Prediction tendency, eye movements, and attention in a unified framework of neural speech tracking

Juliane Schubert[1]*, Quirin Gehmacher[1,3,4]*, Fabian Schmidt[1], Thomas Hartmann[1], Nathan Weisz[1,2]

[1]Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria.
[2]Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria.
[3] Department of Experimental Psychology, University College London, United Kingdom.
[4] Wellcome Centre for Human Neuroimaging, University College London, United Kingdom.
*contributed equally

Address for correspondence
Juliane Schubert
University of Salzburg
Center for Cognitive Neuroscience
Hellbrunnerstr. 34
A-5020 SALZBURG
Austria/Europe
Tel. 0043 662 8044 5135
Email: juliane.schubert@plus.ac.at

ABSTRACT

Auditory speech comprehension is a multi-faceted process in which attention, prediction, and sensorimotor integration (via active sensing) interact with or complement each other. Although different conceptual models that focus on one of these aspects exist, we still lack a unified understanding of their role in speech processing. Here, we first replicated two recently published studies from our lab, confirming 1) a positive relationship between individual prediction tendencies and neural speech tracking, and 2) the phenomenon of ocular speech tracking - the tracking of attended speech by eye movements - and its shared contribution with neural activity to speech processing. In addition, we extended these findings with complementary analyses and investigated these phenomena in relation to each other in a multi-speaker paradigm with continuous, narrative speech. Importantly, prediction tendency and ocular speech tracking seem to be unrelated. In contrast to the shared contributions of oculomotor and neural activity to speech processing over a distributed set of brain regions that are critical for attention, individual prediction tendency and its relation to neural speech tracking seem to be largely independent of attention. Based on these findings, we propose a framework that aims to bridge the gaps between attention, prediction, and active (ocular) sensing in order to contribute to a holistic understanding of neural speech processing. In this speculative framework for listening, auditory inflow is, on a basic level, temporally modulated via active ocular sensing, and incoming information is interpreted based on probabilistic assumptions.

Keywords: active auditory sensing, speech processing, predictive processing, selective attention, eye movements, magnetoencephalography

INTRODUCTION

Listening challenges the brain to infer meaning from vastly overlapping spectrotemporal information. For example, in complex, everyday environments, we need to select and segregate streams of speech from different speakers for further processing. To accomplish this task, (mainly independent) research lines suggested contributions of predictive and attentional processes to speech perception.

Predictive brain accounts (K. Friston, 2010; K. J. Friston et al., 2021; Knill & Pouget, 2004; Yon et al., 2019) suggest an active engagement in speech perception. In this view, experience-based internal models constantly generate and continuously compare top-down predictions with bottom-up input, thus inferring sound sources from neural activity patterns. This idea is supported by the influential role of speech predictability (e.g. semantic context and word surprisal) on speech processing in naturalistic contexts (Broderick et al., 2019; Donhauser & Baillet, 2020; Weissbart et al., 2020).

Selective attention describes the process by which the brain prioritises information to focus our limited cognitive capacities on relevant inputs while ignoring distracting, irrelevant ones. Its beneficial role in the processing of complex acoustic scenes, like the cocktail-party scenario (Zion Golumbic et al., 2013; Zion-Golumbic & Schroeder, 2012), supports the key role of attentional and predictive processes in natural listening. It is important to note that even "normal hearing" individuals vary greatly in everyday speech comprehension and communication (Ruggles et al., 2011), which consequently promotes interindividual variability in predictive processes (Siegelman & Frost, 2015) or selective attention (Oberfeld & Klöckner-Nowotny, 2016) as underlying modulators.

In a recent study (Schubert et al., 2023), we quantified individual differences in predictive processes as the tendency to anticipate low-level acoustic features (i.e. prediction tendency). We established a positive relationship of this "trait" with speech processing, demonstrating that speech tracking is enhanced in individuals with stronger prediction tendency, independent of selective attention. Furthermore, we found an increased tracking of words of high surprisal, demonstrating the importance of predictive processes in speech perception.

Another important aspect with regard to speech processing - that has been vastly overlooked so far in neuroscience - is active auditory sensing. The potential benefit of active sensing - the engagement of covert motor routines in acquiring sensory inputs - was outlined for other sensory modalities (Schroeder et al., 2010). In the auditory domain, and especially for speech perception, motor contributions to an increased weighting of temporal precision were suggested to modulate auditory processing gain in support of (speech) segmentation (Morillon et al., 2015). Along these lines, It has been shown that covert, mostly blink related eye activity aligns with higher-order syntactic structures of temporally predictable, artificial speech (i.e. monosyllabic words). In support of ideas that the motor system is actively engaged in speech

perception (Galantucci et al., 2006; Liberman & Mattingly, 1985), the authors suggest a global entrainment across sensory and (oculo)motor areas which implements temporal attention.
In another recent study from our lab (Gehmacher et al., 2024), we showed that eye movements continuously track intensity fluctuations of attended natural speech, a phenomenon we termed ocular speech tracking. We further established links to increased intelligibility with stronger ocular tracking and provided evidence that eye movements and neural activity share contributions to speech tracking. Considering the aforementioned study by Schubert and colleagues (2023), the two recently uncovered predictors of neural tracking (individual prediction tendency and ocular tracking) raise several questions regarding their potential relationship with speech processing: Are predictive processes similarly related to active ocular sensing as to neural tracking in natural listening, i.e. are stronger prediction tendencies related to increased ocular speech tracking? Or vice versa? Or not at all? To what extent is their relationship to neural speech tracking affected by selective attention? Are they related to behavioural outcomes? While all these concepts have been shown to contribute to successful listening in challenging environments, a thorough investigation of potential interactions and their consequences on speech perception has been lacking.

Here, we set out to answer these questions by a) replicating aforementioned findings in a single experiment, and b) combining prediction tendency, eye movements, and attention into a unified framework of neural speech tracking. We therefore repeated the study protocol of Schubert et al. (2023) with slight modifications: Again, participants performed a separate, passive listening paradigm (also see (Demarchi et al., 2019)) that allowed us to quantify individual prediction tendency. Afterward, they listened to sequences of audiobooks (using the same stimulus material), either in a clear speech (0 distractors) or a multi-speaker (1 distractor) condition. Simultaneously recorded magnetoencephalographic (MEG) and Eye Tracking data confirmed the previous findings of Schubert et al. (2023) and Gehmacher et al. (2024): 1) Individuals with a stronger prediction tendency showed an increased neural speech tracking over left frontal areas, 2) eye movements track acoustic speech in selective attention, and 3) further mediate neural speech tracking effects over widespread, but mostly auditory regions. Additionally, we found an increased neural tracking of semantic violations (compared to their lexically identical controls), indicating that surprisal evoked responses indeed encode information about the stimulus. Interestingly, we could not find this difference in semantic processing for ocular speech tracking. Finally, we behaviorally assessed speech comprehension by probing participants on story content. Responses indicate that weaker performance in comprehension was related to increased ocular speech tracking while we did not find a significant relation to neural speech tracking. The findings suggest a differential role of prediction tendency, eye movements, and attention in speech processing. Behavioural

responses further indicate substantial differences in ocular and neural engagement and perceptual outcomes. Based on these findings, we propose a unified framework of neural speech tracking where anticipatory predictions support the interpretation of auditory input along the perceptual hierarchy while active ocular sensing increases the temporal precision of peripheral auditory responses to facilitate bottom-up processing of selectively attended input.

METHODS

Subjects

In total, 39 subjects were recruited to participate in the experiment. For 3 participants, eye tracker calibration failed and they had to abort the experiment. We further controlled for excessive blinking (> 50% of data samples) and eye movements away from fixation cross (> 50% of data samples exceeded ⅓ of screen size on the horizontal or vertical plane) that suggest a lack of commitment to the experiment instructions. 7 subjects had to be excluded according to these criteria. Thus, a final sample size of 29 subjects (12 female, 17 male; mean age = 25.70, range = 19 - 43) was used for the analysis of brain and gaze data.
All participants reported normal hearing and had normal, or corrected to normal, vision. They gave written, informed consent and reported that they had no previous neurological or psychiatric disorders. The experimental procedure was approved by the ethics committee of the University of Salzburg and was carried out in accordance with the declaration of Helsinki. All participants received either a reimbursement of 10 € per hour or course credits.

Experimental Procedure

The current study is a replication of a previous experiment (for details see Schubert et al., 2023), using the same experimental structure and the same auditory stimuli. The current design only differs in that one condition was dropped and the whole experiment was shortened (participants spent approximately 2 hours in the MEG including preparation time). In addition to the previous study, this time we recorded ocular movements (also see Data acquisition and Preprocessing). Before the start of the experiment, participants' head shapes were assessed using cardinal head points (nasion and pre-auricular points), digitised with a Polhemus Fastrak Digitiser (Polhemus), and around 300 points on the scalp. For every participant, MEG sessions started with a 5-minute resting-state recording, after which the individual hearing threshold was determined using a pure tone of 1043 Hz. This was followed by 2 blocks of passive listening to tone sequences of varying entropy levels to quantify individual prediction tendencies (see Quantification of individual prediction tendency and **Figure 1A-C**).

