Eye movements track prioritized auditory features in selective attention to natural speech

Quirin Gehmacher[1]*, Juliane Schubert[1], Fabian Schmidt[1], Thomas Hartmann[1], Patrick Reisinger[1], Sebastian Rösch[3], Konrad Schwarz[4], Tzvetan Popov[5,6], Maria Chait[7], Nathan Weisz[1,2]

[1]Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria.

[2]Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria.

[3]Department of Otorhinolaryngology, Head and Neck Surgery, Paracelsus Medical University Salzburg, 5020 Salzburg, Austria.

[4]MED-EL GmbH, 6020 Innsbruck, Austria

[5]Methods of Plasticity Research, Department of Psychology, University of Zurich, CH-8050 Zurich, Switzerland

[6]Department of Psychology, University of Konstanz, DE- 78464 Konstanz, Germany

[7]Ear Institute, University College London, London, UK

*Corresponding author: quirin.gehmacher@plus.ac.at

Abstract

Over the last decades, cognitive neuroscience has identified a distributed set of brain regions that are critical for attention - one of the key principles of adaptive behavior. A strong anatomical overlap with brain regions critical for oculomotor processes suggests a joint network for attention and eye movements. However, the role of this shared network in complex, naturalistic environments remains understudied. Here, we investigated eye movements in relation to (un)attended sentences of natural speech in simultaneously recorded eye tracking and magnetoencephalographic (MEG) data. Using temporal response functions (TRF), we show that eye gaze tracks acoustic features (envelope and acoustic onsets) of attended speech, a phenomenon we termed *ocular speech tracking*. Ocular speech envelope tracking even differentiates a target from a distractor in a multi speaker context and is further related to intelligibility. Moreover, we provide evidence for its contribution to neural differences in speech processing, emphasizing the necessity to consider oculomotor activity in future research and in the interpretation of neural differences in auditory cognition. Our results extend previous findings of a joint network of attention and eye movement control as well as motor theories of speech. They provide valuable new directions for research into the neurobiological mechanisms of the phenomenon, its dependence on learning and plasticity, and its functional implications in social communication.

Introduction

The brain is highly efficient in processing a vast amount of information in complex environments, thereby enabling adaptive behavior. A key principle of adaptive behavior is the goal-directed prioritization and selection of relevant events or objects by attention. From a neurobiological perspective, a distributed attention network extending from relevant sensory cortices to temporal, parietal, and frontal regions (Corbetta et al., 2008; Corbetta & Shulman, 2002; Hopfinger et al., 2000; Luo & Maunsell, 2019) shows a strong anatomical overlap with brain regions critical for oculomotor processes, suggesting a joint network for attention and eye movements (Astafiev et al., 2003; Corbetta et al., 1998; Wardak et al., 2006).

Just as eye movements are necessary for the goal-directed exploration of the visual field, gathering and evaluating additional information via omnidirectional hearing is an inevitable requirement for action preparation and adaptive behavior. Studies on monkeys and cats suggest a midbrain level hub of inferior colliculus (IC, an obligatory station of the ascending auditory pathway) and superior colliculus (SC, which controls ocular dynamics) to integrate sounds and visual scenes via eye movements (e.g. Bulkin & Groh, 2012; Lee & Groh, 2012; Porter et al., 2007). This circuit has recently been extended to the auditory periphery in humans (Lovich et al., 2022; Murphy et al., 2022). Accordingly, several studies in humans point towards interactions between eye movements and auditory cognition in sound localization (Getzmann, 2002), spatial discrimination (Maddox et al., 2014), and spatial attention (Pomper & Chait, 2017) with lateralized engagement of the posterior parietal cortex in unison with lateralized gaze direction (Popov et al., 2022). However, the role of a shared network of auditory attention and eye movements in more complex, naturalistic listening situations remains largely unknown.

Speech represents a key component of social communication that requires a highly selective allocation of spatial, temporal, and feature-based attention. In a mixture of spatially distributed speakers (i.e. “cocktail party scenario”), orienting the eyes towards the target source seems to increase intelligibility (Best et al., 2020), and eye blinks are more likely to occur during pauses in target speech compared to distractor speech (Holtze et al., 2022). In addition, Jin and colleagues (2018) showed that blink related eye activity aligned with higher-order syntactic structures of temporally predictable, artificial speech (i.e. monosyllabic words), similar to neural activity. Their results suggest a global neural entrainment across sensory and (oculo)motor areas which implements temporal attention, supporting ideas that the motor system is actively engaged in speech perception (Galantucci et al., 2006; Liberman & Mattingly, 1985). Taken together, the evidence strongly suggests an engagement of the oculomotor system in auditory selective attention even in more complex scenes involving speech. This engagement also seems to support adaptive behavior. However, several important questions that are essential for a comprehensive understanding of a joint network of auditory attention and eye movements remain unanswered:

Firstly, it is unknown whether eye movements (aside from blinking) continuously track ongoing acoustics of speech, especially in naturalistic scenarios where features often overlap in a mixture of target and distracting sources. This ocular speech tracking could be sensitive to selective attention, for example by gaze reorientation in concordance with relevant structures of attended speech streams. Crucially, the absence of any spatial cues or discriminability could additionally provide valuable information on the underlying principles of oculomotor action in auditory selective attention. Secondly, it is unknown whether ocular speech tracking is related to adaptive behavior, as quantification of important markers like intelligibility or effort are, to date, lacking. Thirdly, the contribution of eye movements to neural processes and underlying computations in selective attention to speech has been overlooked completely. In their aforementioned study, Popov et al. (2022) indicated that a partial contribution of goal-driven oculomotor activity to typical cognitive effects in spatial auditory attention was retained even after removing scalp signal variance (e.g. by means of independent component analysis, ICA) related to ocular muscle activity. Based on their findings, it is important to address this possible contribution when evaluating neural responses in selective attention to speech.

To answer these questions, we analyzed simultaneously recorded eye tracking and magnetoencephalographic (MEG) data from participants listening to short sentences of natural speech at phantom center. Critically, we manipulated attention within and across modalities such that sentences were presented as distractors (Condition 1), as targets (Condition 2), or as a mixture of target and distractor in a multi speaker scenario (Condition 3). Using temporal response functions (TRF; Crosse et al., 2016, 2021), we show that attended features of speech (i.e. envelope and acoustic onsets) are, in fact, tracked by eye movements. Crucially, ocular speech envelope tracking is stronger for a target compared to a distractor speaker in the multi speaker condition. Furthermore, this ocular speech envelope tracking is related to intelligibility. Finally, using a mediation analysis approach, we show that eye movements and selective attention to speech share neural mechanisms over temporal, parietal, and frontal sensors, suggesting a partial contribution of ocular speech tracking to brain regions typically involved in speech processing. Taken together, these results extend previous evidence for a joint network of auditory selective attention and eye movements in complex naturalistic environments. They extend motor theories of speech and additionally provide implications on adaptive behavior. Moreover, they suggest a contribution of oculomotor activity to neural responses in auditory selective attention that should be taken into consideration by future research on auditory cognition.

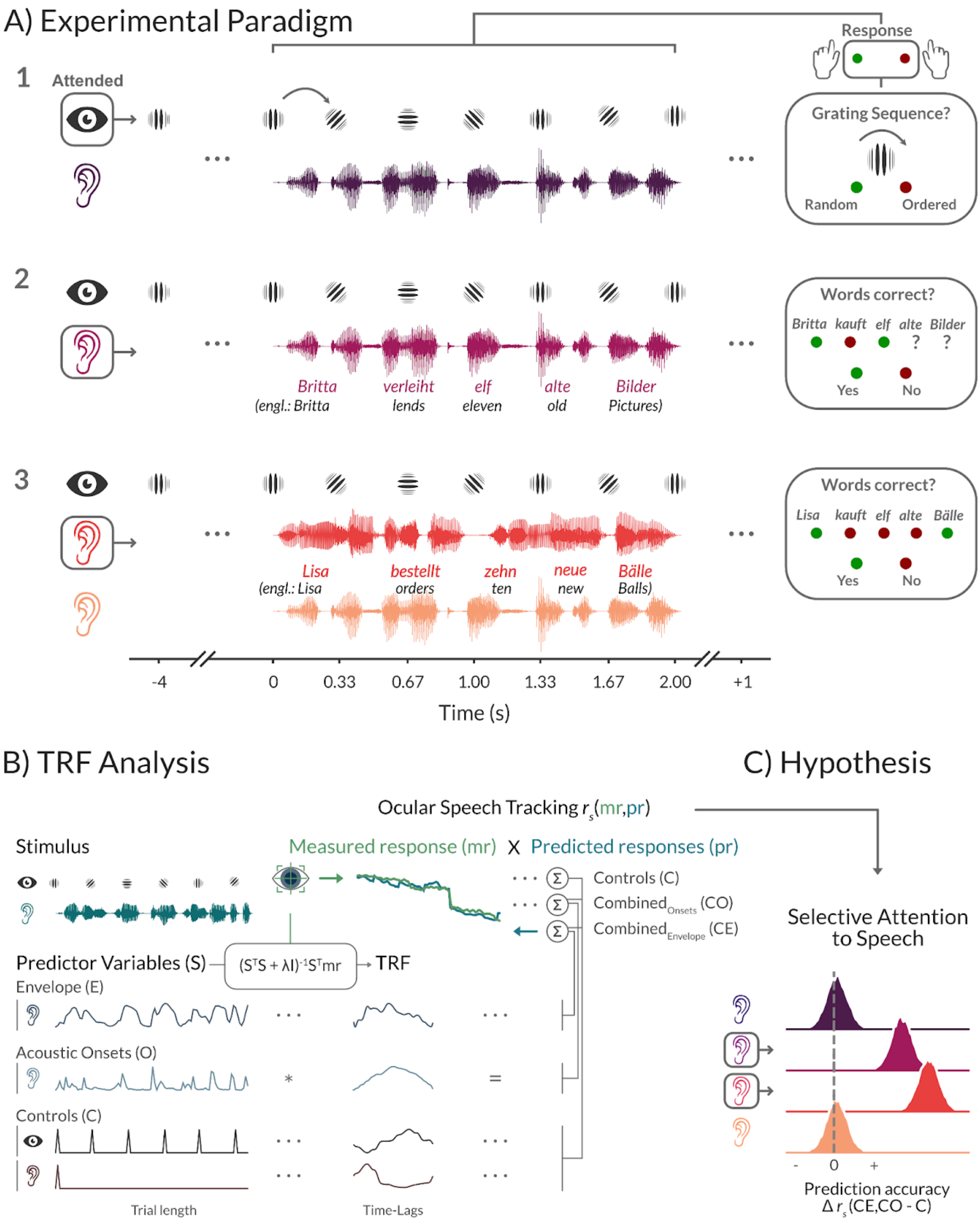
A) Experimental Paradigm
Response
1 Attended
Grating Sequence?
Random
Ordered
2
Words correct?
Britta kauft elf alte Bilder
? ?
Yes No
Britta verleiht elf alte Bilder
(engl.: Britta lends eleven old Pictures)
3
Words correct?
Lisa kauft elf alte Bälle
Yes No
Lisa bestellt zehn neue Bälle
(engl.: Lisa orders ten new Balls)
-4 0 0.33 0.67 1.00 1.33 1.67 2.00 +1
Time (s)
B) TRF Analysis
C) Hypothesis
Ocular Speech Tracking r_s(mr,pr)
Stimulus
Measured response (mr) X Predicted responses (pr)
Controls (C)
$\text{Combined}_{\text{Onsets}}$ (CO)
$\text{Combined}_{\text{Envelope}}$ (CE)
Predictor Variables (S)
$(S^TS + \lambda I)^{-1}S^T mr$
TRF
Envelope (E)
Acoustic Onsets (O)
*
=
Controls (C)
Trial length
Time-Lags
Selective Attention to Speech
- 0 +
Prediction accuracy
Δr_s(CE,CO - C)

