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

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

[1]Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria.
[2]Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria.
*contributed equally

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

ABSTRACT

Auditory speech comprehension is a multi-faceted process in which attention, prediction, and sensorimotor integration (via active sensing) interact with or complement each other. Although different conceptual models that focus on one of these aspects exist, we still lack a unified understanding of their role in speech processing. Here, we replicated and extended two recently published studies from our lab to investigate the influence of prediction tendency on ocular and neural tracking of attended speech. We propose that selective attention guides ocular speech tracking, which in turn mediates neural encoding of speech. In contrast, individual prediction tendency and its relation to neural speech tracking seem to be largely independent of attention. Importantly, prediction tendency and ocular speech tracking seem to be unrelated. With the current work, 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.

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

INTRODUCTION

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

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

Selective attention describes the process by which the brain prioritizes 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 along with 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), consequently promoting 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 an influential role of this “trait” on speech processing independent of selective attention, where speech tracking is enhanced with stronger prediction tendency. Furthermore, we found an increased tracking of words of high surprisal, demonstrating the importance of predictive processes in speech perception.

Another important aspect with regards 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). Even covert (eye)movements have been shown to temporally align with (higher syntactical) speech that was artificially presented in a fixed stimulation rate (Jin et al., 2018).

In another recent study (Gehmacher et al., 2023), we showed that eye movements continuously track intensity fluctuations of attended natural speech, a phenomenon we termed ocular speech tracking. We further established links to increased intelligibility with stronger ocular tracking, and provided evidence for their contribution to the typically observed neural speech tracking effects in selective attention. Taking the aforementioned study of Schubert and colleagues (2023) into account, mediating (ocular) and variance explaining (prediction tendency) effects in neural tracking urge the question how they might be related with regards to speech processing. Are predictive processes similarly related to active sensing as to neural tracking in natural listening, i.e. are stronger prediction tendencies related to increased ocular speech tracking? Or vice versa? Or not at all? Are they differently affected by selective attention? Are they related to behavioral outcomes? While all of these concepts have been shown to contribute to successful listening in challenging environments, a thorough investigation on potential interactions and their consequences on speech perception has been lacking.

Here, we set out to answer these questions by a) replicating aforementioned findings in a single experiment, and b) combining prediction tendency, eye movements, and attention into a unified framework of cortical speech tracking. We therefore repeated the study protocol of Schubert et al. (2023) with slight modifications: Again, participants performed a passive listening paradigm (also see Demarchi et al., 2019) that allowed us to quantify individual prediction tendency. Afterwards, 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 our previous findings: Individuals with a stronger prediction tendency showed an increased neural speech tracking over left frontal areas, eye movements track acoustic speech in selective attention (even for continuous narrative speech), and further mediate neural speech traking effects over widespread, but mostly auditory regions. Additionally, 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 replication findings again suggest a differential role of prediction tendency, ocular tracking, and attention in speech processing. Behavioral responses further indicate substantial differences in ocular and neural engagement and perceptual outcomes. Based on these findings, we propose a unified framework of cortical 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.

