Attentional engagement with target and distractor streams predicts speech comprehension in multitalker environments

Alice Vivien Barchet [1,2,3*], Andrea Bruera [1], Jasmin Wend [1], Johanna M. Rimmele [4], Jonas Obleser [5,6], and Gesa Hartwigsen [1,2]

[1]Research Group Cognition and Plasticity, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[2]Cognitive and Biological Psychology, Leipzig University, Leipzig, Germany

[3]International Max Planck Research School on Cognitive NeuroImaging (IMPRS CoNI), Leipzig, Germany

[4]Department of Cognitive Neuropsychology, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

[5]Department of Psychology, University of Lübeck, Lübeck, Germany

[6]Center of Brain, Behavior, and Metabolism, University of Lübeck, Lübeck, Germany

[*]Corresponding Author: barchet@cbs.mpg.de

Abstract

Understanding speech while ignoring competing speech streams in the surrounding environment is challenging. Previous studies have demonstrated that attention shapes the neural representation of speech features. Attended streams are typically represented more strongly than unattended ones, suggesting either enhancement of the attended or suppression of the unattended stream. However, it is unclear how these complementary processes support attentional filtering and speech comprehension on different hierarchical levels. In this study, we used multivariate temporal response functions to analyze the EEG signals of 43 young adults, examining the relationship between the neural tracking of acoustic and higher-level linguistic features and a fine-grained speech comprehension measure. We show that the neural tracking of word and phoneme onsets and word-level linguistic features in the attended stream predicted comprehension at the individual single-trial level. Moreover, acoustic tracking of the ignored speech stream was positively correlated with comprehension performance, whereas word level linguistic neural tracking of the ignored stream was negatively correlated with comprehension. Collectively, our results suggest that attentional filtering during speech comprehension requires target enhancement as well as distractor suppression at different hierarchical levels.

Significance Statement

In social settings, speech comprehension is often challenged by the presence of multiple speakers talking simultaneously. The ability to focus on a relevant stream while ignoring irrelevant speech information in the background is crucial for successful and efficient interpersonal interactions. However, the precise neural mechanisms underlying this selective filtering process remain unclear. We establish the interplay of acoustic and higher-level information as objective markers of attentional selection and comprehension success.

Introduction

In everyday life, spoken speech streams are often obscured by background noise or competing speech streams, resulting in complex listening conditions. To successfully manage these challenges, listeners rely on information from multiple hierarchical levels to filter the relevant signals and facilitate comprehension. In this EEG study, we examined how the processing of acoustic and higher-level linguistic features of both attended and unattended speech streams contributes to speech comprehension in a multitalker environment.

Speech is hierarchically organized into several interacting levels of information, which have been shown to be represented in the human brain (Berwick et al., 2013; Ding et al., 2016). This organization appears to be implemented by low-frequency cortical activity that temporally aligns with the incoming speech signal (Henry & Obleser, 2012; Obleser & Kayser, 2019; Schroeder et al., 2010). This phenomenon, referred to as neural tracking, has been demonstrated across several hierarchical levels, including acoustic, phonemic, and semantic information (Brodbeck et al., 2018; Broderick et al., 2018; Donhauser & Baillet, 2020; Gillis et al., 2021; Heilbron et al., 2022).

Successful comprehension in challenging listening environments relies on selective attention. Selective attention is required when multiple stimuli compete for further cognitive processing (Moore & Zirnsak, 2017). Thus, attentional filtering is needed to select the relevant stream. This selection can be implemented using two mechanisms: target enhancement and distractor suppression (Orf et al., 2023). In competing speech comprehension, attention has been shown to selectively enhance the neural tracking of acoustic properties of the attended speech signal (Ding & Simon, 2012; Golumbic et al., 2013; Horton et al., 2014; Mesgarani & Chang, 2012; O'sullivan et al., 2015). However, the presence of acoustic neural tracking alone does not indicate comprehension. Acoustic neural tracking has been observed for non-speech sounds, as well as unintelligible speech (Hämäläinen et al., 2012; Karunathilake et al., 2023; Steinschneider et al., 2013; Tierney & Kraus, 2015). Instead, the neural tracking of higher-level linguistic information has been suggested to be a neural marker of speech comprehension (Gillis et al., 2021; Verschueren et al., 2022).

This is supported by evidence showing that the neural representation of higher level speech features is typically absent for unattended speech streams (Brodbeck et al., 2018; Broderick et al., 2018). Previous research has indicated that the attentional selection occurring in competing speech situations relies on higher-order, top-down mechanisms (Golumbic et al., 2013; J. M. Rimmele et al., 2015). Thus, higher level distractor suppression could be an important mechanism supporting the attentional selection in competing speech comprehension. Neural tracking of acoustic information has been observed for ignored speech streams (Ding & Simon, 2012; Golumbic et al., 2013; Horton et al., 2014; Kaufman & Golumbic, 2023; O'sullivan et al., 2015; J. M. Rimmele et al., 2015). However, the extent of neural tracking of acoustic information for unattended speech has been shown to be unrelated (Orf et al., 2023) or even positively related to speech comprehension performance (Fiedler et al., 2019). The role of distractor acoustic tracking in supporting attentional selection therefore remains unclear.

Direct evidence for the interplay of lower- and higher-level neural representations of attended and unattended speech streams in predicting speech comprehension is lacking. Many previous studies used continuous listening paradigms and were unable to capture rapid variations in comprehension success due to the lack of detailed behavioral responses (Brodbeck et al., 2018; Broderick et al., 2018; Gillis et al., 2021; Verschueren et al., 2022). Especially in multitalker situations, speech comprehension varies substantially on a word-by-word basis, due to amplitude fluctuations in the attended and unattended streams. This enables listeners to grasp small segments of speech, even in highly unfavorable signal-to-noise ratios (Cooke, 2006; MacPherson & Akeroyd, 2014). In this study, we aimed to uncover the neural mechanisms underlying comprehension by employing a fine-grained, temporally resolved measure of comprehension (i.e., sentence repetition). This approach provides deeper insights into the dynamics of speech comprehension. Additionally, we quantified acoustic and linguistic speech representations for attended and unattended speech streams in a fine-grained fashion. To derive linguistic representations, we relied on large language models which quantify the linguistic predictability of each word given the previous sentence context.

Given the previous evidence, we expected that target enhancement, as indexed by enhanced processing of the attended speech stream at acoustic and linguistic levels, would be positively cor-

related with comprehension performance. For the unattended stream, we expected that distractor suppression at the acoustic level should not be related to comprehension performance (see Orf et al., 2023). In contrast, we expected that distraction by the unattended stream would manifest in an increased higher-level tracking of the unattended stream. Therefore, linguistic tracking of the distractor stream was expected to be negatively related to comprehension performance.

Materials and Methods

Participants

The sample included N = 43 right handed German native speakers aged 19 to 51 (M = 33.86, SD = 10.57; 24 women). Participants had normal hearing and no history of neurological or psychiatric disorders. Normal hearing abilities were confirmed by a pure tone audiometry prior to the experiment. Normal hearing was defined as having a mean pure tone audiometric threshold of below 25 dB across a frequency range of 250 to 8000 Hz in at least one ear. The mean pure tone threshold for each participants' better ear in our sample was M = 4.31 dB (range = 0 - 15.5 dB; see SI Figure 1 for the full audiometric results). The average difference between the better ear and the other ear was 3.3 dB (range = 0 - 14 dB).