Fig. 1: *The framework for isolating modulations of ocular speech tracking in selective attention.* **A)** The task contained trials of short 5-word sentences of natural speech. Participants' attention was modulated within and across modalities. In Condition 1, a rotating gabor patch was attended in the visual modality, while speech served as a distractor. This was reversed in Condition 2, where speech was the focus of attention. In Condition 3, another speaker of the opposite sex was added to investigate ocular speech tracking of a target speaker with a simultaneously presented distractor. After each trial, participants responded to questions on the screen with a handheld button box: to the

gabor rotations in Condition 1 and to the presented words in the target speaker in Conditions 2 & 3. **B)** A regularized linear regression approach called temporal response functions (TRF) was used to predict how features of speech are encoded in eye movements. The difference in prediction accuracy for a control model (C) and combined models that additionally contained the speech envelope (CE) or acoustic onsets (CO) was used to estimate ocular speech tracking solely related to the acoustic features of interest, i.e. speech envelope and acoustic onsets. Prediction accuracies were calculated by Spearman's rank correlation between measured eye movements (mr) and predicted eye movements (pr). The visualized responses represent actual data of an example trial, highlighting the accuracy of the approach. **C)** We expected ocular speech tracking to be modulated by task induced selective attention. Tracking difference of combined and control models (i.e. pure speech tracking) was expected to be higher whenever sentences were the target in a single speaker or multi speaker condition. For statistical computations we used Bayesian multilevel regression models and illustrated the posterior distributions.

Materials and Methods

Participants

30 healthy participants (19 female, M_{age} = 26.27, SD_{age} = 9.08) were recruited for this study. Participants were compensated either financially or via course credits. All participants were German native speakers, reported normal hearing, and (corrected to) normal vision. Participants gave written, informed consent and reported 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.

Experimental Procedure

The experiment lasted ~ 3.5 hours. Five head position indicator (HPI) coils were first applied to the participants' scalp. Anatomical landmarks (nasion and left/right pre-auricular points), HPI locations, and around 300 additional head shape points were then sampled using a Polhemus FASTTRAK. Recording sessions started with 5 min of resting state data (2.5 min eyes open / closed), followed by two blocks of passive listening to tone sequences of varying entropy level (as in Schubert et al., 2022). Afterwards, one block of 10000 clicks at 60dB sound pressure level was presented to determine individual auditory brainstem responses (as in Schmidt et al., 2020) while participants watched a landscape movie (LoungeV Films, 2017). As these parts of the experiment relate to separate research questions, they are not explained in further detail here. The main task (Fig. 1A) consisted of three conditions split into six blocks of 50 trials, i.e. 100 trials per condition. The order of the blocks and trials was randomized across participants. The purpose of the three conditions was to modulate attention within as well as across modalities. Condition 1 tasked the participants with attending to the visual modality (regularity of gabor rotations) while being distracted by a short sentence spoken by a male voice. Condition 2 reversed the task, requiring participants to allocate attention to the auditory modality (natural spoken sentence) while visual stimulation served as a distractor. In Condition 3, the visual modality and the male speaker distracted the participants from attending to an added female target speaker. Each trial started with a silent 4 s prestimulus interval during which participants had to keep their gaze on a gabor patch presented in the center of the screen. Then, a short sentence was played while simultaneously the gabor patch on the screen tilted to one of four perceptually different angles (0°, 45°, 90°, 135°). For the duration of the sentence, the gabor patch was tilted with a fixed stimulation rate of 3 Hz, each lasting for 100 ms. The tilting either followed a) an ordered, clockwise sequence where the upcoming gabor was tilted 45° with a probability of 75% or stayed at the same angle at 25% probability, or b) was randomly tilted to one of the four predefined angles, all equally likely with 25% (note that the transitional probabilities and stimulation rates were the same as in the passive listening task and chosen to not interfere/co-occurre with common syllable rates in language at ~ 4Hz). Ordered and random sequences were pseudorandomized across trials. Stimulation offset was followed by a 1 s silent poststimulus interval with the gabor patch at its original tilt at 0°. During the whole trial period, participants were instructed to keep their gaze on the gabor patch in the center of the screen - regardless of condition - to allow for valid eye-tracking data (for heat maps with gaze position and comparison across experimental conditions; see *Supplementary Figures*, Fig. 1). Each trial was followed by a behavioral response with a question on the stimulation period presented on the screen. In Condition 1, the response required participants to successfully infer whether the gabor transitions followed an ordered or random sequence. In Conditions 2 & 3 we assessed intelligibility scores, probing participants on every word in the attended sentence. For this, we randomly replaced up to all five words of the sentence during the stimulation period (see *Stimuli*) and presented them on the screen. Participants could then mark every word as 'yes', i.e. correct, or 'no', i.e. false. Every word on the response screen could have potentially been correct or false. At the end of each block, we additionally assessed task engagement and subjectively perceived effort on a 5-point Likert scale. All responses were given on a handheld button box. All auditory stimuli were presented binaurally at phantom center at a comfortable loudness level. The experiment was coded and conducted with Psychtoolbox-3 (D. H. Brainard & Vision, 1997; Kleiner et al., 2007) implemented in Matlab R2020b (The MathWorks, Natick, Massachusetts, USA) with an additional class-based library ('Objective Psychophysics Toolbox', o_ptb; Hartmann & Weisz, 2020).

Stimuli

The visual stimulation was a gabor patch (spatial frequency: 0.01 cycles/pixel, sigma: 60 pixels, phase: 90°). For auditory stimulation, we used 100 sentences from the 'Oldenburger Satztest' (OLSA; Wagener et al., 1999) for the male speaker. We created 100 additional 'surrogate' sentences in the same style for the female speaker. Alongside randomization, this ensured that any effects, especially in the multispeaker Condition 3, could not be attributed to memorization of previous trials. Often used in studies on hearing impairment, the OLSA is a standardized audiometric test to assess speech intelligibility. It features lists of 5-word sentences in a fixed form: Name - Verb - Number - Adjective - Noun (see Fig. 1A). Ten words of each word type are used to create 100 unique sentences through random combinations. For the 'surrogate' sentence list, we substituted the ten words per word type with ten other Names, Verbs, Numbers, Adjectives, and Nouns respectively. This led to 100 unique sentences for the male speaker (target in Condition 2, distractor in Conditions 1 & 3) and 100 unique sentences for the female speaker (target in Condition 3). Unlike commonly used questions on general content or last words in speech tracking designs with longer segments (e.g. audiobooks), the fixed 5-word structure allowed us to probe speech intelligibility on a word-by-word level. To synthesize the 200 extracted sentences into natural-sounding speech, we used the IBM Watson text-to-speech service (TextToSpeechV1 package). We synthesized german text-to-speech at a sampling rate of 44.1 kHz using the implemented voices for the male speaker (voice 'de-DE_DieterV3Voice') with adjusted prosody rate to -10 % in order to match the female speaker's ('de-DE_ErikaV3Voice') syllable rate. This led to slightly different sentence durations for the male and female speaker (M_{male} = 2.02 s, SD_{male} = 0.16, M_{female} = 2.22 s, SD_{female} = 0.13) due to the slightly longer surrogate sentences (as number words were exhausted from 1 - 10 in the original list, for surrogate sentences the number words included thirteen, fourteen, …). However, this was controlled for later in the analysis by cropping the aligned data (see *TRF Model Estimation and Data Analysis*) in all conditions to the respective shorter trials of the multispeaker Condition 3, resulting in equal durations for both speakers. In addition, rare hardware buffer issues during the experiment led to additional noise in the stimulation for some participants. We excluded those trials from later analysis and randomly subsampled the same amount of trials for all other participants. In sum, 98 trials per condition were retained for further analysis.

Data acquisition and preprocessing

MEG data were simultaneously acquired alongside ocular data at a sampling frequency of 10 kHz (hardware filters: 0.1 - 3300 Hz) with a whole head system (102 magnetometers and 204 orthogonally placed planar gradiometers at 102 different positions; Elekta Neuromag Triux, Elekta Oy, Finland) that was placed within a standard passive magnetically shielded room (AK3b, Vacuumschmelze, Germany). For further data processing, a signal space separation algorithm implemented in the Maxfilter program (version 2.2.15) provided by the

MEG manufacturer was used to remove external noise and realign data from different blocks to a common standard head position. Afterwards, we preprocessed the data using Matlab and Fieldtrip. At first, 10 kHz data were resampled to 1000 Hz for further computations using the default implementation in FieldTrip (cutoff frequency = 500, kaiser window FIR filter, order: 200). Then, a bandpass filter between 0.1 - 40 Hz was applied (zero-phase FIR filter, order: 16500, hamming window). To remove ocular (horizontal and vertical) and cardiac artifacts, 50 components were identified from each experimental block using runica independent component analysis (ICA). Components originating from eye movements and heartbeat were then identified by visual inspection and removed. Artifact-free brain data was then cut into epochs from -1 to 4 s around stimulus (i.e. speech) onset and corrected for a 16 ms delay between trigger and stimulus onset generated by sound traveling through pneumatic headphones into the shielded MEG room. Eye-tracking data from both eyes were acquired at a sampling rate of 2 kHz using a Trackpixx3 binocular tracking system (Vpixx Technologies, Canada) with a 50 mm lens. Participants were seated in the MEG at a distance of 82 cm from the screen, with their chin resting on a chinrest to reduce head movements. Each experimental block started with a 13-point calibration and validation procedure that was then used throughout the block. Blinks and saccades were automatically detected by the Trackpixx3 system and excluded from horizontal and vertical eye movement data. Subsequently, data were preprocessed in Matlab R2020b. Position data from left and right eyes were averaged to increase the accuracy of gaze estimation (Cui & Hondzinski, 2006). We then converted data from pixel to visual angle in degrees. Gaps in the data due to blink and saccade removal were interpolated using a piecewise cubic Hermite interpolation. Artifact free gaze data was then imported into the FieldTrip Toolbox (Oostenveld et al., 2011), bandpass filtered between 0.1 - 40 Hz (zero-phase finite impulse response (FIR) filter, order: 33000, hamming window), resampled to 1000 Hz and cut into epochs from -1 to 4 s around stimulus (i.e. speech) onset. Finally, we corrected again for the 16 ms delay between trigger onset and actual stimulation.

