexed Oscillatory Sensory Coding in the Human Brain. PLoS Biol. 11, e1001752.
739 https://doi.org/10.1371/journal.pbio.1001752
740 Haegens, S., Zion Golumbic, E., 2018. Rhythmic facilitation of sensory processing: A critical review.
741 Neurosci. Biobehav. Rev. 86, 150–165. https://doi.org/10.1016/j.neubiorev.2017.12.002
742 Hagerman, B., 1982. Sentences for Testing Speech Intelligibility in Noise. Scand. Audiol. 11, 79–87.
743 https://doi.org/10.3109/01050398209076203
744 Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., Weisz, N., 2018. A Visual Cortical Network for
745 Deriving Phonological Information from Intelligible Lip Movements. Curr. Biol. 28, 1453-1459.e3.
746 https://doi.org/10.1016/j.cub.2018.03.044
747 Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., Schyns, P.G., 2017. A statistical framework
748 for neuroimaging data analysis based on mutual information estimated via a gaussian copula. Hum.
749 Brain Mapp. 38, 1541–1573. https://doi.org/10.1002/hbm.23471
750 Keitel, A., Gross, J., Kayser, C., 2020. Shared and modality-specific brain regions that mediate auditory and
751 visual word comprehension. Elife 9, 1–23. https://doi.org/10.7554/ELIFE.56972
752 Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex
753 reflects distinct linguistic features. PLoS Biol. 16, e2004473.
754 https://doi.org/10.1371/journal.pbio.2004473
755 Keitel, A., Ince, R.A.A., Gross, J., Kayser, C., 2017. Auditory cortical delta-entrainment interacts with
756 oscillatory power in multiple fronto-parietal networks. Neuroimage 147, 32–42.
757 https://doi.org/10.1016/j.neuroimage.2016.11.062
758 Koike, K.J., Hurst, M.K., Wetmore, S.J., 1994. Correlation between the American Academy of
759 Otolaryngology-Head and Neck Surgery Five-Minute Hearing Test and standard audiologic data.
760 Otolaryngol. - Head Neck Surg. 111, 625–632. https://doi.org/10.1016/S0194-5998(94)70531-3
761 Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M.A., Uslar, V., Brand, T., Wagener, K.C., 2015. The
762 multilingual matrix test: Principles, applications, and comparison across languages: A review. Int. J.
763 Audiol. 54, 3–16. https://doi.org/10.3109/14992027.2015.1020971
764 Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations
765 as a mechanism of attentional selection. Science 320, 110–3.
766 https://doi.org/10.1126/science.1154735
767 Lazard, D.S., Giraud, A.L., 2017. Faster phonological processing and right occipito-temporal coupling in deaf
768 adults signal poor cochlear implant outcome. Nat. Commun. 8, 1–9.
769 https://doi.org/10.1038/ncomms14872
770 Ludman, C.N., Summerfield, A.Q., Hall, D., Elliott, M., Foster, J., Hykin, J.L., Bowtell, R., Morris, P.G., 2000.
771 Lip-reading ability and patterns of cortical activation studied using fMRI. Br. J. Audiol. 34, 225–230.
772 https://doi.org/10.3109/03005364000000132
773 Luo, H., Liu, Z., Poeppel, D., 2010. Auditory Cortex Tracks Both Auditory and Visual Stimulus Dynamics Using

774 Low-Frequency Neuronal Phase Modulation. PLoS Biol. 8, e1000445.
775 https://doi.org/10.1371/journal.pbio.1000445
776 Mégevand, P., Mercier, M.R., Groppe, D.M., Golumbic, E.Z., Mesgarani, N., Beauchamp, M.S., Schroeder,
777 C.E., Mehta, A.D., 2020. Crossmodal Phase Reset and Evoked Responses Provide Complementary
778 Mechanisms for the Influence of Visual Speech in Auditory Cortex. J. Neurosci. 405597.
779 https://doi.org/10.1101/405597
780 Meyer, L., Henry, M.J., Gaston, P., Schmuck, N., Friederici, A.D., 2017. Linguistic bias modulates
781 interpretation of speech via neural delta-band oscillations. Cereb. Cortex 27, 4293–4302.
782 https://doi.org/10.1093/cercor/bhw228
783 Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual Prosody and Speech
784 Intelligibility: Head Movement Improves Auditory Speech Perception. Psychol. Sci. 15, 133–137.
785 https://doi.org/10.1111/j.0963-7214.2004.01502010.x
786 Nichols, T., Holmes, A., 2003. Nonparametric Permutation Tests for Functional Neuroimaging. Hum. Brain
787 Funct. Second Ed. 25, 887–910. https://doi.org/10.1016/B978-012264841-0/50048-2
788 Nidiffer, A.R., Cao, C.Z., O'Sullivan, A.E., Lalor, E.C., 2021. A linguistic representation in the visual system
789 underlies successful lipreading. bioRxiv. https://doi.org/10.1101/2021.02.09.430299
790 O'Sullivan, A.E., Crosse, M.J., Di Liberto, G.M., Lalor, E.C., 2017. Visual cortical entrainment to motion and
791 categorical speech features during silent lipreading. Front. Hum. Neurosci. 10, 1–11.
792 https://doi.org/10.3389/fnhum.2016.00679
793 Obleser, J., Kayser, C., 2019. Neural Entrainment and Attentional Selection in the Listening Brain. Trends
794 Cogn. Sci. 23, 913–926. https://doi.org/10.1016/j.tics.2019.08.004
795 Oganian, Y., Chang, E.F., 2019. A speech envelope landmark for syllable encoding in human superior
796 temporal gyrus. Sci. Adv. 5, 1–14. https://doi.org/10.1126/sciadv.aay6279
797 Oldfield, R.C., 1971. The assessment and analysis of handedness: The Edinburgh inventory.
798 Neuropsychologia 9, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4
799 Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: Open source software for advanced
800 analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011, 156869.
801 https://doi.org/10.1155/2011/156869
802 Ozker, M., Yoshor, D., Beauchamp, M.S., 2018. Frontal cortex selects representations of the talker's mouth
803 to aid in speech perception. Elife 7, 1–14. https://doi.org/10.7554/eLife.30387
804 Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., Gross, J., 2018. Representational interactions during
805 audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in
806 left motor cortex. PLOS Biol. 16, e2006558. https://doi.org/10.1371/journal.pbio.2006558
807 Park, H., Kayser, C., Thut, G., Gross, J., 2016. Lip movements entrain the observers' low-frequency brain
808 oscillations to facilitate speech intelligibility. Elife 5. https://doi.org/10.7554/eLife.14521
809 Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N.A., De Giovanni, U., Sensolo, S., Fazio, F., 2003. A