Participants were recruited from the participant database at the Max Planck Institute for Human Cognitive and Brain Sciences and received monetary compensation of 12 € per hour for participation. The study was approved by the ethics committee of the Medical Faculty at Leipzig University (ethics vote: 032/24-ek).

Procedure

Participants listened to 240 trials separated into 12 blocks of 20 sentences each. Each trial consisted of two sentences presented in parallel. One sentence was spoken by a female speaker and the other one was spoken by a male speaker. Participants were instructed to follow the male speaker while ignoring the female speaker. After listening to the sentences and a brief period of 1 second to avoid EEG artifacts, participants had to repeat the target sentence. If participants did not understand the complete sentence, they were instructed to repeat the portion of the sentence they

did understand. If they were unable to understand any part of the sentence, they were asked to say so. The behavioral paradigm is illustrated in figure 1. Experiment control was administered by the Psychophysics Toolbox Version 3 running on Matlab on a Windows machine (Brainard & Vision, 1997; Kleiner et al., 2007). The experiment was conducted in a sound-proof, electromagnetically shielded booth. Stimuli were presented using Koss KSC75 headphones clipped to the participants' ears. Speech was recorded using a high quality condenser microphone (Rode NT55) at 44100 Hz. Target sentences were presented at six individually adjusted target-to-distractor ratios (T/D) [-2, -1, 0, 1, 2, 10 dB relative to their 50% speech reception threshold]. We additionally calculated word-by-word audibility by quantifying the proportion of the glimpsed target signal (see below). The latter condition served as a control condition, in which the target sentence should be well comprehended by every participant and thus serving as a control for attention and working memory. Control trials were excluded from behavioral and neural analyses and mean accuracy in the control condition was added as a control variable in the regression analysis (see below). Accuracy in the control condition was high (M = 97%, range = [93 - 99 %]), indicating that all participants attentively completed the task.

To control behavioral performance for peripheral differences in hearing abilities, speech reception thresholds (SRTs) were measured using an adaptive staircase procedure adapted from (Kollmeier et al., 1988; Rysop et al., 2021) including 20 additional stimuli. Distractor intensity was held constant at 70 dB sound pressure level. Target intensity was controlled by a staircase procedure, with a decrease in intensity if participants correctly repeated the full target sentence, and an increase in intensity if participants incorrectly repeated at least one word. Accuracy was rated by the experimenter. Similar as in (Rysop et al., 2021), the staircase procedure started with an T/D of 5 dB and a step size of 6 dB. Step size was decreased by a factor of 0.85 following each turning point. SRTs were calculated as the average of the T/D ratios preceding the final 5 turning points. The mean SRT across participants was $M = -8.16 (SD = 0.93)$.

Stimuli

Stimuli were recorded at 44,100 Hz using a high-quality condenser microphone (Rode NT55) in a sound-proof booth. The distractor sentences started 0.5 seconds prior and lasted for a minimum

of 0.3 seconds longer than the target sentences to ensure a complete masking of the target stream. Accordingly, the target sentences were comprised by 5-7 words and lasted for 2-4 seconds. The distractor sentences were comprised by 7-10 words and lasted for 3-5 seconds. The sentence material was retrieved from a German sentence corpus (Schiel & Baumann, 2006) after filtering for the suitable sentence length and excluding sentences that contained names or very unusual words. Additionally, similar suitable sentences were created by GPT-3.5. The prompts used for sentence generation can be retrieved from Table S1. All sentences were loudness normalized to -23 dB LUFS using pyloudnorm (Steinmetz & Reiss, 2021).

EEG data acquisition

EEG was recorded in a soundproof, electromagnetically shielded booth at a sampling rate of 1,000 Hz using a REFA8 68-channel amplifier system (TMSi, Oldenzaal, the Netherlands), grounded to the sternum. The recording included 63 Ag/AgCl electrodes positioned according to the 10-20 layout in ANT Neuro waveguard original caps as well as two mastoid external mastoid references (A1,A2). Eye movements were recorded using bipolar electrooculogram (EOG) electrodes placed at the outer sides of both eyes and at the top and bottom of the right eye. The EEG signals were monitored throughout the experiment.

EEG preprocessing

EEG preprocessing was conducted using MNE python (Gramfort et al., 2013) and followed recommendations for temporal response function (TRF) analyses (Crosse et al., 2021). The signal was filtered between 0.5 and 15 Hz and re-referenced to the average of the mastoid references. Bad channels were rejected manually based on a larger variance than neighboring channels. Independent component analysis (ICA) with 15 components was conducted to reduce artifacts from blinking and eye movements. The rejected components were selected automatically based on correlations of at least 0.3 with the EOG electrodes. We finally interpolated bad channels using spherical splines (Perrin et al., 1989) and downsampled the signals to 128 Hz for TRF estimation to reduce computation times. The signal was epoched between target sentence onset and offset.

Features

Acoustic features

Acoustic features included stimulus envelopes and acoustic onsets. Additionally, we investigated the neural tracking of word and phoneme onsets. All features were generated separately for target and distractor streams.

Envelopes and acoustic onsets were derived from auditory spectrograms calculated using naplib (Mischler et al., 2023). Spectrogram calculation involved a cochlear filter bank of 128 logarithmically-spaced constant-Q filters and a hair cell model to approximate the human peripheral auditory system. Envelopes were calculated using the mean across all spectrogram channels. Acoustic onsets were derived using the half-wave rectified derivative of the envelope (Brodbeck et al., 2018).

To derive word and phoneme onsets, stimuli and text transcripts were automatically aligned using forced alignment available in the WebMAUS Basic module of the BAS Web Services (Kisler et al., 2017; Schiel, 1999).

Linguistic features

To quantify the neural tracking of linguistic features, we investigated features at the word and phoneme levels, as previous studies showed that these features are significantly tracked in continuous listening paradigms (Brodbeck et al., 2018; Gillis et al., 2021).

Word level features

Word frequency was derived from the SUBTLEX-DE subtitles corpus (Brysbaert et al., 2011). Word frequency was defined as the logarithm of word occurence per million, as provided in the SUBTLEX-DE database.

Word surprisal is defined as the inverse probability of each word given the preceding context. Surprisal of word i is defined as the negative logarithm of word probability:

$$word\ surprisal_i = -log_2(p(w_i|context))$$

with $(p(w_i|context)$ being the word probability given the preceding sentence context derived from GPT-2 trained on a German text corpus (Schweter, 2020). Since we used isolated sentences, surprisal estimates of the first word in each sentence cannot be interpreted based on prior context. The surprisal values of the first word in each sentence were therefore omitted.

Phoneme level features

Phoneme surprisal reflects probability of each phoneme given the preceding phonemes in the current word. It is based on a probability prior captured by word frequency. Phoneme surprisal was defined as the inverse conditional probability of each phoneme, given the preceding phonemes in the same word:

$$phoneme\ surprisal_i = -log_2\left(\frac{freq(cohort_i)}{freq(cohort_{i-1})}\right)$$

$Cohort_i$ is the activated cohort of words given the current phoneme. This is the amount of words in the vocabulary that could be formed from the given sequence of phonemes. $Freq(cohort)$ is the sum of the word frequencies of all words in the cohort.