Predictor Variables for TRF Models

Controls

We included control predictors for eye responses to visual (gabor) onsets according to the fixed 3 Hz presentation rate throughout the sentence and pure auditory (speech) onsets by adding intercepts (i.e. impulse trains) at respective timings.

Envelope

Both auditory predictors (Envelope & Acoustic Onsets) were based on gammatone spectrograms of the 200 natural-sounding speech sentences. Spectrograms were calculated over 256 frequencies, covering a range of 20 - 5000 Hz in equivalent rectangular bandwidth space (Heeris, 2018), resampled to 1000 Hz and scaled with exponent 0.6 (Biesmans et al., 2016) using Eelbrain (Brodbeck et al., 2021). The 1-Band Envelope was then derived by taking the sum of gammatone spectrograms across all frequency bands (Brodbeck et al., 2021), thus reflecting the broadband acoustic signal.

Acoustic Onsets

Additionally, we derived acoustic onsets by applying a neurally inspired auditory edge detection transformation to the gammatone spectrogram (Brodbeck et al., 2020) using the publicly available 'TRF-Tools'' (https://github.com/christianbrodbeck/TRF-Tools; Brodbeck, 2021) edge detection implementation for python, with default settings and saturation scaling factor of c = 30. Again, 1-Band Acoustic Onset representations were obtained by taking the sum across all frequency bands.

All predictors (see Fig. 1B) were resampled to 1000 Hz for subsequent alignment and analysis, to match with the sampling frequency of eye-tracking data.

Model Comparisons

In order to estimate ocular tracking of the speech *envelope* and *acoustic onsets*, we chose to include a control model (C) in the analysis that uses visual onsets and trial / speech onsets as predictors (see Fig. 1A), as they confounded the responses to the speech features of interest. Using the prediction accuracy of this control model as a basis, we then combined the control predictors with one of the speech features, leading to two combined models controlling for visual and trial onsets and entailing the speech envelope (CE) or acoustic onsets (CO) as predictors. In order to obtain the predictive power solely related to the speech features, we then subtracted the prediction accuracies of the control model from those of the combined models. This was done separately for every participant in each condition, resulting in a 'pure' prediction accuracy value $\mathbf{\Delta}r_s$ as an estimate of ocular speech tracking (see Fig. 1C).

TRF Model Estimation and Data Analysis

Prior to model computations, preprocessed eye-tracking and MEG data were temporally cut and aligned to the corresponding predictor variables (the blink-rate, and therefore the amount of samples that were interpolated for later analysis, was low: *M* = 5.00%, *SD* = 4.27%). Then, aligned trials were downsampled to 50 Hz for TRF model estimation after an antialiasing low-pass filter at 20 Hz was applied (zero-phase FIR filter, order: 662, hamming window). Impulse trains, i.e. control predictors, were then restored by adding "1s" at the nearest timepoints of original sampling rate onsets without applying any filters to avoid artifacts. We chose to downsample to 50 Hz as the most relevant power modulations of speech and attention do not exceed 20 Hz (i.e. 2 ½ * the sampling rate).

To further probe speech encoding in ocular activity under selective attention, we used a system identification technique called temporal response functions (TRF) as implemented in, and provided by, the open-source mTRF-Toolbox (Crosse et al., 2016, 2021) for Matlab. In short, TRFs pose time-resolved model weights to describe a stimulus - response relationship (forward / encoding models), e.g. how features of speech are transformed into responses at multiple time-lags. Whereas this technique is usually used to model neural responses, here we exploited this approach and applied TRF models on eye tracking data to investigate the relationship between speech features and eye movements (see Fig. 1B). In the present study, we used ridge regression, a regularized linear regression approach, at a time-lag window of -100 to 550 ms to compute encoding models, following a leave-one-trial-out cross-validation procedure to control for overfitting. This means we used all but one trial (of the 98 per condition) to estimate TRFs for a set of stimulus features (i.e. predictors) that were in turn applied to those of the left out test-trial to obtain a predicted ocular response. Prediction accuracy was then evaluated by calculating a Spearman's rank correlation between the originally measured, preprocessed response and the predicted response by the model (note that forward models make predictions independently for each channel, which here refers to horizontal and vertical eye movement 'channels' in eye-tracking data, as well as 102 magnetometers in neural data in the next section). Before model estimation, predictors and responses were scaled by their respective ℓ1 norms. After each trial had been the test-trial once, prediction accuracies were averaged over trials. This resulted in one correlation value r_s, for horizontal and vertical eye movements respectively, that describes how well the TRF model could predict an ocular response of a particular participant to speech in differing conditions of attention, which renders prediction accuracy a measure of ocular speech tracking (note that for statistical analysis we averaged prediction accuracies for horizontal and vertical eye movements once the difference between combined and control models were calculated, also see *Model Comparison*). The modeling procedure described above was carried out for every condition of selective attention to speech in the presented paradigm (see Fig. 1A), i.e. when a single speaker was the distractor in Condition 1, a single speaker was the target in Condition 2, a speaker was the target in a dual speaker mixture in Condition 3, a speaker of opposite sex was the distractor in a dual speaker mixture in Condition 3.

To further control for overfitting, ridge regression includes a regularization parameter λ that penalizes large model weights (for a detailed explanation of the ridge parameter, see Crosse et al., 2021). We empirically validated the optimal ridge parameter by using a nested cross-validation procedure (i.e. within every n-1 training set another leave-one-out cross-validation was used to obtain model results for different λ values). This procedure was carried out over a range of λ values of 10^{-5} - 10^{5} (in steps of 10^{1}) for each model (see *Model Comparisons*) over all participants and conditions. The final optimal lambda value for a certain model was then obtained by averaging the mean absolute error of the cross-validation over all trials, channels, conditions, and participants. We consequently chose the lambda value that led to the lowest mean absolute error. Based on this procedure, a single optimal lambda value of $\lambda = 10^{-3}$ was used to estimate speech encoding in ocular activity for all encoding models (see *Model Comparisons*) for all conditions of selective attention.

Following this analysis on ocular speech tracking, we established the behavioral relevance of this effect. As relevant dependent variables for the *Statistical Analysis,* we calculated individual intelligibility and effort scores from participants' behavioral responses. Intelligibility scores were calculated as the sum of all correct responses (i.e. a word on the response screen was correctly marked as *yes*, i.e. *heard,* during the 5-word sentence of a trial) divided by the number of all presented words in a condition (0% - 100%). Effort scores were calculated by averaging a participant's responses on the 5-point Likert scale at the end of each block per condition (1 = low effort, 5 = high effort). Task engagement scores were calculated in the same way as effort scores, and served as a control variable to rule out bias from different task engagements for conditions of selective attention (see *Supplementary Statistics*).

Mediation Analysis

To investigate whether the top-down control of eye movements could partially contribute to neural differences in selective attention to speech, we conducted an additional analysis based on the logic of a mediation analysis, adapted to our time-resolved regression analyses (i.e. encoding models). 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). Based on this assumption we can try to establish the different contributions in the triad relationship of speech, eye movements and neural activity (see Fig. 3A). A very well established finding states that speech acoustics can predict neural activity (e.g. Brodbeck et al., 2018; Di Liberto et al., 2015; Vanthornhout et al., 2018). Given our hypothesis that the speech envelope is encoded in ocular movements, we assume this finding to be mediated to some extent by ocular speech tracking. To test this assumption 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:

$$\textit{neural response} = \textit{TRF(c)} * \textit{speech envelope}$$

The direct (residual) effect (not mediated by eye movements) is obtained from a model including two predictors:

$$\textit{neural response} = \textit{TRF(c')} * \textit{speech envelope} + \textit{TRF(b)} * \textit{eye movements}$$

and represented in the exclusive weights (c’) of the former predictor (i.e. speech envelope). Note that the evaluation of the effect of the speech envelope on eye movements (termed “a” in Fig. 3A) preceded this analysis (see previous section). The two models described above were calculated to predict the neural responses of all 102 magnetometer channels separately. Subsequently, we used a cluster-based permutation dependent t-test to compare TRFs (using absolute values) of both models at each location (note that the polarity of neural responses is not of interest here). 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 (see Fig. 3A). Since no effect of ocular speech tracking was found for the distractor in a single speaker condition (Condition 1), we limited our TRF comparison to the speech envelope encoding of the target speaker in a single speaker (Condition 2), and the target and distractor speaker in a multi speaker context (Condition 3).

Statistical Analysis

To investigate acoustic speech tracking of eye movements under different conditions of attention, 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, the correlation between predicted eye movements from the combined TRF-models, including control predictors, acoustic features of interest (speech envelope and acoustic onsets), and measured eye movements was calculated (see Fig. 1B). We then subtracted the encoding results (i.e. correlation between predicted and measured eye movements) from a model which included only the control variables to isolate the effect of acoustic tracking from potential confounds. This difference was then averaged over horizontal and vertical channels and used as a dependent variable according to the Wilkinson notation (Wilkinson & Rogers, 1973):

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

$$\textit{acoustic onset tracking} \sim 0 + \textit{condition} + (1|\textit{subject})$$

Note that by removing the Intercept from the model, all conditions have been tested against a zero-effect of tracking.

To directly compare the tracking of the target and distractor speech in the multi speaker condition, a post-hoc model was calculated including only these two encoding results.

To additionally investigate whether ocular speech tracking is related to behavioral performance, we further included intelligibility and subjectively rated listening effort (see *TRF Model Estimation and Data Analysis*) into the model. Both variables were zero-centered by subtracting the median across subjects within condition before entering the model:

*envelope tracking ~ condition * intelligibility + condition * effort + (1|subject)*

*acoustic onset tracking ~ condition * intelligibility + condition * effort + (1|subject)*

Note that intelligibility was only probed for attended speech, therefore only these two conditions (multi vs. single speaker) were included in the behavioral model.

