# Minimal background noise enhances neural speech tracking: Evidence of stochastic resonance

Björn Herrmann*

Rotman Research Institute,
Baycrest Academy for Research and Education, M6A 2E1, North York, ON, Canada

Department of Psychology,
University of Toronto, M5S 1A1, Toronto, ON, Canada

*Correspondence concerning this article should be addressed to Björn Herrmann, Rotman Research Institute, Baycrest, 3560 Bathurst St, North York, ON, M6A 2E1, Canada. E-mail: bherrmann@research.baycrest.org

**Keywords:** Electroencephalography, speech encoding, story listening, temporal response function, background noise, attention

**Conflict of interest statement:** The author declares no competing interests.

**Acknowledgments:** We thank Priya Pandey, Tiffany Lao, and Saba Junaid for their help with data collection and story generation. The research was supported by the Canada Research Chair program (CRC-2019-00156) and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant: RGPIN-2021-02602).

## Abstract

Neural activity in auditory cortex tracks the amplitude envelope of continuous speech, but recent work counter-intuitively suggests that neural tracking increases when speech is masked by background noise, despite reduced speech intelligibility. Noise-related amplification could indicate that stochastic resonance – the response facilitation through noise – supports neural speech tracking. However, a comprehensive account of the sensitivity of neural tracking to background noise and of the role cognitive investment is lacking. In five electroencephalography (EEG) experiments (N=109; box sexes), the current study demonstrates a generalized enhancement of neural speech tracking due to minimal background noise. Results show that a) neural speech tracking is enhanced for speech masked by background noise at very high SNRs (~30 dB SNR) where speech is highly intelligible; b) this enhancement is independent of attention; c) it generalizes across different stationary background maskers, but is strongest for 12-talker babble; and d) it is present for headphone and free-field listening, suggesting that the neural-tracking enhancement generalizes to real-life listening. The work paints a clear picture that minimal background noise enhances the neural representation of the speech envelope, suggesting that stochastic resonance contributes to neural speech tracking. The work further highlights non-linearities of neural tracking induced by background noise that make its use as a biological marker for speech processing challenging.

## Introduction

Speech in everyday life is often masked by background sound, such as music or speech by other people, making speech comprehension challenging, especially for older adults (Pichora-Fuller et al., 2016; Herrmann and Johnsrude, 2020). Having challenges with speech comprehension is a serious barrier to social participation (Nachtegaal et al., 2009; Heffernan et al., 2016) and can have long-term negative health consequences, such as cognitive decline (Lin and Albert, 2014; Panza et al., 2019). Understanding how individuals encode speech in the presence of background sound is thus an important area of research and clinical application. One successful approach to characterize speech encoding in the brain is to quantify how well neural activity tracks relevant speech features, such as the amplitude envelope, of continuous speech (Crosse et al., 2016; Crosse et al., 2021). Greater speech tracking has been associated with higher speech intelligibility (Ding et al., 2014; Vanthornhout et al., 2018; Lesenfants et al., 2019), leading to the suggestion that speech tracking could be a useful clinical biomarker, for example, in individuals with hearing loss (Gillis et al., 2022; Palana et al., 2022; Schmitt et al., 2022). Counterintuitively, however, neural speech tracking has been shown to increase in the presence of background masking sound even at masking levels for which speech intelligibility is decreased (Yasmin et al., 2023; Panela et al., 2024). The causes of this increase in speech tracking under background masking are unclear.

In many everyday situations, a listener must block out ambient, stationary background noise, such as multi-talker babble in a busy restaurant. For low signal-to-noise ratios (SNRs) between speech and a masker, neural speech tracking decreases relative to high SNRs (Ding and Simon, 2013; Vanthornhout et al., 2018; Zou et al., 2019; Yasmin et al., 2023), possibly reflecting the decreased speech understanding under speech masking. In contrast, for moderate masking levels, at which listeners can understand most words, if not all, neural speech tracking can be increased relative to clear speech (Yasmin et al., 2023; Panela et al., 2024). This has been interpreted to reflect the increased attention required to understand speech (Hauswald et al., 2022; Yasmin et al., 2023; Panela et al., 2024). However, the low-to-moderate SNRs (e.g., <10 dB) typically used to study neural speech tracking require listeners to invest cognitively and thus do not allow distinguishing a cognitive mechanism from the possibility that the noise per se leads to an increase in neural tracking, for example, through stochastic resonance, where background noise amplifies the response of a system to an input (Kitajo et al., 2007; McDonnell and Ward, 2011; Krauss et al., 2016). Examining neural speech tracking under high SNRs (e.g., >20 dB SNR), for which

individuals can understand speech with ease, is needed to understand the noise-related tracking enhancement.

