eLife

Reviewed Preprint
v2 • May 13, 2025
Revised by authors

Reviewed Preprint
v1 • October 14, 2024

Neuroscience

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

Juliane Schubert ✉, **Quirin Gehmacher, Fabian Schmidt, Thomas Hartmann, Nathan Weisz**

Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria • Department of Experimental Psychology, University College London, London, United Kingdom • Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom • Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria

eLife Assessment

These are **valuable** findings for those interested in how neural signals reflect auditory speech streams, and in understanding the roles of prediction, attention, and eye movements in this tracking. However, the evidence as it stands is **incomplete**. Further analyses are needed to clarify how the observed results relate to the relevant theoretical claims.

https://doi.org/10.7554/eLife.101262.2.sa4

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.

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 "neural speech tracking" is enhanced in individuals with stronger prediction tendency, independent of situative demands on selective attention. Furthermore, we found an increased tracking of words of high surprisal, demonstrating the importance of predictive processes in speech perception. Here (and in the following) "neural speech tracking" refers to a correlation coefficient between actual brain responses and responses predicted from an encoding model based solely on the speech envelope.

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; Jin et al, 2018). In support of ideas that the motor system is actively engaged in speech perception (Limerman et al., 19985; Galantucci et al., 2006), 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. In the present study, we focused on gaze patterns rather than blink-related activity, both to replicate findings from Gehmacher et al. (2024) and because blink activity is

unsuitable for TRF analysis due to its discrete and artifact-prone nature. Hence, “Ocular speech tracking” (similarly to “neural speech tracking” refers to the correlation coefficient between actual eye movements and movements predicted from an encoding model based on the speech envelope.

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. Taking into account the aforementioned study by Schubert and colleagues (2023), the two recently uncovered predictors of neural tracking (individual prediction tendency and ocular tracking) raise several empirical questions regarding the relationship between predictive processes, selective attention, and active ocular sensing in speech processing:

1. Are predictive processes related to active ocular sensing in the same way they are to neural speech tracking? Specifically, do individuals with a stronger tendency to anticipate predictable auditory features, as quantified through pre-stimulus neural representations in an independent tone paradigm, show increased or even decreased ocular speech tracking, measured as the correlation between predicted and actual eye movements? Or is there no relationship at all?
2. To what extent does selective attention influence the relationship between prediction tendency, neural speech tracking, and ocular speech tracking? For example, does the effect of prediction tendency or ocular speech tracking on neural tracking differ between a single-speaker and multi-speaker listening condition?
3. Are individual prediction tendency and ocular speech tracking related to behavioral outcomes, such as comprehension and perceived task difficulty? Speech comprehension is assessed through accuracy on comprehension questions, and task difficulty is measured through subjective ratings.

Although predictive processes, selective attention, and active sensing have been shown to contribute to successful listening, their potential interactions and specific roles in naturalistic speech perception remain unclear. Addressing these questions will help disentangle their contributions and establish an integrated framework for understanding how neural and ocular speech tracking support speech processing.

Here, we set out to answer these questions by a) replicating aforementioned key findings in a single experiment, and b) operationalising prediction tendency, active ocular sensing, and attention in an integrated framework of neural speech tracking. The purpose of this conceptual framework will be to organise these entities according to their assumed function for speech processing and to describe their relationship with each other. 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 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 conceptual 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**). 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, 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 **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 (N = 79) and replaced them with the other half to induce unexpected semantic violations. The swap of nouns happened in the written script before the audio material was recorded in order to avoid any effects of audio clipping. Narrators were aware of the semantic violations and had been instructed to read out the words as normal. Consequently, all target words occurred twice in the text, once in a natural context (serving as "lexical controls") and once in a mismatched context (serving as "semantic violations") within each trial, resulting in two sets of lexically identical words that differed greatly in their contextual probabilities (see **Figure 1F** for an example). Participants were unaware of these semantic violations. 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

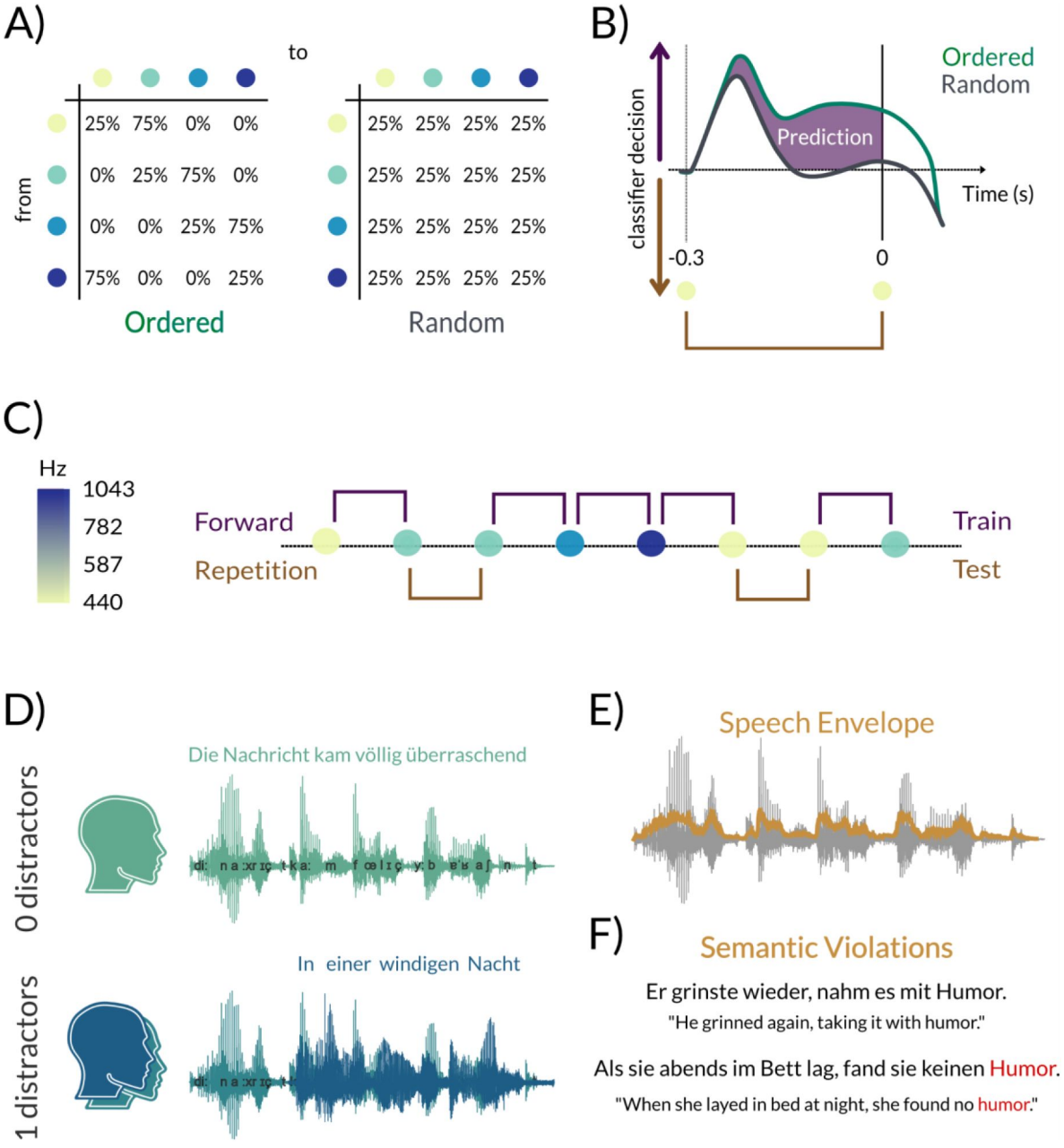


Figure 1

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.

“Winterzauber in der kleinen Keksbäckerei” by Holly Hepburn.
“Die Möwe Jonathan” by Richard Bach.
“Die Eskimos: Geschichte und Schicksal der Jäger im hohen Norden by T. Jeier.
“Sofies Welt” by Jostein Gaarder.

Note: each story was narrated by a target speaker (2 x ~3 min excerpts each), and a distractor speaker (1 x ~3 min different excerpt of the same story).

Table 1

List of all ebooks and short stories that were used as a basis for audio material.

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**). This design gives us the opportunity to directly compare the processing of “predictable” and “unpredictable” sounds of the same frequency in a time-resolved manner.

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. We specifically looked at a prestimulus window in order to capture top-down expectations driven by contextual regularity and temporal predictability in the design. The classification approach enables us to infer on “anticipatory predictions” (i.e. the representation of sound frequencies of high probabilities before their observation). 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., 2021). 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) 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, we used the same TRFs trained on the whole dataset (with 4-fold cross validation) since single word epochs would have been too short to derive meaningful TRFs. Instead, 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 (thus, we conducted the same analysis as for the overall encoding effect, focusing on only part of the data). 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., 2020), a python package built on top of the PyMC3 package (Salvatier et al., 2016), for probabilistic programming. First, as a conceptual 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:

$$\textit{Neural speech envelope tracking} \sim \textit{condition} * \textit{prediction tendency} + (1|\textit{subject})$$

"Neural speech envelope tracking" refers to the correlation coefficients between predicted and true brain responses from the aforementioned encoding model, trained and tested on the whole audio material within condition (single vs. multi-speaker). (Note that 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:

$$\textit{Ocular speech envelope tracking} \sim 0 + \textit{condition} + \textit{prediction tendency} + (1|\textit{subject})$$

"Ocular speech envelope tracking" refers to the correlation coefficients between predicted and true eye movements from the encoding model (again trained and tested on the whole audio material within condition.

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:

$$Speech\ envelope\ tracking \sim word\ type * condition + (1|subject)$$

Here "Speech envelope tracking" refers to the correlation coefficients between predicted and true responses (either neural or ocular) from the encoding model trained on the whole audio material but tested solely on the word-segments.

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):

$$absolute\ weights \sim 0 + time\text{-}lags + (1|subject)$$

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:

$$comprehension \sim neural\ speech\ tracking * condition + prediction\ tendency + (1|subject)$$
$$difficulty \sim neural\ speech\ tracking * condition + prediction\ tendency + (1|subject)$$

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:

$$comprehension \sim ocular\ speech\ tracking * condition + (1|subject)$$
$$difficulty \sim ocular\ speech\ tracking * condition + (1|subject)$$

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., 2020) 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 (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). Thus, this measure is a single value per subject, which comprises a) differences between two contextual probabilities (i.e. ordered vs. random) in b) feature-specific tone representations c) in advance of their observation (summed over a time-window of –0.3 – 0 s). Importantly, this prediction tendency was assessed in an independent entropy modulation paradigm (see **Fig. 1**). On a group level we found an increased tendency to pre-activate a stimulus of high probability (i.e. forward transition) in an ordered context compared to a random context (see Fig, 2A). This effect replicates results from our previous work (Schubert et al., 2023, 2024). Using the summed difference between entropy levels (ordered – random) across pre-stimulus time, one value was extracted per subject (**Fig. 2B**). This value was used as a proxy for "individual prediction tendency" and correlated with encoding of clear speech across different MEG sensors.

Neural speech tracking, quantified as the correlation coefficients between predicted and observed MEG responses to the speech envelope, was used as the dependent variable in Bayesian regression models (**Fig. 3A**; see **Fig. 3B** for TRF weights). These models included condition (single vs. multi-speaker) as a fixed effect, with either prediction tendency or word surprisal as an additional predictor, and random effects for participants. Replicating previous findings (Schubert et al., 2023), we found widespread encoding of clear speech (average over cluster: β = 0.035, 94%HDI = [0.024, 0.046]), predominantly over auditory processing regions (**Fig. 3C**), that was decreased (β = –0.018, 94%HDI = [-0.029, –0.006]) in a multi-speaker condition (**Fig. 3D**). Furthermore, a stronger prediction tendency was associated with increased neural speech tracking (β = 0.014, 94%HDI = [0.004, 0.025]) over left frontal sensors (see **Fig. 3E**). We found no interaction between prediction tendency and condition (see **Fig. 3F**). 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 (β = 0.039, 94%HDI = [0.007, 0.071]) over left temporal areas (see **Fig. 3G**). Furthermore, we found no interaction between word surprisal and speaker condition (see **Fig. 3H**). 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.

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

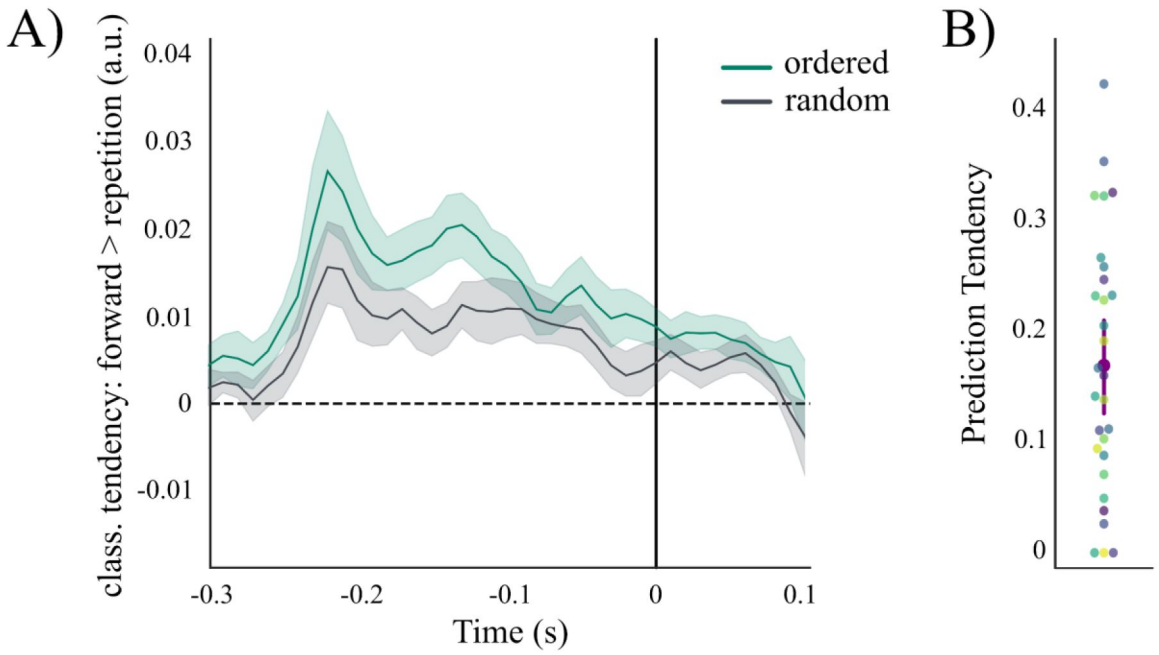


Figure 2

Individual prediction tendency.

A) Time-resolved contrasted classifier decision: forward > repetition for ordered and random repetition trials. Classifier tendencies showing frequency-specific prediction for tones with the highest probability (forward transitions) can be found even before stimulus onset but only in an ordered context (shaded areas always indicate 95% confidence intervals). Using the summed difference across pre-stimulus time, one prediction value was extracted per individual subject. **B)** Distribution of prediction tendency values across subjects (*N* = 29).