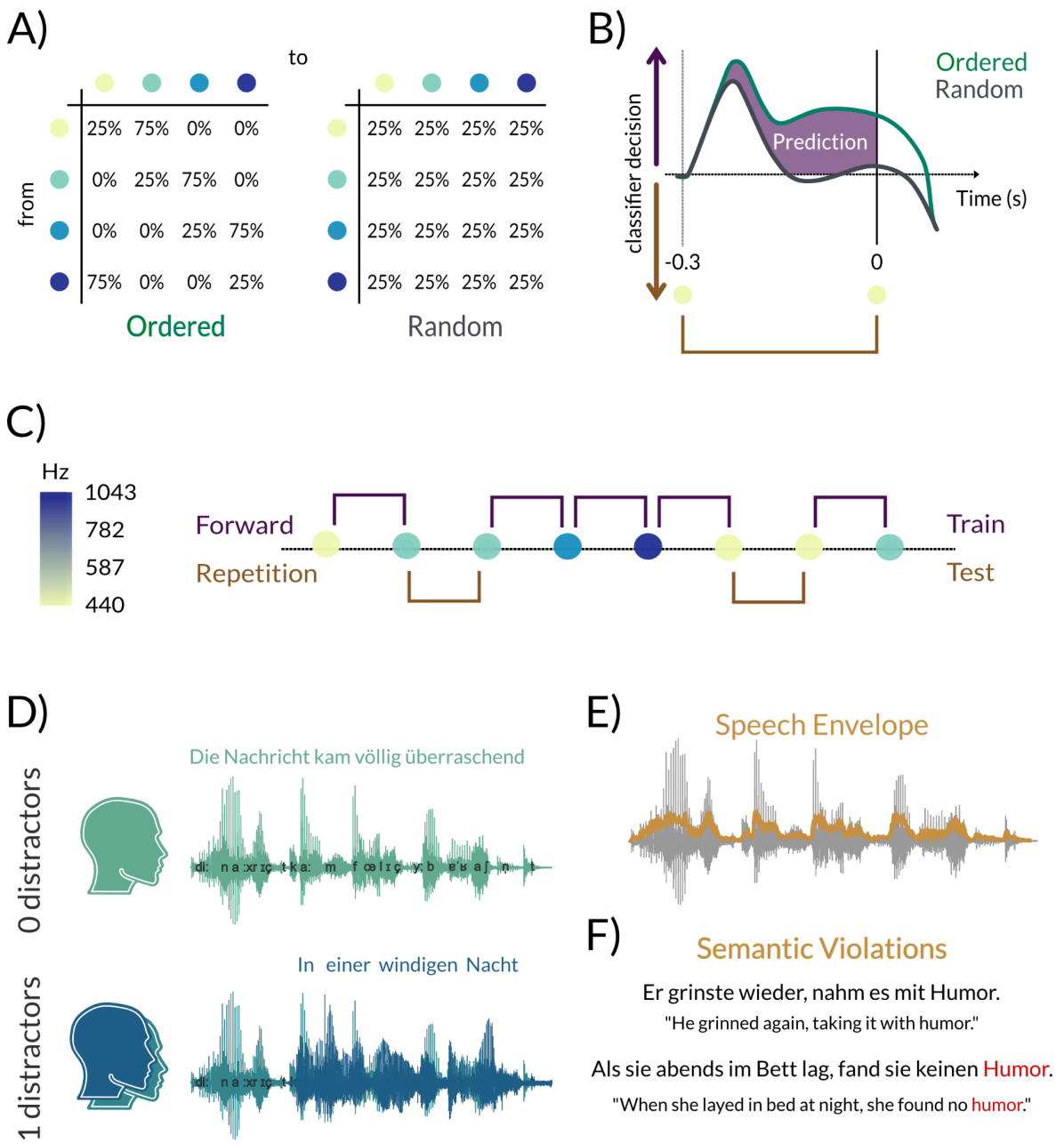

Figure 1: *Quantification of individual prediction tendency and the multi speaker paradigm.* **A)** Participants passively listened to sequences of pure tones in different conditions of entropy (ordered vs. random). Four tones of different fundamental frequencies were presented with a fixed stimulation 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.

METHODS

Subjects

In total, 39 subjects were recruited to participate in the experiment. For 2 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 experiment instructions. 6 subjects had to be excluded according to these criteria. Thus, a final sample size of 31 subjects (mean age = 25.10 , range = 19 - 42) was used for analysis of brain and gaze data.

All participants reported normal hearing and had normal, or corrected to normal, vision. They gave written, informed consent and reported that they had no previous neurological or psychiatric disorders. The experimental procedure was approved by the ethics committee of the University of Salzburg and was carried out in accordance with the declaration of Helsinki. All participants received either a reimbursement of 10 € per hour or course credits.

Experimental Procedure

The current study is a replication of a previous experiment (for details see Schubert et al., 2023), using the same experimental structure and the same auditory stimuli. The current design only differs in that one condition was dropped and the whole experiment was shortened (participants spent approximately 2 hours in the MEG including preparation time). In addition to the previous study, this time we recorded ocular movements (also see Data acquisition and Preprocessing).

Before the start of the experiment, participants' individual 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 level 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 center 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 center). 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. Furthermore, participants indicated their task engagement and their perceived task-difficulty on a 5-point Likert scale at the end of every trial. During audiobook presentation, participants were again instructed to look at the fixcross 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 and replaced them with the other half to induce unexpected semantic violations within each trial, resulting in two sets of lexically identical words (N = 79) that differed greatly in their contextual probabilities (see **Figure 1F** for an example). All trials were recorded twice, narrated by a different speaker (male vs. female). Stimuli were presented in 4 blocks containing 2 trials each (a single and a multi speaker trial), resulting in 2 male and 2 female target speaker blocks for every participant.

Table 1: List of all ebooks and short stories that were used as basis for audio material

"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 ~3min excerpts each), and a distractor speaker (1 x ~3min different excerpt of the same story)

Data Acquisition and Preprocessing

Bran 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, although only magnetometers were selected for later analysis. 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 artifacts, 50 independent components were identified from filtered (0.1 - 100 Hz), downsampled (1000 Hz) continuous data of the recordings from the entropy modulation paradigm, and on average 3 components were removed for every subject. All data was filtered between 0.1 Hz and 30 Hz (Kaiser windowed finite impulse response filter) and downsampled to 100 Hz. Data of the entropy modulation paradigm was epoched into segments of 1200 ms (from 400 ms before sound onset to 800 ms after onset). Multivariate pattern analysis (see quantification of individual prediction tendency) was carried out using the MVPA-Light package (Treder, 2020). Data of the listening task was temporally aligned with the corresponding speech envelope, which was extracted from the audio files using the Chimera toolbox (Smith et al., 2002) over a broadband frequency range of 100 Hz - 10 kHz (in 9 steps, equidistant on the tonotopic map of auditory cortex; see also **Figure 1E**).

Eye tracking data from both eyes were acquired using a Trackpixx3 binocular tracking system (Vpixx Technologies, Canada) at a sampling rate of 2 kHz with a 50 mm lens. Participants were seated in the MEG at a distance of 82 cm from the screen. Their chin rested on a chinrest to reduce head movements. Each experimental block started with a 13-point calibration / validation procedure that was used throughout the block. Blinks were automatically detected by the Trackpixx3 system and excluded from horizontal and vertical eye movement data. We additionally excluded 100 ms of data around blinks to control for additional blink artifacts that were not automatically detected by the eye tracker. We then averaged gaze position data from the left and right eye to increase the accuracy of gaze estimation (Cui & Hondzinski, 2006). Missing data due to blink removal were then interpolated with a piecewise cubic Hermite interpolation. Afterwards, data was imported into the FieldTrip Toolbox, bandpass filtered between 0.1 - 40 Hz (zero-phase finite impulse response (FIR) filter, order: 33000, hamming window) and cut to the length of respective speech segments of a block. Data were then downsampled to 1000 Hz to match the sampling frequency of neural and speech envelope data and further corrected for a 16 ms delay between trigger onset and actual stimulation. Finally, gaze data was temporally aligned with the corresponding speech envelope and downsampled to 100 Hz for TRF analysis after

an antialiasing low-pass filter at 20 Hz was applied (zero-phase FIR filter, order: 662, hamming window).