Cohort entropy reflects the uncertainty about the next phoneme. It is therefore defined as the Shannon entropy of the activated cohort (Brodbeck et al., 2018). Entropy at cohort i is defined by:

$$cohort\ entropy_i = -\sum_{word}^{cohort_{i-1}} p_{word} * log_2(p_{word})$$

with p_{word} being the probability of each word as reflected by its relative frequency. To appropriately separate the word and phoneme levels, we only used phoneme events that did not correspond to word onsets, i.e., omitting the first phonemes in each word. All features were min-max normalized to the range from 0 to 1 to avoid negative values before entering the TRF.

Speaker characteristics

As mentioned above, target and distractor sentences were presented by a male and a female voice, respectively. This was done to prevent confounding effects of the speaker, potentially dampening the effects. To control for the differences introduced by the speakers, the analysis was controlled for fundamental frequency. Fundamental frequency was calculated using the pYIN algorithm

(De Cheveigné & Kawahara, 2002; Mauch & Dixon, 2014) available in the python package librosa (McFee et al., 2024).

Additionally, we controlled the linguistic analyses for a set of variables capturing perception of speaker-specific characteristics. To obtain these speaker characteristics, a subsample of the participants (N = 23) took part in an online experiment after completing the EEG experiment. Here, participants were presented with ten sentences spoken by the same female and male speakers used in the EEG experiment. After each sentence, they were asked to rate the speakers on a set of 7 characteristics: age, educatedness, dominance, attractiveness, health, and professionalism (Lavan et al., 2024). The characteristics were rated on 8 point ratings scales and the ratings were summarized across trials. Using two-sided paired Wilcoxon tests corrected for multiple comparisons using false discovery rate (FDR), we revealed significant differences between the speakers on 3 of the characteristics (age [W = 21.5, p = .001], attractiveness [W = 8.5, p = .001], and dominance [W = 56, p = .031]). The speaker-specific mean ratings of these characteristics were added into the acoustic model containing the respective characteristic at each word onset in order to control the linguistic analysis for the speaker characteristics.

Temporal response function modeling

In order to investigate how comprehension performance is correlated with the interplay of acoustic and linguistic features of the attended and unattended streams, we used multivariate TRFs to quantify how features were neurally tracked. The model fits were then entered into a generalized linear mixed effects model to access the effect of feature tracking on comprehension.

To investigate the neural encoding of acoustic and linguistic features, we used multivariate TRFs. The analysis was implemented in mTRFpy (Bialas et al., 2023) using multiple linear regression with Tikhonov regularization. The TRF is a kernel that describes the linear transformation from an ongoing stimulus to an ongoing neural response for a specified set of time lags. When applying forward modeling, the TRF is used to predict the EEG signal from a set of stimulus features. To obtain the predicted EEG signal, the TRF is convolved with the speech features of the target and distractor sentences. The prediction performance is then obtained by measuring the correlation between the predicted EEG signal and the actual signal. The prediction performance can then be

interpreted as the amount of neural tracking of the speech features. Feature representations were investigated in time lags from -100 to 800 ms relative to event onset.

For the TRF analysis, we divided the trials into correct (i.e., all words of the sentence were repeated correctly) and incorrect trials. The criterion for correct sentences was deliberately chosen to be strict in order to ensure that participants grasped the full sentence context. Separate models for correct and incorrect trials were trained across participants, as previously recommended for small datasets (Crosse et al., 2021; Jessen et al., 2021). This was done to increase the amount of training data, as the reliability of TRF analyses has been shown to benefit from an increase in training data, especially for sparse linguistic features (Mesik & Wojtczak, 2023). Additionally, we aimed to reduce the effect of varying numbers of correct and incorrect trials across participants. We used an 80-20 cross-validation with 50 randomly sampled folds following recent recommendations for evaluating brain decoding models (Varoquaux et al., 2017). For each fold, 8 participants were randomly selected to comprise the test set. Data from these participants was not used for the model training. Models for correct trials were, on average, trained on 3831 trials, models for incorrect trials were trained on a mean number of 3169 trials. TRFs were evaluated for the duration of the target sentences, omitting the initial and the terminal segments where only the distractor sentence was presented.

We investigated two separate models: an acoustic and a linguistic TRF model. In the acoustic model, we included the lower-level acoustic features (i.e., envelope and acoustic onsets) as well as word and phoneme onsets. The linguistic model included all linguistic features on the word and phoneme levels. Naturally, higher-level features are highly correlated to lower-level features in speech, making it important to control analyses investigating higher-level features for lower-level variance (Gillis et al., 2021). We therefore controlled the higher-level analyses for the acoustic features. For this control, we conducted a separate TRF analysis predicting the EEG signals from all acoustic variables and word and phoneme onsets, including the speaker characteristics described above, with no cross-validation. Model predictions from this analysis were then subtracted from the original EEG data to generate residualized EEG responses that do not contain responses resulting from acoustic features. All analyses involving linguistic features were conducted on these residualized EEG responses. To isolate the influence of the distinct feature groups (see 1), we

systematically shuffled the features of interest (see below). The regularization parameter was fitted separately for acoustic and linguistic models, but it was commonly fitted across participants and correct and incorrect model variants to avoid overfitting and ensure comparability between models. Acoustic models were fitted using a regularization parameter of 0.1 and linguistic models were fitted with a regularization of 1e-6.

Statistical Inference

We used a shuffling approach to infer statistical significance of the TRF model fits. This was done to make sure that all models were trained on the same number of variables and improvements in model performance cannot solely be explained by the addition of variables. We grouped the individual features into four groups to avoid multicollinearity issues in regression model fitting. The feature groups and their corresponding features are shown in Table 1. We conducted 50 shuffling iterations for each model variant. In each iteration, we systematically shuffled the features of interest while keeping all other input features and model parameters identical. We used different shuffling methods for acoustic and linguistic models, corresponding to the characteristics of the respective features. For the acoustic models, trial labels were shuffled. For the linguistic models, we employed a more conservative approach of shuffling the values for each feature while preserving the onset times. Following the recommendation for investigating word-level linguistic neural tracking by (Chalehchaleh et al., 2024), we constrained the analyses to time points from 400 to 800 ms after word onset. These latencies correspond to traditionally observed time scales of linguistic effects (Hsin et al., 2023; Kutas & Federmeier, 2011; Lau et al., 2008; Šoškić et al., 2022).

Table 1: Speech features

Feature group	**Shuffled features**
Acoustic	Envelope, acoustic onsets
Onsets	Word onsets, phoneme onsets
Phoneme linguistic	Phoneme surprisal, cohort entropy
Word linguistic	Word surprisal, word frequency

Statistical significance was inferred by subtracting the mean shuffled model fits from the unshuffled model fits. Statistical inference was performed using mass-univariate two tailed one sample t-tests with threshold-free cluster enhancement and a cluster forming threshold of $\alpha = 0.05$. This analysis yields p-values corrected for multiple comparisons (Maris & Oostenveld, 2007). Clusters surrounding the electrode with the highest model fit were selected for further analysis. Clusters consisted of the five closest electrodes surrounding the peak electrode (Graumann et al., 2022).

Neuro-Behavioral Correlations

The speech recordings were automatically transcribed offline using whisper large-v2 (Bain et al., 2023). Subsequently, the transcriptions were manually controlled. We then assessed the overlap between the transcribed response and the target sentence, marking each word as accurately repeated or inaccurate.