A few previous studies suggest that noise per se may be a relevant factor. For example, the neural-tracking increase has been observed for speech masked by multi-talker background babble (Yasmin et al., 2023; Panela et al., 2024), but appears to be less present for noise that spectrally matches the speech signal (Ding and Simon, 2013; Synigal et al., 2023), suggesting that the type of noise is important. Other research indicates that neural responses to tone bursts can increase in the presence of minimal noise (i.e., high SNRs) relative to clear conditions (Alain et al., 2009; Alain et al., 2012; Alain et al., 2014), pointing to the critical role of noise in amplifying neural responses independent of speech.

Understanding the relationship between neural speech tracking and background noise is critical because neural tracking is frequently used to investigate consequences of hearing loss for speech processing (Presacco et al., 2019; Decruy et al., 2020; Van Hirtum et al., 2023). Moreover, older adults often exhibit enhanced neural speech tracking (Presacco et al., 2016; Brodbeck et al., 2018; Broderick et al., 2021; Panela et al., 2024) which is thought to be due to a loss of inhibition and increased internal noise (Zeng, 2013; Auerbach et al., 2014; Krauss et al., 2016; Zeng, 2020; Herrmann and Butler, 2021). The age-related neural tracking enhancement may be harder to understand if external, sound-based noise also drives increases in neural speech tracking.

The current study comprises 5 EEG experiments in younger adults that aim to investigate how neural speech tracking is affected by different degrees of background masking (Experiment 1), whether neural tracking enhancements are due to attention investment (Experiment 2), the generalizability of changes in neural speech tracking for different masker types (Experiments 3 and 4), and whether effects generalize from headphone to free-field listening (Experiment 5). The results point to a highly generalizable enhancement in neural speech tracking at minimal background masking levels that is independent of attention, suggesting that stochastic resonance plays a critical role.

## General methods

### Participants

The current study comprised 5 experiments. Participants were native English speakers or grew up in English-speaking countries (mostly Canada) and have been speaking English since early

childhood (<5 years of age). Participants reported having normal hearing abilities and no neurological disease (one person reported having ADHD, but this did not affect their participation). Detailed demographic information is provided in the respective participant section for each experiment. Participants gave written informed consent prior to the experiment and were compensated for their participation. The study was conducted in accordance with the Declaration of Helsinki, the Canadian Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (TCPS2-2014), and was approved by the Research Ethics Board of the Rotman Research Institute at Baycrest Academy for Research and Education.

**Sound environment and stimulus presentation**

Data collection was carried out in a sound-attenuating booth. Sounds were presented via Sennheiser (HD 25-SP II) headphones and computer loudspeakers (Experiment 5) through an RME Fireface 400 external sound card. Stimulation was run using Psychtoolbox in MATLAB (v3.0.14; MathWorks Inc.) on a Lenovo T480 laptop with Microsoft Windows 7. Visual stimulation was projected into the sound booth via screen mirroring. All sounds were presented at about 65 dB SPL.

**Story materials**

In each of the 5 experiments, participants listened to 24 stories of about 1:30 to 2:30 min duration each. OpenAI's GPT3.5 (OpenAI et al., 2023) was used to generate each story, five corresponding comprehension questions, and four associated multiple-choice answer options (1 correct, 3 incorrect). Each story was on a different topic (e.g., a long-lost friend, a struggling artist). GPT stories, questions, and answer choices were manually edited wherever needed to ensure accuracy. Auditory story files were generated using Google's modern artificial intelligence (AI)-based speech synthesizer using the male "en-US-Neural2-J" voice with default pitch and speed parameters (https://cloud.google.com/text-to-speech/docs/voices). Modern AI speech is experienced as very naturalistic and speech perception is highly similar for AI and human speech (Herrmann, 2023). A new set of 24 stories and corresponding comprehension questions and multiple-choice options were generated for each of the 5 experiments.