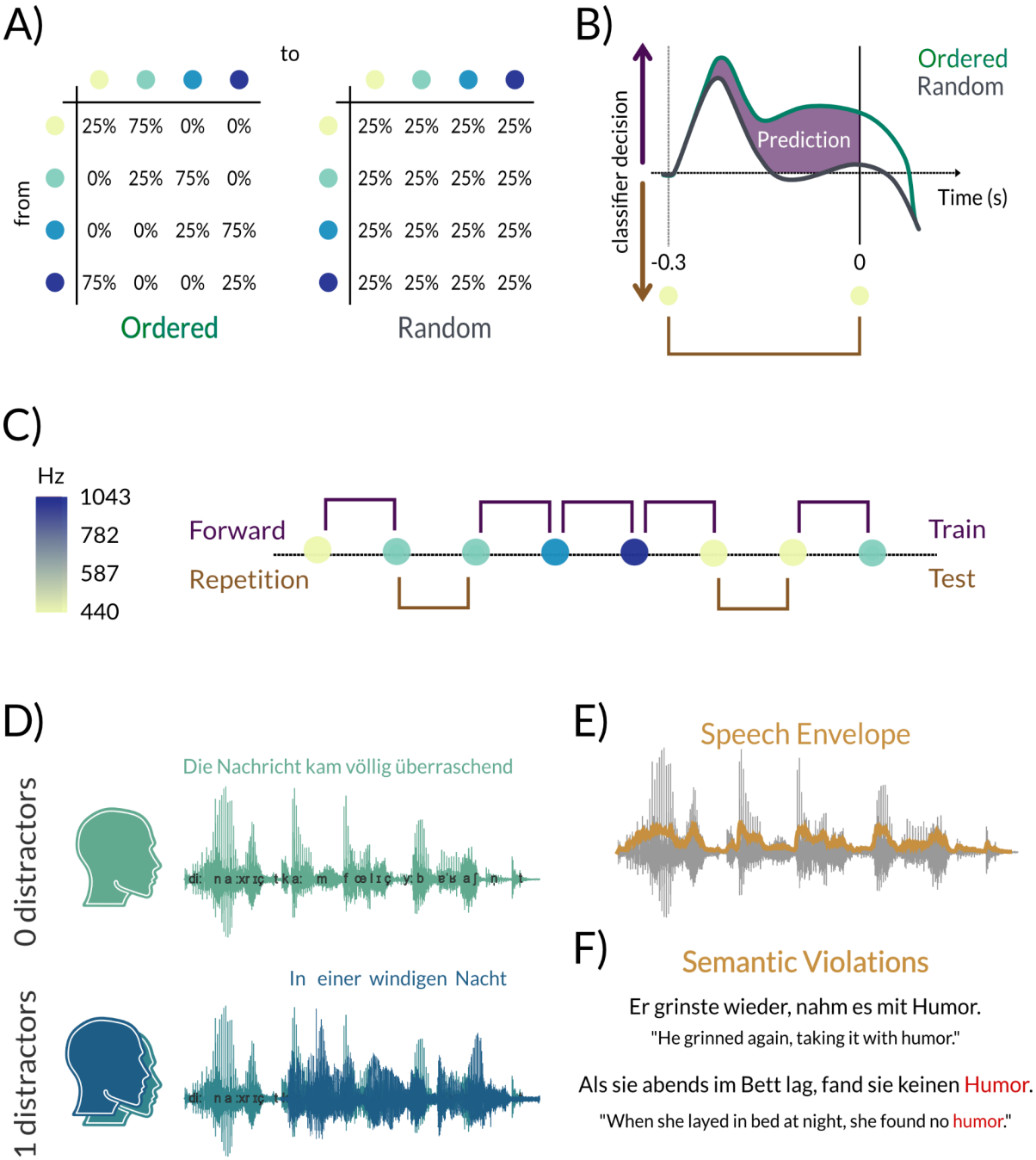

Figure 1: *Quantification of individual prediction tendency and the multi-speaker paradigm.* **A)** Participants passively listened to sequences of pure tones in different conditions of entropy (ordered vs. random). Four tones of different fundamental frequencies were presented with a fixed stimulation rate of 3 Hz, their transitional probabilities varied according to respective conditions. **B)** Expected classifier decision values contrasting the brains' prestimulus tendency to predict a forward transition (ordered vs. random). The purple shaded area represents values that were considered as prediction tendency **C)** Exemplary excerpt of a tone sequence in the ordered condition. An LDA classifier was trained on forward transition trials of the ordered condition (75% probability) and tested on all repetition trials to decode sound frequency from brain activity across time. **D)** Participants either attended to a story in clear speech, i.e. 0 distractor condition, or to a target speaker with a simultaneously presented distractor (blue), i.e. 1 distractor condition. **E)** The speech envelope was used to estimate neural and ocular speech tracking in respective conditions with temporal response functions (TRF). **F)** The last noun of some sentences was replaced randomly with an improbable candidate to measure the effect of envelope encoding on the processing of semantic violations. Adapted from Schubert et al., 2023.

Participants were instructed to look at a black fixation cross at the centre of a grey screen. In the main task, 4 different stories were presented in separate blocks in random order and with randomly balanced selection of the target speaker (male vs. female voice). Each block consisted of 2 trials with a continuous storyline, with each trial corresponding to one of 2 experimental conditions: a single speaker and a multi-speaker condition (see also **Figure 1D**). The distractor speaker was always of the opposite sex of the target speaker (and was identical to the target speaker in a different run). Distracting speech was presented exactly 20 s after target speaker onset, and all stimuli were presented binaurally at equal volume (40db above individual hearing threshold) for the left and right ear (i.e. at phantom centre). Participants were instructed to attend to the first speaker and their understanding was tested using comprehension questions (true vs. false statements) at the end of each trial (e.g.: "Das Haus, in dem Sofie lebt, ist rot" *(The house Sofie lives in is red)*, "Ein gutes Beispiel für unterschiedliche Dialekte sind die Inuit aus Alaska und Grönland" *(A good example of different dialects are the Inuit from Alaska and Greenland)*...). Furthermore, participants indicated their task engagement and their perceived task-difficulty on a 5-point Likert scale at the end of every trial. During the audiobook presentation, participants were again instructed to look at the fixation-cross on the screen and to blink as little as (still comfortably) possible. The experiment was coded and conducted with the Psychtoolbox-3 (Brainard & Vision, 1997; Kleiner et al., 2007), with an additional class-based library ('Objective Psychophysics Toolbox', o_ptb) on top of it (Hartmann & Weisz, 2020).

Stimuli

We used the same stimulus material as in Schubert et al. (2023), recorded with a t.bone SC 400 studio microphone at a sampling rate of 44100 Hz. In total, we used material from 4 different, consistent stories (see **Supplementary Table 1**). These stories were split into 3 separate trials of approximately 3 - 4 min. The first two parts of each story were always narrated by a target speaker, whereas the last part served as distractor material. Additionally, we randomly selected half of the nouns that ended a sentence and replaced them with the other half to induce unexpected semantic violations within each trial, resulting in two sets of lexically identical words (N = 79) that differed greatly in their contextual probabilities (see **Figure 1F** for an example). All trials were recorded twice, narrated by a different speaker (male vs. female). Stimuli were presented in 4 blocks containing 2 trials each (a single and a multi-speaker trial), resulting in 2 male and 2 female target speaker blocks for every participant.

Data Acquisition and Preprocessing

Brain activity was recorded using a whole head MEG system (Elekta Neuromag Triux, Elekta Oy, Finland), placed within a standard passive magnetically shielded room (AK3b, Vacuumschmelze, Germany). We used a sampling frequency of 1 kHz (hardware filters: 0.1 - 330 Hz). The signal was recorded with 102 magnetometers and 204 orthogonally placed planar gradiometers at 102 different positions. In a first step, a signal space separation algorithm, implemented in the Maxfilter program (version 2.2.15) provided by the MEG manufacturer, was used to clean the data from external noise and realign data from different blocks to a common standard head position. Data preprocessing was performed using Matlab R2020b (The MathWorks, Natick, Massachusetts, USA) and the FieldTrip Toolbox (Oostenveld et al., 2011). All data was filtered between 0.1 Hz and 30 Hz (Kaiser windowed finite impulse response filter) and downsampled to 100 Hz. To identify eye-blinks and heart rate artefacts, 50 independent components were identified from filtered (0.1 - 100 Hz), downsampled (1000 Hz) continuous data of the recordings from the entropy modulation paradigm, and on average 3 components were removed for every subject. All data was filtered between 0.1 Hz and 30 Hz (Kaiser windowed finite impulse response filter) and downsampled to 100 Hz. Data of the entropy modulation paradigm was epoched into segments of 1200 ms (from 400 ms before sound onset to 800 ms after onset). Multivariate pattern analysis (see quantification of individual prediction tendency) was carried out using the MVPA-Light package (Treder, 2020). Data of the listening task was temporally aligned with the corresponding speech envelope, which was extracted from the audio files using the Chimera toolbox (Smith et al., 2002) over a broadband frequency range of 100 Hz - 10 kHz (in 9 steps, equidistant on the tonotopic map of auditory cortex; see also **Figure 1E**).