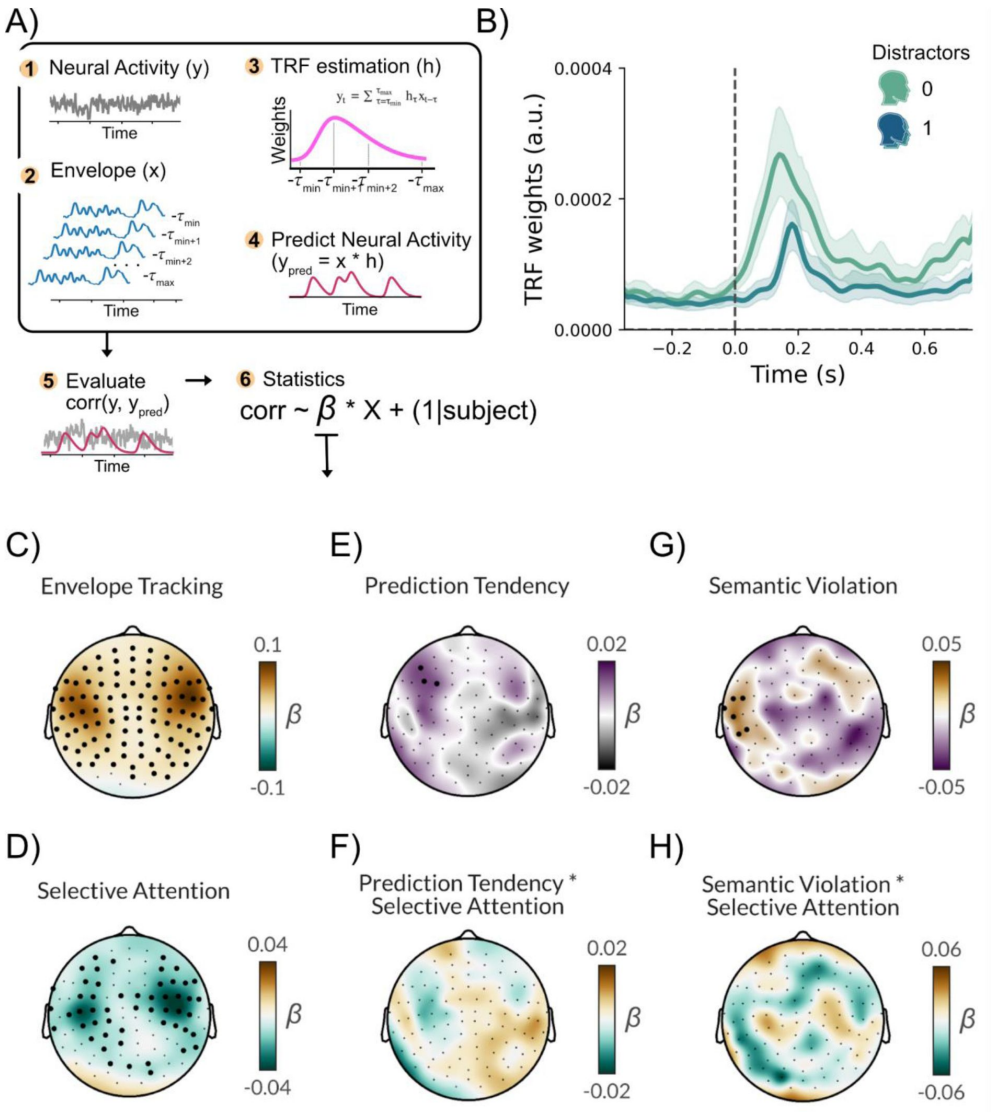


Figure 3

Neural speech tracking is related to prediction tendency and word surprisal, independent of selective attention.

A) Envelope (x) – response (y) relationships are estimated using deconvolution (Boosting). The TRF (filter kernel, h) models how the brain processes the envelope over time. This filter is used to predict neural responses via convolution. Predicted responses are correlated with actual neural activity to evaluate model fit and the TRF's ability to capture response dynamics. Correlation coefficients from these models are then used as dependent variables in Bayesian regression models. (Panel adapted from Gehmacher et al., 2024b). **B)** Temporal response functions (TRFs) depict the time-resolved neural tracking of the speech envelope for the single speaker and multi speaker target condition, shown here as absolute values averaged across channels. Solid lines represent the group average. Shaded areas represent 95% Confidence Intervals. **C–H)** The beta weights shown in the sensor plots are derived from Bayesian regression models in **A)**. For Panel C, this statistical model is based on correlation coefficients computed from the TRF models (further details can be found in the Methods Section). **C)** In a single speaker condition, neural tracking of the speech envelope was significant for widespread areas, most pronounced over auditory processing regions. **D)** The condition effect indicates a decrease in neural speech tracking with increasing noise (1 distractor). **E)** Stronger prediction tendency was associated with increased neural speech tracking over left frontal areas. **F)** However, there was no interaction between prediction tendency and conditions of selective attention. **G)** Increased neural tracking of semantic violations was observed over left temporal areas. H) There was no interaction between word surprisal and speaker condition, suggesting a representation of surprising words independent of background noise. Marked sensors indicate 'significant' clusters, defined as at least two neighboring channels showing a significant result. *N* = 29.

acoustic speech envelope using temporal response functions (TRFs). The resulting model fit (i.e. correlation between true and predicted eye movements) is commonly referred to as "speech tracking". Bayesian regression models were applied to evaluate tracking effects under different conditions of selective attention (single speaker, attended multi-speaker, unattended multi-speaker). Furthermore, we assessed whether individual prediction tendency or semantic word surprisal influenced ocular speech tracking. 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 4A**). 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 4B**). 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 **Table 2** for a summary of ocular speech tracking effects).

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 **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. This model evaluates to what extent gaze behaviour functions as a mediator between acoustic speech input and brain activity. To establish a time-resolved mediation analysis in source-space, we computed the main components of the effect (via PCA, see **Figure 5**).

We found significant mediation effects for all three principal components (PC) for both vertical eye movements in the clear speech condition (see **Figure 5A**) and horizontal eye movements in response to a target in the multi-speaker condition (see **Figure 5B**). 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

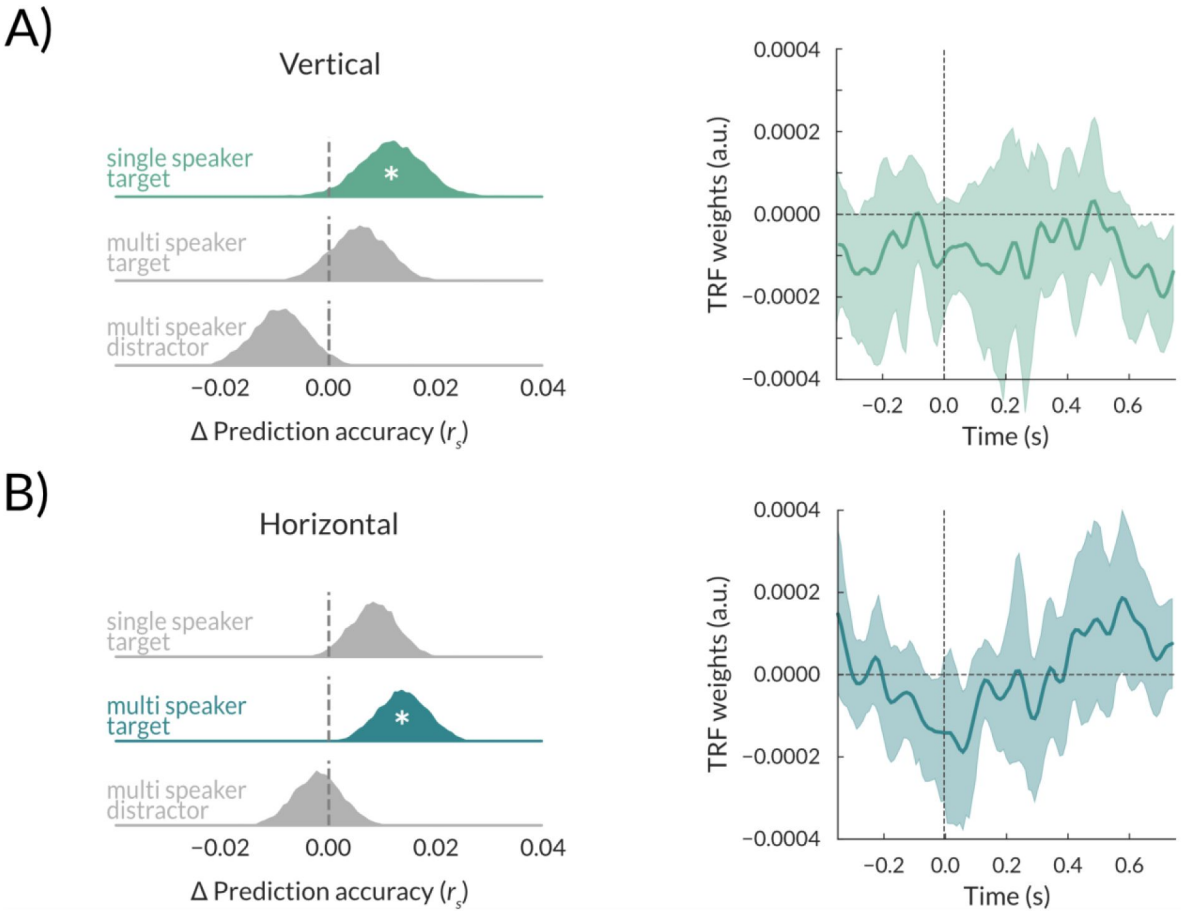


Figure 4

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.

Table 2

Model summary statistics for ocular speech tracking depending on condition and prediction tendency.

	vertical eye movements				horizontal eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
cond1: single speaker - attended	0.012	0.006	0.001	0.023	0.009	0.005	-0.001	0.017
cond2: multi-speaker - attended	0.006	0.006	-0.005	0.016	0.014	0.005	0.005	0.024
cond3: multi-speaker - unattended	-0.008	0.006	-0.020	0.003	-0.002	0.005	-0.011	0.007
prediction tendency	0.001	0.004	-0.005	0.008	-0.001	0.003	-0.007	0.004

Note: Dependent Variable = ocular speech tracking (correlation between true and predicted eye movements).

Table 3

Model summary statistics for ocular speech tracking depending on word type and condition.

	vertical eye movements				horizontal eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
Intercept*	0.053	0.019	0.019	0.092	0.026	0.018	-0.008	0.061
wordtype (semantic violation effect)	-0.024	0.026	-0.074	0.023	0.028	0.025	-0.019	0.075
condition (multi-speaker effect)	-0.005	0.025	-0.052	0.042	-0.007	0.025	-0.053	0.043
wordtype x condition	0.008	0.036	-0.058	0.077	0.001	0.035	-0.065	0.068

Note: Dependent Variable = ocular speech tracking (correlation between true and predicted eye movements).

* Intercept represents speech tracking for lexical control words in a single speaker condition.

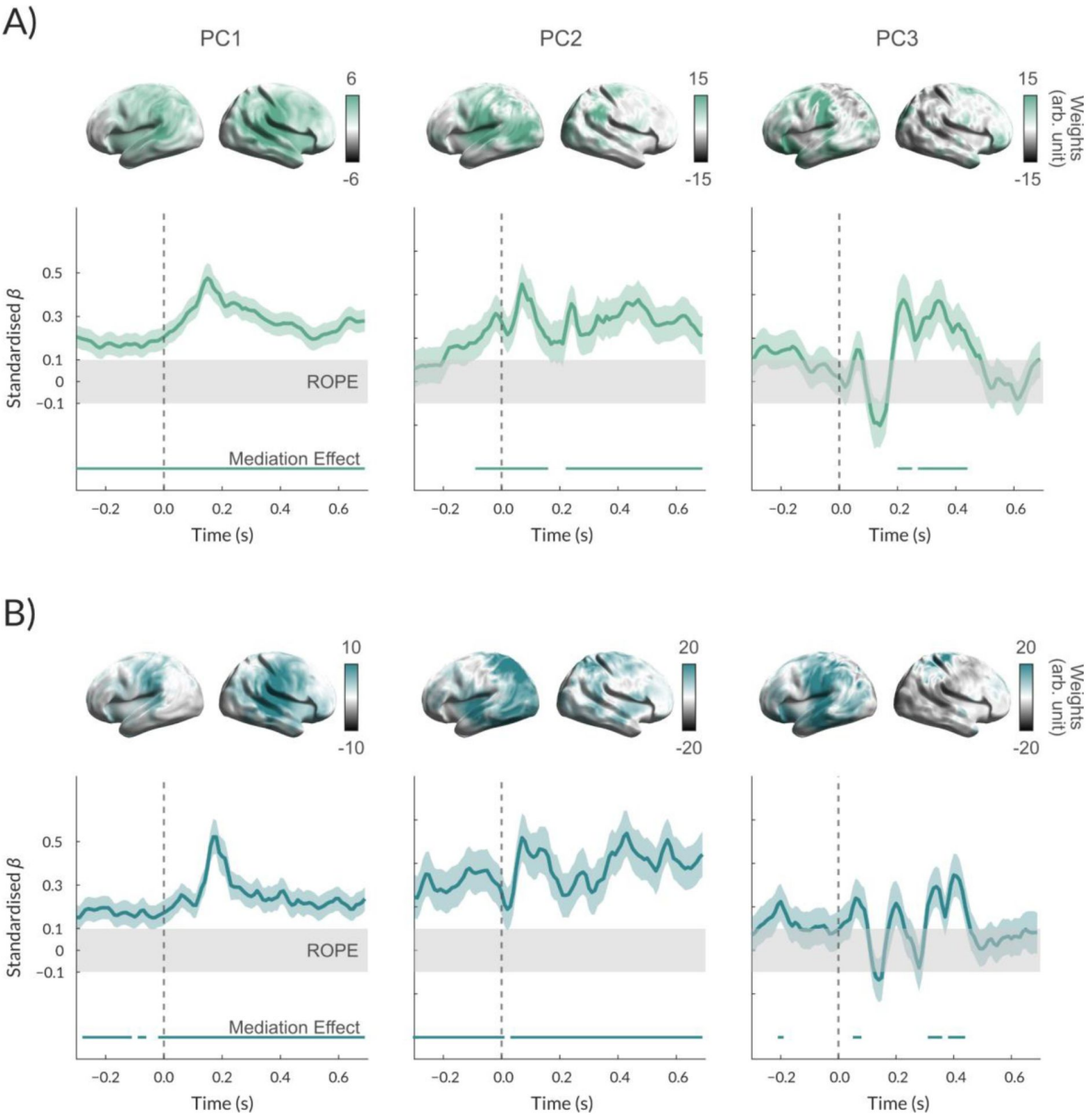


Figure 5

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 (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.

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.

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). Bayesian regression models were used to investigate relationships between neural/ocular speech tracking and comprehension or difficulty. Ocular speech tracking was analysed separately for horizontal and vertical eye movements. 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 6A**). 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 6B-C** and **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 **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).

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

Figure 6

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.

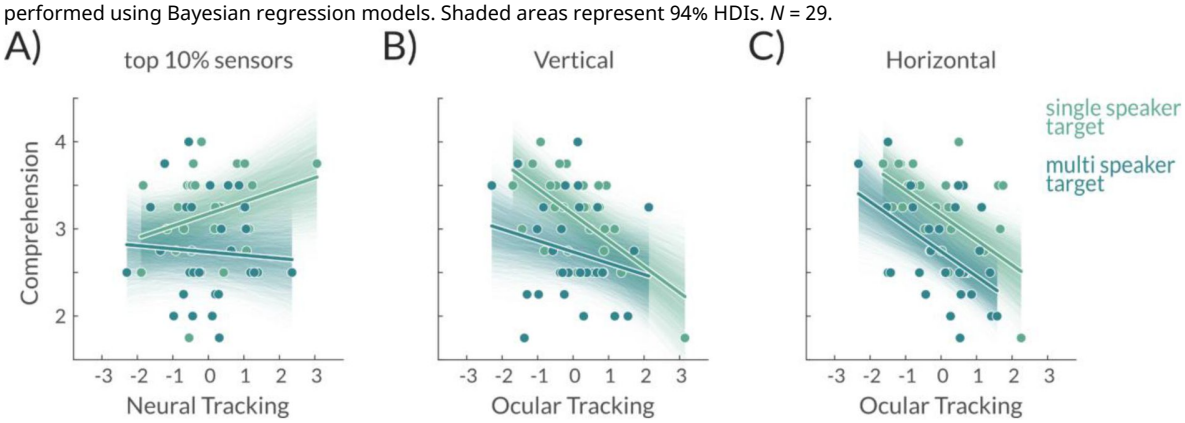


Table 4

Model summary statistics for comprehension depending on ocular speech tracking and condition.

	vertical eye movements				horizontal eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
Intercept*	3.17	0.099	2.986	3.356	3.162	0.095	2.972	3.333
ocular speech tracking	-0.304	0.099	-0.491	-0.121	-0.293	0.101	-0.489	-0.104
condition (multi-speaker effect)	-0.431	0.132	-0.683	-0.195	-0.419	0.125	-0.655	-0.180
speech tracking x condition	0.181	0.140	-0.074	0.456	0.010	0.134	-0.230	0.277

Note: Dependent Variable = comprehension (averaged accuracy for comprehension questions).