Word-by-word repetition performance was predicted from word surprisal and word audibility. The amount of masking varies on a moment-by-moment basis in competing speech stimuli, leading to some speech segments being relatively well audible. This has been described as glimpsing (Cooke, 2006). Glimpsed segments can inform the comprehension of segments that coincide with higher amounts of acoustic masking. To estimate the amount of available glimpses in each target word, we calculated the so-called "glimpse rate" for each target word segmentation (Cooke, 2006). This means that we computed the available amount of acoustic information in each word by comparing target and distractor spectrogram amplitudes. The glimpse rate was defined as the proportion of spectrotemporal events that exceed a certain local T/D threshold within each word. This threshold was set to -5 dB based on previous work (Cooke, 2006).

To estimate the effects of word surprisal, audibility, and their interaction on word repetition performance (correct versus incorrect word repetition) in the experimental conditions, we calculated a generalized linear mixed model with a logistic link function using lme4 in R (Bates, 2014). P-values were obtained using lmerTest (Kuznetsova et al., 2017). The analysis was controlled for word frequency, sentence length (in words), word order, T/D condition, and accuracy in the control condition. We included random intercepts for subjects and sentences. To investigate the

relationship between TRF model fits and behavioral performance, standardized trial-wise model fits for all model variants were entered into the logistic regression model.

To rule out issues of multicollinearity, we calculated variance inflation factors using the R package car (Fox & Weisberg, 2019). The variance inflation factors ranged from 1.00 to 1.40 and we therefore assume no issues of multicollinearity (O'brien, 2007). We additionally inspected the correlations between the model fits and revealed that none of the model fits were strongly correlated. All correlations are displayed in supplementary figure 2. The full model explained 15.8 % of variance in comprehension considering fixed effects only, and 49.0 % of variance if random effects were taken into account.

Audibility control analysis

Due to the competing speech paradigm, speech clarity varies on a word-by-word basis, influencing comprehension. To infer influences of speech clarity on lower-level processing, we calculated separate TRF model fits based on the word audibility of each target word for both conditions (correct and incorrect). Separate model fits were obtained by dividing the observed and the predicted EEG signals into segments of low and high audibility. Predicted EEG signals resulted from the cross-validated predictions of the TRF models trained on all data. Then, the correlation between both signals was measured separately for segments of high and low audibility. The division of the words was done separately for correct and incorrect trials, so that each split would include half of the signal in each condition (Tezcan et al., 2023). Model fits were summarized within the previously defined clusters with maximal model fit (see above). Word audibility is influenced by word-by-word speech clarity, as well as by the T/D condition of the sentences. To combine word-by-word audibility and sentence T/D, we summed the standardized glimpse rate of each target word segmentation (Cooke, 2006) and the T/D condition of the respective sentence. To infer statistical significance, we conducted paired samples t-tests comparing model fits for segments with high and low audibility separately for correct and incorrect trials. Due to the explorative nature of the analysis, we report uncorrected as well as FDR corrected p-values.

Code Accessibility

Custom code for all analyses will be made available upon acceptance and publication of the manuscript.

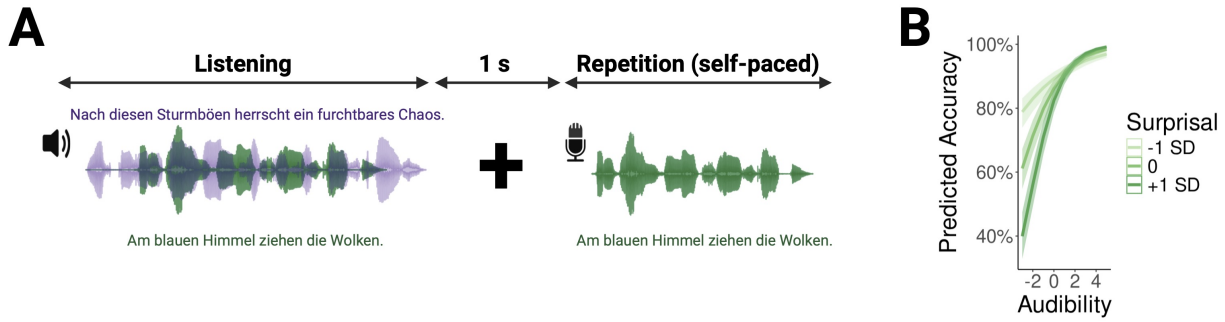
Results

Behavioral Results

Comprehension is modulated by linguistic features particularly at low audibility

In a generalized linear mixed model, we predicted word-by-word comprehension performance from word-level audibility and surprisal. Word-level audibility was derived from the proportion of the target signal exceeding a certain local SNR threshold (e.g., the "glimpse rate"). Results revealed that audibility, surprisal, as well as their interaction predicted word repetition performance. Figure 1 displays the interaction between word audibility and surprisal, indicating that the effect of surprisal was stronger at low levels of audibility (see Table 2 for the full set of results). This means that more predictable words were easier to understand than less predictable words when it was more difficult to hear the attended speech stream.

Figure 1: Behavioral analysis



A: Behavioral Paradigm. Participants listened to 240 trials comprised by one target and one distractor sentence. They subsequently repeated the target sentence. English translations of the sentences: After these squalls there is a terrible chaos (purple), Clouds gather in the blue sky (green). B: Effects of word surprisal and word audibility on word-by-word comprehension accuracy. Surprisal and audibility jointly predicted comprehension accuracy, with a stronger effect of surprisal at low levels of audibility.

Table 2: Generalized Linear Mixed Model outputs predicting comprehension performance

	Estimate	Std. Error	CI	z	p
Audibility	0.43	0.02	[0.39, 0.46]	23.19	<.001**
Surprisal	-0.32	0.02	[-0.36, -0.29]	-18.24	<.001**
Audibility:Surprisal	0.18	0.02	[0.15, 0.22]	9.66	<.001**
Acoustic target	-0.01	0.02	[-0.04, 0.02]	-0.54	.591
Onsets target	0.06	0.02	[0.03, 0.09]	4.03	<.001**
Word linguistic target	0.03	0.01	[0.004, 0.06]	2.22	.026*
Phoneme linguistic target	0.004	0.02	[-0.03, 0.02]	0.24	.81
Acoustic distractor	0.03	0.02	[0.004, 0.07]	2.20	.027*
Onsets distractor	0.01	0.01	[-0.02, 0.04]	0.49	.555
Word linguistic distractor	-0.03	0.01	[-0.06, -0.01]	-2.63	.008**
Phoneme linguistic distractor	-0.004	0.01	[-0.03, 0.02]	-0.27	.788
Target-to-Distractor ratio condition (dB)	0.48	0.01	[0.46, 0.50]	44.59	<.001**
Word frequency	-0.08	0.02	[-0.12, -0.04]	-3.60	<.001**
Word order	-0.11	0.01	[-0.13,-0.09]	-10.75	<.001**
Sentence length	-0.14	0.12	[-0.37, 0.09]	-1.21	.224
Control accuracy	0.20	0.09	[0.02, 0.39]	2.18	.029*
Speech reception threshold	0.27	0.05	[0.17, 0.37]	5.29	<.001**