Eye tracking data from both eyes were acquired using a Trackpixx3 binocular tracking system (Vpixx Technologies, Canada) at a sampling rate of 2 kHz with a 50 mm lens. Participants were seated in the MEG at a distance of 82 cm from the screen. Their chin rested on a chinrest to reduce head movements. Each experimental block started with a 13-point calibration / validation procedure that was used throughout the block. Blinks were automatically detected by the Trackpixx3 system and excluded from horizontal and vertical eye movement data. We additionally excluded 100 ms of data around blinks to control for additional blink artefacts that were not automatically detected by the eye tracker. We then averaged gaze position data from the left and right eye to increase the accuracy of gaze estimation (Cui & Hondzinski, 2006). Missing data due to blink removal were then interpolated with a piecewise cubic Hermite interpolation. Afterwards, data was imported into the FieldTrip Toolbox, bandpass filtered between 0.1 - 40 Hz (zero-phase finite impulse response (FIR) filter, order: 33000, hamming window), and cut to the length of respective speech segments of a block. Data were then

downsampled to 1000 Hz to match the sampling frequency of neural and speech envelope data and further corrected for a 16 ms delay between trigger onset and actual stimulation. Finally, gaze data was temporally aligned with the corresponding speech envelope and downsampled to 100 Hz for TRF analysis after an antialiasing low-pass filter at 20 Hz was applied (zero-phase FIR filter, order: 662, hamming window).

Quantification of Prediction Tendency

The quantification of individual prediction tendency was the same as in Schubert et al. (2023): We used an entropy modulation paradigm where participants passively listened to sequences of 4 different pure tones (f1: 440 Hz, f2: 587 Hz, f3: 782 Hz, f4: 1043 Hz, each lasting 100 ms) during two separate blocks, each consisting of 1500 tones presented with a temporally predictable rate of 3 Hz. Entropy levels (ordered / random) changed pseudorandomly every 500 trials within each block, always resulting in a total of 1500 trials per entropy condition. While in an “ordered” context certain transitions (hereinafter referred to as forward transitions, i.e. f1→f2, f2→f3, f3→f4, f4→f1) were to be expected with a high probability of 75%, self repetitions (e.g., f1→f1, f2→f2,...) were rather unlikely with a probability of 25%. However, in a “random” context all possible transitions (including forward transitions and self repetitions) were equally likely with a probability of 25% (see Figure 1A). To estimate the extent to which individuals anticipate auditory features (i.e. sound frequencies) according to their contextual probabilities, we used a multiclass linear discriminant analyser (LDA) to decode sound frequency (f1 - f4) from brain activity (using data from the 102 magnetometers) between -0.3 and 0 s in a time-resolved manner. Based on the resulting classifier decision values (i.e. d1 - d4 for every test-trial and time-point), we calculated individual prediction tendency. We define individual prediction tendency as the tendency to pre-activate sound frequencies of high probability (i.e. a forward transition from one stimulus to another: f1→f2, f2→f3, f3→f4, f4→f1). In order to capture any prediction-related neural activity, we trained the classifier exclusively on ordered forward trials (see Figure 1B). Afterwards, the classifier was tested on all self-repetition trials, providing classifier decision values for every possible sound frequency, which were then transformed into corresponding transitions (e.g. d1(t) | f1(t-1) “dval for 1 at trial t, given that 1 was presented at trial t-1” → repetition, d2(t) | f1(t-1) → forward,...). The tendency to represent a forward vs. repetition transition was contrasted for both ordered and random trials (see Figure 1C). Using self-repetition trials for testing, we ensured a fair comparison between the ordered and random contexts (with an equal number of trials and the same preceding bottom-up input). Thus, we quantified “prediction tendency” as the classifier's pre-stimulus tendency to a forward transition in an ordered context exceeding the same tendency in a random context (which can be attributed to carry-over processing of the preceding

stimulus). Then, using the summed difference across pre-stimulus times, one value can be extracted per subject (also see **Figure 1B**).

Encoding Models

To quantify the neural representations corresponding to the acoustic envelope, we calculated a multivariate temporal response function (TRF) using the Eelbrain toolkit (Brodbeck et al., 2023). A deconvolution algorithm (boosting; (David et al., 2007)) was applied to the concatenated trials to estimate the optimal TRF to predict the brain response from the speech envelope, separately for each condition (single vs. multi-speaker). Before model fitting, MEG data of 102 magnetometers were normalised by subtracting the mean and dividing by the standard deviation (i.e. z-scoring) across all channels (as recommended by (Crosse et al., 2016, 2021)). Similarly, the speech envelope was also z-scored, however, after the transformation the negative of the minimum value (which naturally would be zero) was added to the time-series to retain zero values ($z' = z+(\min(z)*-1)$). The defined time-lags to train the model were from -0.4 s to 0.8 s. To evaluate the model, the data was split into 4 folds, and a cross-validation approach was used to avoid overfitting (Ying, 2019). The resulting predicted channel responses (for all 102 magnetometers) were then correlated with the true channel responses to quantify the model fit and the degree of speech envelope tracking at a particular sensor location.

To investigate the effect of semantic violations, the true as well as the predicted data was segmented into single word epochs of 2 seconds starting at word onset (using a forced-aligner; (Kisler et al., 2017; Schiel & Ohala, 1999)). We selected semantic violations as well as their lexically identical controls and correlated true with predicted responses for every word. We then averaged the result within each condition (i.e. single vs. multi-speaker) and word type (i.e. high vs. low suprisal).

The same TRF approach was also used to estimate ocular speech tracking, separately predicting eye movements in the horizontal and vertical direction using the same time-lags (from -0.4 s to 0.8 s). The same z-scoring was applied to the speech envelope. However, horizontal and vertical eye channel responses were normalised within channels.

Mediation Analysis

To investigate the contribution of eye movements to neural speech tracking, we approached a mediation analysis similar to Gehmacher et al. (2024). The TRFs that we obtained from these encoding models can be interpreted as time-resolved weights for a predictor variable that aims to explain a dependent variable (very similar to beta-coefficients in classic regression analyses). We simply compared the plain effect of the speech envelope on neural activity to

its direct (residual) effect by including an indirect effect via eye movements into our model. In order to account for a reduction in speech envelope weights simply due to the inclusion of this additional (eye-movement) predictor, we obtained a control model by including a time-shuffled version of the eye-movement predictor in addition to the unchanged speech envelope. Thus, the plain effect (i.e. speech envelope predicting neural responses) is represented in the absolute weights (i.e. TRFs) obtained from this control model with the speech envelope and shuffled eye movement data as the predictor of neural activity. The direct (residual) effect (not mediated by eye movements) is obtained from the model including the speech envelope as well as true eye movements and is represented in the exclusive weights (c') of the former predictor (i.e. speech envelope). If model weights are significantly reduced by the inclusion of true eye movements into the model in comparison to a model with a time-shuffled version of the same predictor (i.e. c' < c), this indicates that a meaningful part of the relationship between the speech envelope and neural responses was mediated by eye movements (for further details see also Gehmacher et al., 2024).

Source and Principal Component Analysis

In order to estimate the location along with the temporal profile of this mediation effect and at the same time minimise the number of comparisons, we computed the main components of the effect, projected into source space. For this, we used all 306 MEG channels for our models as described in the previous section (note that here MEG responses were z-scored within channel type, i.e. within magnetometers and gradiometers separately). The resulting envelope model weights were then projected into source-space using an LCMV beamforming approach (Van Veen et al., 1997). Spatial filters were computed by warping anatomical template images to the individual head shape and further brought into a common space by co-registering them based on the three anatomical landmarks (nasion, left and right preauricular points) with a standard brain from the Montreal Neurological Institute (MNI, Montreal, Canada; (Mattout et al., 2007)). For each participant, a single-shell head model (Nolte, 2003) was computed. Finally, as a source model, a grid with 1 cm resolution and 2982 voxels based on an MNI template brain was morphed into the brain volume of each participant. This allowed group-level averaging and statistical analysis as all the grid points in the warped grid belong to the same region across subjects.

Afterwards, envelope TRF edges of both plain and direct models were cropped to time-lags from -0.3 - 0.7 s to exclude potential regression artefacts. We then subtracted the absolute direct from the absolute plain TRF to obtain the 'abs' indirect (mediation) effect. We transformed the 2982 voxel space into an orthogonal component space with a principal component analysis (PCA) based on the grand average mediation effect. The number of

components for further analysis was visually determined by plotting the ranked cumulative explained variance and estimating the 'elbow'. Based on this inspection, we extracted weight matrices of the first three components and multiplied individual 'abs' plain TRFs (from the control model with time-shuffled eye movements) and 'abs' direct TRFs (from the test model with true eye-movements) with this weight matrix to obtain individual source location and temporal profile of the mediation components. The projected, single subject data was then used for statistical analysis.