functional-anatomical model for lipreading. J. Neurophysiol. 90, 2005–2013. https://doi.org/10.1152/jn.00926.2002

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Tarkiainen, A., Sams, M., 2005. Primary auditory cortex activation by visual speech: An fMRI study at 3 T. Neuroreport 16, 125–128. https://doi.org/10.1097/00001756-200502080-00010

Plass, J., Brang, D., Suzuki, S., Grabowecky, M., 2020. Vision perceptually restores auditory spectral dynamics in speech. Proc. Natl. Acad. Sci. U. S. A. 117, 16920–16927. https://doi.org/10.1073/pnas.2002887117

Rauschecker, J.P., 2012. Ventral and dorsal streams in the evolution of speech and language. Front. Evol. Neurosci. 4, 5–8. https://doi.org/10.3389/fnevo.2012.00007

Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32, 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., 2008. Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. 12, 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Scott, S.K., 2019. From speech and talkers to the social world: The neural processing of human spoken language. Science (80-.). https://doi.org/10.1126/science.aax0288

Stefanics, G., Hangya, B., Hernadi, I., Winkler, I., Lakatos, P., Ulbert, I., 2010. Phase Entrainment of Human Delta Oscillations Can Mediate the Effects of Expectation on Reaction Speed. J. Neurosci. 30, 13578–13585. https://doi.org/10.1523/JNEUROSCI.0703-10.2010

Suess, N., Hauswald, A., Reisinger, P., Rösch, S., Keitel, A., Weisz, N., 2022. Cortical Tracking of Formant Modulations Derived from Silently Presented Lip Movements and Its Decline with Age. Cereb. Cortex 1–16. https://doi.org/10.1093/cercor/bhab518

Teng, X., Tian, X., Doelling, K., Poeppel, D., 2018. Theta band oscillations reflect more than entrainment: behavioral and neural evidence demonstrates an active chunking process. Eur. J. Neurosci. 48, 2770–2782. https://doi.org/10.1111/ejn.13742

Teoh, E.S., Cappelloni, M.S., Lalor, E.C., 2019. Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. Eur. J. Neurosci. 50, 3831–3842. https://doi.org/10.1111/ejn.14510

Tivadar, R.I., Gaglianese, A., Murray, M.M., 2020. Auditory Enhancement of Illusory Contour Perception. Multisens. Res. 34, 1–15. https://doi.org/10.1163/22134808-bja10018

Tivadar, R.I., Retsa, C., Turoman, N., Matusz, P.J., Murray, M.M., 2018. Sounds enhance visual completion processes. Neuroimage 179, 480–488. https://doi.org/10.1016/j.neuroimage.2018.06.070

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

van Bree, S., Sohoglu, E., Davis, M.H., Zoefel, B., 2020. Sustained neural rhythms reveal endogenous oscillations supporting speech perception, PLoS Biology. https://doi.org/10.1101/2020.06.26.170761

Van Veen, B.D., Van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44, 867–880. https://doi.org/10.1109/10.623056

Velleman, P.F., Welsch, R.E., 1981. Efficient computing of regression diagnostics. Am. Stat. 35, 234–242. https://doi.org/10.1080/00031305.1981.10479362

Yu, C., Zhou, Y., Liu, Y., Jiang, T., Dong, H., Zhang, Y., Walter, M., 2011. Functional segregation of the human cingulate cortex is confirmed by functional connectivity based neuroanatomical parcellation. Neuroimage 54, 2571–2581. https://doi.org/10.1016/j.neuroimage.2010.11.018

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a “Cocktail Party.” J. Neurosci. 33, 1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013

Zuk, N.J., Murphy, J.W., Reilly, R.B., Lalor, E.C., 2021. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies, PLOS Computational Biology. https://doi.org/10.1371/journal.pcbi.1009358