For all models 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. 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).

After establishing ocular speech tracking effects and their relations to behavior, we further quantified the extent to which this tracking contributes to speech encoding in ICA-cleaned neural responses. Using a mediation analysis approach, we compared model weights of the speech envelope for predicting neural responses with model weights from an encoding model, including eye movements as an additional predictor (see *Mediation Analysis*). To establish whether there is a significant difference we used a cluster-based randomization approach (Maris & Oostenveld, 2007) on all 102 magnetometers, averaging over time lags (from -50 to 500 ms to exclude possible regression edge artifacts). We computed the randomization distribution of t-values after 10000 permutations with a cluster alpha threshold of 0.05 that was then compared against the original contrast at an alpha level of 0.05, Bonferroni corrected. This procedure was carried out for three one-sided contrasts (see Fig. 3B), where we compared the plain (c) and direct (residual, c') effects (i.e. absolute TRF weights) for target speech in the single speaker condition (c' < c), for target speech in the multi speaker condition (c' < c), and finally where speech was the distractor in the multi speaker condition (c' < c). Subsequently, we report *p*-values of clusters and effect size Cohen's *d* as average over sensors within a cluster.

Data Visualization

Individual plots were generated in python (3.9.12) using matplotlib (Hunter, 2007), seaborn (Waskom, 2021), and mne-python (Gramfort et al., 2013). Plots were then arranged as cohesive figures with affinity designer (https://affinity.serif.com/en-us/designer/).

Data and Code Availability

Preprocessed Data and Code are available at the corresponding author's GitLab repository (https://gitlab.com/qubitron).

Results

Eye movements track prioritized acoustic features of natural speech

We first answered the question as to whether eye movements track natural speech, and how this tracking is modulated by selective attention. Participants listened to short sentences of natural speech in different conditions of selective attention. Sentences featured either a single speaker as a distractor to visual attention, a single speaker as target of the auditory modality, a target in a multi speaker condition, and consequently a distractor in a multi speaker condition (see Fig. 1A). Using TRFs (see Fig. 1B), a regularized linear regression approach, we evaluated ocular speech tracking based on the model's ability to predict held-out eye movement data in a nested cross-validation procedure. Eye-tracking and MEG data were simultaneously recorded. We subsequently provide evidence that ocular speech tracking prioritizes acoustic features of a target speaker, even in the presence of a simultaneously presented distractor. Bayesian multilevel models with prediction accuracies of ocular speech tracking as dependent variables revealed that eye movements only track the envelope of a single speaker when it was presented as the target of attention (β = 0.011, 94%HDI = [0.006, 0.015]), not when it served as a distractor to the visual modality (β = 0.002, 94%HDI = [-0.002, 0.006]). We observed a similar effect when using acoustic onsets as a predictor, indicating ocular tracking of the target (β = 0.014, 94%HDI = [0.010, 0.015]) but not the distractor sentences (β = 0.001, 94%HDI = [-0.004, 0.005]). For the multi speaker condition, direct post-hoc comparison between target and distractor speech revealed that speech envelope tracking (β = -0.006, 94%HDI = [-0.009, -0.002]) was weaker for the distractor speaker compared to the target speaker. The same comparison for acoustic onset tracking points towards a similar effect (β = -0.004, 94%HDI = [-0.007, -0.000]), however, as the HDI’s upper limit is zero, conclusions should be drawn more cautiously. Overall, the results show stronger ocular speech tracking for attended speech compared to ignored speech in a single speaker *and* in a multi speaker context (see Fig. 2A). They further indicate that in a multi speaker context, eye movements track a target speaker more strongly compared to a distractor speaker, with the effect being more pronounced for speech envelope tracking. A summary of the statistics can be found in *Supplementary Tables*, Table 1.

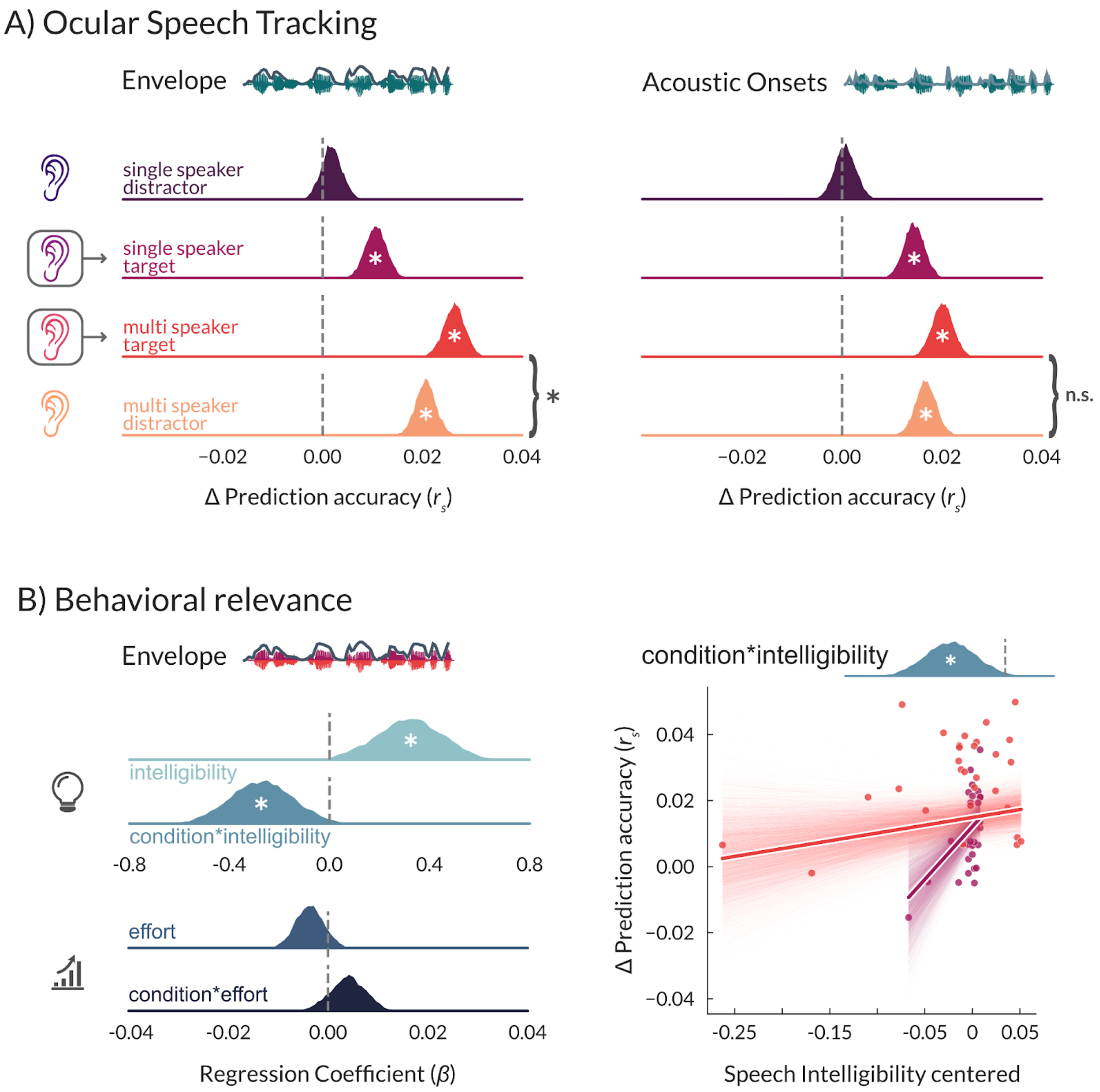

Fig. 2: *The effect of selective attention on ocular speech tracking and its relation to speech intelligibility and subjective listening effort.* **A)** Differences in prediction accuracies (**Δ*r*$_s$**) between models that additionally included the speech envelope and a control model indicate a significant tracking whenever speech was attended in a single speaker and multi speaker context (left panel). For acoustic onsets (right panel), we found evidence for ocular tracking of a target but not a distractor in a single speaker context. In a multi speaker context, no significant difference was found between the target and distractor speaker. **B)** Intelligibility was probed only for attended speech. We therefore used intelligibility and subjective listening effort scores for targeted speech in the single and multi speaker context to assess its relation to ocular speech tracking. While intelligibility was related to ocular speech envelope tracking with an interaction effect, subjective listening effort scores were not (left panel). The interaction effect (right panel) indicates a stronger influence of intelligibility on speech envelope tracking in the single speaker condition (shaded areas represent the 94%HDI, dots represent participants). 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). Curly brackets indicate post-hoc comparisons: ‘*’ = significant, ’n.s.’ = not significant. *N* = 30.

Ocular speech envelope tracking is related to intelligibility

In response to a second question, we addressed the behavioral relevance of ocular speech tracking in terms of intelligibility and subjectively perceived listening effort. As intelligibility was probed for attended speech, i.e. where it was the target in a single speaker and a multi speaker condition (see Fig. 1A), only these two conditions (multi vs. single speaker) were included into the Bayesian multilevel model. Again, prediction accuracies of ocular speech tracking were included as the dependent variable. We find a positive effect for intelligibility on the encoding of the speech envelope (β = 0.322, 94%HDI = [0.056, 0.594]), indicating that, in the single speaker condition, higher intelligibility is reflected in stronger ocular speech envelope tracking (see Fig. 2B). Additionally, we find a negative interaction for this effect with the condition of a multi vs. single speaker (β = -0.274, 94%HDI = [-0.531, -0.026]). This indicates a stronger influence of intelligibility on envelope tracking in the single speaker condition. No effect was found for intelligibility with regards to acoustic onset tracking (β = 0.172, 94%HDI = [-0.057, 0.404]). Similarly, subjectively perceived effort had no effect on neither ocular speech envelope (β = -0.004, 94%HDI = [-0.009, 0.002]) nor acoustic onset tracking (β = -0.003, 94%HDI = [-0.007, 0.002]). A summary of the statistics can be found in *Supplementary Tables*, Table 2.