Quantification of Prediction Tendency

The quantification of individual prediction tendency was the same as in Schubert et al. (2023): We used an entropy modulation paradigm where participants passively listened to sequences of 4 different pure tones (f1: 440 Hz, f2: 587 Hz, f3: 782 Hz, f4: 1043 Hz, each lasting 100 ms) during two separate blocks, each consisting of 1500 tones presented with a temporally predictable rate of 3 Hz. Entropy levels (ordered / random) changed pseudorandomly every 500 trials within each block, always resulting in a total of 1500 trials per entropy condition. While in an “ordered” context certain transitions (hereinafter referred to as forward transitions, i.e. f1→f2, f2→f3, f3→f4, f4→f1) were to be expected with a high probability of 75%, self repetitions (e.g., f1→f1, f2→f2,...) were rather unlikely with a probability of 25%. However, in a “random” context all possible transitions (including forward transitions and self repetitions) were equally likely with a probability of 25% (see **Figure 1A**). To estimate the extent to which individuals anticipate auditory features (i.e. sound frequencies) according to their contextual probabilities, we used a multiclass linear discriminant analyser (LDA) to decode sound frequency (f1 - f4) from brain activity between -0.3 and 0 s in a time-resolved manner. Based on 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). The defined time-lags to train the model were from -0.1 s to 0.5 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 in order to quantify the model fit and the degree of speech envelope tracking at a particular sensor location.

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

The same TRF approach was also used to estimate ocular speech tracking, separately predicting eye movements in the horizontal and vertical direction. Here, we used longer time-lags (from -0.4 s to 0.8 s) in order to account for potentially longer latencies of ocular movements (compared to neural responses) and to investigate the temporal dynamics of the effect more closely.

To investigate the contribution of eye movements to neural speech tracking, we calculated a mediation analysis as in Gehmacher et al. (2023): The TRFs that we obtained from our 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. Thus, the plain effect (i.e. speech envelope predicting neural responses) is represented in the absolute weights (i.e. TRFs) obtained from a simple model with speech envelope as the sole predictor of neural activity. The direct (residual) effect (not mediated by eye movements) is obtained from a model including the speech envelope as well as eye movements and is represented in the exclusive weights (c') of the former predictor (i.e. speech envelope). Separate models were calculated for horizontal and vertical eye movements and to predict the neural responses of all 102 magnetometer channels using the same time-lags as for ocular speech tracking (-0.4 s - 0.8 s). If model weights are significantly reduced by the

inclusion of eye movements into the model (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 details see also Gehmacher et al., 2023).

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, 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:

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

(Note that “envelope tracking” refers to the correlation between predicted and true responses from the aforementioned encoding model and prediction tendency was always zscored 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), separately for horizontal and vertical envelope tracking:

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

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

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

For the mediation analysis, we compared weights (TRFs) for speech envelope between a plain model and a residual model that included eye movements as a second predictor (i.e. c’ < c). In order to investigate the temporal dynamics of the mediation, we averaged absolute TRFs over 4 time-bins of interest: from -0.2 s to 0 s, from 0 s to 0.2 s, from 0.2 s to 0.4 s and from 0.4 s to 0.6 s. The first as well as the last 200 ms of the weights were excluded in order to ensure that the mediation analysis is not biased by typical edge artifacts resulting from the encoding model. Additionally, we checked whether the averaged weights within each time

window were different from 0 (since only then a decrease in weights is actually meaningful) using the following model for every time window, condition, channel and gaze direction:

absolute weights ~ 0 + model + (1|subject)

Only channels for which we found that absolute weights of the plain model were greater than 0 in the particular time window were considered for the mediation analysis. To investigate potential null effects (neural speech tracking that is definitely independent of eye movements), we used a region of practical equivalence (ROPE) approach. The dependent variable (absolute weights) was zscored across channels, timewindows and models to get standardized betas for the mediation effect (i.e. c' < c):

*absolute weights ~ model * prediction tendency + (1|subject)*

This model was calculated for every channel (magnetometer), gaze direction (horizontal vs. vertical), time window and condition (single vs. multi speaker). As suggested by Kruschke (2018), a null effect was considered for betas ranging between -0.1 and 0.1. Conversely, it was considered a significant mediation effect if betas were below -0.1.
Finally, we investigated the relationship between neural as well as ocular speech tracking and behavioral 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 (zscored within condition before entering the model) as predictor:

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

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

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

For all models (as in Gehmacher et al, 2023 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 summarize 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 below the ROPE). Furthermore, we ensured the absence of divergent transitions ($\hat{r} < 1.05$ for all relevant parameters) and an effective sample size > 400 for all models (an exhaustive summary of Bayesian model diagnostics can be found in Vehtari et al., 2021). Finally, when we estimated an effect on brain sensor level (using all 102 magnetometers), we defined clusters for which an effect was only considered as significant if at minimum two neighbouring channels also showed a significant result.

RESULTS

Individual prediction tendency is related to neural speech tracking

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

Additionally, we wanted to investigate how semantic violations affect neural speech tracking. For this reason, we introduced rare words of high surprisal into the story by randomly replacing half of the nouns at the end of a sentence with the other half. In a direct comparison with lexically identical controls, we found an increased neural tracking of semantic violations over bilateral temporal areas (see **Fig. 2E**). Furthermore, we found a small occipital cluster indicating a negative interaction between word surprisal and speaker

condition (see **Fig. 2F**). These findings indicate an increased representation of surprising words which dissolves with increased background noise.