** p < .01, * p < .05

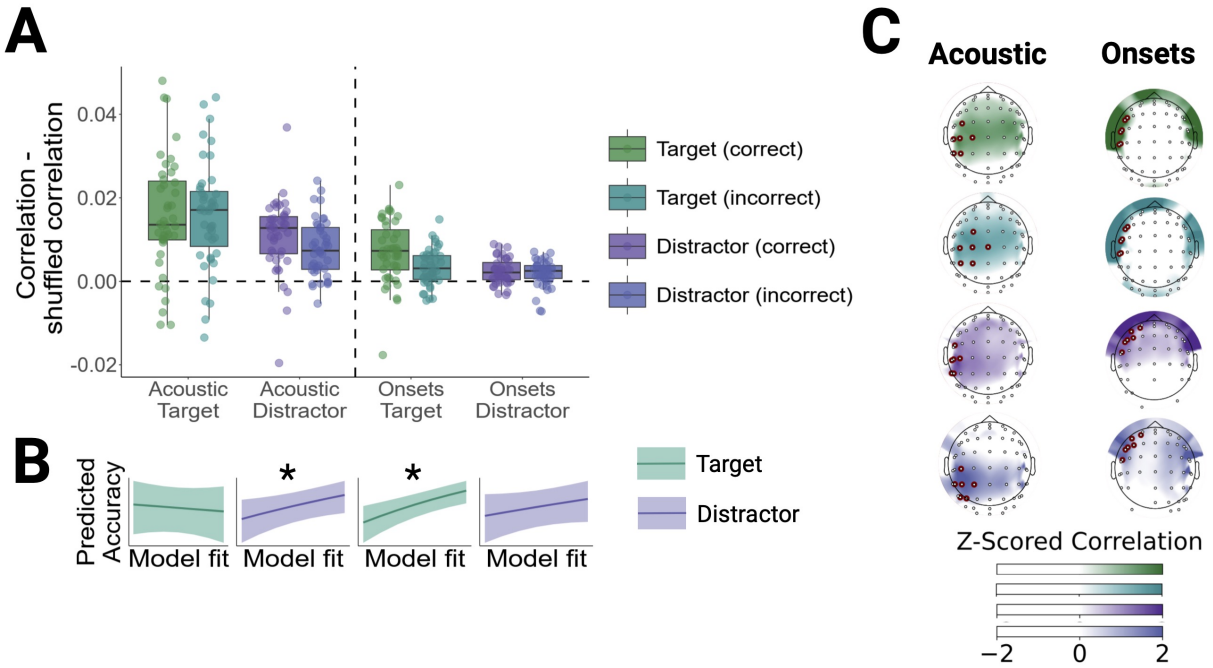
Temporal response function analysis

Acoustic features are significantly tracked for target and distractor sentences

To determine regions with significant model fits, we conducted mass-univariate two tailed one sample t-tests on the difference between TRF model fits and shuffled model fits. Acoustic features were significantly tracked in target and distractor sentences. As displayed in Figure 2 C, the model fit was significant at all electrodes. The strongest model fits were observed in left temporal electrodes. Similarly, word and phoneme onsets were significantly tracked for target and distractor

sentences. For the target sentences, model fits were significant across all sensors, but maximal model fits were observed in left temporal electrodes. For the distractor sentences, the fits were significant across frontal electrodes, and the strongest fits were observed in left frontal electrodes. The subsequent analyses of model fits and model weights were constrained to clusters showing the best prediction performance. These clusters were selected as the five electrodes surrounding the electrode with the best model fit and are displayed in 2 C.

Figure 2: Acoustic TRF model fits, clusters, and model predictions



A: Fits for the acoustic features. Each dot represents the fit for one participant. Correlations indicate the correlations between the EEG signals predicted by the temporal response functions and the actual EEG signals. B: Generalized linear mixed model predictions for the acoustic features. Correct and incorrect trials were combined in the generalized linear mixed effects model. * $p < .05$ C: Clusters used in the analyses. Colors indicate the magnitude of the correlation between observed and predicted EEG signal minus shuffled correlations. The model fits were z-scored within conditions to facilitate the visual inspection of their spatial organization. Electrodes in clusters with significant effects ($p < .05$ at the cluster level) are shown in white, selected cluster electrodes are marked by red circles. Acoustic refers to the model fits for the envelope and the acoustic onsets features. Onsets refers to the model fits for word and phoneme onsets.

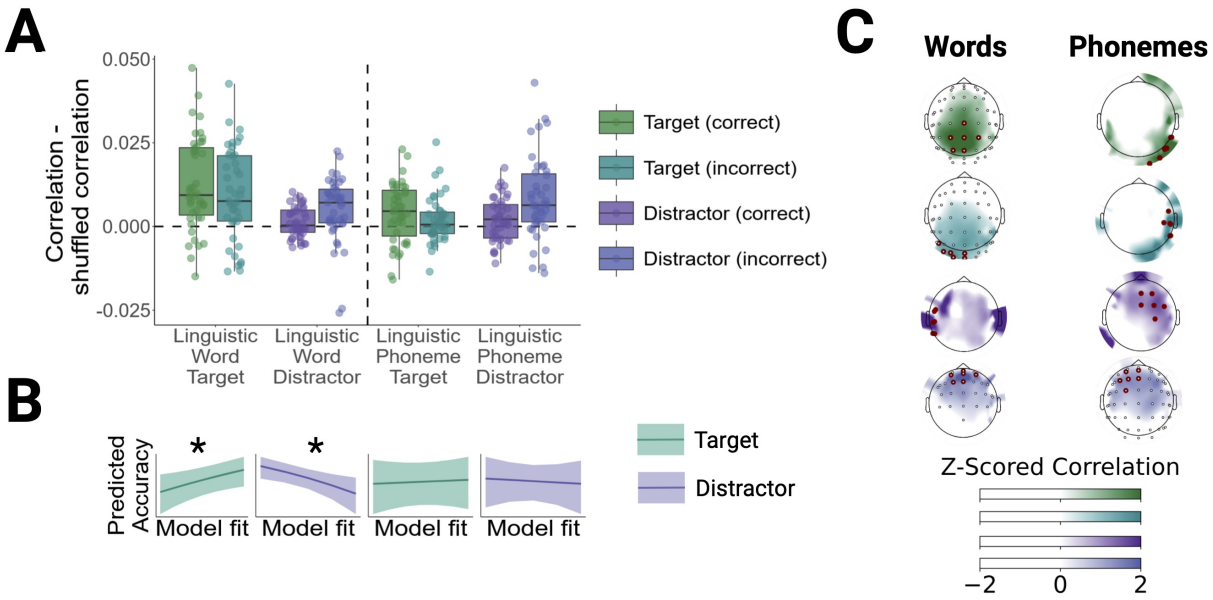
Linguistic features are significantly tracked for target sentences.

Mass-univariate statistics revealed significant clusters for word level linguistic features in the target sentences for correct and incorrect trials. As shown in Figure 3, the prediction performance was significantly different from the shuffled null distribution across all sensors. Prediction performance peaked in central posterior electrodes in correct trials for the word level linguistic features. In incorrect trials, significant prediction performance for the word-level features was widely spread across almost all electrodes. The best prediction performance was observed in more posterior sensors compared to correct trials. Phoneme level linguistic features were not significantly tracked for the target sentences.

Linguistic features are only tracked for distractor sentences in incorrect trials.