* Intercept represents comprehension for average speech tracking in a single speaker condition.

	vertical eye movements				horizontal eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
Intercept*	2.186	0.123	1.971	2.429	2.192	0.122	1.971	2.430
ocular speech tracking	0.099	0.106	-0.094	0.304	0.101	0.116	-0.115	0.317
condition (multi-speaker effect)	1.356	0.127	1.102	1.583	1.355	0.135	1.095	1.606
speech tracking x condition	-0.077	0.144	-0.342	0.194	-0.099	0.157	-0.406	0.185

Note: Dependent Variable = rated difficulty (averaged subjective ratings on a 5-point likert scale).

* Intercept represents rated difficulty for average speech tracking in a single speaker condition.

Table 5

Model summary statistics for rated difficulty depending on ocular speech tracking and condition.

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 sound 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. It is assumed that predictions are directly linked to the interpretation of sensory information. This interpretation a) is likely to occur simultaneously at different places in the cognitive (and anatomical) hierarchy and b) describes a bidirectional process in which internal models influence and are influenced by sensory input (Knill & Pouget, 2004). It should be noted that our measure of individual prediction tendency makes no claim on the relative weighting of top-down or bottom-up processing; it does, however, infer an absolute tendency to generate anticipatory (top-down) predictions. This type of prediction is relevant for acoustic processing such as speech and music, whose predictability unfolds over time.

The replicable association between individual differences in this prediction tendency and speech processing 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., 2024), 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) conceptual 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. Importantly, in contrast to Gehmacher et al. (2024), we did not observe ocular tracking of the multi-speaker distractor in this study. This difference is likely attributable to the simplistic single-trial, 5-word task structure in Gehmacher et al., which resulted in high temporal overlap between the target and distractor speech streams and likely drove the significant distractor-tracking effects observed in that study. The absence of such an effect during continuous listening in our study suggests that ocular tracking is indeed more specific to selective attention. However, 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 (e.g. the classic cocktailparty situation; Cherry, 1953). 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. Relatedly, it remains an open question whether microsaccades are a key feature driving ocular speech tracking. However, our current study does not analyze microsaccades due to methodological constraints: microsaccades are binary response vectors, which are incompatible with TRF analyses used here. Addressing this would require adapting models to handle time-continuous binary response data or potentially exploring alternative approaches, such as regression-based ERFs (e.g., as in Heilbron et al.,

2022). While these limitations preclude microsaccade analysis in the current study, we hypothesize that they could enhance temporal precision and selectively amplify relevant sensory input, supporting auditory perception. Future studies should explore this possibility to uncover the specific contributions of microsaccades to speech tracking.

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. Our choice of ocular movements as a mediator was motivated by the fact that the relationship between acoustic speech and neural activity is well established, as well as previous results indicating that oculomotor activity contributes to cognitive effects in auditory attention (Popov et al., 2022). However, an alternative model may suggest that neural activity mediates the effect of ocular speech tracking. 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. In fact, when rejecting the notion of a single bottom-up flow of information and replacing it with a model of distributed parallel and dynamic processing, it seems only reasonable to assume that the direction of communication (between our eyes and our brain) will depend on where (within the brain) as well as when we look at the effect. Thus, the regions and time-windows reported here should be taken as an illustration of oculo-neural communication during speech processing rather than an attempt to "explain" neural speech processing by ocular movements.

Despite the finding that eye movements mediate neural speech tracking, the behavioural relevance for semantic comprehension appears to differ between ocular and neural speech tracking. Specifically, we found a negative association between ocular speech tracking and comprehension, indicating that participants with lower comprehension performance exhibited increased ocular speech tracking. Interestingly, no significant relationship was observed between neural tracking and comprehension.

In this context, the negative association between ocular tracking and comprehension might reflect individual differences in how participants allocate cognitive resources. Participants with lower comprehension may rely more heavily on attentional mechanisms to process acoustic features, as evidenced by increased ocular tracking. This reliance could represent a compensatory strategy when higher-order processes, such as semantic integration or memory retrieval, are less effective. Importantly, our comprehension questions (see Experimental Procedure) targeted a broad range of processes, including intelligibility and memory, suggesting that this relationship reflects a trade-off in resource allocation between low-level acoustic focus and integrative cognitive tasks.

Rather than separating eye and brain responses conceptually, our analysis highlights their complementary contributions. Eye movements may enhance neural processing by increasing sensitivity to acoustic properties of speech, while neural activity builds on this foundation to integrate information and support comprehension. Together, these systems form an interdependent mechanism, with eye and brain responses working in tandem to facilitate different aspects of speech processing.

This interpretation is consistent with the absence of a difference in ocular tracking for semantic violations (e.g., words with high surprisal versus lexically matched controls), reinforcing the view that ocular tracking primarily reflects attentional engagement with acoustic features rather than direct involvement in semantic processing. This aligns with previous findings that attention

modulates auditory responses to acoustic features (e.g., Forte et al., 2017), further supporting the idea that ocular tracking reflects mechanisms of selective attention rather than representations of linguistic content.

Future research should investigate how these systems interact and explore how ocular tracking mediates neural responses to linguistic features, such as lexical or semantic processing, to better understand their joint contributions to comprehension.

Interestingly, we were not able to relate individual prediction tendency to ocular speech tracking. This raises the question whether ocular speech tracking reflects attentional mechanisms rather than predictive 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 motivated 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, predictions encode probabilistic representations of a feature and "prediction tendency" refers to the individual reliance on these representations, 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 shifts of gaze at exact points in time aligned with 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., 2022). 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 ocular sensing work (independently) together to support auditory speech perception (see **Figure 7** for

a schematic illustration). We suggest that anticipatory predictions about a feature help to interpret auditory information by prioritizing representational content of high probability at different levels along the perceptual hierarchy. Accordingly, these predictions are formed in parallel and carry high feature-specificity but low temporal precision (as they are anticipatory in nature). This idea is supported by our finding that pure-tone anticipation is visible over a widespread prestimulus interval, instead of being locked to sound onset. It should be noted that the representational content is likely to be different at different levels of the perceptual hierarchy (e.g. encoding of phonemes, words, semantics or even more abstract speaker intentions). Even though the terminology is suggestive of a fixed sequence (similar to a multi storey building) with levels that must be traversed one after each other (and even the more spurious idea of a rooftop, where the final perceptual experience is formed and stored into memory), we distance ourselves from these (possibly unwarranted) ideas. Our usage of "higher" or "lower" simply refers to the observation that the probability of a feature at a higher (as in more associative) level affects the interpretation (and thus the representation and prediction) of a feature at lower (as in more segregated) levels (Caucheteux et al., 2023). Our own findings suggest that individuals differ in their general tendency to create such anticipatory predictions, which leads to differences in neural speech tracking.

Active ocular 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). This assumption is supported by our finding that semantic violations are not differentially encoded in gaze behaviour than lexical controls. 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).

With this speculative framework we attempt to describe and relate three important phenomena with respect to their relevance for speech processing: 1) "Anticipatory predictions" that are created in absence of attentional demands and contain probabilistic information about stimulus features (here, inferred from frequency-specific pre-activations during passive listening to sound sequences). 2) "Selective attention" that allocates resources towards relevant (whilst suppressing distracting) information (which was manipulated by the presence or absence of a distractor speaker). And finally 3) "active ocular sensing", which refers to gaze behaviour that is temporally aligned to attended (but not unattended) acoustic speech input (inferred from the discovered phenomenon of ocular speech tracking). We propose that auditory inflow is, at a basic level, temporally modulated via active ocular sensing, which "opens the gates" in the sensory periphery at relevant timepoints. How exactly this mechanism is guided (for example where the information about crucial timepoints comes from, if not from prediction, and whether it requires habituation to a speech stream etc.) is yet unclear. Unlike predictive tendencies, active ocular sensing appears to reflect selective attention, manifesting as a mechanism that optimizes sensory precision. Individual differences with respect to anticipatory predictions on the other hand, seem to be independent from the other two entities, but nevertheless relevant for speech processing. We therefore support the notion that representational content is interpreted based on prior probabilistic assumptions. If we consider the idea that "a percept" of an (auditory) object is actually temporally and spatially distributed (across representational spacetime – see **Fig. 7**), the content of information depends on where and when it is probed (see for example Dennett, 1991 for similar ideas on consciousness). Having to select from multiple interpretations across space and time 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.

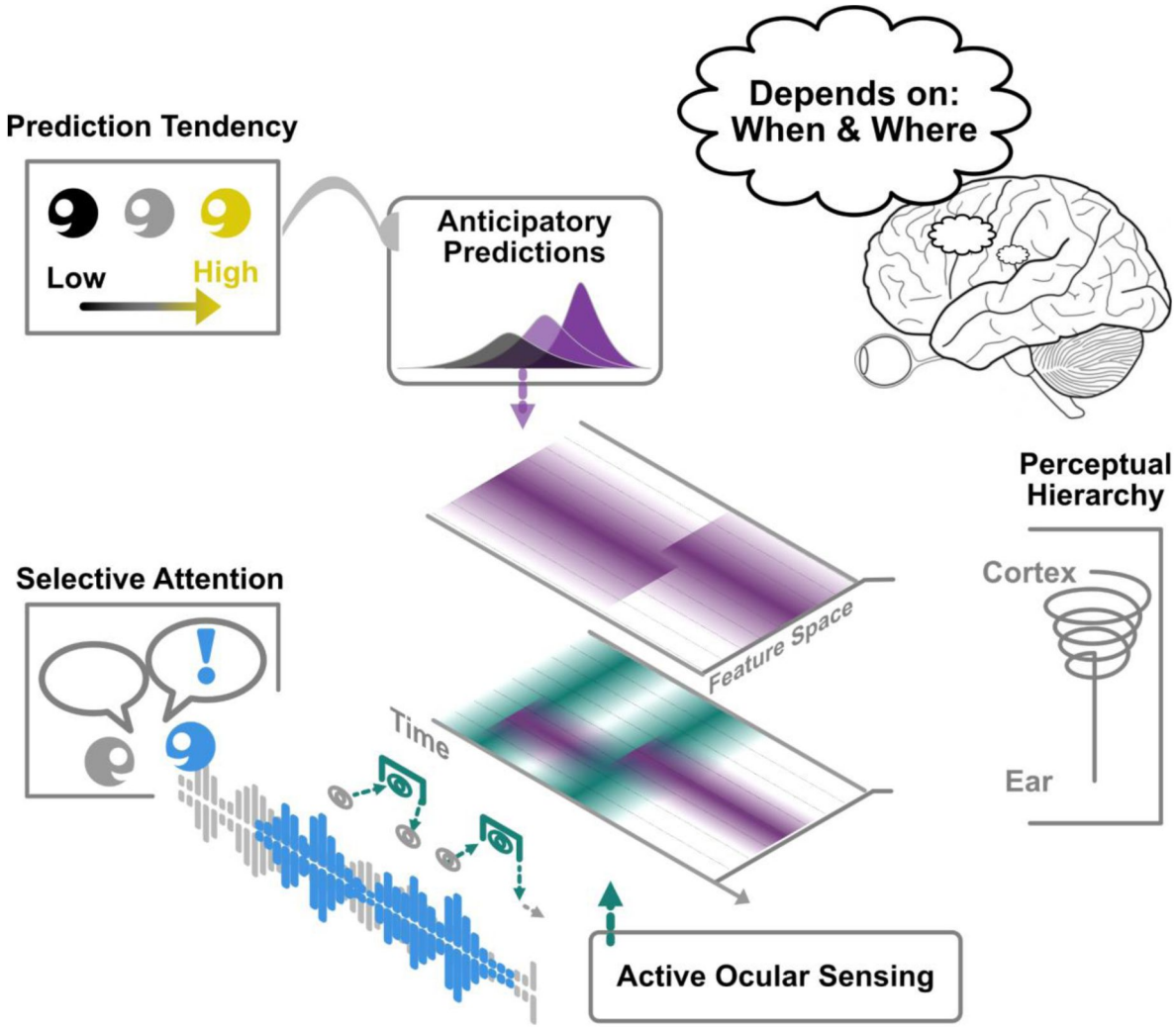


Figure 7

A schematic illustration of the framework:

A speech percept is processed in representational (feature-) spacetime (X and Y-axes) at different levels of the cognitive hierarchy (Z-axis), which ranges from highly selective to more associative representations (without the notion of an "apex" in this hierarchy). The temporal and spatial characteristics of a representation depends on where and when (in the brain) it is probed. Anticipatory predictions (purple) help to interpret auditory information at different levels in parallel with high feature-specificity but low temporal precision. These anticipatory predictions reflect to some extent individual tendencies (and differences) that generalise across listening situations. In contrast, active ocular sensing (green) increases the temporal precision already at lower stages of the auditory system to facilitate bottom–up processing at specific timescales (similar to neural oscillations). It does not necessarily convey feature-specific information, but is more likely used to boost (or filter) information around relevant time windows. Our results suggest that this mechanism is motivated by selective attention (blue) rather than predictive 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.