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

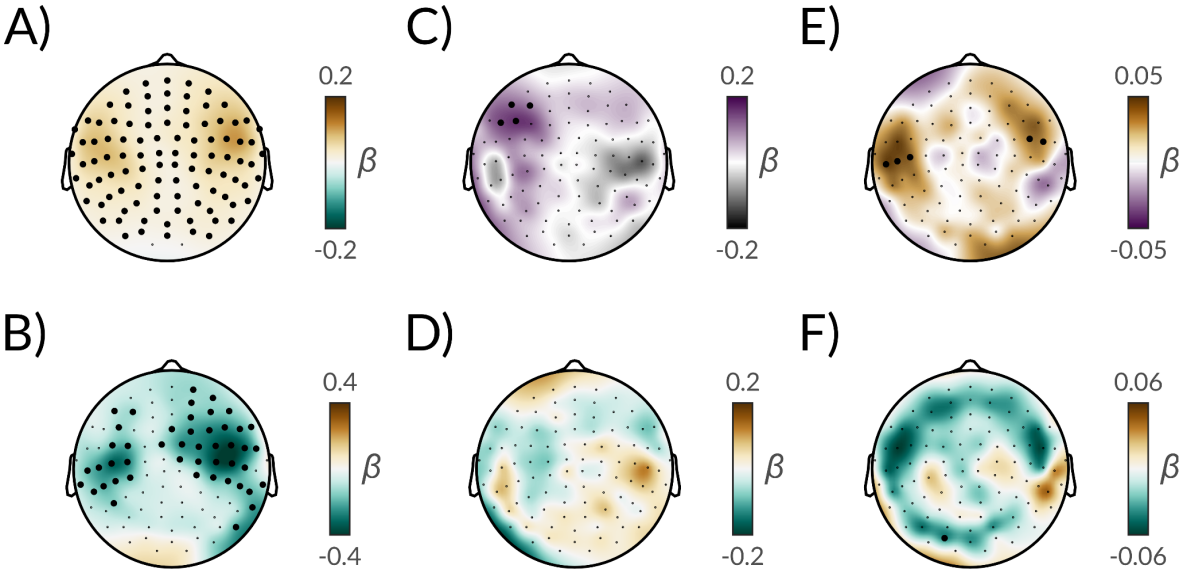

Figure 2: *Neural speech tracking in relation to attention, prediction tendency, and word surprisal.* **A)** In a single speaker condition, neural tracking of the speech envelope was significant for widespread areas, most pronounced over auditory processing regions. **B)** The condition effect indicates a decrease of neural speech tracking with increasing noise (1 distractor). **C)** Stronger prediction tendency was associated with increased neural speech tracking over left frontal areas. **D)** However, there was no interaction between prediction tendency and conditions of selective attention. **E)** There was an increased neural tracking of semantic violations over bilateral temporal areas. **F)** A small occipital cluster indicates a negative interaction between word surprisal and speaker condition, suggestive of an increased representation of surprising words that dissolves with increased background noise. Statistics were performed using Bayesian regression models. Marked sensors show 'significant' clusters where at minimum two neighbouring channels showed a significant result. *N* = 31.

Eye movements track acoustic speech in selective attention

In a second step we aimed to replicate previous finding from Gehmacher and colleagues, (2023) showing that eye movements track acoustic features of speech in absence of visual information. For this we separately predicted horizontal and vertical eye movements from the acoustic speech envelope. We found that horizontally directed eye movements selectively track attended (β = 0.014, 94%HDI = [0.005, 0.023]), but not unattended (β = -0.003, 94%HDI = [-0.011, 0.006]) acoustic speech in a multi speaker condition (see **Figure 3A**). Speech tracking in a single speaker condition did not reach significance (β = 0.007, 94%HDI = [-0.001, 0.016]) for horizontal eye movements. For vertical eye movements, on the other

hand, we found evidence for attended speech tracking in a single speaker condition (β = 0.014, 94%HDI = [0.004, 0.025]) but not in a multi speaker condition (β = 0.005, 94%HDI = [-0.005, 0.015]; see **Figure 3B**). Again, there was no evidence for tracking of the distracting speech stream (β = -0.009, 94%HDI = [-0.020, 0.001]). 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).