After each story in Experiments 1, 3, 4, and 5, participants answered the 5 comprehension questions about the story. Each comprehension question comprised four response options (chance level = 25%). Participants further rated the degree to which they understood the gist of what was said in the story, using a 9-point scale that ranged from 1 (strongly disagree) to 9 (strongly agree; the precise wording was: "I understood the gist of the story" and "(Please rate this statement

independently of how you felt about the other stories)"). Gist ratings were linearly scaled to range between 0 and 1 to facilitate interpretability similar to the proportion correct responses (Mathiesen et al., 2023; Panela et al., 2024). For short sentences, gist ratings have been shown to highly correlated with speech intelligibility scores (Davis and Johnsrude, 2003; Ritz et al., 2022). In Experiment 2, participants performed a visual n-back task (detailed below) while stories were presented and no comprehension questions nor the gist rating were administered.

**Electroencephalography recordings and preprocessing**

Electroencephalographical signals were recorded from 16 scalp electrodes (Ag/Ag–Cl-electrodes; 10-20 placement) and the left and right mastoids using a BioSemi system (Amsterdam, The Netherlands). The sampling frequency was 1024 Hz with an online low-pass filter of 208 Hz. Electrodes were referenced online to a monopolar reference feedback loop connecting a driven passive sensor and a common-mode-sense (CMS) active sensor, both located posteriorly on the scalp.

Offline analysis was conducted using MATLAB software. An elliptic filter was used to suppress power at the 60-Hz line frequency. Data were re-referenced by averaging the signal from the left and right mastoids and subtracting the average separately from each of the 16 channels. Re-referencing to the averaged mastoids was calculated to gain high signal-to-noise ratio for auditory responses at fronto-central-parietal electrodes (Ruhnau et al., 2012; Herrmann et al., 2013). Data were filtered with a 0.7-Hz high-pass filter (length: 2449 samples, Hann window) and a 22-Hz low-pass filter (length: 211 samples, Kaiser window).

EEG data were segmented into time series time-locked to story onset and down-sampled to 512 Hz. Independent components analysis was used to remove signal components reflecting blinks, eye movement, and noise artifacts (Bell and Sejnowski, 1995; Makeig et al., 1995 Oostenveld et al., 2011). After the independent components analysis, remaining artifacts were removed by setting the voltage for segments in which the EEG amplitude varied more than 80 μV within a 0.2-s period in any channel to 0 μV (cf. Dmochowski et al., 2012; Dmochowski et al., 2014; Cohen and Parra, 2016; Irsik et al., 2022; Yasmin et al., 2023; Panela et al., 2024). Data were low-pass filtered at 10 Hz (251 points, Kaiser window) because neural signals in the low-frequency range are most sensitive to acoustic features (Di Liberto et al., 2015; Zuk et al., 2021; Yasmin et al., 2023).

**Calculation of amplitude-onset envelopes**

For each clear story (i.e., without background noise or babble), a cochleogram was calculated using a simple auditory-periphery model with 30 auditory filters (McDermott and Simoncelli, 2011). The resulting amplitude envelope for each auditory filter was compressed by 0.6 to simulate inner ear compression (McDermott and Simoncelli, 2011). Such a computationally simple peripheral model has been shown to be sufficient, as compared to complex, more realistic models, for envelope-tracking approaches (Biesmans et al., 2017). Amplitude envelopes were averaged across auditory filters and low-pass filtered at 40-Hz filter (Butterworth filter). To obtain the amplitude-onset envelope, the first derivative was calculated and all negative values were set to zero (Hertrich et al., 2012; Fiedler et al., 2017; Fiedler et al., 2019; Yasmin et al., 2023; Panela et al., 2024). The onset-envelope was down-sampled to match the sampling of the EEG data.

**EEG temporal response function and prediction accuracy**

A forward model based on the linear temporal response function (TRF; Crosse et al., 2016; Crosse et al., 2021) was used to quantify the relationship between the amplitude-onset envelope of a story and EEG activity (note that cross-correlation led to very similar results; cf. Hertrich et al., 2012). The ridge regularization parameter lambda (λ), which prevents overfitting, was set to 10 based on previous work (Fiedler et al., 2017; Fiedler et al., 2019; Yasmin et al., 2023; Panela et al., 2024). Pre-selection of λ based on previous work avoids extremely low and high λ on some cross-validation iterations and avoids substantially longer computational time. Pre-selection of λ also avoids issues if limited data per condition are available, as in the current study (Crosse et al., 2021).