Additional information

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

1. Brainard D. H (1997) **The Psychophysics Toolbox** *Spatial Vision* **10**:433–436 https://doi.org/10.1163/156856897X00357Google Scholar

2. Brodbeck C., Das P., Kulasingham J. P., Bhattasali S., Gaston P., Resnik P., Simon J. Z (2021) **Eelbrain: A Python toolkit for time-continuous analysis with temporal response functions** *BioRxiv* :2021.08.01.454687 https://doi.org/10.1101/2021.08.01.454687Google Scholar

3. Brodbeck C., Simon J. Z (2020) **Continuous speech processing** *Current Opinion in Physiology* **18**:25–31 https://doi.org/10.1016/j.cophys.2020.07.014Google Scholar

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

5. Capretto T., Piho C., Kumar R., Westfall J., Yarkoni T., Martin O. A (2020) **Bambi: A simple interface for fitting Bayesian linear models in Python** *arXiv* Google Scholar

6. Caucheteux C., Gramfort A., King J. R (2023) **Evidence of a predictive coding hierarchy in the human brain listening to speech** *Nature human behaviour* **7**:430–441 Google Scholar

7. Cherry E. C (1953) **Some experiments on the recognition of speech, with one and with two ears** *Journal of the acoustical society of America* **25**:975–979 Google Scholar

8. 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 Google Scholar

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

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

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

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

13. Feldman H., Friston K. J (2010) **Attention, uncertainty, and free-energy** *Frontiers in Human Neuroscience* **4**:215 https://doi.org/10.3389/fnhum.2010.00215Google Scholar

14. 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.27203Google Scholar

15. Friston K (2010) **The free-energy principle: A unified brain theory?** *Nature Reviews Neuroscience* **11** https://doi.org/10.1038/nrn2787Google Scholar

16. 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.2020.107998Google Scholar

17. Galantucci B., Fowler C. A., Turvey M. T (2006) **The motor theory of speech perception reviewed** *Psychonomic bulletin & review* **13**:361–377 Google Scholar

18. 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**:3692 Google Scholar

19. Gehmacher Q., Schubert J., Kaltenmaier A., Weisz N., Press C. (2024b) **The "Ocular Response Function" for encoding and decoding oculomotor related neural activity** *bioRxiv* Google Scholar

20. Hartmann T., Weisz N (2020) **An Introduction to the Objective Psychophysics Toolbox** *Frontiers in Psychology* **11**:585437 https://doi.org/10.3389/fpsyg.2020.585437Google Scholar

21. Heilbron M., Armeni K., Schoffelen J.-M., Hagoort P., de Lange F. P. (2022) **A hierarchy of linguistic predictions during natural language comprehension** *Proceedings of the National Academy of Sciences* **119**:e2201968119 https://doi.org/10.1073/pnas.2201968119Google Scholar

22. Heilbron M., Armeni K., Schoffelen J.-M., Hagoort P., de Lange F. P. (2022) **A hierarchy of linguistic predictions during natural language comprehension** *Proceedings of the National Academy of Sciences* **119**:e2201968119 https://doi.org/10.1073/pnas.2201968119Google Scholar

23. Jin P., Zou J., Zhou T., Ding N (2018) **Eye activity tracks task-relevant structures during speech and auditory sequence perception** *Nature Communications* **9** https://doi.org/10.1038/s41467-018-07773-yGoogle Scholar

24. 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.2017.01.005Google Scholar

25. Kleiner M., Brainard D. H., Pelli D., Ingling A., Murray R., Broussard C (2007) **What's new in Psychtoolbox-3** *Perception* **36**:1–16 https://doi.org/10.1068/v070821Google Scholar

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

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

28. 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**:1185–1195 https://doi.org/10.1016/j.cub.2023.02.018Google Scholar

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

30. Mattout J., Henson R. N., Friston K. J (2007) **Canonical source reconstruction for MEG** *Computational Intelligence and Neuroscience* **2007**:067613 Google Scholar

31. 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.2014.12.005Google Scholar

32. 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**:3637 Google Scholar

33. 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.16747Google Scholar

34. 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** Google Scholar

35. 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**:15516–15521 https://doi.org/10.1073/pnas.1108912108Google Scholar

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

37. Schiel F., Ohala J. J (1999) **Automatic Phonetic Transcription of Non-Prompted Speech. 14th International Congress of Phonetic Sciences, San Francisco, USA, 1. - 7. August 1999. Ohala, John J. (ed.)** In: *Proceedings of the XIVth International Congress of Phonetic Sciences : ICPhS 99* pp. 607–610 https://doi.org/10.5282/ubm/epub.13682Google Scholar

38. 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**:172–176 https://doi.org/10.1016/j.conb.2010.02.010Google Scholar

39. Schubert J., Schmidt F., Gehmacher Q., Bresgen A., Weisz N (2023) **Cortical speech tracking is related to individual prediction tendencies. Cerebral Cortex** *bhac* **528** https://doi.org/10.1093/cercor/bhac528Google Scholar

40. 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.2015.02.001Google Scholar

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

42. Lovich Stephanie N, King Cynthia D, Murphy David LK, Landrum Rachel, Shera Christopher A, Groh Jennifer M (2022) **Parametric information about eye movements is sent to the ears**

BioRxiv :2022.11.27.518089 https://doi.org/10.1101/2022.11.27.518089Google Scholar

43. Summerfield C., de Lange F. P. (2014) **Expectation in perceptual decision making: Neural and computational mechanisms** *Nature Reviews Neuroscience* **15** https://doi.org/10.1038/nrn3838Google Scholar

44. 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.2020.00289Google Scholar

45. 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**:867–880 Google Scholar

46. Weissbart H., Kandylaki K. D., Reichenbach T (2020) **Cortical Tracking of Surprisal during Continuous Speech Comprehension** *Journal of Cognitive Neuroscience* **32**:155–166 https://doi.org/10.1162/jocn_a_01467Google Scholar

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

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

49. 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**:980–991 https://doi.org/10.1016/j.neuron.2012.12.037Google Scholar

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

Author information

Juliane Schubert*
Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria
ORCID iD: 0000-0002-2536-6522

For correspondence: juliane.schubert@plus.ac.at
*contributed equally

Quirin Gehmacher*
Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria, Department of Experimental Psychology, University College London, London, United Kingdom, Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom
ORCID iD: 0000-0002-9407-2763

*contributed equally

Fabian Schmidt
Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria
ORCID iD: 0000-0002-9839-1614

Thomas Hartmann
Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria
ORCID iD: 0000-0002-8298-8125

Nathan Weisz
Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria, Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria
ORCID iD: 0000-0001-7816-0037

Editors

Reviewing Editor
Roberto Bottini
University of Trento, Trento, Italy

Senior Editor
Barbara Shinn-Cunningham
Carnegie Mellon University, Pittsburgh, United States of America

Reviewer #1 (Public review):

Summary:

This study aimed at replicating two previous findings that showed (1) a link between prediction tendencies and neural speech tracking, and (2) that eye movements track speech. The main findings were replicated which supports the robustness of these results. The authors also investigated interactions between prediction tendencies and ocular speech tracking, but the data did not reveal clear relationships. The authors propose a framework that integrates the findings of the study and proposes how eye movements and prediction tendencies shape perception.

Strengths:

This is a well-written paper that addresses interesting research questions, bringing together two subfields that are usually studied in separation: auditory speech and eye movements. The authors aimed at replicating findings from two of their previous studies, which was overall successful and speaks for the robustness of the findings. The overall approach is convincing, methods and analyses appear to be thorough, and results are compelling.

Weaknesses:

Eye movement behavior could have presented in more detail and the authors could have attempted to understand whether there is a particular component in eye movement behavior (e.g., blinks, microsaccades) that drives the observed effects.

https://doi.org/10.7554/eLife.101262.2.sa3

Reviewer #2 (Public review):

Summary

Schubert et al. recorded MEG and eye tracking activity while participants were listening to stories in single-speaker or multi-speaker speech. In a separate task, MEG was recorded while the same participants were listening to four types of pure tones in either structured (75% predictable) or random (25%) sequences. The MEG data from this task was used to quantify individual 'prediction tendency': the amount by which the neural signal is modulated by whether or not a repeated tone was (un)predictable, given the context. In a replication of earlier work, this prediction tendency was found to correlate with 'neural speech tracking' during the main task. Neural speech tracking is quantified as the multivariate relationship between MEG activity and speech amplitude envelope. Prediction tendency did not correlate with 'ocular speech tracking' during the main task. Neural speech tracking was further modulated by local semantic violations in the speech material and by whether or not a distracting speaker was present. The authors suggest that part of the neural speech tracking is mediated by ocular speech tracking. Story comprehension was negatively related with ocular speech tracking.

Strengths

This is an ambitious study, and the authors' attempt to integrate the many reported findings related to prediction and attention in one framework is laudable. The data acquisition and analyses appear to be done with great attention to methodological detail. Furthermore, the experimental paradigm used is more naturalistic than was previously done in similar setups (i.e.: stories instead of sentences).

Weaknesses

While the analysis pipeline is outlined in much detail, some analysis choices appear ad-hoc and could have been more uniform and/or better motivated (other than: this is what was done before).

https://doi.org/10.7554/eLife.101262.2.sa2

Reviewer #3 (Public review):

I thank the authors for their extensive revision of this paper, and I found some elements greatly improved.
In particular, the authors do embrace a somewhat more speculative tone in the current version, which I think is fitting for this work, as the data seem (to me) to be not fully conclusive. The data set collected here is clearly valuable and unique (and I would encourage the authors to make it publicly available!), however, my overall impression is that the specific analyses reported here might not fully

Despite the revised description of methods, results and figures, I still have trouble understanding many of the results and the authors conclusive interpretation of them. These are my main reservations:

(1) Regarding "individual prediction tendency" - thank you for adding clarifying methodological details and showing the data in a new Figure (#2). Honestly, however, I still can't say that I fully understand the result. For example, why is there also a significant response in the random condition as well? And how do you interpret the interesting time-

course (with a peak ~200ms prior to the stimulus, and a reduction overtime from there? Also (I may have missed this, but..) what neural data was used to train the classifier and derive the "prediction tendency" index? Was it just the broadband neural response? Is there a way to know which sensors contributed to this metric (e.g., are they predominantly auditory? Frontal?)? And is there a way to establish the statistical significance of this metric (e.g., how good the decoder actually was in predicting behavioral sensitivity?). I don't see any statistics in the results section describing the individual prediction tendency.

(2) Regarding the TRF analysis - Thanks for clarifying the approach used to obtain 2-second long "segments" of speech tracking. This is an interesting approach, however I think quite new(?) , and for me it raises a whole new set of questions, as well as additional controls and data that I would have liked to see, to be convinced that results are significant. I will elaborate:

- Do I understand correctly that you segment the real and predicted neural response into 2-second long segments and then calculate the Pearsons' correlation between them to assess the goodness of the model? This is very unclear, since in the methods section you state only that "the same" analysis was performed as for the full data - but what exactly? Clearly, values will be very different when using such short segments. I feel that additional details are still required (and perhaps data shown) to fully understand the "semantic violation" analysis of TRFs.

- I would like to reiterate my previous comment regarding the use of permutation tests to verify the validity of TRF-based measures derived. This would be especially important when using new approaches (such as the segmentation used here). The authors argue that this is not needed since this was not done in their previously published study. However, this sounds a bit like "two wrongs make a right" argument... why not just do it, and let us know that this 2-second segmentation approach allows estimating reliable speech tracking?

- Following up on my previous comment that defining "clusters" as at least two neighboring channels (Figure 3) - the fact that this is a default in Fieldtrip is by no means sufficient justification!. This seems quite liberal to me, especially given the many comparisons performed. Here too, permutations can help to determine the necessary data-driven threshold for corrections. This is of course critical for interpreting the result shown in Figures 3E&G that are critical "take home messages" of the paper - i.e., that the prediction-index from the first part of the experiment is related to speech tracking in the second part of the experiment. To my eyes, this does not look extremely convincing, but perhaps the authors can show more conclusive data to support this (e.g., scatter plots of the betas across participant?).
- A similar point can be made for the effect of semantic violations (though here the scalp-level result is somewhat more clustered). The authors point out that the semantic effect is a "replication" of their result reported in Schubert et al. 2023, but if I am not mistaken the results there were somewhat different (as was the manipulation). It would be nice to explicitly discuss the similarity/difference between these effects.

(3) Regarding the ocular-TRFs -

- Maybe this is just me, but I believe that effects that are robust should be clearly visible in the data, without the need for fancy "black-box" statistical models. In the case of the ocular TRFs, it is hard for me to see how these time-courses are not just noise (and, again, a permutation test would have helped to convince me..). The inconsistent results for horizontal and vertical eye-movements vis a vis the experimental conditions (single vs. multi-speaker conditions) don't help either, despite the authors argument that these are "independent" - but why should this be the case, especially if there is nothing really to look at in this task?
- I remain with this scepticism for the mediation-portion of the analysis as well... But perhaps

replications from other groups or making the data public will help shed further light on this in the future.

Minor
- Thanks for adding information about the creation of semantic-violation stimuli. Since the violations and lexical-controls were taken from different audio recordings, it would have been nice to verify that differences between neural responses cannot be attributed to differences in articulations (e.g., by comparing their spectro-temporal properties).

https://doi.org/10.7554/eLife.101262.2.sa1

Author response:

The following is the authors' response to the original reviews

> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *This study aimed at replicating two previous findings that showed (1) a link between prediction tendencies and neural speech tracking, and (2) that eye movements track speech. The main findings were replicated which supports the robustness of these results. The authors also investigated interactions between prediction tendencies and ocular speech tracking, but the data did not reveal clear relationships. The authors propose a framework that integrates the findings of the study and proposes how eye movements and prediction tendencies shape perception.*
>
> *Strengths:*
>
> *This is a well-written paper that addresses interesting research questions, bringing together two subfields that are usually studied in separation: auditory speech and eye movements. The authors aimed at replicating findings from two of their previous studies, which was overall successful and speaks for the robustness of the findings. The overall approach is convincing, methods and analyses appear to be thorough, and results are compelling.*
>
> *Weaknesses:*
>
> *Linking the new to the previous studies could have been done in more detail, and the extent to which results were replicated could have been discussed more thoroughly.*
>
> *Eye movement behavior could have been presented in more detail and the authors could have attempted to understand whether there is a particular component in eye movement behavior (e.g., microsaccades) that drives the observed effects.*

We would like to thank you for your time and effort in reviewing our work and we appreciate the positive comments!

We extended our manuscript, now providing intermediate results on individual prediction tendency, which can be compared to our results from Schubert et al., (2023).

Furthermore, we expanded our discussion now detailing the extent to which our results (do not) replicate the previous findings (e.g. differences in horizontal vs. vertical ocular speech tracking, lack of distractor tracking, link between ocular speech tracking and behavioral outcomes).

While we agree with the reviewer that it is an important and most interesting question, to what extent individual features of gaze behavior (such as microsaccades, blinks etc.) contribute to the ocular speech tracking effect, it is beyond the scope of the current manuscript. It will be methodologically and conceptually challenging to distinguish these features from one another and to relate them to diverse cognitive processes. We believe that a separate manuscript is needed to give these difficult questions sufficient space for new methodological approaches and control analyses. The primary goal of this manuscript was to replicate the findings of Gehmacher et al. (2024) using similar methods and to relate them to prediction tendencies, attention, and neural speech tracking.

> ***Reviewer #2 (Public review):***
>
> *Summary*
>
> *Schubert et al. recorded MEG and eye-tracking activity while participants were listening to stories in single-speaker or multi-speaker speech. In a separate task, MEG was recorded while the same participants were listening to four types of pure tones in either structured (75% predictable) or random (25%) sequences. The MEG data from this task was used to quantify individual 'prediction tendency': the amount by which the neural signal is modulated by whether or not a repeated tone was (un)predictable, given the context. In a replication of earlier work, this prediction tendency was found to correlate with 'neural speech tracking' during the main task. Neural speech tracking is quantified as the multivariate relationship between MEG activity and speech amplitude envelope. Prediction tendency did not correlate with 'ocular speech tracking' during the main task. Neural speech tracking was further modulated by local semantic violations in the speech material, and by whether or not a distracting speaker was present. The authors suggest that part of the neural speech tracking is mediated by ocular speech tracking. Story comprehension was negatively related to ocular speech tracking.*
>
> *Strengths*
>
> *This is an ambitious study, and the authors' attempt to integrate the many reported findings related to prediction and attention in one framework is laudable. The data acquisition and analyses appear to be done with great attention to methodological detail (perhaps even with too much focus on detail-see below). Furthermore, the experimental paradigm used is more naturalistic than was previously done in similar setups (i.e.*

stories instead of sentences).

> *Weaknesses*
>
> *For many of the key variables and analysis choices (e.g. neural/ocular speech tracking, prediction tendency, mediation) it is not directly clear how these relate to the theoretical entities under study, and why they were quantified in this particular way. Relatedly, while the analysis pipeline is outlined in much detail, an overarching rationale and important intermediate results are often missing, which makes it difficult to judge the strength of the evidence presented. Furthermore, some analysis choices appear rather ad-hoc and should be made uniform and/or better motivated.*

We would like to thank you very much for supporting our paper and your thoughtful feedback!

To address your concerns, that our theoretical entities as well as some of our analytical choices lack transparency, we expanded our manuscript in several ways:

(1) We now provide the intermediate results of our prediction tendency analysis (see new Figure 2 of our manuscript). These results are comparable to our findings from Schubert et al. (2023), demonstrating that on a group level there is a tendency to pre-activate auditory stimuli of high probability and illustrating the distribution of this tendency value in our subject population.

(2) We expanded our methods section in order to explain our analytical choices (e.g. why this particular entropy modulation paradigm was used to measure individual prediction tendency).

(3) We now provide an operationalisation of the terms “neural speech tracking” and “ocular speech tracking” at their first mention, to make these metrics more transparent to the reader.

(4) We are summarizing important methodological information ahead of each results section, in order to provide the reader with a comprehensible background, without the necessity to read through the detailed methods section.