Eye movements contribute to neural differences in selective attention to speech

Thirdly, we asked whether eye movements and auditory selective attention share neural processes. As it was shown by Popov et al. (2022), a partial contribution of goal-driven oculomotor activity to typical cognitive effects in spatial auditory attention was retained even after removing scalp signal variance (e.g. by means of independent component analysis, ICA) related to ocular muscle activity. Based on their findings, it is important to address this possible contribution when evaluating neural responses in selective attention to speech, especially since our design did not entail any spatial discriminability or cues. With a mediation analysis approach (see Fig. 3A), we evaluated the influence of eye movements on neural speech tracking with a cluster-based permutation test, contrasting the plain (c) against the direct (c') effects, i.e. model weights of the speech envelope for predicting neural responses on the one hand with model weights from an encoding model also including eye movements as an additional predictor on the other hand (see Fig. 3B, also see *Mediation Analysis* and *Statistical Analysis*). The tests revealed a significant difference (c' < c) for all three conditions with a bilateral topographic pattern suggestive of auditory processing areas. When speech was the target in a single speaker condition, eye movements contributed to neural speech processing mostly over left temporal sensors ($p < 0.05$, Cohen's $d < -0.46$). For both target and distractor speech in the multi speaker condition, this influence extended to left parietal ($p < 0.05$, Cohen's $d_{\text{target}} < -0.49$, Cohen's $d_{\text{distractor}} < -0.52$) and right temporal sensors ($p < 0.01$, Cohen's $d_{\text{target}} < -0.58$, Cohen's $d_{\text{distractor}} < -0.58$, see Fig. 3B).

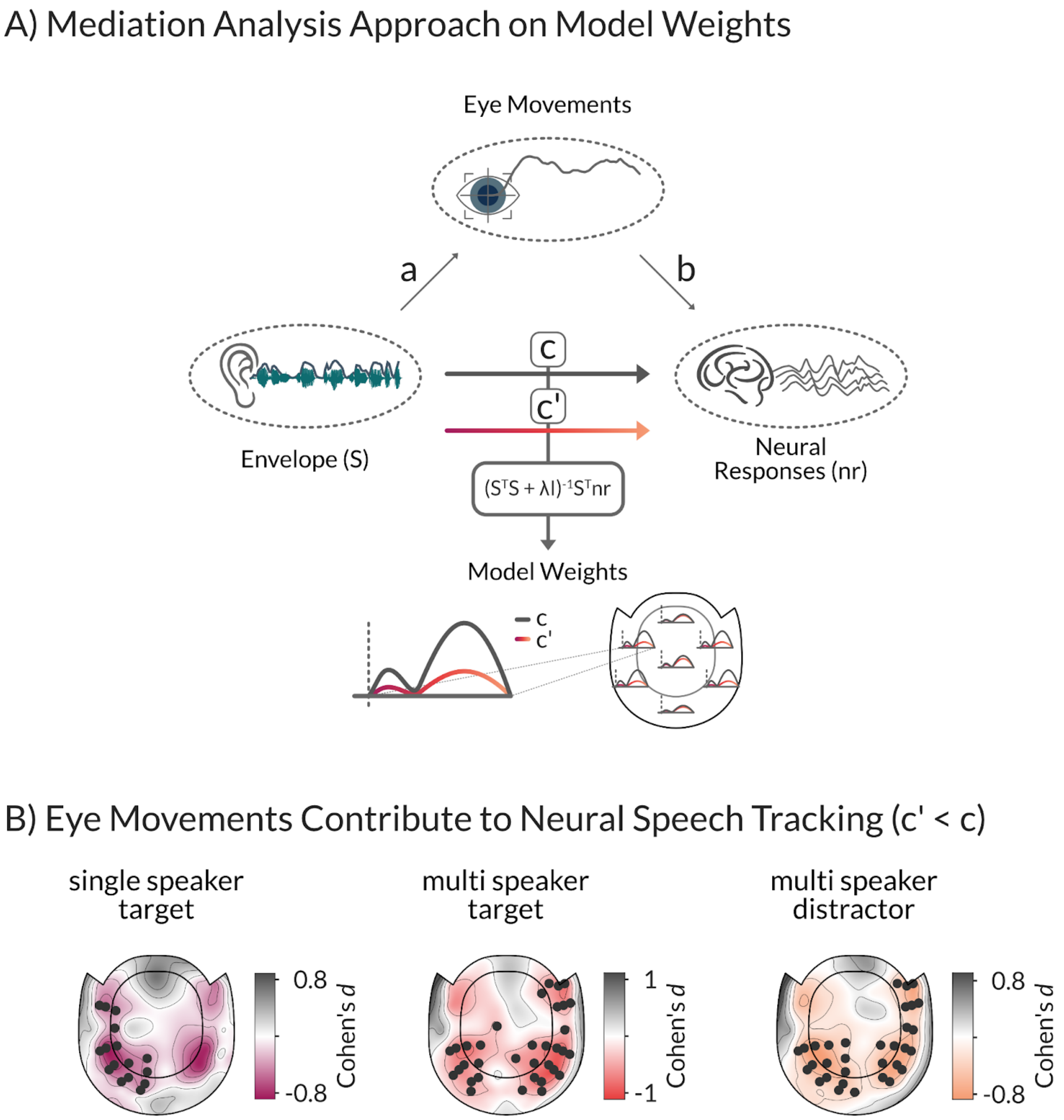

Fig. 3: *The framework to establish contributions of eye movements to neural speech tracking*. **A)** With a mediation analysis approach, we investigated the partial contribution of eye movements on the speech envelope encoding in neural responses. For this, we compared the plain effect (c) of the speech envelope on neural activity to its direct (residual) effect (c') by including an indirect effect via eye movements. The two models described above were calculated to predict neural responses of all 102 magnetometer channels separately. **B)** We used a cluster-based permutation dependent t-test to compare TRFs (using absolute values) of both models at each sensor and report effect size Cohen's *d* averaged over sensors within significant clusters. Contrasts (c' < c) revealed a small mediation effect by eye movements for the relationship between the speech envelope and neural responses in the single speaker condition over left temporal sensors ($p < 0.05$, Cohen's $d < -0.46$). Eye movements had a stronger and more widespread mediation effect in the multi speaker condition over left parietal ($p < 0.05$, Cohen's $d_{target} < -0.49$, Cohen's $d_{distractor} < -0.52$) and right temporal sensors ($p < 0.01$,

Cohen's d_{target} < -0.58, Cohen's $d_{distractor}$ < -0.58). Marked sensors in topographies belong to sensor clusters on the basis of which the null hypothesis of no difference was rejected. $N = 30$.

Discussion

Previous research established fundamental evidence for a joint network of attention and eye movements. In the auditory domain, several studies point towards interactions of the oculomotor system and selective processing. The generalizability and validity of such interactions in complex, naturalistic environments has, to date, not been quantified. Here, we aimed to establish a direct link between ocular movements, selective attention to speech, and adaptive behavior. We further investigated the contribution of this ocular speech tracking to underlying neural processes. Using the sampled signal of continuous horizontal and vertical eye gaze activity in combination with TRFs, we show that eye movements track prioritized auditory features (i.e. envelope and acoustic onsets) in selective attention to speech. Crucially, ocular speech envelope tracking differentiates between a simultaneously presented target and distractor speaker in the absence of any spatial discriminability and is further related to intelligibility. Moreover, using simultaneously recorded MEG data, we demonstrate that ocular speech envelope tracking contributes to the neural tracking effects of speech over sensors suggestive of auditory processing regions. Our findings provide new insights into the encoding of speech in a joint network of auditory selective attention and eye movement control, as well as their contribution to the neural representations of speech perception.

Potential Principles of Ocular Speech Tracking

Our results show that gaze activity continuously tracks acoustic features of attended natural speech. This high attentional selectivity of gaze in relation to speech provides new insights into the dynamics of a shared network processing auditory attention and eye movement control in humans. Hereafter, we discuss several implications for potential underlying principles.

Gaze and Prioritization of Spectrotemporal Acoustic Information

In complex, naturalistic environments, eye movements could aid the auditory system in unraveling the vastly overlapping spectrotemporal information that reaches the ears. Recent evidence in humans suggests that eye movements contribute to the computation of sound locations in relation to the visual scene at the very first stages of sound processing (Lovich et al., 2022; Murphy et al., 2022). Similar studies with monkeys and cats suggest a midbrain hub of inferior and superior colliculus (IC, SC) that affects auditory processing based on eye positions (e.g. Bulkin & Groh, 2012; Lee & Groh, 2012; Porter et al., 2007). Barn owls engage the IC to create auditory space maps based on frequency maps and interaural time and level differences, integrating visual maps with cohesive sensory space maps in the optic tectum (the avian homologue of the SC; Brainard & Knudsen, 1998; Pena & Gutfreund, 2014) under top-down gaze control (Winkowski & Knudsen, 2006). It has been shown that barn owls recalibrate sound localization based on vision during their development (Knudsen & Knudsen, 1989), suggesting a learned alignment of auditory and visual stimuli based on a common source. Natural gaze behavior under the control of auditory attention could thus play an important role in the alignment of visual and auditory space maps across species to navigate and interact with the environment, filtering and matching events or objects based on shared audiovisual spectrotemporal information. In humans, gaze activity could align to the acoustic features of attended speech to infer information about speaker identity and location (e.g. azimuth and distance) and match it with visual input. Speech, or verbal communication in general, has gained a central role as an advantageous survival strategy of social groups. Humans could exploit the ocular system during development to associate certain sound patterns with certain speakers, developing specific frequency and space maps, associating lip movements with sound and meaning, and ultimately guiding the development of speech in infants. Possibly, selective attention and gaze support the prioritization of relevant acoustic information already at the cochlea via learned association of activation patterns caused by a sound (e.g. a female compared to a male voice). Auditory frequency and space maps could aid the differentiation between speakers of opposite sex and support the temporal alignment of attention observed in stronger ocular speech tracking for a target in the multi speaker condition. This idea is supported by recent evidence in humans demonstrating a top-down modulation of the auditory periphery by selective attention (e.g. Gehmacher et al., 2022; Köhler et al., 2021; Köhler & Weisz, 2022) and could also explain why, in the current study, we observed the reported ocular speech tracking effects even without any meaningful visual information. Further studies could investigate potential effects / benefits of matching visual input (videos), e.g. lip movements, on the phenomenon.

Temporal Alignment and Predictive Processes

The idea of an active sampling strategy of spatiotemporal information is further supported by the temporal dynamics of ocular speech tracking. TRFs show a first, initial peak around zero lag (see *Supplementary Figures*, Fig. 2B and 3B), potentially indicative of a supportive mechanism of ocular speech tracking at the very first stages of sound processing to aid prioritization of overlapping spatiotemporal information. This would also align with the results on the contribution of ocular speech tracking to neural responses at auditory processing areas (see Fig. 3B). Such an immediate engagement of the ocular system would further suggest a complementary predictive processing account. Anticipation and accurate allocation of events in time have been found for language processing (Dikker & Pylkkänen, 2013) and eye movements in motion perception (e.g. Damasse et al., 2018; Kowler & Steinman, 1979a, 1979b; Pasturel et al., 2020). Predictive mechanisms should lead to a reduction in processing costs, thus interindividual differences in anticipatory TRF peaks could be related to subjectively perceived listening effort. However, we would like to point out that the presented study design limited the analysis and interpretation of TRFs due to its short 5-word sentence structure. For one, anticipatory effects could be biased by the highly predictable syntactic structure of the sentences. Presumably, this could also be the reason why we did not find a differentiation of target and distractor in a multi speaker context for acoustic onsets, as they are temporarily more correlated than the envelopes. Secondly, sentences were too short to allow for wider TRF windows that could give more detailed information about later dynamics > 500 ms where a clear differentiation of target and distractor in a dual speaker mixture seems to take place. Future studies should investigate the precise temporal dynamics of ocular speech tracking and potential predictive processes in continuous designs (e.g. with audiobooks).