As shown in Figure 3 C, word and phoneme level linguistic features for unattended distractor sentences were only significantly tracked in incorrect trials. This means that no significant tracking of higher-level features in the unattended speech stream emerged if comprehension of the target stream was intact. In contrast, higher-level features for the unattended stream were neurally tracked when comprehension of the attended stream collapsed. Mass-univariate statistics revealed significant prediction performance for word level linguistic features in frontal and central sensors in incorrect trials. The best prediction performance was observed in frontal sensors. Model fits for linguistic phoneme features were significant in frontal and central sensors, and the best prediction performance was observed in right temporal sensors. In correct trials, no higher-level tracking for distractor sentences was observed.

Figure 3: Linguistic TRF model fits, clusters, and model predictions



A: Fits for the linguistic features. Each dot represents the fit for one participant. Correlations indicate the correlations between the EEG signals predicted by the temporal response functions and the residualized EEG signals. B: Generalized linear mixed model predictions for the linguistic features. Correct and incorrect trials were combined in the generalized linear mixed effects model. * $p < .05$ C: Clusters used in the analyses. Colors indicate the magnitude of the correlation between observed and predicted EEG signal minus shuffled correlations. The model fits were z-scored within conditions to facilitate the visual inspection of their spatial organization. Electrodes in clusters with significant effects ($p < .05$ at the cluster level) are shown in white, selected cluster electrodes are marked by red circles. Linguistic word model fits refer to the model fits for word surprisal and word frequency. Linguistic phoneme model fits refer to the model fits for phoneme surprisal and cohort entropy.

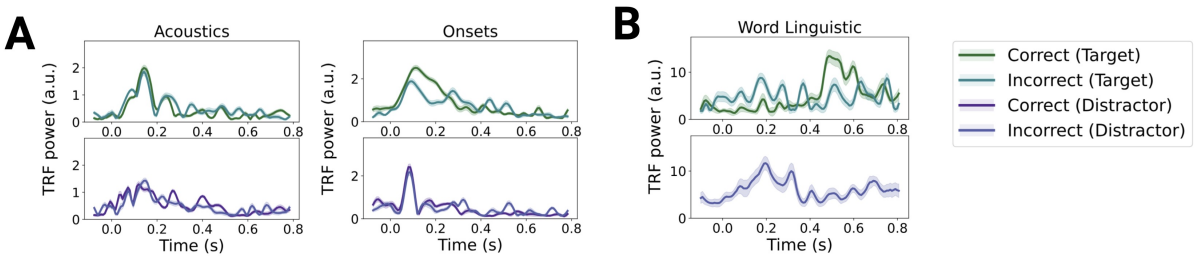
Neuro-Behavioral Correlations

Enhanced tracking of the target stream supports comprehension

To assess effects of the neural tracking of lower-level and higher-level features on comprehension, we predicted word-by-word comprehension accuracy from the trial-wise TRF model fits. We summarized model fits within the clusters displayed in Figures 2 C and 3 C. Results revealed that comprehension accuracy was positively correlated with the neural tracking of target onsets and word level linguistic features. This means that the neural tracking of onsets and word level information was stronger when the sentences were comprehended. Model parameters can be retrieved from Table 2 and Figures 2 B and 3 B.

As a result of training the TRF models across participants, the model weights do not reflect subject-specific, independent observations, which complicates statistical inference. Nevertheless, descriptive analysis of the weights shown in Figure 4 reveals similar results. Weights for the word linguistic features in correct trials are characterized by stronger responses starting from 400 ms. This corresponds to the typically observed time range of linguistic word level effects (Hsin et al., 2023; Kutas & Federmeier, 2011; Lau et al., 2008; Van Petten & Luka, 2012). In contrast, weights for incorrect trials showed smaller amplitudes in late windows. Weights for the onsets reveal stronger early responses in correct trials than in incorrect trials. Fits for the phoneme linguistic features only reached significance in incorrect trials for the distractor sentence. Weights for these trials can be retrieved from Figure S3.

Figure 4: TRF weights



A: TRF weights for the acoustic models. B: TRF weights for the linguistic word level model. Fits for the linguistic features did not reach significance for the distractor in correct trials and were therefore omitted.

Distractor processing is beneficial at lower and detrimental at higher levels

By predicting target comprehension accuracy from the trial-wise TRF model fits for the distractor stream, we revealed that acoustic tracking of the distractor was positively correlated with comprehension performance. In contrast, word level distractor tracking was negatively correlated with comprehension performance. Full results can be retrieved from 2. This shows that, only at higher levels, distractor processing showed a distracting effect, i.e., performance for target comprehension was negatively impacted when distractor information was overrepresented.

As mentioned above, interpretation of the TRF model weights is difficult given the low variance across iterations. However, we observed that distractor weights at the word level for incorrect trials were primarily characterized by early responses, as shown in figure 4 B. This differs from the time scale of the strongest effects in the target streams.

Influences of Audibility on TRF Model Fits

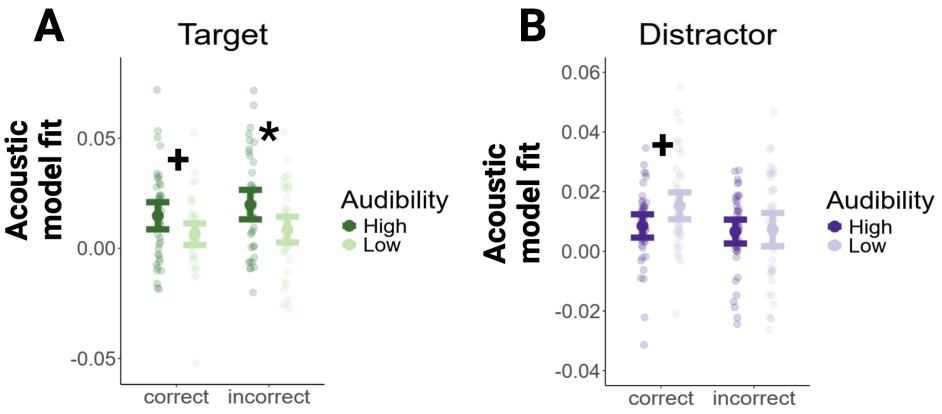
Lower level processing is influenced by speech clarity.

To infer bottom-up acoustic influences on the encoding of different features, we calculated separate model fits in sections with high and low word audibility. We then conducted paired samples t-tests comparing high and low audibility model fits seperately for incorrect and correct trials. Significant results can be visually inspected in Figure 5. Results revealed that target acoustic tracking was higher in segments of high audibility compared to segments of low audibility in correct trials ($t = 2.53, df = 42, p_{uncorrected} = .015, p_{corrected} = .091$), as well as in incorrect trials ($t = 3.32, df = 42, p_{uncorrected} = .001, p_{corrected} = .022$). Additionally, distractor acoustics were tracked more strongly at low compared to high target audibility in correct trials ($t = -2.32, df = 42, p_{uncorrected} = .025, p_{corrected} = .101$).