(5) We expanded our discussion section, with a special emphasis on relating the key variables of the current investigation to theoretical entities.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *In this paper, the authors measured neural activity (using MEG) and eye gaze while individuals listened to speech from either one or two speakers, which sometimes contained semantic incongruencies.*
>
> *The stated aim is to replicate two previous findings by this group: (1) that there is "ocular speech tracking" (that eye-movements track the audio of the speech), (2) that individual differences in neural response to tones that are predictable vs. not-predictable in their pitch is linked to neural response to speech. In addition, here they try to link the above two effects to each other, and to link "attention, prediction, and active sensing".*
>
> *Strengths:*
>
> *This is an ambitious project, that tackles an important issue and combines different sources of data (neural data, eye-movements, individual differences in another task) in order to obtain a comprehensive "model" of the involvement of eye-movements in sensory processing.*
>
> *The authors use many adequate methods and sophisticated data-analysis tools (including MEG source analysis and multivariate statistical models) in order to achieve this.*
>
> *Weaknesses:*
>
> *Although I sympathize with the goal of the paper and agree that this is an interesting and important theoretical avenue to pursue, I am unfortunately not convinced by the results and find that many of the claims are very weakly substantiated in the actual data.*
>
> *Since most of the analyses presented here are derivations of statistical models and very little actual data is presented, I found it very difficult to assess the reliability and validity of the results, as they currently stand. I would be happy to see a thoroughly revised version, where much more of the data is presented, as well as control analyses and rigorous and well-documented statistical testing (including addressing multiple comparisons).*

We thank you for your thoughtful feedback. We appreciate your concerns and will address them below in greater detail.

> *These are the main points of concern that I have regarding the paper, in its current format.*
>
> *(1) Prediction tendencies - assessed by listening to sequences of rhythmic tones, where the pitch was either "predictable" (i.e., followed a fixed pattern, with 25% repetition) or "unpredictable" (no particular order to the sounds). This is a very specific type of prediction, which is a general term that can operate along many different dimensions. Why was this specific design selected? Is there theoretical reason to believe that this type of prediction is also relevant to "semantic" predictions or other predictive aspects of speech processing?*

Theoretical assumptions and limitations of our quantification of individual prediction tendency are now shortly summarized in the first paragraph of our discussion section. With this paradigm we focus on anticipatory “top-down” predictions, whilst controlling for possibly confounding “bottom-up” processes. Since this study aimed to replicated our previous work we chose the same entropy-modulation paradigm as in other studies from our group (e.g. Demarchi et al. 2019, Schubert et al. 2023;2024, Reisinger et al. 2024), which has proven to give reproducible findings of feature-specific preactivations of sounds in a context of low entropy. One advantage of this design is that it gives us the opportunity to directly compare the processing of “predictable” and “unpredictable” sounds of the same frequency in a time-resolved manner (this argument is now also included in the Methods section).

Regarding the question to what extent this type of prediction might also be relevant to “semantic” predictions we would like to refer to our previous study (Schubert et al., 2023), where we explicitly looked at the interaction between individual prediction tendency and encoding of semantic violations in the cortex. (In short, there we found a spatially dissociable interaction effect, indicating an increased encoding of semantic violations that scales with prediction tendency in the left hemisphere, as well as a disrupted encoding of semantic violations for individuals with stronger prediction tendency in the right hemisphere.) We did not aim to replicate all our findings in the current study, but instead we focused on merging the most important results from two independent phenomena in the domain of speech processing and bringing them into a common framework. However, as now stated in our discussion, we believe that “predictions are directly linked to the interpretation of sensory information. This interpretation is likely to occur at different levels along the cognitive (and anatomical) hierarchy…” and that “this type of prediction is relevant for acoustic processing such as speech and music, whose predictability unfolds over time.”

> *(2) On the same point - I was disappointed that the results of "prediction tendencies" were not reported in full, but only used later on to assess correlations with other metrics. Even though this is a "replication" of previous work, one would like to fully understand the results from this independent study. On that note, I would also appreciate a more detailed explanation of the method used to derive the "prediction tendency" metric (e.g, what portion of the MEG signal is used? Why use a pre-stimulus and not a post-stimulus time window? How is the response affected by the 3Hz steady-state response that it is riding on? How are signals integrated across channels? Can we get a sense of what this "tendency" looks like in the actual neural signal, rather than just a single number derived per participant (an illustration is provided in Figure 1, but it would be nice to see the actual data)? How is this measure verified statistically? What is its distribution across the sample? Ideally, we would want enough information for others to be able to replicate this finding).*

We now included a new figure (similar to Schubert et al. 2023) showing the interim results of the “prediction tendency” effect as well as individual prediction tendency values of all subjects.

Furthermore we expanded the description of the “prediction tendency” metric in the Methods section, where we explain our analytical choices in more detail. In particular we used a pre-stimulus time window in order to capture “anticipatory predictions”. The temporally predictably design gives us the opportunity to capture this type of predictions. The integration across channels is handled by the multivariate pattern analysis (MVPA), which inherently integrates multidimensional data (as mentioned in the methods section we used data from 102 magnetometers) and links it to (in this case) categorical information.

> *(3) Semantic violations - half the nouns ending sentences were replaced to create incongruent endings. Can you provide more detail about this - e.g., how were the words selected? How were the recordings matched (e.g., could they be detected due to audio editing?)? What are the "lexically identical controls that are mentioned"? Also, is there any behavioral data to know how this affected listeners? Having so many incongruent sentences might be annoying/change the nature of listening. Were they told in advance about these?*

We expanded the Methods section and included the missing information:

“We randomly selected half of the nouns that ended a sentence (N = 79) and replaced them with the other half to induce unexpected semantic violations. The swap of nouns happened in the written script before the audio material was recorded in order to avoid any effects of audio clipping. Narrators were aware of the semantic violations and had been instructed to read out the words as normal. Consequently all target words occurred twice in the text, once in a natural context (serving as lexical controls) and once in a mismatched context (serving as semantic violations) within each trial, resulting in two sets of lexically identical words that differed greatly in their contextual probabilities (see Figure 1F for an example). Participants were unaware of these semantic violations.” Since we only replaced 79 words with semantic violations in a total of ~ 24 minutes of audio material we believe that natural listening was not impaired. In fact none of the participants mentioned to have noticed the semantic violations during debriefing (even though they had an effect on speech tracking in the brain).

> *(4) TRF in multi-speaker condition: was a univariate or multivariate model used? Since the single-speaker condition only contains one speech stimulus - can we know if univariate and multivariate models are directly comparable (in terms of variance explained)? Was any comparison to permutations done for this analysis to assess noise/chance levels?*

For mTRF models it depends on the direction (“encoding” vs. “decoding”) whether or not the model is comparable to a univariate model. In our case of an encoding model the TRFs are fitted to each MEG channel independently. This gives us the possibility to explore the effect over different areas (whereas a multivariate “decoding” model would result in only one speech reconstruction value).

In both conditions (single and multi speaker) a single input feature (the envelope of the attended speech stream) was used. Of course it would be possible to fit the model to use a multivariate encoding model, predicting the brain’s response to the total input of sounds. This would, however, target a slightly different question than ours as we aimed to investigate how much of the attended speech is tracked.

Regarding your suggestion of a comparison to permutations to assess noise levels we would like to point out that we chose the same methodological approach as in our previous studies, that we aimed to replicate here. Indeed in these original studies no permuted versions (with

exception of the mediation analysis where comparing a model with an additional input predictor to a single predictor model would not result in a fair comparison) have been used. We conducted the mTRF approach considering the guidelines of Crosse et al. (2016) to the best of our knowledge and in accordance with similar studies in this field.

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.

> *(5) TRF analysis at the word level: from my experience, 2-second segments are insufficient for deriving meaningful TRFs (see for example the recent work by Mesik & Wojtczak). Can you please give further details about how the analysis of the response to semantic violations was conducted? What was the model trained on (the full speech or just the 2-second long segments?) Is there a particular advantage to TRFs here, relative - say - to ERPs (one would expect a relatively nice N400 response, not)? In general, it would be nice to see the TRF results on their own (and not just the modulation effects).*

We fully agree with the reviewers statement that 2-second segments would have been too short to derive meaningful TRFs. To investigate the effect of semantic violations, we used the same TRFs trained on the whole dataset (with 4-fold cross validation). The resulting true as well as the predicted data was segmented into single word epochs of 2 seconds. We selected semantic violations as well as their lexically identical controls and correlated true with predicted responses for every word. Thus, we conducted the same analysis as for the overall encoding effect, focusing on only part of the data. We have reformulated the Methods section accordingly to clear up this misunderstanding. Since the TRFs are identical to the standard TRFs from the overall neural speech tracking, they are not informative to the semantic violation effect. However, since the mTRF approach is the key method throughout the manuscript (and our main focus is not on the investigations of brain responses to semantic violations) we have favoured this approach over the classical ERF analysis.

> *(6) Another related point that I did not quite understand - is the dependent measure used for the regression model "neural speech envelope tracking" the r-value derived just from the 2sec-long epochs? Or from the entire speech stimulus? The text mentions the "effect of neural speech tracking" - but it's not clear if this refers to the single-speaker vs. twospeaker conditions or to the prediction manipulation. Or is it different in the different analyses? Please spell out exactly what metric was used in each analysis.*

As suggested we now provide a clear definition of each dependent metric for each analysis.

“Neural speech tracking” refers to the correlation coefficients between predicted and true brain responses from the aforementioned encoding model, trained and tested on the whole audio material within condition (single vs. multi-speaker).

> ***Recommendations for the authors:***
>
> ***Reviewing Editor Comments:***
>
> *The reviewers have provided a number of recommendations to improve the manuscript, particularly requesting that more data be reported, with an emphasis on the measurements themselves (eye movements and TRFs) rather than just the numerical outputs of mathematical models.*

We appreciate all the reviewers' and editor’s comments and effort to improve our manuscript. In the revised version we provide interim findings and missing data, updated figures that include an intuitive illustration of the metrics (such as TRFs), and a thoroughly

revised discussion section where we focus on the relationship between our observed quantities and theoretical entities. We now offer operationalized definitions of the relevant concepts (“prediction tendency”, “active ocular sensing” and “selective attention”) and suggest how these entities might be related in the context of speech processing, based on the current findings. We are confident that this revision has improved the quality of our paper a lot and we are grateful for all the feedback and suggestions.

> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) Participants had to fixate throughout the tasks. How did the authors deal with large eye movements that violated the instructed fixation?*

As described in the Methods section: “Participants were instructed to look at a black fixation cross at the center of a grey screen.” This instruction was not intended to enforce strict fixation but rather to provide a general reference point, encouraging participants to keep their gaze on the grey screen and avoid freely scanning the room or closing their eyes. Unlike trial-based designs, where strict fixation is feasible due to shorter trial durations, this approach did not impose rigid fixation requirements. Consequently, the threshold for "instruction violation" was inherently more flexible, and no additional preprocessing was applied to the gaze vectors.

> *Fixating for such an extended period of time (1.5 hours?) is hard. Did fixation behavior change over time? Could (fixation) fatigue affect the correlations between eye movements and speech tracking? For example, fatigued participants had to correct their fixation more often and this drives, in part, the negative correlation with comprehension?*

Yes, participants spent approximately 2 hours in the MEG, including preparation time (~30 minutes). However, participants were given opportunities to rest their eyes between different parts and blocks of the experiment (e.g., resting state, passive listening, and audiobook blocks), which should help mitigate fatigue to some extent.

That said, we agree that it is an intriguing idea that fatigue could drive the ocular speech tracking effect, with participants potentially needing to correct their gaze more as the experiment progresses. However, our analysis suggests this is unlikely for several reasons:

(1) Cross-validation in encoding models: Ocular speech tracking effects were calculated using a 4-fold cross-validation approach (this detail has now been added to the Methods section; please see our response to public review #3). This approach reduces the influence of potential increases in gaze corrections over time, as the models are trained and validated on independent data splits. Moreover, if there were substantial differences in underlying response magnitudes between folds - for instance, between the first and fourth fold - this would likely compromise the TRF's ability to produce valid response functions for predicting the left-out data. Such a scenario would not result in significant tracking, further supporting the robustness of the observed effects.

(2) TRF time-course stability: If fatigue were driving increased gaze corrections, we would expect this to be reflected in a general offset (capturing the mean difference between folds) in the TRF time-courses shown in Figure 4 (right panel). However, no such trend / offset is evident.

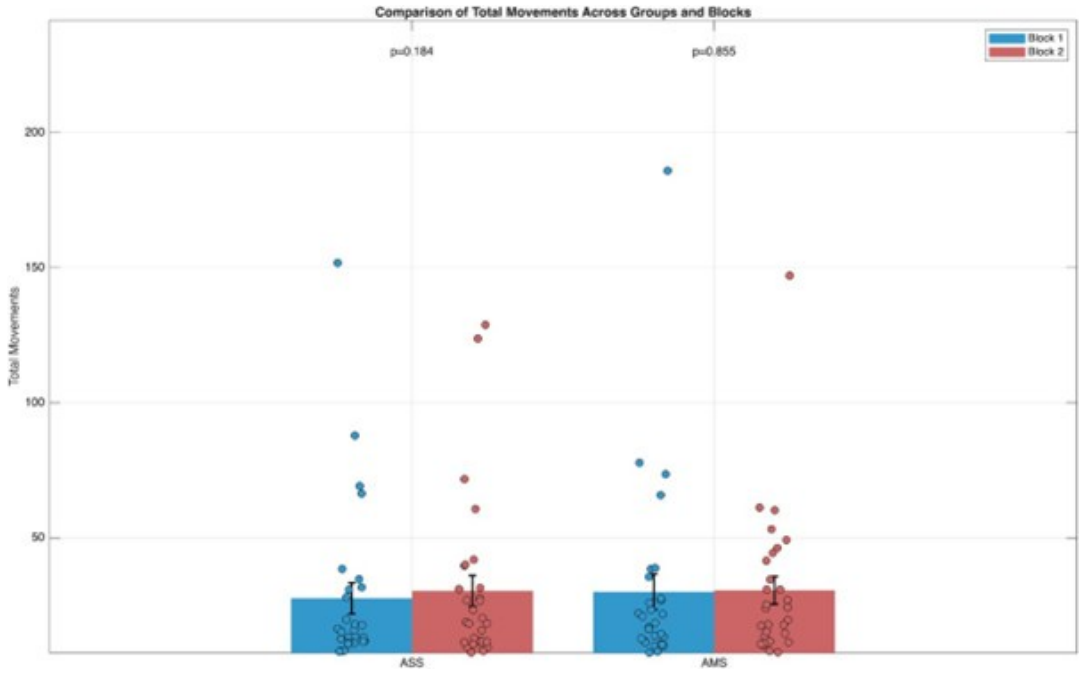
(3) Comparison of eye movement data: To directly investigate this possibility, we compared the amount of total eye movements between the first and last blocks for both the single and multi-speaker conditions. Total movement was calculated by first calculating the differences in pixel values between consecutive eye positions on both the x- and y-axes. The Euclidean distance was then computed for each difference, providing a measure of movement between successive time points. Summing these distances yielded the total movement for each block.

Statistical analysis was performed separately for the single speaker (ASS) and multi-speaker (AMS) conditions. For each condition, paired comparisons were made between the first and last blocks (we resorted to non-parametric tests, if assumptions of normality were violated):

For the single speaker condition (ASS), the normality assumption was not satisfied ($p \leq 0.05p$, Kolmogorov-Smirnov test). Consequently, a Wilcoxon signedrank test was conducted, which revealed no significant difference in total movements between the first and last blocks ($z=-1.330$, $p=0.184$). For the multi-speaker condition (AMS), the data met the normality assumption ($p>0.05$), allowing the use of a paired t-test. The results showed no significant difference in total movements between the first and last blocks ($t=-0.184$, $p=0.855$).