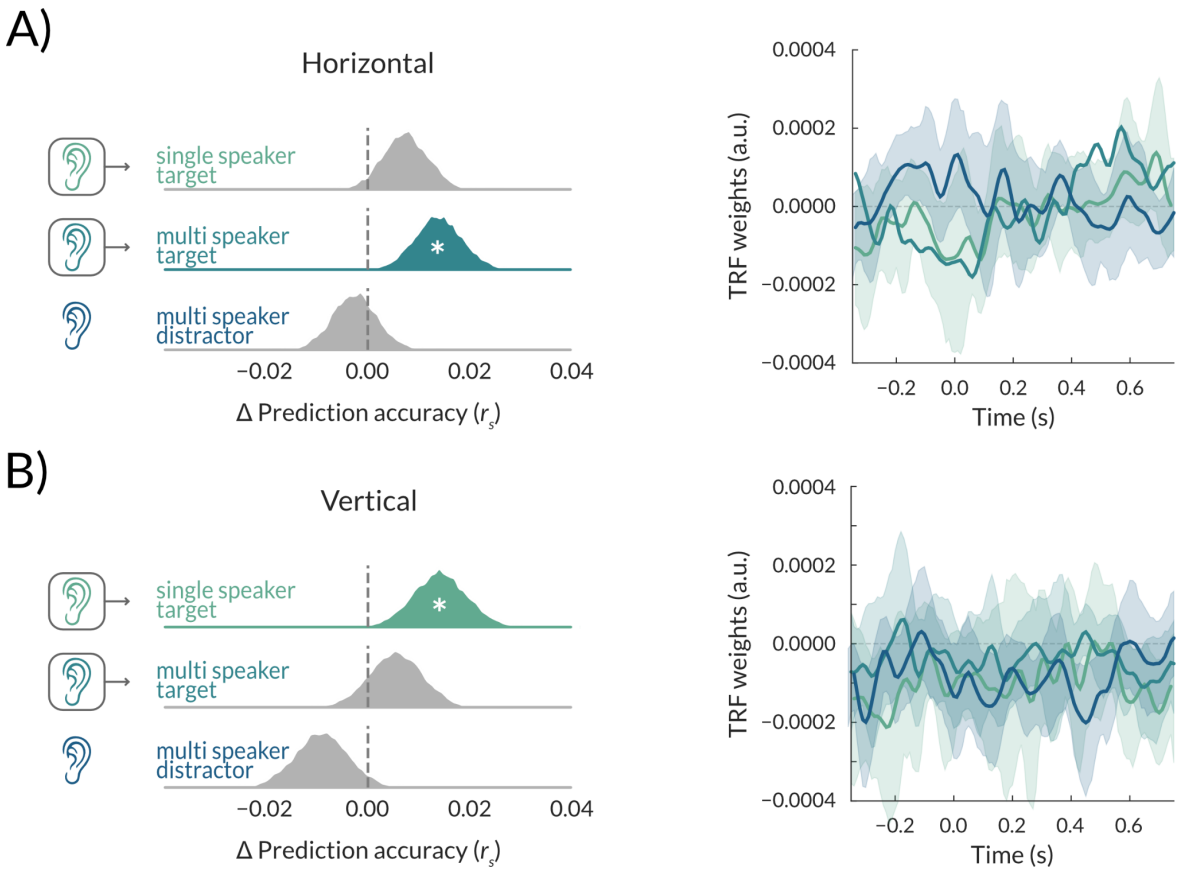

Figure 3: *The effect of selective attention on ocular speech tracking.* **A)** Horizontal eye movements 'significantly' track attended speech in a multi speaker condition. Temporal profiles of this effect show a left-rightwards (negative to positive TRF weights) pattern, suggestive of even anticipatory sensing of speech. **B)** Vertical eye movements 'significantly' track attended clear speech, but not in a multi speaker condition. Temporal profiles of this effect show a downwards pattern (negative TRF weights), suggestive of highly anticipatory clear speech tracking via vertical eye movements. 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 = 31.

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 horizontal (β = -0.001, 94%HDI = [-0.006, 0.005]) or vertical (β = 0.001, 94%HDI = [-0.005, 0.008]) 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 exert in ocular speech tracking independent of their individual prediction tendency or overall semantic probability.

Table 2: Model summary statistics for ocular speech tracking depending on condition and prediction tendency

	horizontal eye movements				vertical eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
cond1: single speaker - attended	0.007	0.005	-0.001	0.016	0.014	0.006	0.004	0.025
cond2: multi speaker - attended	0.014	0.005	0.005	0.023	0.005	0.005	-0.005	0.015
cond3: multi speaker - unattended	-0.003	0.005	-0.011	0.006	-0.009	0.006	-0.020	0.001
prediction tendency	-0.001	0.003	-0.006	0.005	0.001	0.003	-0.005	0.008

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

Table 3: Model summary statistics for ocular speech tracking depending on word type and condition

	horizontal eye movements				vertical eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
Intercept*	-0.022	0.016	-0.009	0.050	0.016	0.020	-0.020	0.053
wordtype (semantic violation effect)	0.003	0.022	-0.038	0.046	0.005	0.026	-0.045	0.053
condition (multi speaker effect)	-0.018	0.023	-0.062	0.023	-0.006	0.026	-0.054	0.044
wordtype x condition	0.016	0.032	-0.046	0.075	-0.004	0.036	-0.072	0.063

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

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

Neural speech tracking is mediated by eye movements

Additionally, we performed a similar, but more detailed mediation analysis compared to Gehmacher and colleagues (2023) 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 as the single predictor for neural responses) to its direct (residual) effect (c') by including horizontal or vertical eye movements as a second predictor into the stimulus model. The decrease in predictor weights from the plain to the residual stimulus model indicates the extent of the mediation effect. To establish a time-resolved mediation analysis, we averaged absolute TRF values across four time windows of interest (see **Figure 4**).