Discussion

In this study, we examined how acoustic and linguistic processing contribute to speech comprehension within a competing speech paradigm. As a main finding, we show that the attentional engagement with the target stream positively contributed to comprehension at lower and higher

Figure 5: Acoustic TRF model fits by Audibility



Significant effects of word-by-word audibility on the TRF model fits. A: Increased acoustic target tracking in high audibility. B: Increased acoustic distractor tracking in correct trials in low audibility. Points show the means and their 95% confidence intervals. Transparent dots show the raw individual subject data. * $p < .05$ (FDR corrected for multiple comparisons), + $p < .05$ (uncorrected for multiple comparisons)

hierarchical levels. In contrast, the neural tracking of distractor information was positively correlated with comprehension at lower levels, whereas higher level distractor engagement had a negative effect on comprehension. These results suggest that target enhancement and distractor suppression jointly support comprehension. However, increased attention to distractor features only impedes comprehension at higher hierarchical levels, but not at lower acoustic processing levels.

Acoustic and linguistic features influence comprehension

On the behavioral level, we showed that word predictability and word audibility jointly and interactively predicted comprehension performance. The effect of word predictability was strongest when audibility was low and the target words were therefore more difficult to hear. This is consistent with previous literature showing influences of word predictability on comprehension performance (Kalikow et al., 1977), as well as on behavioral indicators of processing demands, such as reading times (Smith & Levy, 2013).

To obtain a fine-grained behavioral comprehension measure, we adopted the widely used method of sentence repetition (Kalikow et al., 1977; Rysop et al., 2021; Stickney & Assmann, 2001). To infer the effect of working memory constraints on sentence repetition, we assessed task-specific working memory performance in a control condition comprised by sentences that were well comprehended by the participants. The analysis was controlled for the accuracy in this control condition. Accuracy in the control condition amounted to an almost perfect accuracy of 97% of words across participants. Thus, we assume that comprehension scores in the remaining conditions truly reflect comprehension instead of working memory constraints.

Target enhancement predicts comprehension

Consistent with previous studies (Etard & Reichenbach, 2019; Iotzov & Parra, 2019; Lesenfants et al., 2019; Tune et al., 2021; Vanthornhout et al., 2018), we found that the stronger neural representation of the target stream positively predicted comprehension performance. This effect was present for word and phoneme onsets, suggesting that successful word and phoneme segmentation is crucial for comprehension. This is congruent with a recent study revealing that the neural representation of word onsets differentiates between intelligible and unintelligible speech while controlling for acoustic differences (Karunathilake et al., 2023). Additionally, the neural representation of word level linguistic information was positively predictive of comprehension performance. This effect is unsurprising given the previous literature (Broderick et al., 2018; Holcomb, 1993; Karunathilake et al., 2023; Obleser & Kotz, 2011). However, extending previous work, we provide direct evidence that increased attention towards lower- and higher-level neural representations of the attended stream is a neural indicator of comprehension success in a challenging multitalker situation. Importantly, higher level tracking was predictive of comprehension performance while controlling for lower level acoustic neural tracking, underlining its importance for the prediction of successful language comprehension.

Interestingly, the neural representation of acoustic target information (i.e., the envelope and acoustic onsets) was not predictive of comprehension performance. This seems to contrast previous studies demonstrating that acoustic tracking is sensitive to speech clarity (Ahissar et al., 2001; Doelling et al., 2014; Etard & Reichenbach, 2019; Iotzov & Parra, 2019; Lesenfants et al.,

2019; Vanthornhout et al., 2018). In our audibility control analysis, we show that target envelope tracking was indeed sensitive to word-by-word speech clarity. However, we did not find an effect of target envelope tracking on speech comprehension. Similarly, several previous studies showed that acoustic tracking is not directly related to speech comprehension (Kösem et al., 2023; Verschueren et al., 2022). This is consistent with a recent study showing that higher level representations (i.e., word onsets and word level linguistic features) differentiate intelligible from unintelligible speech, while controlling for speech acoustics (Karunathilake et al., 2023). Collectively, the previous and present findings suggest that the neural tracking of higher-level linguistic information is crucial for higher-level comprehension while lower-level acoustic target tracking is related to speech clarity.

Distractor attention is differentially related to comprehension

In contrast to the representation of target information, previous findings on the neural representation of distractor information have been more heterogeneous. By disentangling acoustic and linguistic processing levels, we show that lower level distractor representation does not interfere with comprehension of the target speech stream. In fact, acoustic tracking of the distractor stream was weakly positively predictive of comprehension, suggesting that stronger distractor tracking at the acoustic level may facilitate comprehension. This may be attributed to the dominance of the distractor in the acoustic speech mixture, with stronger distractor tracking reflecting greater overall listening engagement. Additionally, improved acoustic tracking of the distractor stream may support comprehension by aiding stream segregation (Fiedler et al., 2019; J. Rimmele et al., 2012). Consistent with previous studies, we show that the attentional competition between attended and unattended signals does not primarily operate at the acoustic level (Golumbic et al., 2013; Orf et al., 2023), as acoustic tracking persists even for the unattended speech stream (Ding & Simon, 2012; Horton et al., 2014; Kaufman & Golumbic, 2023; O'sullivan et al., 2015).

In contrast, higher level distractor tracking was negatively correlated with comprehension, providing direct evidence for a distraction effect. Thus, the presence of higher-level distractor tracking indicates that more attention was allocated to the unattended stream, leading to lower comprehension of the attended stream. Only higher level distractor representation interfered with target comprehension, indicating that, only at this higher level, distractor disengagement becomes

perceptually adaptive. Additionally, we replicate the finding that higher level tracking typically does not occur for unattended speech (Brodbeck et al., 2018; Broderick et al., 2018), as long as listeners successfully allocate selective attention. We did not reveal any significant representation of word or phoneme level linguistic features for the distractor sentences in correct trials. Crucially, these representations were present for incorrect trials, suggesting that significant neural tracking of higher level distractor information only emerged when attention allocation to the target stream failed.

Interestingly, higher level representations of the distractor sentences were present in different sensors and latencies than higher level representations of the target sentences. Word level linguistic representations of correctly reproduced target sentences were characterized by late responses in central-posterior sensors, which is consistent with previous results using temporal response functions (Gillis et al., 2021) and traditional evidence in event-related potentials, including the N400 as well as later positive components (Aurnhammer et al., 2021; Hsin et al., 2023; Kutas & Federmeier, 2011; Lau et al., 2008; Šoškić et al., 2022; Van Petten & Luka, 2012). Thus, this response profile seems to reflect successful linguistic processing of the target sentences. In contrast, the highest prediction performance for distractor word level features was located in frontal sensors. Additionally, the weights for distractor word level features showed earlier peaks than the weights for target word level features. This response profile does not correspond to typically observed signatures of linguistic processing. Instead, the responses may reflect suppression processes in response to higher level distractor information. The potentially different mechanisms in response to target and distractor linguistic features should be further uncovered in future investigations.

Conclusion

This study provides evidence that target enhancement as well as distractor suppression at different hierarchical levels are predictive of comprehension performance. While target processing at lower and higher hierarchical levels were positively correlated with comprehension, distractor influences were positively correlated at lower and negatively correlated at higher levels. This underlines the relevance of the distinct roles of acoustic and linguistic representations in attended and unattended

streams when attempting to unravel the neural mechanisms contributing to selective attention and distraction in multitalker environments.

Competing interests

The authors declare no competing interest.

Acknowledgments

We thank Heike Boethel for her support in data acquisition.