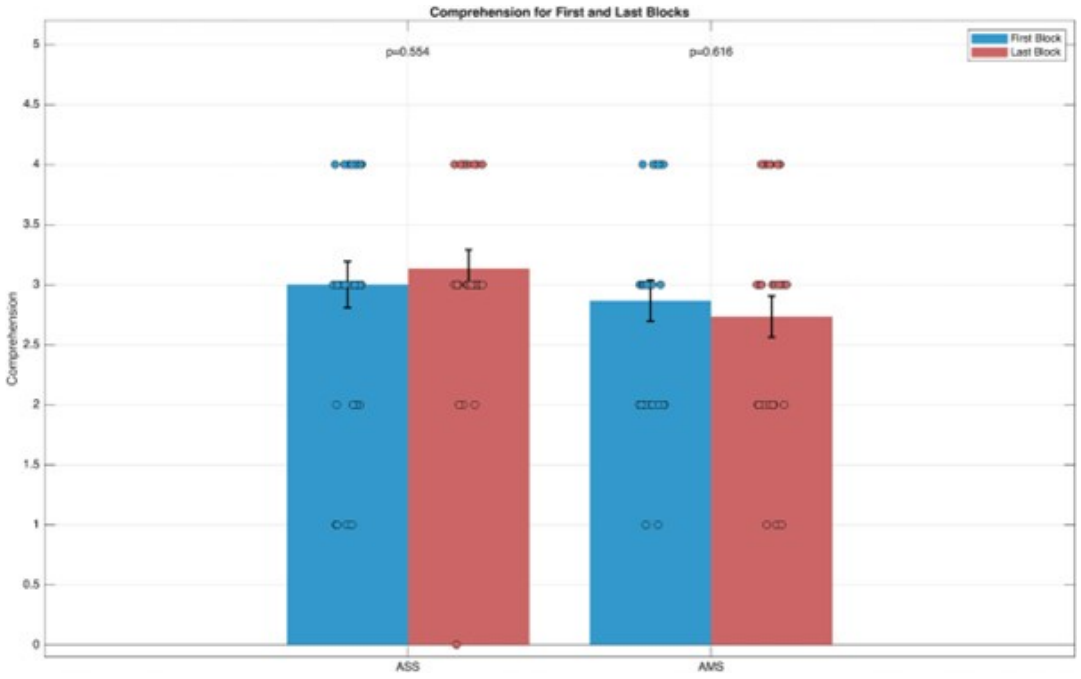
The results are visualized in a bar plot (see below), where individual data points are displayed alongside the mean and standard error for each block. Statistical annotations indicate that neither condition demonstrated significant differences between the blocks. These findings suggest that total eye movements remained stable across the experimental conditions, regardless of whether participants were exposed to a single or multiple speakers.

Author response image 1.



(4) Behavioral responses: Participants' behavioral responses did not indicate any decrease in comprehensibility for later blocks compared to earlier ones. Specifically, a comparison of comprehension scores between the first and last blocks revealed no significant difference in either the single-speaker condition (ASS; Wilcoxon signed-rank test $Z=-0.5911$, $p=0.5545$) or the multi-speaker condition (AMS; Wilcoxon signed-rank test: $Z=0.5018$, $p=0.6158$). These findings suggest that participants maintained consistent levels of comprehension throughout the experiment, regardless of the condition or block order. The results are visualized in a bar plot (see below), where individual data points are displayed alongside the mean and standard error for each block. Statistical annotations indicate that neither condition demonstrated significant differences between the blocks.

Author response image 2.



Together, these factors suggest that fatigue is unlikely to be a significant driver of the ocular speech tracking effects observed in this study.

> *(2) The authors should provide descriptive statistics of fixation behavior /fixational eye movements. What was the frequency and mean direction of microsaccades, do they follow the main sequence, etc., quantify drift and tremor?*

Thank you for their suggestion regarding descriptive statistics. To address this, we computed the rates of microsaccades (which were extracted using the microsaccade detection algorithm as proposed in Liu, B., Nobre, A. C. & van Ede, F. Functional but not obligatory link between microsaccades and neural modulation by covert spatial attention. Nat. Commun. 13, 3503 (2022)) and fixations as these metrics are directly relevant to our study and the requests above.

Microsaccade Rates:

- Single speaker Condition: Mean = 2.306 Hz, SD = 0.363 Hz. ○ Multi speaker: Mean = 2.268 Hz, SD = 0.355 Hz.

Fixation Rates:

- Single speaker Condition: Mean = 2.858 Hz, SD = 1.617 Hz. ○ Multi speaker Condition: Mean = 2.897 Hz, SD = 1.542 Hz.

These values fall within the expected ranges reported in the literature (fixation rates: 2– 4 Hz, microsaccade rates: ~0.5–2.5 Hz) and serve as a sanity check, confirming the plausibility of our eye-tracking data. Regarding the reviewer's request for additional metrics (e.g., microsaccade direction, main sequence analysis, drift, and tremor), extracting these features would require advanced algorithms and analyses not supported by our current preprocessing pipeline or dataset. We hope that the provided metrics, which were the main focus of this study, serve as a sufficient sanity check and highlight the robustness of our data.

> *Related to this, I am wondering whether microsaccades are the feature that drives speech tracking.*

This is an important and pressing question that we aim to address in future publications. Currently, our understanding - and the reason microsaccades and blinks are not analysed in this manuscript - is limited by methodological constraints. Specifically, microsaccades are binary response vectors, which are not compatible with TRF analyses. Addressing this would require adapting future models to handle timecontinuous binary response data or exploring alternative approaches, such as regression-based ERFs (for example as in Heilbron et al. 2022). As the primary goal of this manuscript was to replicate the findings of Gehmacher et al. (2024) using similar methods and to integrate these findings into an initial unified framework, we did not investigate additional eye movement features here. However, we agree that microsaccades (and also blinks, see below) likely contribute, at least in part, to the observed ocular speech tracking effects, and we now suggest this in the Discussion:

“Relatedly, it remains an open question whether microsaccades are a key feature driving ocular speech tracking. However, our current study does not analyze microsaccades due to methodological constraints: microsaccades are binary response vectors, which are incompatible with TRF analyses used here. Addressing this would require adapting models to handle time-continuous binary response data or potentially exploring alternative approaches, such as regression-based ERFs (e.g., as in Heilbron et al., 2022). While these limitations preclude microsaccade analysis in the current study, we hypothesize that they could enhance temporal precision and selectively amplify relevant sensory input, supporting auditory perception. Future studies should explore this possibility to uncover the specific contributions of microsaccades to speech tracking.”

> *(3) Can the authors make sure that interpolated blinks did not drive any of the effects? Can interpolated blink trials be excluded?*

Using continuous audiobooks as stimuli meant that we could not exclude blink periods from the analysis without introducing substantial continuation artifacts in the TRF analysis. Importantly, the concept of covert motor routines and active sensing suggests that participants engage more strongly in motor routines - including ocular behaviors such as microsaccades and blinks - during tasks like speech tracking. These motor routines are inherently tied to individual gaze patterns, making microsaccades and blinks correlated with other ocular behaviors. This complicates efforts to disentangle their individual contributions to the observed ocular speech tracking effects.

Engagement in these motor routines, as posited by active sensing, would naturally load onto various viewing behaviors, further intertwining their roles.

Even if we were to examine correlations, such as the amount of blinks with the ocular speech tracking effect, it is unlikely to provide a clearer understanding due to these inherent overlaps. The methodological and conceptual challenge lies in distinguishing these features from one another and understanding their respective roles in driving the observed effects.

However, the aim of this manuscript was not to dissect the ocular speech tracking effect in greater detail, but rather to relate it - based on similar analytical choices as in Gehmacher et al - to prediction tendencies, attention, and neural speech tracking. While it will be crucial in future work to differentiate these patterns and their connections to diverse cognitive processes, it is beyond the scope of this study to address all these questions comprehensively.

We acknowledge that eye movements, including microsaccades and blinks (however, see challenges for this in response 2), remain underexplored in many experimental paradigms.

Their interplay with cognitive processes - such as attention, prediction, and sensory integration - will undoubtedly be an important focus for future studies.

> *(4) Could the authors provide more details on how time shuffling was done for the eyemovement predictor, and include a circularly shifted version (or a version that does not destroy temporal contiguity) in their model comparisons? Some types of shuffling can result in unrealistic time series, which would end up in an unfair comparison with the model that has the real eye movement traces as predictors.*

We thank the reviewer for their insightful question regarding the time-shuffling procedure for the eye-movement predictor and for suggesting the inclusion of a circularly shifted version in our model comparisons. Below, we provide further details about our approach and the rationale behind it:

(1) Random Shuffling: In our analysis, the eye-movement predictor was randomly shuffled over time, meaning that individual samples were randomly replaced. This method completely disrupts the temporal structure of the signal, providing a null model that directly tests whether the temporal mediation observed is due to the specific temporal relationship between ocular movements and envelope tracking.

(2) Circular Shifting: While circular shifting maintains temporal contiguity, it introduces certain challenges in the context of TRF analysis. Specifically:

- Adaptation to Shifts: The TRF model could adapt to the introduced shift, potentially reducing the validity of the null comparison.

- Similarity due to Repetition: The broadband envelope exhibits strong repetitive patterns over time, such as rhythms inherent to speech. Circular shifting can therefore produce predictors that are very similar to the original signal. As a result, this similarity may lead to null distributions that do not adequately disrupt the temporal mediation we aim to test, making it less robust as a control.

(3) Rationale for Random Shuffling: The primary goal of our mediation analysis is to determine whether there is a temporal mediation of envelope tracking by ocular movements. By deliberately destroying the temporal structure through random shuffling, we ensure that the null model tests for the specific temporal relationship that is central to our hypothesis. Circularly shifted predictors, on the other hand, may partially preserve temporal dependencies, making them less suitable for this purpose.

In summary, while circular shifting is a valuable approach in other contexts, it is less appropriate for the specific goals of this study. We hope this explanation clarifies our methodological choices and demonstrates their alignment with the aims of our analysis.

> *(5) Replication: I want to point out that it is great that the previous findings were in principle replicated. However, I would like to suggest a more nuanced evaluation of the replication:*
>
> *a) Instead of a (direct) replication, the present study should be called a 'conceptual replication', since modifications in design and procedure were made.*

Thank you very much for this suggestion! We now use the term ‘conceptual replication’ throughout the manuscript.

> *b) Not all the findings from the Gehmacher et al., 2024 study were replicated to a full extent:*

> *Did the authors find indications of a vertical vs. horizontal tracking difference in the Gehmacher 2024 data? Could they check this in the Gehmacher 2024 data?*

The findings for horizontal and vertical gaze tracking in Gehmacher et al. (2024) are detailed in the supplementary material of that publication. Both single-speaker and multi-speaker target conditions showed significant speech tracking effects in both horizontal and vertical directions. However, there was a slightly stronger tracking effect for the single-speaker condition in the vertical direction. Due to the highly predictable structure of words in Gehmacher et al. effects here were probably overall boosted as compared to continuous audiobook listening, likely leading to the differentiation of horizontal and vertical gaze. See figures in Gehmacher et al. supplementary file for reference.

> *c) Another difference between their previous and this study is the non-existent tracking of the multi-speaker distractor in this study. The authors should point this out clearly in the discussion and potentially provide an explanation.*

Thank you for highlighting this point! We now address this in the discussion:

“Importantly, in contrast to Gehmacher et al. (2024), we did not observe ocular tracking of the multi-speaker distractor in this study. This difference is likely attributable to the simplistic single-trial, 5-word task structure in Gehmacher et al., which resulted in high temporal overlap between the target and distractor speech streams and likely drove the significant distractor-tracking effects observed in that study. The absence of such an effect during continuous listening in our study suggests that ocular tracking is indeed more specific to selective attention.”

> *Minor:*
>
> *(1) I was a little surprised to not see an indication of eyes/eye movements in Figure 6. The intention of the authors might have been to create a general schematic illustration, but I find this a bit misleading. This paper provides nice evidence for a specific ocular effect in speech tracking. There is, to my knowledge, no indication that speech would be influenced by different kinds of active sensing (if there are, please include them in the discussion). Given that the visuomotor system is quite dominant in humans, it might actually be the case that the speech tracking the authors describe is specifically ocular.*

Taking into account all the reviewers' remarks on the findings and interpretations, we have updated this figure (now Fig. 7) in the manuscript to make it more specific and aligned with the revised discussion section. Throughout the manuscript, we now explicitly refer to active ocular sensing in relation to speech processing and have avoided the broader term 'active sensing' in this context. We hope these revisions address the concerns raised.

> *(2) I find the part in the discussion (page 2, last paragraph) on cognitive processes hard to follow. I don't agree that 'cognitive processes' are easily separable from any of the measured responses (eye and brain). Referring to the example they provide, there is evidence that eye movements are correlated with brain activity that is correlated with memory performance. How, and more importantly, why would one separate those?*

Thank you for raising this important point. We have carefully considered your comments, particularly regarding the interplay between cognitive processes and measured responses (eye and brain), as well as the challenge of conceptually separating them. Additionally, we have incorporated Reviewer #2's query (13) into a unified and complementary reasoning. In response, we have rewritten the relevant paragraph in the discussion to provide a clearer and more detailed explanation of how ocular and neural responses contribute to speech

processing in an interdependent manner. We hope this revision addresses your concerns and offers a more precise and coherent discussion on this topic:

“Despite the finding that eye movements mediate neural speech tracking, the behavioural relevance for semantic comprehension appears to differ between ocular and neural speech tracking. Specifically, we found a negative association between ocular speech tracking and comprehension, indicating that participants with lower comprehension performance exhibited increased ocular speech tracking. Interestingly, no significant relationship was observed between neural tracking and comprehension.

In this context, the negative association between ocular tracking and comprehension might reflect individual differences in how participants allocate cognitive resources. Participants with lower comprehension may rely more heavily on attentional mechanisms to process acoustic features, as evidenced by increased ocular tracking. This reliance could represent a compensatory strategy when higher-order processes, such as semantic integration or memory retrieval, are less effective. Importantly, our comprehension questions (see Experimental Procedure) targeted a broad range of processes, including intelligibility and memory, suggesting that this relationship reflects a trade-off in resource allocation between low-level acoustic focus and integrative cognitive tasks.

Rather than separating eye and brain responses conceptually, our analysis highlights their complementary contributions. Eye movements may enhance neural processing by increasing sensitivity to acoustic properties of speech, while neural activity builds on this foundation to integrate information and support comprehension. Together, these systems form an interdependent mechanism, with eye and brain responses working in tandem to facilitate different aspects of speech processing.

This interpretation is consistent with the absence of a difference in ocular tracking for semantic violations (e.g., words with high surprisal versus lexically matched controls), reinforcing the view that ocular tracking primarily reflects attentional engagement with acoustic features rather than direct involvement in semantic processing. This aligns with previous findings that attention modulates auditory responses to acoustic features (e.g., Forte et al., 2017), further supporting the idea that ocular tracking reflects mechanisms of selective attention rather than representations of linguistic content.

Future research should investigate how these systems interact and explore how ocular tracking mediates neural responses to linguistic features, such as lexical or semantic processing, to better understand their joint contributions to comprehension.”.

> *(3) Attention vs. predictive coding. I think the authors end up with an elegant description of the observed effects, "as an "active sensing" mechanism that implements the attentional optimization of sensory precision." However, I feel the paragraph starts with the ill-posed question "whether ocular speech tracking is modulated not by predictive, but other (for example attentional) processes". If ocular tracking is the implementation of a process (optimization of sensory precision, aka attention), how could it be at the same time modulated by that process? In my opinion, adding the notion that there is a modulation by a vague cognitive concept like attention on top of what the paper shows does not improve our understanding of how speech tracking in humans works.*

Thank you for raising this point. We agree that it is critical to clarify the relationship between ocular speech tracking, attention, and predictive processes, and we appreciate the opportunity to refine this discussion.