Statistical Analysis and Bayesian Models

To calculate the statistics, we used Bayesian multilevel regression models with Bambi (Capretto et al., 2022), a python package built on top of the PyMC3 package (Salvatier et al., 2016), for probabilistic programming. First, in replication of Schubert et al. (2023), we investigated the effect of neural speech tracking and its relation to individual prediction tendency as well as the influence of increasing noise (i.e. adding a distractor speaker). Separate models were calculated for all 102 magnetometers using the following model formula:

*Neural speech envelope tracking ~ condition * prediction tendency + (1|subject)*

(Note that "speech envelope tracking" refers to the correlation between predicted and true responses from the aforementioned encoding model and prediction tendency was always z-scored before entering the models). Similarly, we investigated the effect of ocular speech tracking under different conditions of attention (i.e. attended single speaker, attended multi-speaker and unattended multi-speaker). In addition to replicating the findings from Gehmacher et al. (2024), we extended this analysis for a detailed investigation of horizontal and vertical ocular speech envelope tracking and further included prediction tendency as a predictor:

Ocular speech envelope tracking ~ 0 + condition + prediction tendency + (1|subject)

To investigate the effect of semantic violations, we compared envelope tracking between target words (high surprisal) and lexically matched controls (low surprisal), both for neural as well as ocular speech tracking:

*Speech envelope tracking ~ word type * condition + (1|subject)*

For the mediation analysis, we compared weights (TRFs) for the speech envelope between a plain (control) model that included shuffled eye-movements as second predictor and a residual (test) model that included true eye movements as a second predictor (i.e. c' < c). In order to investigate the temporal dynamics of the mediation, we included time-lags (-0.3 - 0.7 s) as a fixed effect into the model. To investigate potential null effects (neural speech tracking that is definitely independent of eye movements), we used a region of practical equivalence (ROPE) approach (see for example Kruschke, 2018). The dependent variable (absolute weights) was z-scored across time-lags and models to get standardised betas for the mediation effect (i.e. c' < c):

absolute weights ~ 0 + time-lags + (1|subject)

As suggested by (Kruschke, 2018), a null effect was considered for betas ranging between -0.1 and 0.1. Accordingly, it was considered a significant mediation effect if betas were above 0.1 and at minimum two neighbouring time-points also showed a significant result.
Finally, we investigated the relationship between neural as well as ocular speech tracking and behavioural data using the averaged accuracy from the questions on story content that were asked at the end of each trial (in the following: "comprehension") and averaged subjective ratings of difficulty. In order to avoid calculating separate models for all magnetometers again, we selected 10% of the channels that showed the strongest speech encoding effect and used the averaged speech tracking (z-scored within condition before entering the model) as predictor:

*comprehension ~ neural speech tracking * condition + prediction tendency + (1|subject)*
*difficulty ~ neural speech tracking * condition + prediction tendency + (1|subject)*

To investigate the relation between ocular speech tracking and behavioural performance, we used the following model (again speech tracking was z-scored within condition) separately for horizontal and vertical gaze direction:

*comprehension ~ ocular speech tracking * condition + (1|subject)*
*difficulty ~ ocular speech tracking * condition + (1|subject)*

For all models (as in Gehmacher et al, 2024 and Schubert et al., 2023), we used the weakly- or non-informative default priors of Bambi (Capretto et al., 2022) and specified a more robust Student-T response distributions instead of the default gaussian distribution. To summarise model parameters, we report regression coefficients and the 94% high density intervals (HDI)

of the posterior distribution (the default HDI in Bambi). Given the evidence provided by the data, the prior and the model assumptions, we can conclude from the HDIs that there is a 94% probability that a respective parameter falls within this interval. We considered effects as significantly different from zero if the 94%HDI did not include zero (with the exception of the mediation analysis where the 94%HDI had to fall above the ROPE). Furthermore, we ensured the absence of divergent transitions ($\hat{r} < 1.05$ for all relevant parameters) and an effective sample size > 400 for all models (an exhaustive summary of Bayesian model diagnostics can be found in (Vehtari et al., 2021)). Finally, when we estimated an effect on brain sensor level (using all 102 magnetometers), we defined clusters for which an effect was only considered as significant if at minimum two neighbouring channels also showed a significant result.

RESULTS

Individual prediction tendency is related to neural speech tracking

In the first step, we wanted to investigate the relationship between individual prediction tendency and neural speech tracking under different conditions of noise. We quantified prediction tendency as the individual tendency to represent auditory features (i.e. pure tone frequency) of high probability in advance of an event (i.e. pure tone onset). Importantly, this prediction tendency was assessed in an independent entropy modulation paradigm (see **Fig. 1**). Replicating previous findings (Schubert et al., 2023), we found widespread encoding of clear speech, predominantly over auditory processing regions (**Fig. 2A**), that was decreased in a multi-speaker condition (**Fig. 2B**). Furthermore, a stronger prediction tendency was associated with increased neural speech tracking over left frontal sensors (see **Fig. 2C**). We found no interaction between prediction tendency and condition (see **Fig. 2D**). These findings indicate that the relationship between individual prediction tendency and neural speech tracking is largely unaffected by demands on selective attention.

Additionally, we wanted to investigate how semantic violations affect neural speech tracking. For this reason, we introduced rare words of high surprisal into the story by randomly replacing half of the nouns at the end of a sentence with the other half. In a direct comparison with lexically identical controls, we found an increased neural tracking of semantic violations over left temporal areas (see **Fig. 2E**). Furthermore, we found no interaction between word surprisal and speaker condition (see **Fig. 2F**). These findings indicate an increased representation of surprising words independent of background noise.

In sum, we found that individual prediction tendency as well as semantic predictability affect neural speech tracking.

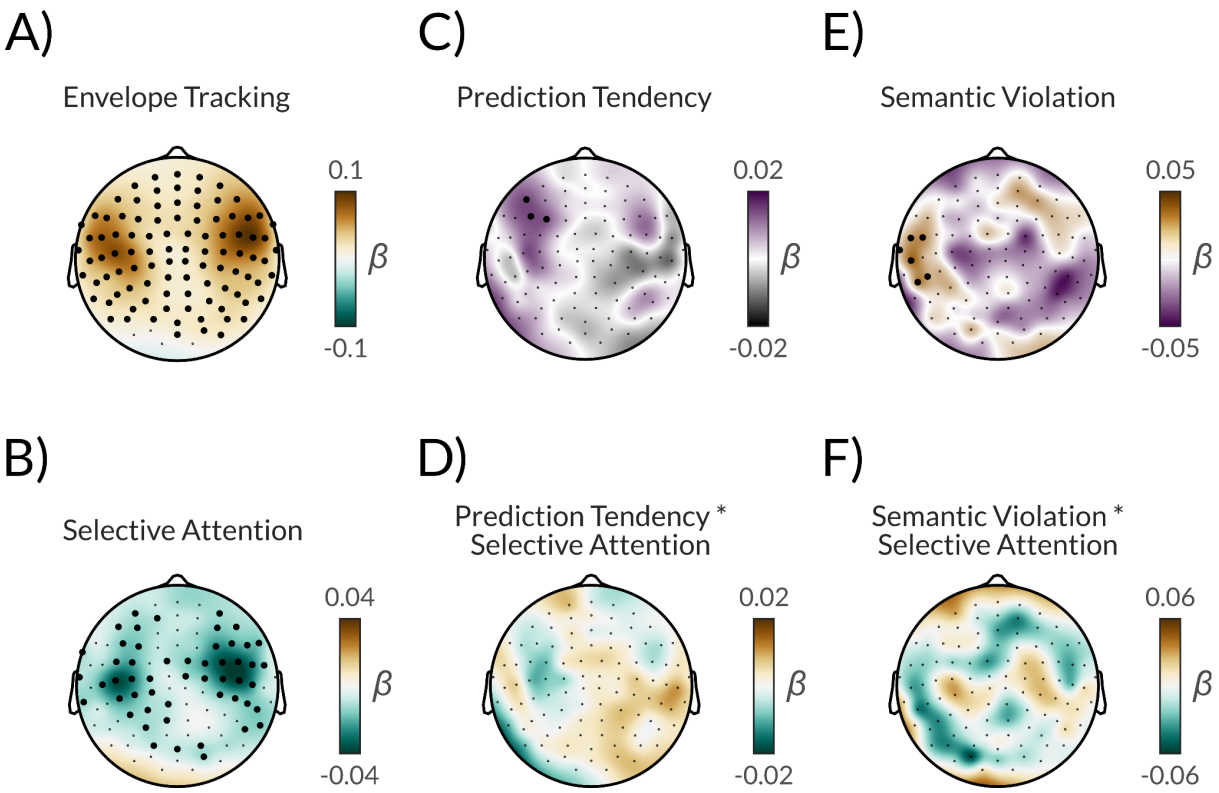

Figure 2: *Neural speech tracking is related to prediction tendency and word surprisal independent of selective attention.* **A)** In a single speaker condition, neural tracking of the speech envelope was significant for widespread areas, most pronounced over auditory processing regions. **B)** The condition effect indicates a decrease of neural speech tracking with increasing noise (1 distractor). **C)** Stronger prediction tendency was associated with increased neural speech tracking over left frontal areas. **D)** However, there was no interaction between prediction tendency and conditions of selective attention. **E)** There was an increased neural tracking of semantic violations over left temporal areas. **F)** There was no interaction between word surprisal and speaker condition, suggestive of a representation of surprising words independent of background noise. Statistics were performed using Bayesian regression models. Marked sensors show 'significant' clusters where at minimum two neighbouring channels showed a significant result. N = 29.