Engagement of the Motor System

Ocular speech tracking could also relate more directly to motor theories of speech, suggesting that the motor system is engaged in support of speech perception (Galantucci et al., 2006; Liberman & Mattingly, 1985). Recent findings suggest a link between rates of eye movements during text reading and typical speech production / perception rates (Gagl et al., 2022). We observed a general right lateralized bias of gaze whenever the auditory modality (i.e. speech) was attended (see *Supplementary Figures*, Fig. 1). Since ocular speech tracking effects suggest a temporal alignment of this right lateralization with speech features, it could be argued that our eyes move with the speech streams as if the words were read as text. If this was the case, we would expect a shift of gaze towards the left side for cultures that read text from right to left. This would render ocular tracking specific to 1) humans, 2) speech, and 3) cultural context. Future studies should address this idea by applying similar designs 1) with animals (that also use verbal communication, e.g. birds), 2) tone sequences, 3) across cultures.

Visual Disengagement and Allocation of Processing Resources

Another explanation for the observed ocular speech tracking effects could be a general push-pull process of task dis- / engagement, i.e. a disengagement from the visual modality to free up processing resources for auditory information. Thus, whenever the task is to listen closely, or if listening becomes increasingly difficult, we move our eyes away to attenuate interfering visual input (note that complete eye closure seems to increase alpha modulations by auditory attention, but does not, however, improve listening behavior; Wöstmann et al., 2020. Also, internal attention in insight problem solving seems to relate to increased blinking and off-center gaze shifts; Salvi et al., 2015). Heat maps of eye gaze during stimulation periods (see *Supplementary Figures*, Fig. 1) show a slight gaze shift off-center as well as higher variance in conditions where participants had to attend the auditory modality. This supports the assumption of a more general disengagement from the visual modality for auditory processing, arguably to free up processing resources. However, participants also could shift their gaze slightly away from the distracting visual stimulus to free up resources for more precise evaluation of temporal speech features. This could support the processing of overlapping spatiotemporal features, especially in the multi speaker condition, since the visual stimulus was meaningless for auditory processing.

Taken together, we propose several potential principles of ocular speech tracking that need to be evaluated in greater detail by future research. Continuous speech designs (e.g. audiobooks) could be utilized to replicate the present findings and further investigate the precise temporal dynamics of the reported effects. In turn, potential interactions with predictive processes as well as interactions with behavioral markers like effort and intelligibility could be quantified. Neurophysiological evidence in animals for similar interactions of eye movements and auditory processing further urge the question whether ocular speech tracking displays a learned association in beings with complex verbal communication structures like humans, or whether it represents a general ocular tracking of the acoustic environment as adaptive behavior across species. It will thus be important for future research to identify the underlying mechanisms of the observed effects. Potential links to alpha modulations (Liu, Nobre, & Ede, 2022; Liu, Nobre, & van Ede, 2022; Popov et al., 2021, 2022) or arousal states with regards to (sub)cortical neurotransmitter-dependent modulations (Schuerman et al., 2022) could be investigated.

Ocular Speech Tracking and its relation to adaptive behavior

We did not find a relationship between ocular speech tracking and subjectively perceived effort, which questions the previous assumption of visual disengagement for reallocation of processing resources. Either effort is not reflected in the engagement of the ocular system in speech envelope tracking, or the measure of effort was not sensitive enough (alternatively, neural measures of effort like alpha modulations (e.g. Haegens et al., 2014; Wöstmann et al., 2017) or pupillometry could be used to estimate effort in more detail). Also, no difference in general task engagement was found (see *Supplementary Statistics*). Instead, we found that ocular speech envelope tracking is related to intelligibility, with a stronger influence of intelligibility on envelope tracking in the single speaker condition. Our results are supported by findings on increased intelligibility for a spatially discriminable target speaker in a multi speaker mixture when the eyes point towards the target location (Best et al., 2020). It therefore seems likely that eye movements and intelligibility of speech are, in fact, related. Taking the possible interpretations for the ocular speech tracking effects into consideration, 1) improved prioritization of relevant spatiotemporal information could improve intelligibility of a target speaker in the multispeaker condition, and is 2) further supported by predictive processes. 3) An engagement of the motor system could support phonological transformation processes to increase speech perception, and 4) disengagement from the visual modality could free resources for speech processing to improve intelligibility. Further research on the topic is needed to get a better understanding of the interaction between ocular speech tracking and intelligibility, also on a neural level.

Contributions of Eye Movements to Neural Speech Tracking

As a final step, we investigated the potential contribution of eye movements to neural differences typically observed in selective attention to speech. Our assumption that ocular speech tracking and selective attention to speech share underlying neural computations was based on recent findings by Popov et al. (2022) who demonstrated a partial contribution of goal-driven oculomotor activity to typical cognitive effects in spatial auditory attention. It was therefore important to address a possible contribution in the present work since we previously established the effects of ocular speech tracking. Using a mediation analysis approach (see *TRF Model Estimation and Data Analysis*), we show that eye movements contribute to neural activity over sensors indicative of auditory processing areas. We find a cluster over left temporal sensors in a single speaker context that extends to clusters over parietal and right temporal sensors in a multi speaker context. We thus observe a general contribution to auditory processing areas, which gains importance when considering recent evidence on the ocular modulation of neural excitability of cortical auditory regions (Leszczynski et al., 2022). To this end, we would like to specify that this exploratory analysis needs to be confirmed by future studies in a more detailed and methodologically tailored manner. Following studies that solely use continuous speech could focus on relating eye movements and alpha oscillations in selective attention to speech and establishing concrete evidence on the temporal dynamics of this interaction on a source level. As we used a multisensory single-trial design, we believe a continuous unisensory approach would be more suitable for this kind of analysis. Here, we provide a first step towards this direction, highlighting a contribution on a sensor level that needs to be taken into consideration in future research on auditory cognition.

Conclusion

The present report establishes a hitherto unknown phenomenon, *ocular speech tracking*, which enables the monitoring of prioritized auditory features in selective attention to natural speech. Crucially, ocular speech envelope tracking is stronger for a target compared to a distractor in a multi speaker condition. Furthermore, ocular speech envelope tracking is related to intelligibility. Moreover, our results suggest a contribution of oculomotor activity to neural responses in speech processing that needs to be taken into consideration in future research on auditory cognition. The present work offers valuable new research directions towards the neurobiological mechanisms of the phenomenon, its dependence on learning and plasticity, as well as its functional implications in social communication.

Acknowledgements

Q.G., J.S., and P.R. are supported by the Austrian Science Fund (FWF; Doctoral College "Imaging the Mind"; W 1233-B). Q.G. and P.R. are 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 measurements. Special thanks to Claudia Contadini-Wright for proofreading the manuscript.

Author Contributions

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

Conflicts of interest

The authors declare no competing interests.

References

Astafiev, S. V., Shulman, G. L., Stanley, C. M., Snyder, A. Z., Essen, D. C. V., & Corbetta, M. (2003). Functional Organization of Human Intraparietal and Frontal Cortex for Attending, Looking, and Pointing. *Journal of Neuroscience*, *23*(11), 4689–4699.

https://doi.org/10.1523/JNEUROSCI.23-11-04689.2003

Best, V., Jennings, T. R., & Kidd Jr, G. (2020). An effect of eye position in cocktail party listening. *Proceedings of Meetings on Acoustics 179ASA*, *42*(1), 050001.

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2016). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(5), 402–412.

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

Brainard, M. S., & Knudsen, E. I. (1998). Sensitive Periods for Visual Calibration of the Auditory Space Map in the Barn Owl Optic Tectum. *The Journal of Neuroscience*, *18*(10), 3929–3942. https://doi.org/10.1523/JNEUROSCI.18-10-03929.1998

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*.

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLoS Biology*, *18*(10), e3000883.

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, *172*, 162–174. https://doi.org/10.1016/j.neuroimage.2018.01.042

Bulkin, D. A., & Groh, J. M. (2012). Distribution of eye position information in the monkey inferior colliculus. *Journal of Neurophysiology*, *107*(3), 785–795.

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*.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., Linenweber, M. R., Petersen, S. E., Raichle, M. E., Van Essen, D. C., & Shulman, G. L. (1998). A Common Network of Functional Areas for Attention and Eye Movements. *Neuron*, *21*(4), 761–773. https://doi.org/10.1016/S0896-6273(00)80593-0

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, *58*(3), 306–324. https://doi.org/10.1016/j.neuron.2008.04.017

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), Article 3. https://doi.org/10.1038/nrn755

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

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

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

Damasse, J.-B., Perrinet, L. U., Madelain, L., & Montagnini, A. (2018). Reinforcement effects in anticipatory smooth eye movements. *Journal of Vision*, *18*(11), 14. https://doi.org/10.1167/18.11.14

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*(1), 55–64. https://doi.org/10.1016/j.bandl.2012.08.004

Gagl, B., Gregorova, K., Golch, J., Hawelka, S., Sassenhagen, J., Tavano, A., Poeppel, D.,

& Fiebach, C. J. (2022). Eye movements during text reading align with the rate of speech production. *Nature Human Behaviour*, *6*(3), Article 3. https://doi.org/10.1038/s41562-021-01215-4

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

Gehmacher, Q., Reisinger, P., Hartmann, T., Keintzel, T., Rösch, S., Schwarz, K., & Weisz, N. (2022). Direct Cochlear Recordings in Humans Show a Theta Rhythmic Modulation of Auditory Nerve Activity by Selective Attention. *Journal of Neuroscience*, *42*(7), 1343–1351. https://doi.org/10.1523/JNEUROSCI.0665-21.2021

Getzmann, S. (2002). The effect of eye position and background noise on vertical sound localization. *Hearing Research*, *169*(1–2), 130–139.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., & Parkkonen, L. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 267.