To avoid the potential confusion that active ocular sensing represents on the one hand an implementation of selective attention on the other it seems to be modulated by it, we now use

the formulation “ocular speech tracking reflects attentional mechanisms rather than predictive processes.”

To address your concern that the conceptualization of attention seems rather vague, we have revised the whole paragraph in order to redefine the theoretical entities in question (including selective attention) and to provide a clearer and more precise picture (see also our revised version of Fig. 6, now Fig. 7). We now focus on highlighting the distinct yet interdependent roles of selective attention and individual prediction tendencies for speech tracking.:

“With this speculative framework we attempt to describe and relate three important phenomena with respect to their relevance for speech processing: 1) “Anticipatory predictions” that are created in absence of attentional demands and contain probabilistic information about stimulus features (here, inferred from frequency-specific pre-activations during passive listening to sound sequences). 2) “Selective attention” that allocates resources towards relevant (whilst suppressing distracting) information (which was manipulated by the presence or absence of a distractor speaker). And finally 3) “active ocular sensing”, which refers to gaze behavior that is temporally aligned to attended (but not unattended) acoustic speech input (inferred from the discovered phenomenon of ocular speech tracking). We propose that auditory inflow is, at a basic level, temporally modulated via active ocular sensing, which “opens the gates” in the sensory periphery at relevant timepoints. How exactly this mechanism is guided (for example where the information about crucial timepoints comes from, if not from prediction, and whether it requires habituation to a speechstream etc.) is yet unclear. Unlike predictive tendencies, active ocular sensing appears to reflect selective attention, manifesting as a mechanism that optimizes sensory precision. Individual differences with respect to anticipatory predictions on the other hand, seem to be independent from the other two entities, but nevertheless relevant for speech processing. We therefore support the notion that representational content is interpreted based on prior probabilistic assumptions. If we consider the idea that “a percept” of an (auditory) object is actually temporally and spatially distributed (across representational spacetime - see Fig. 7), the content of information depends on where and when it is probed (see for example Dennett, 1991 for similar ideas on consciousness). Having to select from multiple interpretations across space and time 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.”

> ***Reviewer #2 (Recommendations for the authors):***
>
> *My main recommendation is outlined in the Weaknesses above: the overarching rationale for many analysis choices should be made explicit, and intermediate results should be shown where appropriate, so the reader can follow what is being quantified and what the results truly mean. Specifically, I recommend to pay attention to the following (in no particular order):*
>
> *(1) Define 'neural speech tracking' early on. (e.g.: 'The amount of information in the MEG signal that can multivariately be explained by the speech amplitude envelope.' (is that correct?))*

Thank you for pointing out that this important definition is missing. It is now defined at the first mention in the Introduction as follows: “Here (and in the following) “neural speech tracking” refers to a correlation coefficient between actual brain responses and responses predicted from an encoding model based solely on the speech envelope”.

> *(2) Same for 'ocular speech tracking'. Here even reading the Methods does not make it unambiguous how this term is used.*

It is now defined at the first mention in the Introduction as follows: “Ocular speech tracking” (similarly to “neural speech tracking” refers to the correlation coefficient between actual eye movements and movements predicted from an encoding model based on the speech envelope”.

In addition also define both (neural and ocular speech tracking) metrics in the Methods Section.

> *(3) Related to this: for ocular speech tracking, are simply the horizontal and vertical eye traces compared to the speech envelope? If so, this appears somewhat strange: why should the eyes move more rightward/upward with a larger envelope? And the direction here depends on the (arbitrary) sign of right = positive, etc. (It would make more sense to quantify 'amount of movement' in some way, but if this is done, I missed it in Methods.)*

Thank you for your insightful comments. You are correct that the horizontal and vertical traces were used for ocular speech tracking, and no additional details were included in the Methods. While we agree that the observed rightward/upward movement may seem unusual, this pattern is consistent with previous findings, including those reported in Gehmacher et al. (2024). In that study, we discussed how ocular speech tracking could reflect a broader engagement of the motor system during speech perception. For example, we observed a general right-lateralized gaze bias when participants attended to auditory speech, which we hypothesized might resemble eye movements during text reading, with a similar temporal alignment (~200 ms). We also speculated that this pattern might differ in cultures that read text from right to left.

We appreciate your suggestion to explore alternative methods for quantifying gaze patterns, such as the "amount of movement" or microsaccades. While these approaches hold promise for future studies, our primary aim here was to replicate previous findings using the same signal and analysis methods to establish a basis for further exploration.

> *(4) In the Introduction, specifically blink-related ocular activity is mentioned as being related to speech tracking (for which a reference is, incidentally, missing), while here, any blink-related activity is excluded from the analysis. This should be motivated, as it appears in direct contradiction.*

Thank you for pointing this out. The mention of blink-related ocular activity in the Introduction refers to findings by Jin et al. (2018), where such activity was shown to align with higher-order syntactic structures in artificial speech. We have now included the appropriate reference for clarity.

While Jin et al. focused on blink-related activity, in the present study, we focused on gaze patterns to investigate ocular speech tracking, replicating findings from

Gehmacher et al. (2024). This approach was motivated by our goal to validate previous results using the same methodology. Importantly to this point, the exclusion of blinks in our analysis was due to methodological constraints of TRF analysis, which requires a continuous response signal; blinks, being discrete and artifact-prone, are incompatible with this approach.

To address your concern, we revised the Introduction to clarify this distinction and provide explicit motivation for focusing on gaze patterns. It now reads:

"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; Jin et al, 2018). 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. In the present study, we focused on gaze patterns rather than blink-related activity, both to replicate findings from

Gehmacher et al. (2024) and because blink activity is unsuitable for TRF analysis due to its discrete and artifact-prone nature. Hence, "Ocular speech tracking" (similarly to "neural speech tracking" refers to the correlation coefficient between actual eye movements and movements predicted from an encoding model based on the speech envelope."

Jin, P., Zou, J., Zhou, T., & Ding, N. (2018). Eye activity tracks task-relevant structures during speech and auditory sequence perception. Nature communications, 9(1), 5374.

> *(5) The rationale for the mediation analysis is questionable. Let speech envelope = A, brain activity = B, eye movements = C. The authors wish to claim that A -> C -> B. But it is equally possible that A -> B -> C. They reflect on this somewhat in Discussion, but throughout the rest of the paper, the mediation analysis is presented as specifically testing whether A -> B is mediated by C, which is potentially misleading.*

Indeed we share your concern regarding the directionality of the relationships in the mediation analysis. Our choice of ocular movements as a mediator was motivated by the fact that the relationship between acoustic speech and neural activity is well established, as well as previous results indicating that oculomotor activity contributes to cognitive effects in auditory attention (Popov et al., 2022).

Indeed, here we treat both interpretations ("ocular movements contribute to neural speech tracking" versus "neural activity contributes to ocular speech tracking") as equal. We now emphasise this point in our discussion quite thoroughly:

"It is important to note that our current findings do not allow for inference on directionality. Our choice of ocular movements as a mediator was motivated by the fact that the relationship between acoustic speech and neural activity is well established, as well as previous results indicating that oculomotor activity contributes to cognitive effects in auditory attention (Popov et al., 2022). However, an alternative model may suggest that neural activity mediates the effect of ocular speech tracking. 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. In fact, when rejecting the notion of a single bottom-up flow of information and replacing it with a model of distributed parallel and dynamic processing, it seems only reasonable to assume that the direction of communication (between our eyes and our brain) will depend on where (within the brain) as well as when we look at the effect. Thus, the regions and time-windows reported here should be taken as an illustration of oculo-neural communication during speech processing rather than an attempt to "explain" neural speech processing by ocular movements."

> *(6) The mediation analysis can be improved by a proper quantification of the effect (sizes or variance explained). E.g. how much % of B is explained by A total, and how much of*

> *that can in turn be explained by C being involved? For drawing directional conclusions perhaps Granger causality could be used.*

In Figure 4 (now Figure 5) of our manuscript we use standardized betas (which correspond to effect sizes) to illustrate the mediation effect. With the current mTRF approach it is however not possible (or insightful) to compare the variance explained. It is reasonable to assume that variance in neural activity will be explained better when including oculomotor behavior as a second predictor along with acoustic simulation. However this increase gives no indication to what extent this oculomotor behavior was task relevant or irrelevant (since all kinds of "arbitrary" movements will be captured with brain activity and therefore lead to an increase in variance explained). For this reason we chose to pursue the widely accepted framework of mediation (Baron & Kenny, 1986). This (correlational) approach is indeed limited in its interpretations (see prev. response), however the goal of the current study was to replicate and illustrate the triad relationship of acoustic speech input, neural activity and ocular movements with no particular hypotheses on directionality.

> *(7) Both prediction tendency and neural speech tracking depend on MEG data, and thus on MEG signal-to-noise ratio (SNR). It is possible some participants may have higher SNR recordings in both tasks, which may result in both higher (estimated) prediction tendency and higher (estimated) speech tracking. This would result in a positive correlation, as the authors observe. This trivial explanation should be ruled out, by quantifying the relative SNR and testing for the absence of a mediation here.*

We agree that for both approaches (MVPA and mTRF models) individual MEG SNR plays an important role. This concern has been raised previously and addressed in our previous manuscript (Schubert et al., 2023). First, it should be noted that our prediction tendency value is the result of a condition contrast (rather than simple decoding accuracy) which compensates for the influence of subject specific signal-to-noise ratio (as no vacuous difference in SNR is to be expected between conditions). Second, in our previous study we also used frequency decoding accuracy as a control variable to correlate with speech tracking variables of interest and found no significant effect.

> *(8) Much of the analysis pipeline features temporal response functions (TRFs). These should be shown in a time-resolved manner as a key intermediate step.*

We now included the Neural Speech tracking TRFs into the Figure (now Figure 3).

> *(9) Figure 2 shows much-condensed results from different steps in the pipeline. If I understand correctly, 2A shows raw TRF weights (averaged over some time window?), while 2B-F shows standardized mean posterior regressor weights after Bayesian stats? It would be very helpful to make much more explicit what is being shown here, in addition to showing the related TRFs.*

Thank you for pointing this out! The figure description so far has been indeed not very insightful on this issue. We now adapted the caption and hope this clarifies the confusion: " Neural speech tracking is related to prediction tendency and word surprisal, independent of selective attention. A) Envelope (x) - response (y) relationships are estimated using deconvolution (Boosting). The TRF (filter kernel, h) models how the brain processes the envelope over time. This filter is used to predict neural responses via convolution. Predicted responses are correlated with actual neural activity to evaluate model fit and the TRF's ability to capture response dynamics. Correlation coefficients from these models are then used as dependent variables in Bayesian regression models. (Panel adapted from Gehmacher et al., 2024b). B) Temporal response functions (TRFs) depict the time-resolved neural tracking of the speech envelope for the single speaker and multi speaker target condition, shown here as

absolute values averaged across channels. Solid lines represent the group average. Shaded areas represent 95% Confidence Intervals. C–H) The beta weights shown in the sensor plots are derived from Bayesian regression models in A). For Panel C, this statistical model is based on correlation coefficients computed from the TRF models (further details can be found in the Methods Section). C) In a single speaker condition, neural tracking of the speech envelope was significant for widespread areas, most pronounced over auditory processing regions. D) The condition effect indicates a decrease in neural speech tracking with increasing noise (1 distractor). E) Stronger prediction tendency was associated with increased neural speech tracking over left frontal areas. F) However, there was no interaction between prediction tendency and conditions of selective attention. G) Increased neural tracking of semantic violations was observed over left temporal areas. H) There was no interaction between word surprisal and speaker condition, suggesting a representation of surprising words independent of background noise. Marked sensors indicate 'significant' clusters, defined as at least two neighboring channels showing a significant result. N = 29."

Gehmacher, Q., Schubert, J., Kaltenmaier, A., Weisz, N., & Press, C. (2024b). The "Ocular Response Function" for encoding and decoding oculomotor related neural activity. bioRxiv, 2024-11.

> *(10) Bayesian hypothesis testing is not done consistently. Some parts test for inclusion of 0 in 94% HDI, while some parts adopt a ROPE approach. The same approach should be taken throughout. Additionally, Bayes factors would be very helpful (I appreciate these depend on the choice of priors, but the default Bambi priors should be fine).*

Our primary aim in this study was to replicate two recent findings: (1) the relationship between individual prediction tendencies and neural speech tracking, and (2) the tracking of the speech envelope by eye movements. To maintain methodological consistency with the original studies, we did not apply a ROPE approach when analyzing these replication effects. Instead, we followed the same procedures as the original work, focusing on the inclusion of 0 in the HDI for the neural effects and using the same methods for the ocular effects. Additionally, we were not specifically interested in potential null effects in these replication analyses, as our primary goal was to test whether we could reproduce the previously reported findings.

For the mediation analysis, however, we chose to extend the original approach by not only performing the analysis in a time-resolved manner but also applying a ROPE approach. This decision was motivated by our interest in gaining more comprehensive insights — beyond the replication goals — by also testing for potential null effects, which can provide valuable information about the presence or absence of mediation effects.

We appreciate your thoughtful feedback and hope this clarifies our rationale for the differing approaches in our Bayesian hypothesis testing.

Regarding Bayes Factors,

We understand that some researchers find Bayes Factors appealing, as they offer a seemingly simple and straightforward way to evaluate the evidence in favor of/ or against H0 in relation to H1 (e.g. BF10 > 102 = Decisive; according to the Jeffreys Scale). However, in practice Bayes Factors are often misunderstood e.g. by interpreting Bayes Factor as posterior odds or not acknowledging the notion of relative evidence in the Bayes Factor (see Wong et al. 2022). Instead of using Bayes Factors, we prefer to rely on estimating and reporting the posterior distribution of parameters given the data, prior and model assumptions (in form of the 94% HDI). This allows for a continuous evaluation of evidence for a given hypothesis that is in our eyes easier to interpret as a Bayes Factor.

Jeffreys, Harold (1998) [1961]. The Theory of Probability (3rd ed.). Oxford, England. p. 432. ISBN 9780191589676.

Wong, T. K., Kiers, H., & Tendeiro, J. (2022). On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations. Collabra: Psychology, 8(1), 36357. https://doi.org/10.1525/collabra.36357

> *(11) It would be helpful if Results could be appreciated without a detailed read of Methods. I would recommend a recap of each key methodological step before introducing the relevant Result. (This may also help in making the rationale explicit.)*

In addition to the short recaps of methods that were already present, and information on quantifications of neural and ocular tracking and bayes statistics (see responses 1, 2, 9), we now added the following parts below to the results sections. Please refer to them in the context of the manuscript where they should now complement a key recap of methodological steps necessary to readily understand each analysis and rational that led to the results:

Individual prediction tendency is related to neural speech tracking:

"Thus, this measure is a single value per subject, which comprises a) differences between two contextual probabilities (i.e. ordered vs. random) in b) feature-specific tone representations c) in advance of their observation (summed over a time-window of -0.3 - 0 s). Importantly, this prediction tendency was assessed in an independent entropy modulation paradigm (see Fig. 1). On a group level we found an increased tendency to pre-activate a stimulus of high probability (i.e. forward transition) in an ordered context compared to a random context (see Fig, 2A). This effect replicates results from our previous work (Schubert et al., 2023, 2024). Using the summed difference between entropy levels (ordered - random) across pre-stimulus time, one value was extracted per subject (Fig. 2B). This value was used as a proxy for "individual prediction tendency" and correlated with encoding of clear speech across different MEG sensors. [...]

Neural speech tracking, quantified as the correlation coefficients between predicted and observed MEG responses to the speech envelope, was used as the dependent variable in Bayesian regression models. These models included condition (single vs. multi-speaker) as a fixed effect, with either prediction tendency or word surprisal as an additional predictor, and random effects for participants."