Eye movements track acoustic speech in selective attention

In a second step, we aimed to replicate previous findings from Gehmacher and colleagues (2024), showing that eye movements track the acoustic features of speech in absence of visual information. For this, we separately predicted horizontal and vertical eye movements from the acoustic speech envelope. For vertical eye movements, we found evidence for attended speech tracking in a single speaker condition (β = 0.012, 94%HDI = [0.001, 0.0023]) but not in a multi-speaker condition (β = 0.006, 94%HDI = [-0.005, 0.016]; see **Figure 3A**). There was no evidence for tracking of the distracting speech stream (β = -0.008, 94%HDI = [-0.020, 0.003]). On the contrary, horizontally directed eye movements selectively track attended (β = 0.014, 94%HDI = [0.005, 0.024]), but not unattended (β = -0.002, 94%HDI = [-0.011, 0.007]) acoustic speech in a multi-speaker condition (see **Figure 3B**). Speech tracking

in a single speaker condition did not reach significance (β = 0.009, 94%HDI = [-0.001, 0.017]) for horizontal eye movements. These findings indicate that eye movements selectively track attended, but not unattended acoustic speech. Furthermore, there seems to be a dissociation between horizontal and vertical ocular speech tracking, indicating that horizontal movements track attended speech in a multi-speaker condition, whereas vertical movements track attended speech in a single speaker condition (see **Supplementary Table 2** for a summary of ocular speech tracking effects).

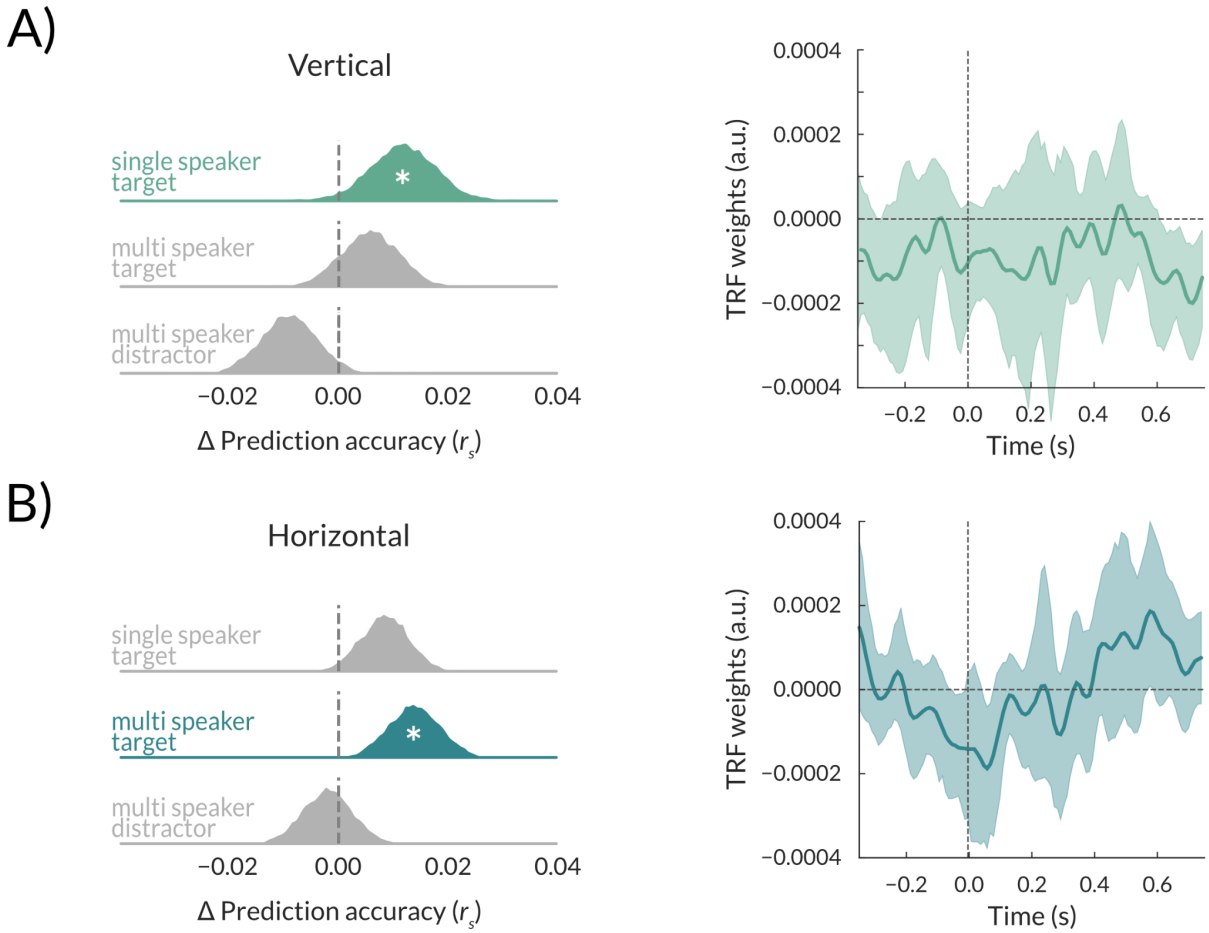

Figure 3: *Ocular speech tracking is dependent on selective attention.* **A)** Vertical eye movements 'significantly' track attended clear speech, but not in a multi-speaker condition. Temporal profiles of this effect show a downward pattern (negative TRF weights). **B)** Horizontal eye movements 'significantly' track attended speech in a multi-speaker condition. Temporal profiles of this effect show a left-rightwards (negative to positive TRF weights) pattern. Statistics were performed using Bayesian regression models. A '*' within posterior distributions depicts a significant difference from zero (i.e. the 94%HDI does not include zero). Shaded areas in TRF weights represent 95% confidence intervals. *N* = 29.

Additionally, we wanted to investigate if predictive processes are related to ocular speech tracking using the same approach as for neural speech tracking (see previous section). Crucially, we found no evidence for a relationship between individual prediction tendency and

ocular speech tracking on a vertical (β = 0.001, 94%HDI = [-0.005, 0.008]) or horizontal (β = -0.001, 94%HDI = [-0.007, 0.004]) plane. Similarly, we found no difference in ocular speech tracking between words of high surprisal and their lexically matched controls (see **Supplementary Table 3**). These findings indicate that individuals engage in ocular speech tracking independent of their individual prediction tendency or overall semantic probability.

Neural speech tracking is mediated by eye movements

Additionally, we performed a similar, but more detailed mediation analysis compared to Gehmacher and colleagues (2024) in order to separately investigate the contribution of horizontal and vertical eye movements to neural speech tracking across different time lags. Following mediation analysis requirements, only significant ocular speech tracking effects were further considered, i.e. vertical eye movements in the clear speech condition and horizontal eye movements in response to a target in the multi-speaker condition. We compared the plain effect (c) of neural speech tracking (using a simple stimulus model with speech envelope and a time-shuffled version of the respective eye movements as the predictors for neural responses) to its direct (residual) effect (c') by including true horizontal or vertical eye movements as a second predictor into the stimulus model. The decrease in predictor weights from the plain to the residual stimulus model indicates the extent of the mediation effect. To establish a time-resolved mediation analysis in source-space, we computed the main components of the effect (via PCA, see **Figure 4**). We found significant mediation effects for all three principal components (PC) for both vertical eye movements in the clear speech condition (see **Figure 4A**) and horizontal eye movements in response to a target in the multi-speaker condition (see **Figure 4B**). PC1 reached significance (nearly) across the whole time-window from -0.3 - 0.7 s over widespread auditory regions in both conditions with a right-hemispheric dominance, peaking at ~ 0.18 s. Interestingly, PC2 instead loaded mostly on left auditory regions, with a slight anticipation effect for both vertical and horizontal eye movements. Additionally, PC2 loaded over left parietal areas for the mediation via horizontal eye movements in the multi-speaker condition. In both conditions, PC2 showed another early peak at ~80 ms. The mediation effect remained significant almost entirely over positive time-lags for both conditions. We found less contributions of vertical eye movements to neural clear speech tracking for PC3 with significant clustering ~ 0.2 - 0.4 s over scattered cortical regions. For horizontal eye movements, PC3 still showed an effect in left auditory areas at several short peaks (~ -0.2 s, ~ 0.05 s, and ~ 0.4 s). Taken together, this suggests that eye movements contribute considerably to neural speech tracking over widespread cortical areas that are commonly related to speech processing and attention.