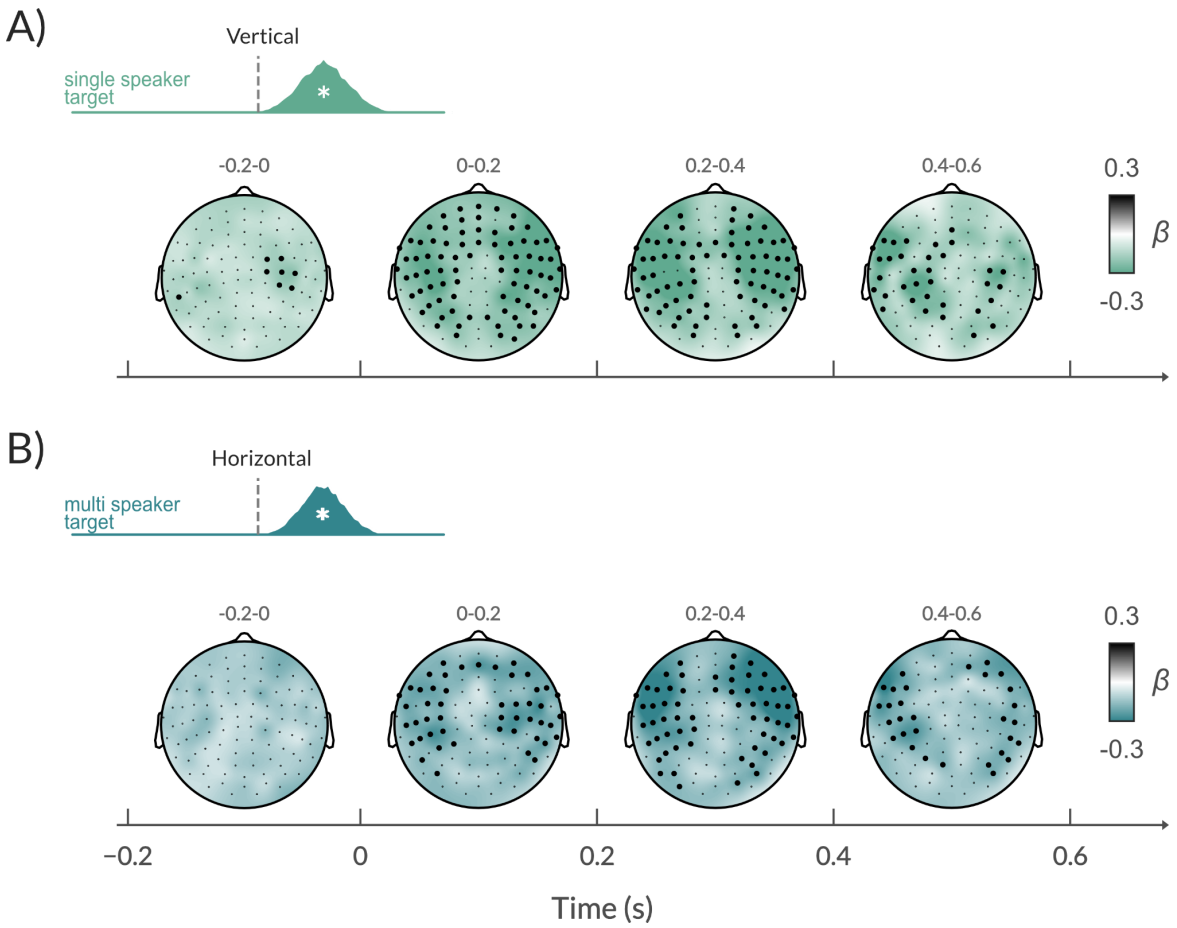

Figure 4: *Contributions of eye movements to neural speech tracking.* **A)** Vertical eye movements significantly mediate neural clear speech tracking throughout the time windows of interest. An anticipatory contribution from -0.2 - 0 s over auditory processing regions leads to a more widespread contribution from 0 - 0.4 s, still most pronounced over auditory sensors. **B)** Horizontal eye movements similarly contribute to neural speech tracking of a target in a multi speaker condition over auditory processing regions. However, we did not find any anticipatory contributions in addition to more distinct and focal effects. Statistics were performed using Bayesian regression models. Marked sensors show 'significant' clusters where at minimum two neighbouring channels showed a significant result. *N* = 31.

When using vertical eye movements in the clear speech condition, we found significant mediation effects in each of the time windows of interest (see **Figure 4A**). Topographies even suggest anticipatory mediation of neural speech tracking over auditory processing regions from -0.2 - 0 s. We then observed a widespread contribution of vertical eye

movements to neural clear speech tracking from 0 - 0.4 s, still mostly contributing to auditory processing regions, before the picture becomes less distinct later on from 0.4 - 0.6 s. For horizontal eye movements in response to a target in the multi speaker condition (see **Figure 4B**), we found similar contribution patterns to neural speech tracking mostly over auditory regions. Interestingly, however, we did not find any anticipatory contributions in addition to more distinct and focal effects for later time lags from 0 - 0.6 s. Furthermore, we did not find any significant null effects (94%HDI falling in a region of practical equivalence) at any of the channels or time-lags for horizontal or vertical eye movements. This suggests that eye movements contribute considerably to neural speech tracking over widespread cortical areas.

Neural and ocular speech tracking are differently related to comprehension

In a final step, we addressed the behavioral 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 speech comprehension. Additionally, we evaluated averaged subjective ratings of difficulty (which were given on a 5-point likert scale).

In order to avoid calculating separate models for all 102 magnetometers, neural encoding was averaged over selected channels (10% showing the strongest encoding effect). We found no significant relationship between neural speech tracking and comprehension (β = 0.144, 94%HDI = [-0.043, 0.337]), but a significant effect for condition (β = -0.363, 94%HDI = [-0.631, -0.099]), indicating that comprehension was decreased in the multi speaker condition (see **Figure 5A**). Similarly, we found no effect of prediction tendency on comprehension (β = -0.008, 94%HDI = [-0.157, 0.133]). When investigating subjective ratings of task difficulty, we also found no effect for neural speech tracking (β = 0.162, 94%HDI = [-0.079, 0.424]) or individual prediction tendency (β = 0.133, 94%HDI = [-0.080, 0.336]). There was, however, a significant difference between conditions (β = 1.253, 94%HDI = [1.009, 1.520]), indicating that the multi speaker condition was rated more difficultly than the single speaker condition.

In contrast to neural findings, we found a negative relationship between ocular speech tracking and comprehension for horizontal as well as vertical eye movements irrespective of condition (see **Figure 5B-C** and **Table 4**). This suggests that participants with weaker performance in 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.

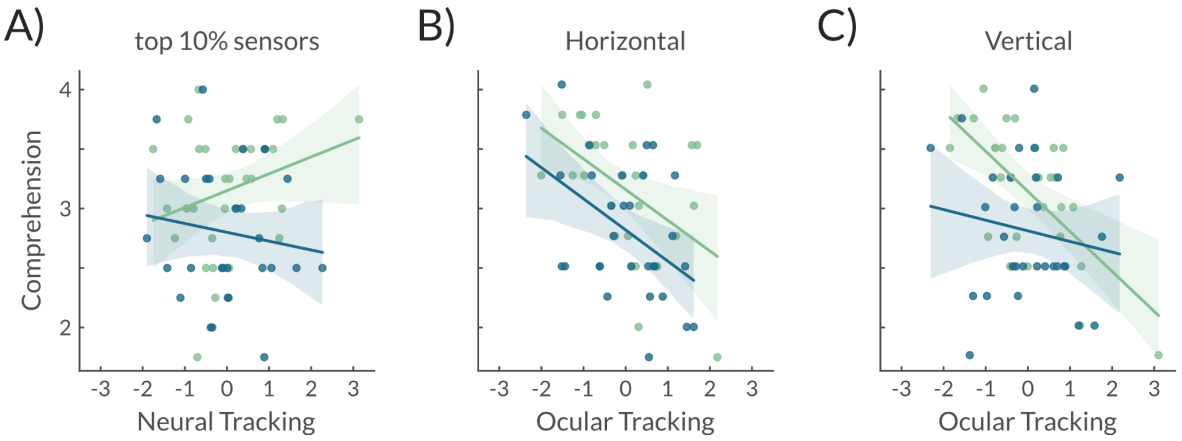

Figure 5: *Relevance of neural and ocular speech tracking for speech comprehension.* **A)** There was no significant relationship between neural speech tracking (10% sensors with strongest encoding effect) and comprehension, however, a condition effect indicated that comprehension was generally decreased in the multi speaker condition. **B & C)** A 'significant' negative relationship between comprehension and horizontal as well as vertical ocular speech tracking shows that participants with weaker comprehension increasingly engaged in ocular speech tracking. For illustration purposes, we depict shaded areas of standard deviations. *N* = 31.

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).