Haegens, S., Cousijn, H., Wallis, G., Harrison, P. J., & Nobre, A. C. (2014). Inter- and intra-individual variability in alpha peak frequency. *NeuroImage*, *92*, 46–55. https://doi.org/10.1016/j.neuroimage.2014.01.049

Hartmann, T., & Weisz, N. (2020). An introduction to the objective psychophysics toolbox. *Frontiers in Psychology*, *11*, 585437.

Heeris, J. (2018). *Gammatone Filterbank Toolkit*.

Holtze, B., Rosenkranz, M., Bleichner, M. G., & Debener, S. (2022). *Eye-blink patterns reflect attention to continuous speech*. PsyArXiv. https://doi.org/10.31234/osf.io/n86yp

Hopfinger, J. B., Buonocore, M. H., & Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature Neuroscience*, *3*(3), Article 3. https://doi.org/10.1038/72999

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(03), 90–95.

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

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

Knudsen, E. I., & Knudsen, P. F. (1989). Vision calibrates sound localization in developing barn owls. *Journal of Neuroscience*, *9*(9), 3306–3313. https://doi.org/10.1523/JNEUROSCI.09-09-03306.1989

Köhler, M. H. A., Demarchi, G., & Weisz, N. (2021). Cochlear activity in silent cue-target intervals shows a theta-rhythmic pattern and is correlated to attentional alpha and theta modulations. *BMC Biology*, *19*(1), 48. https://doi.org/10.1186/s12915-021-00992-8

Köhler, M. H. A., & Weisz, N. (2022). *Cochlear theta activity oscillates in phase opposition during interaural attention* (p. 2022.02.21.481289). bioRxiv. https://doi.org/10.1101/2022.02.21.481289

Kowler, E., & Steinman, R. M. (1979a). The effect of expectations on slow oculomotor control—I. Periodic target steps. *Vision Research*, *19*(6), 619–632. https://doi.org/10.1016/0042-6989(79)90238-4

Kowler, E., & Steinman, R. M. (1979b). The effect of expectations on slow oculomotor control—II. Single target displacements. *Vision Research*, *19*(6), 633–646. https://doi.org/10.1016/0042-6989(79)90239-6

Lee, J., & Groh, J. M. (2012). Auditory signals evolve from hybrid-to eye-centered coordinates in the primate superior colliculus. *Journal of Neurophysiology*, *108*(1), 227–242.

Leszczynski, M., Bickel, S., Nentwich, M., Russ, B. E., Parra, L., Lakatos, P., Mehta, A., & Schroeder, C. E. (2022). *Saccadic modulation of neural excitability in auditory areas of the neocortex* (p. 2022.05.24.493336). bioRxiv. https://doi.org/10.1101/2022.05.24.493336

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised.

Cognition, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Liu, B., Nobre, A. C., & Ede, F. van. (2022). *Microsaccades transiently lateralise EEG alpha activity* (p. 2022.09.02.506318). bioRxiv. https://doi.org/10.1101/2022.09.02.506318

Liu, B., Nobre, A. C., & van Ede, F. (2022). Functional but not obligatory link between microsaccades and neural modulation by covert spatial attention. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31217-3

Lovich, S. N., King, C. D., Murphy, D. L., Landrum, R., Shera, C. A., & Groh, J. M. (2022). *Parametric information about eye movements is sent to the ears* (p. 2022.11.27.518089). bioRxiv. https://doi.org/10.1101/2022.11.27.518089

Luo, T. Z., & Maunsell, J. H. R. (2019). Attention can be subdivided into neurobiological components corresponding to distinct behavioral effects. *Proceedings of the National Academy of Sciences*, *116*(52), 26187–26194. https://doi.org/10.1073/pnas.1902286116

Maddox, R. K., Pospisil, D. A., Stecker, G. C., & Lee, A. K. (2014). Directing eye gaze enhances auditory spatial cue discrimination. *Current Biology*, *24*(7), 748–752.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Murphy, D. L., King, C. D., Lovich, S. N., Landrum, R. E., Shera, C. A., & Groh, J. M. (2022). *Evidence for a system in the auditory periphery that may contribute to linking sounds and images in space* (p. 2020.07.19.210864). bioRxiv. https://doi.org/10.1101/2020.07.19.210864

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*.

Pasturel, C., Montagnini, A., & Perrinet, L. U. (2020). Humans adapt their anticipatory eye movements to the volatility of visual motion properties. *PLOS Computational Biology*, *16*(4), e1007438. https://doi.org/10.1371/journal.pcbi.1007438

Pena, J. L., & Gutfreund, Y. (2014). New perspectives on the owl's map of auditory space. *Current Opinion in Neurobiology*, *0*, 55–62. https://doi.org/10.1016/j.conb.2013.08.008

Pomper, U., & Chait, M. (2017). The impact of visual gaze direction on auditory object tracking. *Scientific Reports*, *7*(1), 1–16.

Popov, T., Gips, B., Weisz, N., & Jensen, O. (2022). Brain areas associated with visual spatial attention display topographic organization during auditory spatial attention. *Cerebral Cortex*, bhac285. https://doi.org/10.1093/cercor/bhac285

Popov, T., Miller, G. A., Rockstroh, B., Jensen, O., & Langer, N. (2021). *Alpha oscillations link action to cognition: An oculomotor account of the brain's dominant rhythm* (p. 2021.09.24.461634). bioRxiv. https://doi.org/10.1101/2021.09.24.461634

Porter, K. K., Metzger, R. R., & Groh, J. M. (2007). Visual- and saccade-related signals in the primate inferior colliculus. *Proceedings of the National Academy of Sciences*, *104*(45), 17855–17860. https://doi.org/10.1073/pnas.0706249104

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

Salvi, C., Bricolo, E., Franconeri, S. L., Kounios, J., & Beeman, M. (2015). Sudden insight is associated with shutting out visual inputs. *Psychonomic Bulletin & Review*, *22*(6), 1814–1819. https://doi.org/10.3758/s13423-015-0845-0

Schmidt, F., Demarchi, G., Geyer, F., & Weisz, N. (2020). A backward encoding approach to recover subcortical auditory activity. *NeuroImage*, *218*, 116961.

Schubert, J., Schmidt, F., Gehmacher, Q., Bresgen, A., & Weisz, N. (2022). Individual prediction tendencies facilitate cortical speech tracking. *BioRxiv*.

Schuerman, W. L., Chandrasekaran, B., & Leonard, M. K. (2022). Arousal States as a Key Source of Variability in Speech Perception and Learning. *Languages*, *7*(1), Article 1. https://doi.org/10.3390/languages7010019

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *Journal of* *the Association for Research in Otolaryngology*, *19*(2), 181–191. https://doi.org/10.1007/s10162-018-0654-z

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

Wagener, K., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache. I-III: Design, Optimierung und Evaluation des Oldenburger Satztests (Development and evaluation of a sentence test for the German language. I-III: Design, optimization and evaluation of the Oldenburg sentence test). *Zeitschrift Für Audiologie (Audiological Acoustics)*, *38*, 4–15.

Wardak, C., Ibos, G., Duhamel, J.-R., & Olivier, E. (2006). Contribution of the Monkey Frontal Eye Field to Covert Visual Attention. *Journal of Neuroscience*, *26*(16), 4228–4235. https://doi.org/10.1523/JNEUROSCI.3336-05.2006

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *22*(3), 392–399.

Winkowski, D. E., & Knudsen, E. I. (2006). Top-down gain control of the auditory space map by gaze control circuitry in the barn owl. *Nature*, *439*(7074), Article 7074. https://doi.org/10.1038/nature04411

Wöstmann, M., Lim, S.-J., & Obleser, J. (2017). The Human Neural Alpha Response to Speech is a Proxy of Attentional Control. *Cerebral Cortex*, *27*(6), 3307–3317. https://doi.org/10.1093/cercor/bhx074

Wöstmann, M., Schmitt, L.-M., & Obleser, J. (2020). Does Closing the Eyes Enhance Auditory Attention? Eye Closure Increases Attentional Alpha-Power Modulation but Not Listening Performance. *Journal of Cognitive Neuroscience*, *32*(2), 212–225. https://doi.org/10.1162/jocn_a_01403

Supplementary Information

Supplementary Tables

Table 1: Model summary statistics for encoding of acoustic features depending on condition

	speech envelope				acoustic onsets			
	b	*sd*	hdi 3%	hdi 97%	*b*	*sd*	hdi 3%	hdi 97%
single speaker - distractor	0.002	0.002	-0.002	0.006	0.001	0.002	-0.004	0.005
single speaker - target	0.011	0.002	0.006	0.015	0.014	0.002	0.010	0.019
multi speaker - target	0.026	0.002	0.022	0.031	0.020	0.002	0.016	0.025
multi speaker - distractor	0.021	0.002	0.017	0.025	0.017	0.002	0.013	0.021

Note: Dependent Variable = encoding results: encoding model - control model (average over channels)

Table 2: Model summary statistics for encoding of acoustic features depending on behavioral performance

	speech envelope				acoustic onsets			
	b	*sd*	hdi 3%	hdi 97%	*b*	*sd*	hdi 3%	hdi 97%
Intercept (single speaker)	0.012	0.002	0.008	0.017	0.015	0.002	0.011	0.020
Condition (multi vs. single speaker)	0.015	0.003	0.010	0.020	0.005	0.002	0.001	0.009
Intelligibility (single speaker)	0.322	0.142	0.056	0.594	0.172	0.123	-0.057	0.404
Intelligibility x Condition	-0.274	0.135	-0.531	-0.026	-0.125	0.115	-0.348	0.084
Effort (single speaker)	-0.004	0.003	-0.009	0.002	-0.003	0.002	-0.007	0.002
Effort x Condition	0.004	0.004	-0.003	0.011	-0.001	0.003	-0.007	0.005

Note: Dependent Variable = encoding results: encoding model - control model (average over channels)

Supplementary Figures

A) Gaze statistics

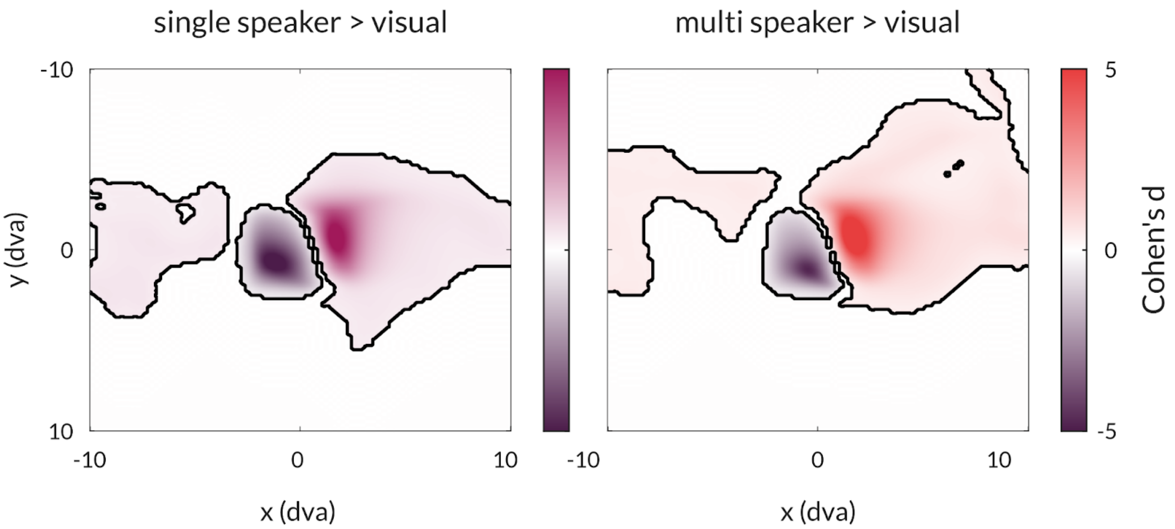

Fig. 1: *Cluster-based permutation tests on eye gaze during speech presentation.* We observe a slight shift of gaze to the top-right and more distributed gaze patterns whenever the auditory modality is attended. Heat maps illustrate gaze positions along the horizontal (x) and vertical (y) plane in degrees of visual angle (dva). Note that the area of the presented gabor patch on the screen, which we instructed participants to keep their gaze on during sentence presentations, was about 8 dva.