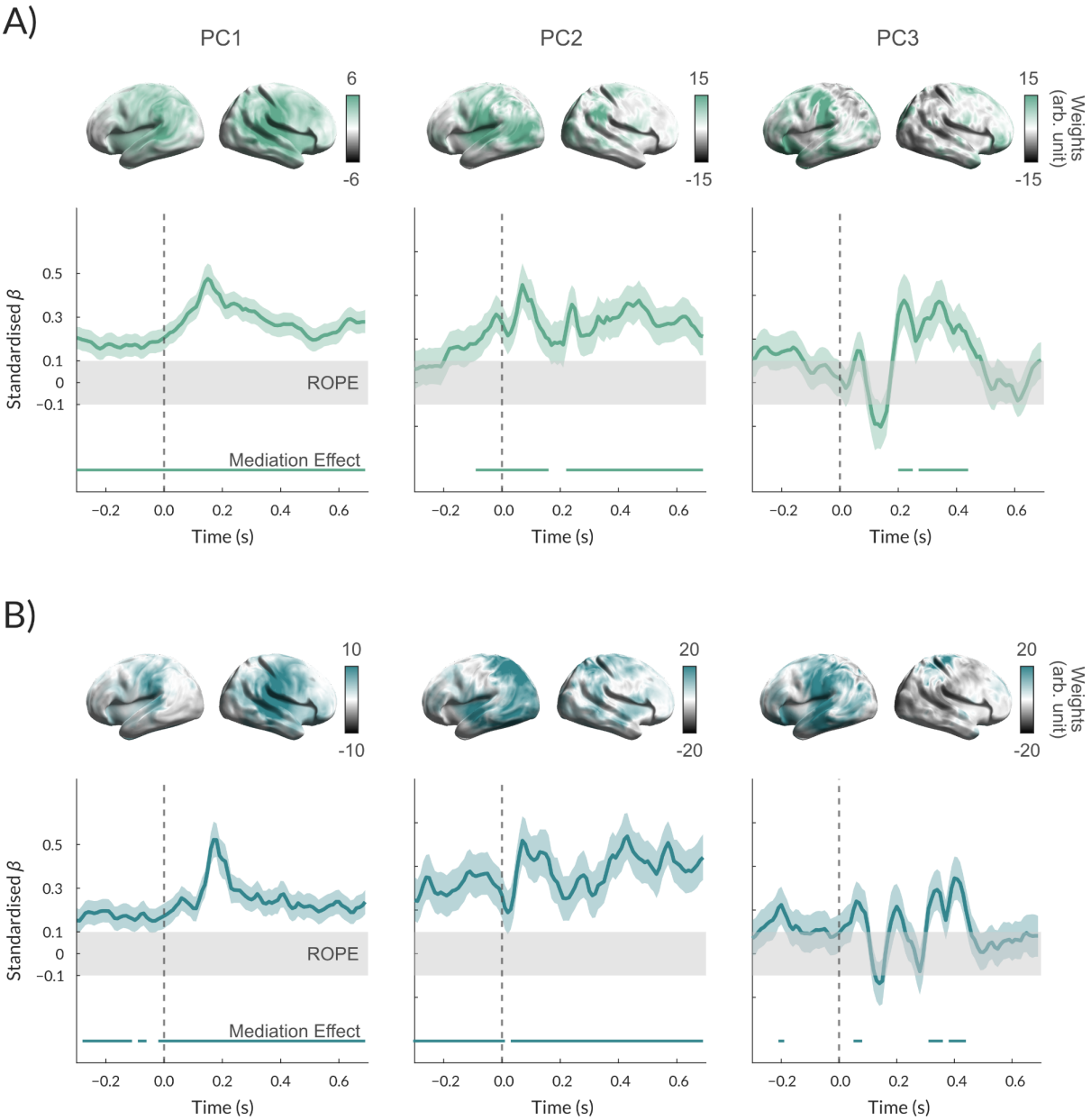

Figure 4: *Ocular speech tracking and selective attention to speech share underlying neural computations.* **A)** Vertical eye movements significantly mediate neural clear speech tracking throughout the time-lags from -0.3 - 0.7 s for principal component 1 (PC1) over right-lateralized auditory regions.This mediation effect propagates to more leftwards lateralized auditory areas over later time-lags for PC2 and PC3. **B)** Horizontal eye movements similarly contribute to neural speech tracking of a target in a multi-speaker condition over right-lateralized auditory processing regions for PC1, also with significant anticipatory contributions and a clear peak at ~ 0.18 s. PC2 shows a clear left-lateralization, however not only over auditory, but also parietal areas almost entirely throughout the time-window of interest with a clear anticipatory effect starting at -0.3 s. For PC3, there still remained a small anticipatory cluster ~ -0.2 s again over mostly left-lateralized auditory regions. Colour bars represent PCA weights for the group-averaged mediation effect. Shaded areas on time-resolved model-weights represent regions of practical equivalence (ROPE) according to Kruschke et al. (2018). Solid lines show 'significant' clusters where at minimum two neighbouring time-points showed a significant mediation effect. Statistics were performed using Bayesian regression models. N = 29.s

Neural and ocular speech tracking are differently related to comprehension

In a final step, we addressed the behavioural relevance of neural as well as ocular speech tracking respectively. At the end of every trial, participants were asked to evaluate 4 different true or false statements about the target story. The accuracy of these responses was averaged within condition (single vs. multi-speaker) and served as an approximation for semantic speech comprehension. Additionally, we evaluated the averaged subjective ratings of difficulty (which were given on a 5-point likert scale).

To avoid calculating separate models for all 102 magnetometers, neural encoding was averaged over selected channels (10% showing the strongest encoding effect). We found no significant relationship between neural speech tracking and comprehension (β = 0.138, 94%HDI = [-0.050, 0.330]), no interaction between neural speech tracking and condition (β = -0.088, 94%HDI = [-0.390, 0.201]), but a significant effect for condition (β = -0.438, 94%HDI = [-0.714, -0.178]), indicating that comprehension was decreased in the multi-speaker condition (see **Figure 5A**). Similarly, we found no effect of prediction tendency on comprehension (β = 0.050, 94%HDI = [-0.092, 0.197]). When investigating subjective ratings of task difficulty, we also found no effect for neural speech tracking (β = 0.156, 94%HDI = [-0.079, 0.382]), no interaction between neural speech tracking and condition (β = 0.041, 94%HDI = [-0.250, 0.320]), nor individual prediction tendency (β = -0.058, 94%HDI = [-0.236, 0.127]). There was, however, a significant difference between conditions (β = 1.365, 94%HDI = [1.128, 1.612]), indicating that the multi-speaker condition was rated more difficult than the single speaker condition.

In contrast to neural findings, we found a negative relationship between ocular speech tracking and comprehension for vertical as well as horizontal eye movements irrespective of condition (see **Figure 5B-C** and **Supplementary Table 4**). This suggests that participants with weaker performance in semantic comprehension increasingly engaged in ocular speech tracking. There was, however, no significant relationship between subjectively rated difficulty and ocular speech tracking (see **Supplementary Table 5**). Presumably, subjective ratings are less comparable between participants, which might be one reason why we did find an effect for objective but not subjective measures. In sum, the current findings show that ocular speech tracking is differently related to comprehension than neural speech tracking, suggesting that they might not refer to the same underlying concept (e.g. improved representations vs. increased attention). A mediation analysis, however, suggests that they are related and that ocular speech tracking contributes to neural speech tracking (or vice versa).

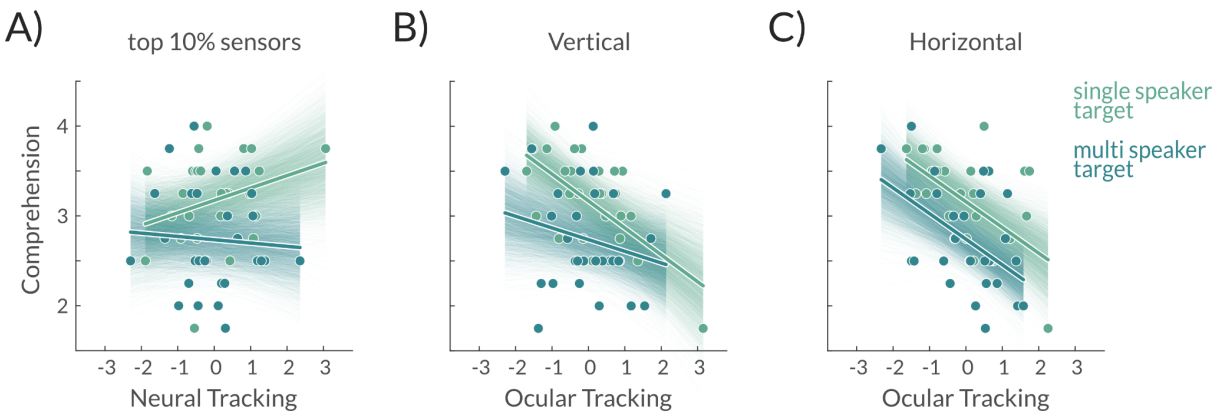

Figure 5: *Ocular, but not neural speech tracking is related to semantic speech comprehension.* **A)** There was no significant relationship between neural speech tracking (10% sensors with strongest encoding effect) and comprehension, however, a condition effect indicated that comprehension was generally decreased in the multi-speaker condition. **B & C)** A 'significant' negative relationship between comprehension and vertical as well as horizontal ocular speech tracking shows that participants with weaker comprehension increasingly engaged in ocular speech tracking. Statistics were performed using Bayesian regression models. Shaded areas represent 94% HDIs. *N* = 29.