Eye movements track acoustic speech in selective attention:

"For this, we separately predicted horizontal and vertical eye movements from the acoustic speech envelope using temporal response functions (TRFs). The resulting model fit (i.e. correlation between true and predicted eye movements) is commonly referred to as "speech tracking". Bayesian regression models were applied to evaluate tracking effects under different conditions of selective attention (single speaker, attended multi-speaker, unattended multi-speaker). Furthermore, we assessed whether individual prediction tendency or semantic word surprisal influenced ocular speech tracking."

Neural speech tracking is mediated by eye movements:

"This model evaluates to what extent gaze behaviour functions as a mediator between acoustic speech input and brain activity."

Neural and ocular speech tracking are differently related to comprehension: "Bayesian regression models were used to investigate relationships between neural/ocular speech tracking and comprehension or difficulty. Ocular speech tracking was analyzed separately for horizontal and vertical eye movements."

> *(12) The research questions in the Introduction should be sharpened up, to make explicit when a question concerns a theoretical entity, and when it concerns something concretely measured/measurable.*

We sharpened them up:

“Taking into account the aforementioned study by Schubert and colleagues (2023), the two recently uncovered predictors of neural tracking (individual prediction tendency and ocular tracking) raise several empirical questions regarding the relationship between predictive processes, selective attention, and active ocular sensing in speech processing:

(1) Are predictive processes related to active ocular sensing in the same way they are to neural speech tracking? Specifically, do individuals with a stronger tendency to anticipate predictable auditory features, as quantified through prestimulus neural representations in an independent tone paradigm, show increased or even decreased ocular speech tracking, measured as the correlation between predicted and actual eye movements? Or is there no relationship at all?

(2) To what extent does selective attention influence the relationship between prediction tendency, neural speech tracking, and ocular speech tracking? For example, does the effect of prediction tendency or ocular speech tracking on neural tracking differ between a single-speaker and multi-speaker listening condition?

(3) Are individual prediction tendency and ocular speech tracking related to behavioral outcomes, such as comprehension and perceived task difficulty? Speech comprehension is assessed through accuracy on comprehension questions, and task difficulty is measured through subjective ratings.

Although predictive processes, selective attention, and active sensing have been shown to contribute to successful listening, their potential interactions and specific roles in naturalistic speech perception remain unclear. Addressing these questions will help disentangle their contributions and establish an integrated framework for understanding how neural and ocular speech tracking support speech processing.”

> *(13) The negative relationship between story comprehension and ocular speech tracking appears to go against the authors' preferred interpretation, but the reflection on this in the Discussion is very brief and somewhat vague.*

Thank you for pointing this out. We have taken your comments into careful consideration and also incorporated Reviewer #1's query (Minor point 2) into a unified and complementary reasoning. We have rewritten the relevant paragraph in the discussion to provide a clearer and more detailed explanation. We hope this revision offers a more precise and less vague discussion on this important point.

“Despite the finding that eye movements mediate neural speech tracking, the behavioural relevance for semantic comprehension appears to differ between ocular and neural speech tracking. Specifically, we found a negative association between ocular speech tracking and comprehension, indicating that participants with lower comprehension performance exhibited increased ocular speech tracking. Interestingly, no significant relationship was observed between neural tracking and comprehension.

In this context, the negative association between ocular tracking and comprehension might reflect individual differences in how participants allocate cognitive resources. Participants with lower comprehension may rely more heavily on attentional mechanisms to process acoustic features, as evidenced by increased ocular tracking. This reliance could represent a

compensatory strategy when higher-order processes, such as semantic integration or memory retrieval, are less effective. Importantly, our comprehension questions (see Experimental Procedure) targeted a broad range of processes, including intelligibility and memory, suggesting that this relationship reflects a trade-off in resource allocation between low-level acoustic focus and integrative cognitive tasks.

Rather than separating eye and brain responses conceptually, our analysis highlights their complementary contributions. Eye movements may enhance neural processing by increasing sensitivity to acoustic properties of speech, while neural activity builds on this foundation to integrate information and support comprehension. Together, these systems form an interdependent mechanism, with eye and brain responses working in tandem to facilitate different aspects of speech processing.

This interpretation is consistent with the absence of a difference in ocular tracking for semantic violations (e.g., words with high surprisal versus lexically matched controls), reinforcing the view that ocular tracking primarily reflects attentional engagement with acoustic features rather than direct involvement in semantic processing. This aligns with previous findings that attention modulates auditory responses to acoustic features (e.g., Forte et al., 2017), further supporting the idea that ocular tracking reflects mechanisms of selective attention rather than representations of linguistic content.

Future research should investigate how these systems interact and explore how ocular tracking mediates neural responses to linguistic features, such as lexical or semantic processing, to better understand their joint contributions to comprehension.”.

> *(14) Page numbers would be helpful.*

We added the page numbers.

> ***Reviewer #3 (Recommendations for the authors):***
>
> *Results*
>
> *(1) Figure 2 - statistical results are reported in this figure, but they are not fully explained in the text, nor are statistical values provided for any of the analyses (as far as I can tell).*
>
> *Also, how were multiple comparisons dealt with (the choice of two neighboring channels seems quite arbitrary)? Perhaps for this reason, the main result - namely the effect of "prediction tendency" and "semantic violations" - is quite sparse and might not survive more a rigorous statistical criterion. I would feel more comfortable with these results if the reporting of the statistical analysis had been more thorough (ideally, including comparison to control models).*

We would like to thank you again for your detailed queries, comments, and questions on our work. We first of all adapted this figure (now Figure 3 in the manuscript, please see responses 8 and 9 to Reviewer #2) to help readers understand the metrics and values within each statistical analysis. In addition, we indeed did not include the detailed statistics in the text! We now added the missing statistic reports calculated as averages over ‘clusters’:

“Replicating previous findings (Schubert et al., 2023), we found widespread encoding of clear speech (average over cluster: $\beta = 0.035$, 94%HDI = [0.024, 0.046]), predominantly over auditory processing regions (Fig. 3C), that was decreased ($\beta = -0.018$, 94%HDI = [0.029, -0.006]) in a multi-speaker condition (Fig. 3D). Furthermore, a stronger prediction tendency was associated with increased neural speech tracking ($\beta = 0.014$, 94%HDI = [0.004, 0.025]) over left frontal sensors (see Fig. 3E). We found no interaction between prediction tendency and condition (see Fig. 3F).” [...] “In a direct comparison with lexically identical controls, we found

an increased neural tracking of semantic violations (β = 0.039, 94%HDI = [0.007, 0.071]) over left temporal areas (see Fig. 3G). Furthermore, we found no interaction between word surprisal and speaker condition (see Fig. 3H)."

Regarding the "prediction tendency" effect, it is important to note that this finding replicates a result from Schubert et al. (2023). The left frontal location of this effect is also consistent over studies, which convinces us of the robustness of the finding. Furthermore, testing this relationship properly requires a mixed-effects model in order to account for the variability across subjects that is not explained by fixed effects and the repeated measures design. For this reason a random Intercept had to be fitted for each subject (1|subject in the respective model formula). This statistical requirement motivated our decision to use bayesian statistics as (at least to our knowledge) there is no implementation of a cluster-based permutation mixed effects model (yet). In order to provide a more conservative criterion (as bayesian statistics don't require a multiple comparison correction) we chose to impose in addition the requirement of a "clustered" effect.

The choice of using two neighboring channels is consistent with the default parameter settings in FieldTrip's cluster-based permutation testing (cfg.minnbchan = 2). This parameter specifies the minimum number of neighboring channels required for a sample to be included in the clustering algorithm, ensuring spatial consistency in the identified clusters. This alignment ensures that our methodology is comparable to numerous prior studies in the field, where such thresholds are standard. While it is true that all statistical analyses involve some degree of arbitrariness in parameter selection (e.g., alpha levels or clustering thresholds), our approach reflects established conventions and ensures comparability with previous findings.

While the original study utilized source space analyses, we replicated this effect using only 102 magnetometers. This choice was made for computational simplicity, demonstrating that the effect is robust even without source-level modeling. Similarly, the "semantic violation" effect, while perceived as sparse, is based solely on magnetometer data and - in our opinion - should not be viewed as overly sparse given the methods employed. This effect aligns with the two-neighbor clustering approach, ensuring spatial consistency across magnetometers. The results reflect the robustness of the effects within the constraints of magnetometer-level analyses.

Overall, the methodological choices, including the choice of a bayesian linear mixed effects model, the use of two neighboring channels and the reliance on magnetometers, are grounded in established practices and methodological considerations. While stricter thresholds or alternative approaches might yield different results, our methods align with best practices in the field and ensure the robustness, comparability, and replicability of our findings.

> *(2) Figure 3 - the difference between horizontal and vertical eye-movements. This result is quite confusing and although the authors do suggest a possible interpretation for this in the discussion, I do wonder how robust this difference is or whether the ocular signal (in either direction) is simply too noisy or the effect too small to be detected consistently across conditions. Also, the ocular-TRFs themselves are not entirely convincing in suggesting reliable response/tracking of the audio - despite the small-but-significant increase in prediction accuracy.*

The horizontal versus vertical comparison was conducted to explore potential differences in how these dimensions contribute to ocular tracking of auditory stimuli (please also see our response to Reviewer #1, Response 5b, that includes the vertical vs. horizontal effects of Gehmacher at al. 2024). It would indeed be interesting to develop a measure that combines the two directions into a more natural representation of 'viewing,' such as a combined vector.

However, this approach would require the use of complex numbers to represent both magnitude and direction simultaneously, hence the development of novel TRF algorithms capable of modeling this multidimensional signal. While beyond the scope of the current study, this presents an exciting avenue for future research and would allow us to move closer to understanding ocular speech tracking and the robustness of these effects, above and beyond the already successful replication.

It is also important to emphasize that ocular-TRFs are derived from (viewing) behavioral data rather than neural signals, and are thus inherently subject to greater variability across participants and time. This higher variability does not necessarily indicate a small or unreliable effect but reflects the dynamic and task-dependent nature of eye movement behavior. The TRFs with shaded error margins represent this variability, highlighting how eye movements are influenced by both individual differences and moment-to-moment changes in task engagement.

Despite this inherent variability, the significant prediction accuracy improvements confirm that ocular-TRFs reliably capture meaningful relationships between eye movements and auditory stimuli. The observed differences between horizontal and vertical TRFs further support the hypothesis that these dimensions are differentially involved in the task, possibly driven by the specific roles they play in sensorimotor coupling.

> *(3) Figure 4 - this figure shows source distribution of 3 PCA components, derived from the results of the mediation effect of eye movements on the speech-tracking. Here too I am having difficulty in interpreting what the results actually are. For one, all three components are quite widespread and somewhat overlapping, so although they are statistically "independent" it is hard to learn much from them about the brain regions involved and whether they truly represent separable contributions. Similarly difficult to interpret are the time courses, which share some similarities with the known TRFs to speech (especially PC3). I would have expected to find a cleaner "auditory" response, and clearer separation between sensory regions and regions involved in the control of eye movements. I also wonder why the authors chose not to show the sourcelocalization of the neural and ocular speech-tracking responses alone - this could have helped us between understand what "mediation" of the neural response might look like.*

We appreciate the reviewer's interest in better understanding the source distribution and time courses of the PCA components. While we acknowledge that the widespread and overlapping nature of the components may make a more fine grained interpretation challenging, it is important to emphasize that our analysis simply reflects the data, hence we can only present and interpret what the analysis revealed.

Regarding your suggestion to show the source localization of ocular speech tracking and neural speech tracking alone, we would like to point out that ocular tracking is represented by only one channel for vertical and one channel for horizontal eye movements. Thus, in this case the estimated source of the effect are the eyes themselves. We believe that the source localization of neural speech tracking has been a thoroughly studied topic in research so far (locating it to perisylvian, auditory areas with a stronger preference for the left hemisphere) and can also be seen in Schubert et al., (2023). Nevertheless, we believe the observed PCA components still provide valuable, and most importantly novel insights into the interplay between eye movements and neural responses in speech tracking.

> *Discussion/interpretation*
>
> *(1) Although I appreciate the authors' attempt to propose a "unified" theoretical model linking predictions about low-level features to higher features, and the potential involvement of eye movements in 'active sensing' I honestly think that this model is*

> *overambitious, given the data presented in the current study. Moreover, there is very little discussion of past literature and existing models of active sensing and hierarchical processing of speech, that could have helped ground the discussion in a broader theoretical context. The entire discussion contains fewer than 20 citations (some of which are by these authors) and needs to be substantially enriched in order to provide context for the authors' claims.*

Thank you very much for your thoughtful feedback and for appreciating our approach. We hope that the revised manuscript addresses your concerns. Specifically, we now emphasize that our proposal is a conceptual framework, with the main goal to operationale “prediction tendency”, “active ocular sensing”, and “selective attention” and to “organise these entities according to their assumed function for speech processing and to describe their relationship with each other.” We did this by thoroughly revising our discussion section with a clear emphasis on the definition of terms, for example:

“With this speculative framework we attempt to describe and relate three important phenomena with respect to their relevance for speech processing: 1) “Anticipatory predictions” that are created in absence of attentional demands and contain probabilistic information about stimulus features (here, inferred from frequency-specific pre-activations during passive listening to sound sequences). 2) “Selective attention” that allocates resources towards relevant (whilst suppressing distracting) information (which was manipulated by the presence or absence of a distractor speaker). And finally 3) “active ocular sensing”, which refers to gaze behavior that is temporally aligned to attended (but not unattended) acoustic speech input (inferred from the discovered phenomenon of ocular speech tracking).”

Our theoretical proposals are now followed by a recap of our results that support the respective idea, for example:

“...these predictions are formed in parallel and carry high feature-specificity but low temporal precision (as they are anticipatory in nature). This idea is supported by our finding that pure-tone anticipation is visible over a widespread prestimulus interval, instead of being locked to sound onset”

“....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). This assumption is supported by our finding that semantic violations are not differentially encoded in gaze behaviour than lexical controls.”

And we put a strong focus on highlighting the boundaries of these ideas, in order to avoid theoretical confusion, misunderstandings or implicit theoretical assumption that are not grounded in data, in particular:

“In fact, when rejecting the notion of a single bottom-up flow of information and replacing it with a model of distributed parallel and dynamic processing, it seems only reasonable to assume that the direction of communication (between our eyes and our brain) will depend on where (within the brain) as well as when we look at the effect. Thus, the regions and time-windows reported here should be taken as an illustration of oculo-neural communication during speech processing rather than an attempt to "explain" neural speech processing by ocular movements.”

“Even though the terminology [“hierarchy”] is suggestive of a fixed sequence (similar to a multi storey building) with levels that must be traversed one after each other (and even the more spurious idea of a rooftop, where the final perceptual experience is formed and stored into memory), we distance ourselves from these (possibly unwarranted) ideas. Our usage of “higher” or “lower” simply refers to the observation that the probability of a feature at a

higher (as in more associative) level affects the interpretation (and thus the representation and prediction) of a feature at lower (as in more segregated) levels (Caucheteux et al., 2023)."

Additionally, we have made substantial efforts to present complementary results (see response to Reviewer #2, point 8) to further substantiate our interpretation. Importantly, we have updated the illustration of the model (see response to Reviewer #, minor point 1) and refined both our interpretations and the conceptual language in the Discussion. Furthermore, we have included additional citations where appropriate to strengthen our argument.

We would also like to briefly note that this section of the Discussion aimed to highlight existing literature that bridges the gap our model seeks to address. However, as this is a relatively underexplored area, the references available are necessarily limited.

> *(2) Given my many reservations about the data, as presented in the current version of the manuscript, I find much of the discussion to be an over-interpretation of the results. This might change if the authors are able to present more robust results, as per some of my earlier comments.*

We sincerely hope that our comprehensive revisions have addressed your concerns and improved the manuscript to your satisfaction.

https://doi.org/10.7554/eLife.101262.2.sa0