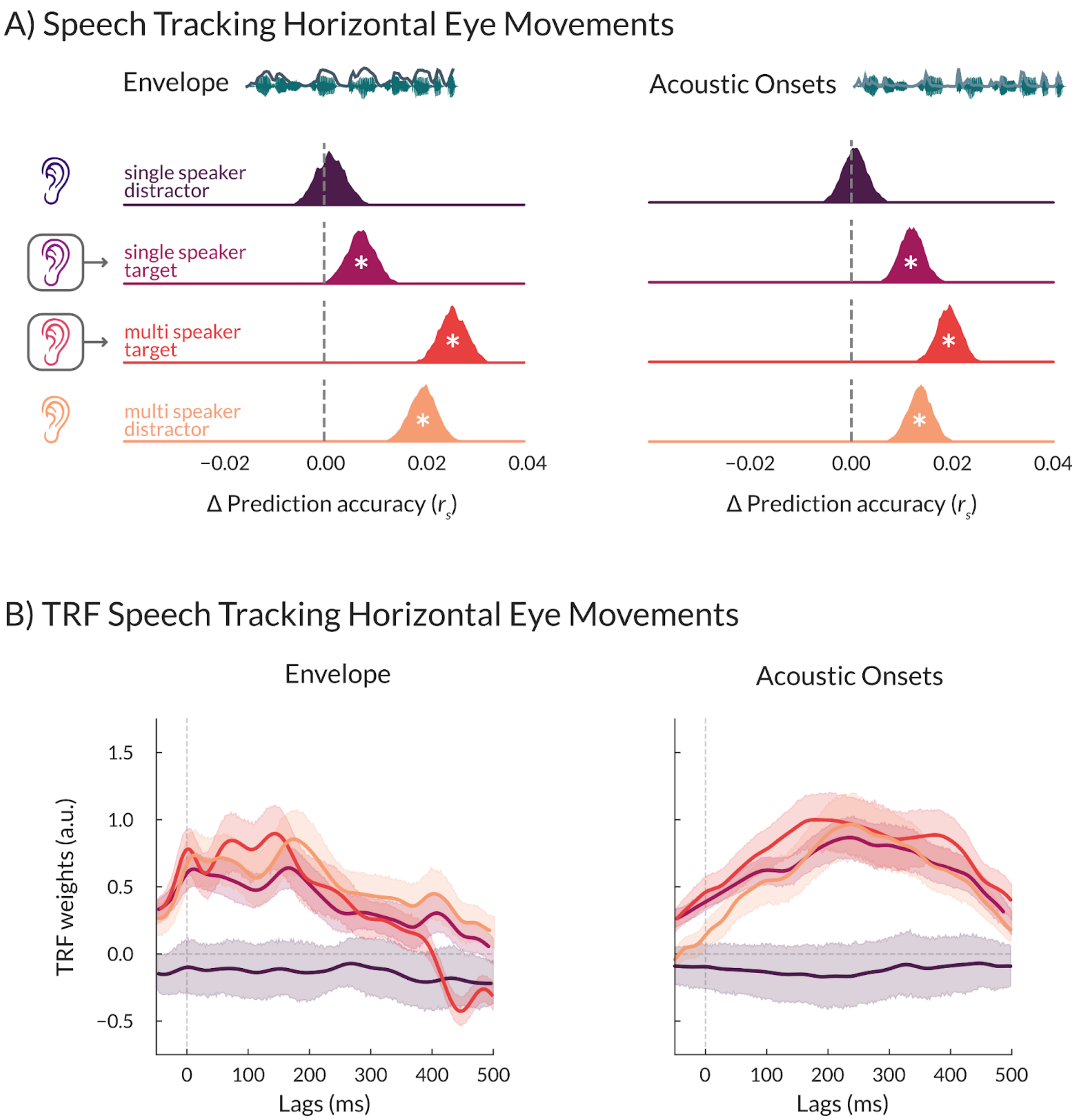

Fig. 2: *The effect of selective attention on speech tracking by horizontal eye movements.* **A)** Differences in prediction accuracies (**Δ**r_s) between models that additionally included the speech envelope and a control model for envelope (left panel) and acoustic onsets tracking (right panel) by horizontal 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). **B)** The temporal response functions (TRF) for speech envelope (left panel) and acoustic onsets tracking (right panel). TRFs were resampled to 500 Hz for visualization. Shaded areas represent 95% confidence intervals. N = 30.

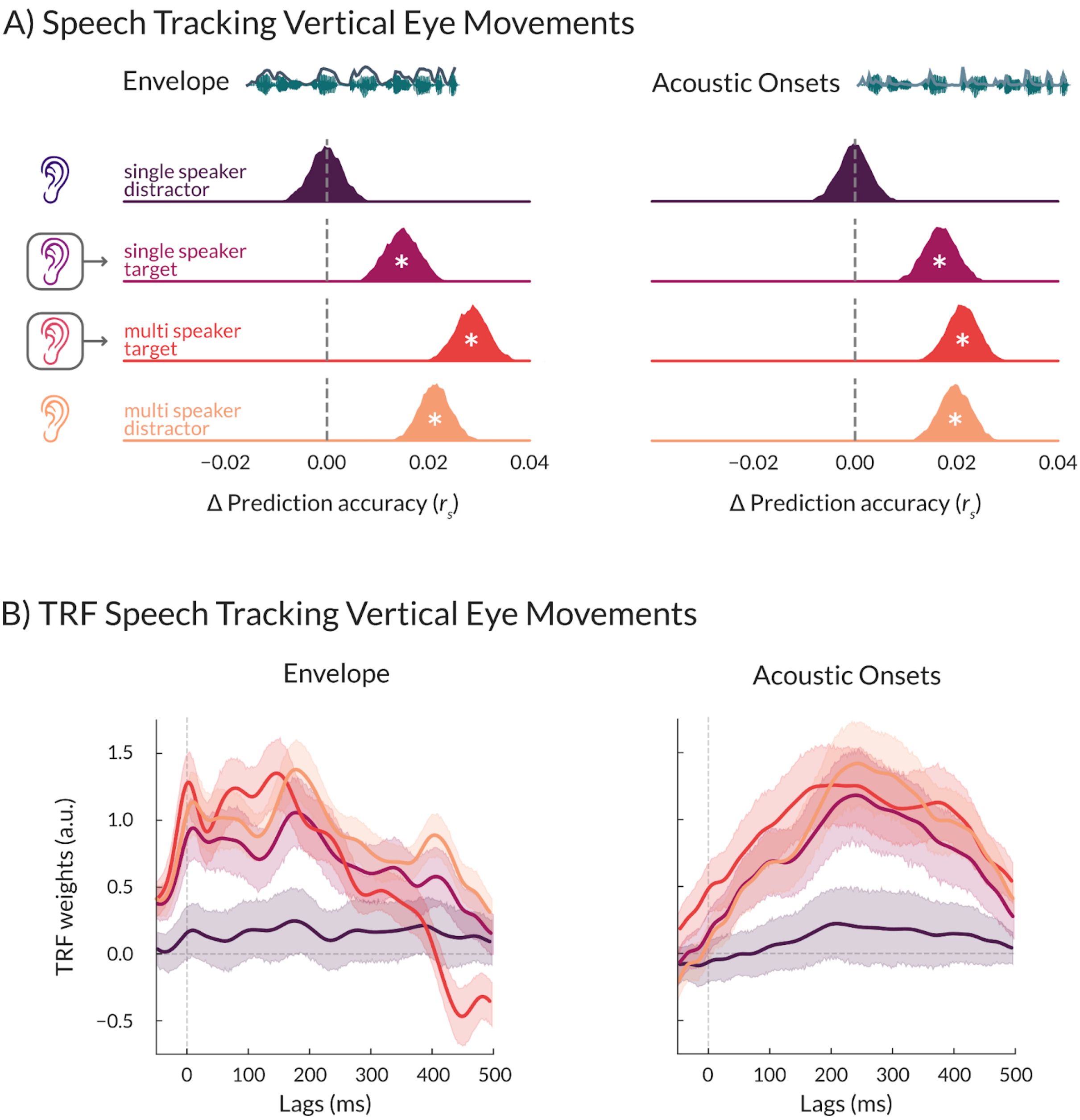

Fig. 3: *The effect of selective attention on speech tracking by vertical eye movements.* **A)** Differences in prediction accuracies (**Δ**r_s) between models that additionally included the speech envelope and a control model for envelope (left panel) and acoustic onsets tracking (right panel) by vertical 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). **B)** The temporal response functions (TRF) for speech envelope (left panel) and acoustic onsets tracking (right panel). TRFs were resampled to 500 Hz for visualization. Shaded areas represent 95% confidence intervals. *N* = 30.

Supplementary Statistics

To investigate whether subjective ratings of listening effort and task engagement differed between conditions, we calculated two additional models including only the attended speech conditions (multi vs. single speaker):

effort ~ condition + (1|subject)

engagement ~ condition + (1|subject)

Listening effort was rated higher in the multispeaker condition compared to the single speaker condition (*b* = 1.709, 94%HDI = [1.350, 2.074]). However, we found no difference in task engagement between the two conditions (*b* = 0.085, 94%HDI = [-0.122, 0.294]).