Funding

AVB was supported by a PhD fellowship from the International Max Planck Research School (IMPRS) on Cognitive Neuroimaging. GH was supported by the Lise Meitner Excellence Program of the Max Planck Society, the German Research Foundation (DFG, HA 6314/4-2; Research Unit 5429/1 (467143400), HA 6314/10-2), and the European Research Council (ERC-2021-COG101043747). JO was supported by the German Research Foundation (DFG, OB 352/2-2). JMR was supported by the Max Planck Institute for Empirical Aesthetics.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367–13372.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (n400) and integration (p600) in expectation-based comprehension. *Plos one*, *16*(9), e0257430.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Bates, D. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, *17*(2), 89–98.

Bialas, O., Dou, J., & Lalor, E. C. (2023). Mtrfpy: A python package for temporal response function analysis. *Journal of Open Source Software*, *8*(89), 5657.

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

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

Chalehchaleh, A., Winchester, M., & Di Liberto, G. (2024). Robust assessment of the cortical encoding of word-level expectations using the temporal response function. *Journal of Neural Engineering*.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573.

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*, 705621.

De Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, *19*(1), 158–164.

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, *85*, 761–768.

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385–393.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience*, *39*(29), 5750–5759.

Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage*, *186*, 33–42.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Sage. https://www.john-fox.ca/Companion/

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring eeg tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, *41*(50), 10316–10329.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, *77*(5), 980–991.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG data analysis

with MNE-Python. *Frontiers in Neuroscience*, *7*(267), 1–13. https://doi.org/10.3389/fnins.2013.00267

Graumann, M., Ciuffi, C., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). The spatiotemporal neural dynamics of object location representations in the human brain. *Nature human behaviour*, *6*(6), 796–811.

Hämäläinen, J. A., Rupp, A., Soltész, F., Szücs, D., & Goswami, U. (2012). Reduced phase locking to slow amplitude modulation in adults with dyslexia: An meg study. *Neuroimage*, *59*(3), 2952–2961.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences*, *109*(49), 20095–20100.

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the n400 in language processing. *Psychophysiology*, *30*(1), 47–61.

Horton, C., Srinivasan, R., & D'Zmura, M. (2014). Envelope responses in single-trial eeg indicate attended speaker in a 'cocktail party'. *Journal of neural engineering*, *11*(4), 046015.

Hsin, C.-H., Chao, P.-C., & Lee, C.-Y. (2023). Speech comprehension in noisy environments: Evidence from the predictability effects on the n400 and lpc. *Frontiers in Psychology*, *14*, 1105346.

Iotzov, I., & Parra, L. C. (2019). Eeg can predict speech intelligibility. *Journal of Neural Engineering*, *16*(3), 036008.

Jessen, S., Obleser, J., & Tune, S. (2021). Neural tracking in infants–an analytical tool for multisensory social processing in development. *Developmental Cognitive Neuroscience*, *52*, 101034.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the acoustical society of America*, *61*(5), 1337–1351.

Karunathilake, I. D., Kulasingham, J. P., & Simon, J. Z. (2023). Neural tracking measures of speech intelligibility: Manipulating intelligibility while keeping acoustics unchanged. *Proceedings of the National Academy of Sciences*, *120*(49), e2309166120.

Kaufman, M., & Golumbic, E. Z. (2023). Listening to two speakers: Capacity and tradeoffs in neural speech tracking during selective and distributed attention. *NeuroImage*, *270*, 119984.

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

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

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *The Journal of the Acoustical Society of America*, *83*(5), 1852–1862.

Kösem, A., Dai, B., McQueen, J. M., & Hagoort, P. (2023). Neural tracking of speech envelope does not unequivocally reflect intelligibility. *NeuroImage*, *272*, 120040.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*(1), 621–647.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of statistical software*, *82*, 1–26.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the n400. *Nature reviews neuroscience*, *9*(12), 920–933.

Lavan, N., Rinke, P., & Scharinger, M. (2024). The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences*, *121*(26), e2318361121.

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., & Francart, T. (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations. *Hearing research*, *380*, 1–9.

MacPherson, A., & Akeroyd, M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in hearing*, *18*, 2331216514537722.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Mauch, M., & Dixon, S. (2014). Pyin: A fundamental frequency estimator using probabilistic threshold distributions. *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, 659–663.

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., ... Pimenta, W. (2024). *Librosa/librosa: 0.10.2* (Version 0.10.2). Zenodo. https://doi.org/10.5281/zenodo.4923181

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Mesik, J., & Wojtczak, M. (2023). The effects of data quantity on performance of temporal response function analyses of natural speech processing. *Frontiers in neuroscience*, *16*, 963629.

Mischler, G., Raghavan, V., Keshishian, M., & Mesgarani, N. (2023). Naplib-python: Neural acoustic data processing and analysis tools in python. *Software impacts*, *17*, 100541.

Moore, T., & Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annual review of psychology*, *68*(1), 47–72.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences*, *23*(11), 913–926.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage*, *55*(2), 713–723.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, *41*, 673–690.

Orf, M., Wöstmann, M., Hannemann, R., & Obleser, J. (2023). Target enhancement but not distractor suppression in auditory neural tracking during continuous speech. *IScience*, *26*(6).

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, *25*(7), 1697–1706.

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, *72*(2), 184–187.

Rimmele, J., Schröger, E., & Bendixen, A. (2012). Age-related changes in the use of regular patterns for auditory scene analysis. *Hearing research*, *289*(1-2), 98–107.

Rimmele, J. M., Golumbic, E. Z., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, *68*, 144–154.

Rysop, A. U., Schmitt, L.-M., Obleser, J., & Hartwigsen, G. (2021). Neural modelling of the semantic predictability gain under challenging listening conditions. *Human Brain Mapping*, *42*(1), 110–127.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. *Proc. of the ICPhS*, 607–610.

Schiel, F., & Baumann, A. (2006). Phondat 1, corpus v. 3.4. *Bavarian Archive for Speech Signals (BAS), Tech. Rep.*

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.

Schweter, S. (2020). *German gpt-2 model* (Version 1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4275046

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Šoškić, A., Jovanović, V., Styles, S. J., Kappenman, E. S., & Ković, V. (2022). How to do better n400 studies: Reproducibility, consistency and adherence to research standards in the existing literature. *Neuropsychology Review*, *32*(3), 577–600.

Steinmetz, C. J., & Reiss, J. D. (2021). Pyloudnorm: A simple yet flexible loudness meter in python. *150th AES Convention*.

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: Is it special? *Hearing research*, *305*, 57–73.

Stickney, G. S., & Assmann, P. F. (2001). Acoustic and linguistic factors in the perception of bandpass-filtered speech. *The Journal of the Acoustical Society of America*, *109*(3), 1157–1165.

Tezcan, F., Weissbart, H., & Martin, A. E. (2023). A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension. *Elife*, *12*, e82386.

Tierney, A., & Kraus, N. (2015). Neural entrainment to the rhythmic structure of music. *Journal of Cognitive Neuroscience*, *27*(2), 400–408.

Tune, S., Alavash, M., Fiedler, L., & Obleser, J. (2021). Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nature Communications*, *12*(1), 4533.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International journal of psychophysiology*, *83*(2), 176–190.

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*, 181–191.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179.

Verschueren, E., Gillis, M., Decruy, L., Vanthornhout, J., & Francart, T. (2022). Speech understanding oppositely affects acoustic and linguistic neural tracking in a speech rate manipulation paradigm. *Journal of Neuroscience*, *42*(39), 7442–7453.