Table 4: Model summary statistics for comprehension depending on ocular speech tracking and condition

	horizontal eye movements				vertical eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
Intercept*	3.132	0.094	2.959	3.311	3.141	0.095	2.966	3.324
ocular speech tracking	-0.270	0.097	-0.451	-0.089	-0.353	0.094	-0.528	-0.174
condition (multi speaker effect)	-0.339	0.127	-0.573	-0.094	-0.346	0.131	-0.597	-0.109
speecht tracking x condition	0.003	0.136	-0.254	0.256	0.235	0.136	-0.026	0.485

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

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

Table 5: Model summary statistics for rated difficulty depending on ocular speech tracking and condition

	horizontal eye movements				vertical eye movements			
	b	sd	hdi 3%	hdi 97%	b	sd	hdi 3%	hdi 97%
Intercept*	2.184	0.117	1.962	2.402	2.181	0.120	1.943	2.397
ocular speech tracking	0.137	0.109	-0.063	0.349	0.102	0.106	-0.097	0.300
condition (multi speaker effect)	1.369	0.129	1.133	1.623	1.374	0.130	1.140	1.629
speecht tracking x condition	-0.138	0.148	-0.414	0.139	-0.09	0.143	-0.359	0.171

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

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

DISCUSSION

In the current study, we aimed to replicate and extend findings from two previous studies in order to integrate them into a corporate view on speech tracking. In Schubert and colleagues (2023), we found that individual prediction tendencies are related to cortical speech tracking. In Gehmacher and colleagues (2023), we found that eye movements track acoustic speech in selective attention (a phenomenon which we termed “ocular speech tracking”). In the present study, we were able to replicate both findings, providing further details and insight into these novel phenomena.

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

In a second step, we were able to support the assumption that eye movements track prioritized acoustic modulations in a continuous speech design without visual input. With the current (extended) replication we were able to address two important considerations from Gemacher and colleagues (2023): First, ocular speech tracking is not restricted to a simplistic 5-word sentence structure but can also be found for continuous, narrative speech. Second, the previous results from the original study cannot be explained by confounds from the visual stimulation. 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 (and naturally horizontal) segregation was increased compared to the single speaker condition where spatial source (phantom center) corresponded to the target.

Irrespective of the potential differences in gaze direction and condition, we found that ocular movements contribute to neural speech tracking over widespread areas. Previous research indicates that attention temporally modulates the alignment of ocular and neural activity with speech features (Jin et al., 2018). It is important to note that our current findings do not allow for inference on directionality. Hence, it is possible that ocular mediation of speech tracking may reflect a) active sensing for information driven by (top-down) selective attention or b) improved neural representations as a consequence of increased sensory gain in the periphery or c) (not unlikely) both. Despite the finding that eye movements mediate neural speech tracking, the behavioral relevance seems to differ between ocular and neural speech tracking. To be more specific, we found a negative association between ocular speech encoding and comprehension, indicating that subjects with weaker performance in comprehension increasingly engaged in ocular speech tracking. This suggests that speech (envelope) tracking may reflect different underlying (cognitive) processes depending on where the response was measured (eyes vs. brain). To be more precise, the extent to which speech tracking can be used as a measure to infer speech comprehension may be different for ocular and neural speech tracking. Regarding neural speech (envelope) tracking, it has been argued that it “is used as a measure to assess whether the pre-conditions for speech intelligibility are satisfied” (Brodbeck & Simon, 2020). Here, we did not find a relationship between neural tracking and comprehension, potentially because our questions on story content target a broad range of cognitive processes (from intelligibility to memory). Crucially, we argue that ocular speech tracking may not be indicative of “qualia” in higher level (linguistic or semantic) stimulus representations. Instead, we propose that it reflects an attentional (and maybe even compensatory) mechanism. For example, it has been found that attention modulates phase in the auditory brainstem response, locking it to the pitch structure of attended but not unattended speech (Forte et al., 2017). Similarly, it is possible that increased ocular tracking indicates increased selective attention in the auditory periphery, rather than improved representations of stimulus content on a higher level.

Interestingly, we were not able to relate individual prediction tendency to ocular speech tracking. This raises the question whether ocular speech tracking is modulated not by predictive, but other (for example attentional) processes. Indeed, the current findings suggest that prediction tendencies and active ocular sensing are related to different aspects of neural speech processing. We propose a perspective in which active sensing is guided by selective attention, whereas anticipatory prediction tendency is an independent correlate of neural speech tracking. Even though predictive processing as well as attention might be considered as the two pillars on which perception rests, research investigating their selective contributions and interactions is rare. Summerfield and de Lange (2014) have argued that predictions and attention are distinctive mechanisms as the former refers to the probability and the latter to the relevance of an event, arguing that events can be conditionally probable but irrelevant for current goals and vice versa. Similarly, it has been proposed that attention can be integrated into the predictive coding framework in reference to the optimization of sensory precision (Feldman & Friston, 2010). In this view, prediction tendency encodes probabilistic representations of a feature, whereas selective attention determines the precision of sensory inflow that leads to internal model updating. This interpretation is line with our finding that a) anticipatory 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) and d) contributes to neural speech processing over widespread areas with a potential overlap with attentional networks. For this reason, we refer to ocular speech tracking as an “active sensing” mechanism. Instead of passively transducing any input into neural activity, we actively engage with our environment to maximize information, i.e. reduce uncertainty. It has been suggested that motor routines contribute to temporal precision of selective attention, largely determining sensory inflow and hence perception (Morillon et al., 2015; Schroeder et al., 2010). In the current framework, we refer to active sensing via eye movements as temporally aligned shifts of gaze at exact points in time aligned with its intensity (information content) fluctuations. In particular, active ocular sensing may be used to increase sensory gain in the auditory periphery at specific intervals, synchronized with the temporal modulation of attended (but not unattended) speech. As a consequence of this “sensory gating” in the auditory periphery, 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. 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 center), we

argue that active ocular sensing increased temporal precision (as speech is composed of rhythmic, temporally predictable patterns within a particular speaker).