DISCUSSION

In the current study, we aimed to replicate and extend findings from two recent studies from our lab in order to integrate them into a comprehensive view of speech processing. In Schubert and colleagues (2023), we found that individual prediction tendencies are related to cortical speech tracking. In Gehmacher and colleagues (2024), we found that eye movements track acoustic speech in selective attention (a phenomenon that we termed "ocular speech tracking"). In the present study, we were able to replicate both findings, providing further details and insight into these phenomena.

In a first step, we confirmed that individuals with a stronger prediction tendency (which was inferred from anticipatory probabilistic representations in an independent paradigm) showed an increased neural speech tracking over left frontal areas. Thus, the finding that prediction tendencies generalise across different listening situations seems to be robust. This stresses the importance of research focusing on the beneficial aspects of individual "traits" in predictive processing. As suggested in a comparative study (Schubert et al., 2023), these predictive tendencies or traits do not generalise across modalities but seem to be reliable within the auditory modality. We suggest that they could serve as an independent predictor of linguistic abilities, complementing previous research on statistical learning (e.g. (Siegelman & Frost, 2015)).

In a second step, we were able to support the assumption that eye movements track prioritised acoustic modulations in a continuous speech design without visual input. With the current (extended) replication, we were able to address an important consideration from Gemacher and colleagues (2024): Ocular speech tracking is not restricted to a simplistic 5-word sentence

structure but can also be found for continuous, narrative speech. In line with previous results, we found that horizontal ocular speech tracking is increased in a multi-speaker (compared to a single speaker) condition in response to an attended target but not a distractor speaker. In contrast, we found that vertical eye movements solely track auditory information in a single speaker condition. The differential contribution of vertical vs. horizontal eye movements to auditory processing is not a widely studied topic in neuroscience and further replications are necessary to establish the robustness of this dissociation. One possible explanation for these findings is, that in the multi-speaker condition, the demand for spatial segregation was increased compared to the single speaker condition. Although speech was always presented at phantom centre even for both speakers during the multi-speaker condition, multiple speakers are distributed mostly horizontally in natural human environments. The observed effect might resemble a residual of this learned association between a spatially segregated target speaker's acoustic output and a lateral shift of gaze (and attention) on the horizontal plane towards the target's source location, maximising information and thereby minimising uncertainty about the targeted auditory object.

Irrespective of the potential differences in gaze direction and condition, we found that ocular movements contribute to neural speech tracking over widespread auditory and parietal areas, replicating the sensor-level findings of Gehmacher et al. (2024). In addition, time-resolved analyses of this mediation effect further extend these findings and suggest even anticipatory contributions. It is important to note that our current findings do not allow for inference on directionality. Hence, it is possible that ocular mediation of speech tracking may reflect a) active (ocular) sensing for information driven by (top-down) selective attention or b) improved neural representations as a consequence of temporally aligned increase of sensory gain or c) (not unlikely) both.

Despite the finding that eye movements mediate neural speech tracking, the behavioural relevance for semantic comprehension seems to differ between ocular and neural speech tracking. To be more specific, we found a negative association between ocular speech tracking and comprehension, indicating that subjects with weaker performance in comprehension increasingly engaged in ocular speech tracking. Interestingly, we did not find a relationship between neural tracking and comprehension. This suggests that speech envelope tracking may reflect different underlying (cognitive) processes depending on where the response was measured (eyes vs. brain). It should be noted that our questions on story content (see section Experimental Procedure for an example) targeted a broad range of cognitive processes from intelligibility to memory encoding / retrieval. Thus, it can be argued that the inverse relationship between ocular tracking and semantic comprehension might be the result of individual differences in the trade-off between the focus on lower level acoustics (with temporal sharpening) and the engagement in higher-level cognitive processes such as memory

(requiring the integration of information over longer timescales). This interpretation would also explain why we did not find a relationship with neural speech tracking, as sensors that show the highest acoustic tracking might not be relevant for more extensive semantic processing. However, a more detailed investigation would require carefully graduated measures of intelligibility and comprehension, which was beyond the scope of the current analysis.

Crucially, as we found no difference in ocular speech tracking between words of high surprisal and their lexically matched control, we argue that ocular speech tracking may not be indicative of “qualia” in higher level (linguistic or semantic) stimulus representations. Instead, we propose that it reflects an attentional (and - also in relation to the point above - maybe even a compensatory) mechanism. For example, it has been found that attention modulates phase in the auditory brainstem response, locking it to the pitch structure of attended but not unattended speech (Forte et al., 2017). Similarly, it is possible that increased ocular tracking indicates increased selective attention in the auditory periphery, rather than improved representations of stimulus content on a higher level.

Interestingly, we were not able to relate individual prediction tendency to ocular speech tracking. This raises the question whether ocular speech tracking is modulated not by predictive, but other (for example attentional) processes. Indeed, the current findings suggest that prediction tendencies and active ocular sensing are related to different aspects of neural speech processing. We propose a perspective in which active sensing is guided by selective attention, whereas anticipatory prediction tendency is an independent correlate of neural speech tracking. Even though predictive processing as well as attention might be considered as the two pillars on which perception rests, research investigating their selective contributions and interactions is rare. Summerfield and de Lange (2014) have argued that predictions and attention are distinctive mechanisms as the former refers to the probability and the latter to the relevance of an event, arguing that events can be conditionally probable but irrelevant for current goals and vice versa. Similarly, it has been proposed that attention can be integrated into the predictive coding framework in reference to the optimization of sensory precision (Feldman & Friston, 2010). In this view, prediction tendency encodes probabilistic representations of a feature, whereas selective attention determines the precision of sensory inflow that leads to internal model updating. This interpretation is in line with our finding that a) anticipatory feature-specific predictions can be found in a passive listening task, b) prediction tendencies are not increasingly linked to speech tracking with increasing demands on selective attention, whereas on the other hand c) ocular speech tracking seems to increase with selective attention (at least on a horizontal plane), d) remains unaffected by semantic probability and e) contributes to neural speech processing over widespread auditory and

parietal areas with a potential overlap with attentional networks. For this reason, we refer to ocular speech tracking as an “active sensing” mechanism that implements the attentional optimization of sensory precision. Instead of passively transducing any input into neural activity, we actively engage with our environment to maximise information, i.e. reduce uncertainty. It has been suggested that motor routines contribute to temporal precision of selective attention, largely determining sensory inflow and hence perception (Morillon et al., 2015; Schroeder et al., 2010). In the current framework, we refer to active sensing via eye movements as temporally aligned shifts of gaze at exact points in time aligned with its intensity (information content) fluctuations. In particular, active ocular sensing may even help to increase sensory gain already in the auditory periphery at specific intervals, synchronised with the temporal modulation of attended (but not unattended) speech. As a consequence of this “sensory gating” already at early stages, eye movements potentially affect auditory processing from the ear to the cortex (also see (Leszczynski et al., 2023) for saccadic modulations of cortical excitability in auditory areas), contributing to neural speech representations rather than encoding them. Again, this idea is supported by our finding that ocular speech tracking seems to be unaffected by semantic violations.
Indeed, similar active ocular sensing mechanisms have already been suggested to facilitate sound localization (Lovich et al., 2023). Since the current paradigm did not allow for spatial segregation (as competing speech streams have all been presented at phantom centre), it required a different strategy such as (spectro-)temporal differentiation. We argue that active ocular sensing increases temporal precision of complex acoustic input (in order to parse speech into its components of rhythmic, temporally predictable patterns within a particular speaker).

Based on the joint findings of the present as well as its preceding studies (Gehmacher et al., 2024, Schubert et al., 2023), we propose a unified working model in which anticipatory predictions as well as active sensing work (independently) together to support auditory speech perception (see **Figure 6** for a schematic illustration). We suggest that anticipatory predictions about a feature help to interpret auditory information at different levels along the perceptual hierarchy. Accordingly, these predictions carry high feature-specificity but low temporal precision (as they are anticipatory in nature). It should be noted that the representational content is likely to be different at different levels of the perceptual hierarchy (e.g. encoding of acoustics vs. semantics). However, the probability of a feature at a higher level affects the interpretation (and thus the representation and prediction) of a feature at lower levels. Furthermore, individuals differ in their general tendency to create such anticipatory predictions, which leads to differences in neural speech tracking. Active sensing, on the other hand, increases temporal precision - potentially already at the early stages of sound transduction -

to facilitate bottom-up processing of selectively attended input. Crucially, we suggest that active (ocular) sensing does not necessarily convey feature- or content-specific information, it is merely used to boost (and conversely filter) sensory input at specific timescales (similar to neural oscillations). Our research suggests that this active ocular sensing mechanism or "filter" is driven by internal (attentional) goals rather than prior beliefs (stressing the distinction to other active sensing mechanisms).

In this speculative framework for listening, auditory inflow is, at a basic level, temporally modulated via active ocular sensing, and incoming information is interpreted based on probabilistic assumptions. Optimal speech processing requires a careful balance between the weighting of internal models and the allocation of resources based on current goals. We suggest that in the case of speech processing, this challenge results in an independent adaptation of feature-based precision-weighting by predictions on the one hand and temporal precision-weighting by selective attention on the other.