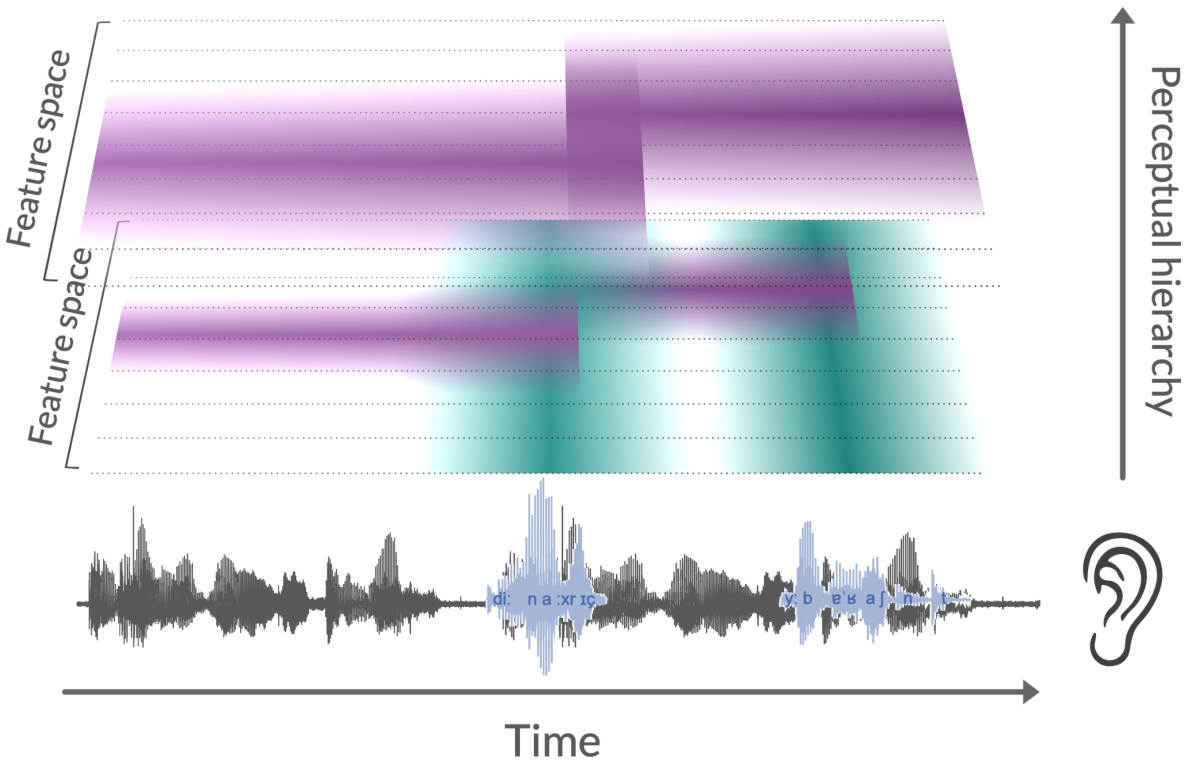

Figure 6: *A schematic illustration of the framework.* Anticipatory predictions help to interpret auditory information at different levels along the perceptual hierarchy (purple) with high feature-specificity but low temporal precision. Active sensing (green) increases the temporal precision of peripheral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

Based on the joint findings of the present as well as its preceding studies (Gehmacher et al., 2023, Schubert et al., 2023), we propose a unified working model in which anticipatory predictions as well as active sensing work (independently) together to support auditory speech perception. We suggest that feature-specific anticipatory predictions help to interpret auditory information at different levels along the perceptual hierarchy. Accordingly, these predictions carry high feature-specificity but low temporal precision (as they are anticipatory in nature). Furthermore, individuals differ in their tendency to create such anticipatory predictions, which leads to differences in neural speech tracking. Active sensing, on the other hand, increases temporal precision at the sensory periphery to facilitate bottom-up processing of selectively attended input (see **Figure 6** for a schematic illustration). 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). 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.

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 behavior) for neural speech tracking. Our interpretation suggests that increased tracking of unexpected input (i.e. semantic violations) should be affected by active ocular sensing if sensory gain in the periphery is indeed dependent on this mechanism. Currently, the findings propose that active ocular sensing, with its substantial contribution to neural speech tracking, is driven by selective attention, and not by individual differences in prediction tendency.

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

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

REFERENCES

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

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/10.1101/2021.08.01.454687

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

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

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). Bambi: A simple interface for fitting Bayesian linear models in Python. *ArXiv Preprint ArXiv:2012.10754*.

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

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, *18*(3), 191–212. https://doi.org/10.1080/09548980701609235

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

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

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

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

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

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2021). Active listening. *Hearing Research*, *399*, 107998. https://doi.org/10.1016/j.heares.2020.107998

Gehmacher, Q., Schubert, J., Schmidt, F., Hartmann, T., Reisinger, P., Rösch, S., Schwarz, K., Popov, T., Chait, M., & Weisz, N. (2023). *Eye movements track prioritized auditory features in selective attention to natural speech* (p. 2023.01.23.525171). bioRxiv. https://doi.org/10.1101/2023.01.23.525171

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

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), Article 1. https://doi.org/10.1038/s41467-018-07773-y

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

Kleiner, M., Brainard, D. 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/v070821

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

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

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

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

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

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

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

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

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

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of Active Sensing and Perceptual Selection. *Current Opinion in Neurobiology*, *20*(2), 172–176. https://doi.org/10.1016/j.conb.2010.02.010

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

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical

perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

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

Stephanie N Lovich, Cynthia D King, David LK Murphy, Rachel Landrum, Christopher A Shera, & Jennifer M Groh. (2022). Parametric information about eye movements is sent to the ears. *BioRxiv*, 2022.11.27.518089. https://doi.org/10.1101/2022.11.27.518089

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

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

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

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

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

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

Zion-Golumbic, E., & Schroeder, C. E. (2012). Attention modulates ‘speech-tracking’ at a cocktail party. *Trends in Cognitive Sciences*, *16*(7), 363–364.

https://doi.org/10.1016/j.tics.2012.05.004