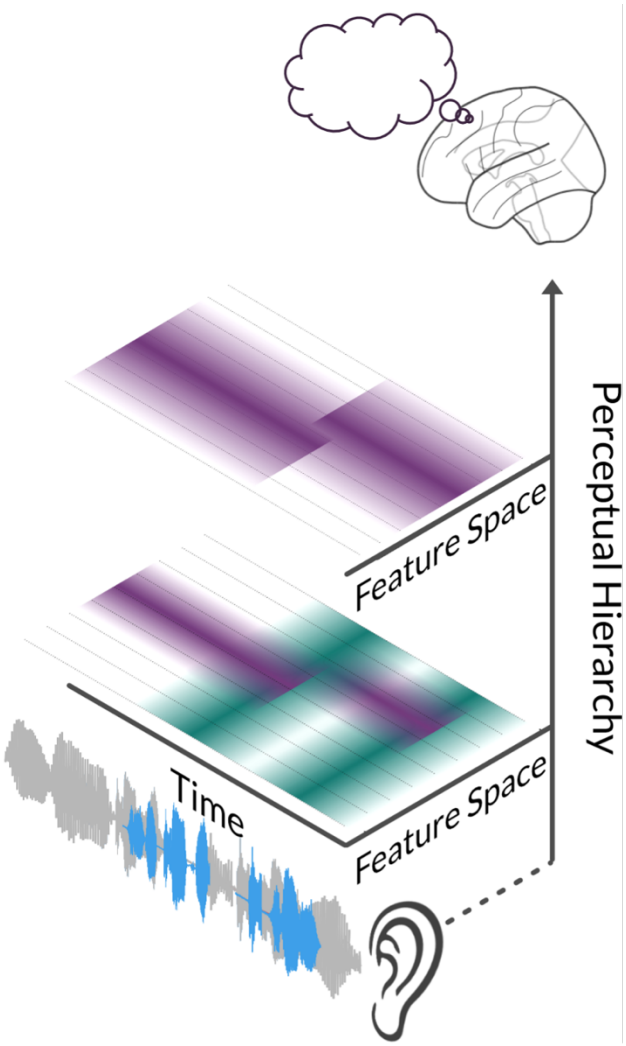

Figure 6: *A schematic illustration of the framework.* Anticipatory predictions help to interpret auditory information at different levels along the perceptual hierarchy (purple) with high feature-specificity but low temporal precision. Active sensing (green) increases the temporal precision already at early stages of the auditory system to facilitate bottom-up processing of selectively attended input (blue). We suggest that auditory inflow, on a basic level, and speech segmentation, on a more complex level, is temporally modulated via active ocular sensing, and incoming information is interpreted based on probabilistic assumptions.

We suggest that future research on auditory perception should integrate conceptual considerations on predictive processing, active (crossmodal) sensing, and selective attention. In particular, it would be interesting whether ocular speech tracking can be observed for unfamiliar languages with unpredictable prosodic rate. Furthermore, the relationship between neural oscillations in selective attention and active sensing should be further investigated using experimental modulations to address the important, pending question of causality. Brain

stimulation (such as tACS; transcranial alternating current stimulation) could be used in an attempt to alter temporal processing frames and / or ocular speech tracking. With regards to the latter, future studies should focus on the potential consequences of inhibited active sensing (e.g. actively disrupting natural gaze behaviour) for neural speech tracking. Our interpretation suggests that increased tracking of unexpected input (i.e. semantic violations) should be affected by active ocular sensing if sensory gain in the periphery is indeed dependent on this mechanism. Currently, the findings propose that active ocular sensing, with its substantial contribution to neural speech tracking, is driven by selective attention, and not by individual differences in prediction tendency.

ACKNOWLEDGEMENTS

J.S. and Q.G. are supported by the Austrian Science Fund (FWF; Doctoral College "Imaging the Mind"; W 1233-B). Q.G. is also supported by the Austrian Research Promotion Agency (FFG; BRIDGE 1 project “SmartCIs”; 871232) and F.S. is supported by WS Audiology. Thanks to the whole research team. Special thanks to Manfred Seifter for his support in conducting the MEG measurements.

AUTHOR CONTRIBUTIONS

J.S. and Q.G. designed the experiment, analysed the data, generated the figures, and wrote the manuscript. T.H. recruited participants and supported the data analysis. F.S. supported the data analysis and edited the manuscript. N.W. acquired the funding, supervised the project, and edited the manuscript.

REFERENCES

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., Resnik, P., & Simon, J. Z. (2023). Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *eLife*, *12*, e85012. https://doi.org/10.7554/eLife.85012

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *Journal of Neuroscience*, *39*(38), 7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2022). Bambi: A

Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, *103*, 1–29. https://doi.org/10.18637/jss.v103.i15

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Frontiers in Neuroscience*, *15*.

Cui, Y., & Hondzinski, J. M. (2006). Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience Letters*, *396*(3), 257–262.

David, S., Mesgarani, N., & Shamma, S. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, *18*, 191–212. https://doi.org/10.1080/09548980701609235

Demarchi, G., Sanchez, G., & Weisz, N. (2019). Automatic and feature-specific prediction-related neural activity in the human auditory system. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-11440-1

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019

Feldman, H., & Friston, K. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00215

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife*, *6*, e27203. https://doi.org/10.7554/eLife.27203

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), Article 2. https://doi.org/10.1038/nrn2787

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2021).

Active listening. *Hearing Research*, *399*, 107998. https://doi.org/10.1016/j.heares.2020.107998

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Gehmacher, Q., Schubert, J., Schmidt, F., Hartmann, T., Reisinger, P., Rösch, S., Schwarz, K., Popov, T., Chait, M., & Weisz, N. (2024). Eye movements track prioritized auditory features in selective attention to natural speech. *Nature Communications*, *15*(1), 3692. https://doi.org/10.1038/s41467-024-48126-2

Hartmann, T., & Weisz, N. (2020). An introduction to the Objective Psychophysics Toolbox. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.585437

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, 1–16.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Leszczynski, M., Bickel, S., Nentwich, M., Russ, B. E., Parra, L., Lakatos, P., Mehta, A., & Schroeder, C. E. (2023). Saccadic modulation of neural excitability in auditory areas of the neocortex. *Current Biology*, *33*(7), 1185-1195.e6. https://doi.org/10.1016/j.cub.2023.02.018

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lovich, S. N., King, C. D., Murphy, D. L. K., Landrum, R. E., Shera, C. A., & Groh, J. M.

(2023). Parametric information about eye movements is sent to the ears. *Proceedings of the National Academy of Sciences*, *120*(48), e2303562120. https://doi.org/10.1073/pnas.2303562120

Mattout, J., Henson, R. N., & Friston, K. J. (2007). Canonical Source Reconstruction for MEG. *Computational Intelligence and Neuroscience*, *2007*, e67613. https://doi.org/10.1155/2007/67613

Morillon, B., Hackett, T. A., Kajikawa, Y., & Schroeder, C. E. (2015). Predictive motor control of sensory dynamics in auditory active sensing. *Current Opinion in Neurobiology*, *31*, 230–238. https://doi.org/10.1016/j.conb.2014.12.005

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Physics in Medicine & Biology*, *48*(22), 3637. https://doi.org/10.1088/0031-9155/48/22/002

Oberfeld, D., & Klöckner-Nowotny, F. (2016). Individual differences in selective attention predict speech identification at a cocktail party. *eLife*, *5*, e16747. https://doi.org/10.7554/eLife.16747

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*.

Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences*, *108*(37), 15516–15521. https://doi.org/10.1073/pnas.1108912108

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Schiel, F., & Ohala, J. J. (1999). *Automatic Phonetic Transcription of Non-Prompted Speech* (J. J. Ohala, Ed.; pp. 607–610). https://doi.org/10.5282/ubm/epub.13682

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010).

Dynamics of Active Sensing and perceptual selection. *Current Opinion in Neurobiology*, *20*(2), 172–176. https://doi.org/10.1016/j.conb.2010.02.010

Schubert, J., Schmidt, F., Gehmacher, Q., Bresgen, A., & Weisz, N. (2023). Cortical speech tracking is related to individual prediction tendencies. *Cerebral Cortex*, *33*(11), 6608–6619. https://doi.org/10.1093/cercor/bhac528

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), Article 6876. https://doi.org/10.1038/416087a

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. https://doi.org/10.1038/nrn3838

Treder, M. S. (2020). MVPA-Light: A Classification and Regression Toolbox for Multi-Dimensional Data. *Frontiers in Neuroscience*, *14*. https://www.frontiersin.org/articles/10.3389/fnins.2020.00289

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 Transactions on Biomedical Engineering*, *44*(9), 867–880. https://doi.org/10.1109/10.623056

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2), 667–718.

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. https://doi.org/10.1162/jocn_a_01467

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Yon, D., Lange, F. P. de, & Press, C. (2019). The Predictive Brain as a Stubborn Scientist. *Trends in Cognitive Sciences*, *23*(1), 6–8. https://doi.org/10.1016/j.tics.2018.10.003

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party”. *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zion-Golumbic, E., & Schroeder, C. E. (2012). Attention modulates ‘speech-tracking’ at a cocktail party. *Trends in Cognitive Sciences*, *16*(7), 363–364. https://doi.org/10.1016/j.tics.2012.05.004