For each story, 50 25-s data snippets (Crosse et al., 2016; Crosse et al., 2021) were extracted randomly from the EEG data and corresponding onset-envelope (Panela et al., 2024). Each of the 50 EEG and onset-envelope snippets were held out once as a test dataset, while the remaining non-overlapping EEG and onset-envelope snippets were used as training datasets. That is, for each training dataset, linear regression with ridge regularization was used to map the onset-envelope onto the EEG activity to obtain a TRF model for lags ranging from 0 to 0.4 s (Hoerl and Kennard, 1970; Crosse et al., 2016; Crosse et al., 2021). The TRF model calculated for the training data was used to predict the EEG signal for the held-out test dataset. The Pearson correlation between the predicted and the observed EEG data of the test dataset was used as a measure of EEG prediction accuracy (Crosse et al., 2016; Crosse et al., 2021). Model estimation and prediction accuracy were calculated separately for each of the 50 data snippets per story, and prediction accuracies were averaged across the 50 snippets.

To investigate the neural-tracking response directly, we calculated TRFs for each training dataset for a broader set of lags, ranging from -0.15 to 0.5 s, to enable similar analyses as for traditional event-related potentials (Yasmin et al., 2023; Panela et al., 2024). TRFs were averaged across the 50 training datasets and the mean in the time window -0.15 to 0 s was subtracted from the data at each time point (baseline correction).

Data analyses focused on a fronto-central electrode cluster (F3, Fz, F4, C3, Cz, C4) known to be sensitive to neural activity originating from auditory cortex (Näätänen and Picton, 1987; Picton et al., 2003; Herrmann et al., 2018; Irsik et al., 2021). Prediction accuracies and TRFs were averaged across the electrodes of this fronto-central electrode cluster prior to further analysis.

Analyses of the TRF focused on the P1-N1 and the P2-N1 amplitude differences. The amplitude of individual TRF components (P1, N1, P2) was not analyzed because the TRF time courses for the clear condition had an overall positive shift (see also Yasmin et al., 2023; Panela et al., 2024) that could bias analyses more favorably towards response differences which may, however, be harder to interpret. The P1, N1, and P2 latencies were estimated from the averaged time courses across participants, separately for each SNR. P1, N1, and P2 amplitudes were calculated for each participant and condition as the mean amplitude in the 0.02 s time window centered on the peak latency. The P1-minus-N1 and P2-minus-N1 amplitude differences were calculated.

### Statistical analyses

All statistical analyses were carried out using MATLAB (MathWorks) and JASP software (JASP, 2023; version 0.18.3.0). Details about the specific tests used are provided in the relevant sections for each experiment.

## Experiment 1: Enhanced neural speech tracking due to minimal background babble

### Methods and materials

#### *Participants*

Twenty-two adults (median: 23.5 years; range: 18–35 years; 12 male, 9 female, 1 transgender) participated in Experiment 1.

### *Stimuli and procedures*

Participants listened to 24 stories in 6 blocks (4 stories per block). Three of the 24 stories were played under clear conditions (i.e., without background noise). Twelve-talker babble was added to the other 21 stories (Bilger, 1984; Bilger et al., 1984; Wilson et al., 2012b). Twelve-talker babble is a standardized masker in speech-in-noise tests (Bilger, 1984; Bilger et al., 1984) that simulates a crowded restaurant, while not permitting the identification of individual words in the masker (Mattys et al., 2012). The babble masker was added at SNRs ranging from +30 to –2 dB in 21 steps of 1.6 dB SNR. Speech in background babble at 15 to 30 dB SNR is highly intelligible (Holder et al., 2018; Spyridakou et al., 2020; Irsik et al., 2022). No difference in speech intelligibility during story listening has been found between clear speech and speech masked by a 12-talker babble at +12 dB SNR (Irsik et al., 2022). Intelligibility typically drops below 90% of correctly reported words for +7 dB and lower SNR levels (Irsik et al., 2022). Hence, listeners have no trouble understanding speech at the highest SNRs used in the current study. All speech stimuli were normalized to the same root-mean-square amplitude and presented at about 65 dB SPL. Participants listened to each story, and after each story rated gist understanding and answered comprehension questions. Stories were presented in random order. Assignment of speech-clarity levels (clear speech and SNRs) to specific stories was randomized across participants.

### *Analyses*

Behavioral data (comprehension accuracy, gist ratings), EEG prediction accuracy, and TRFs for the three clear stories were averaged. For the stories in babble, a sliding average across SNR levels was calculated for behavioral data, EEG prediction accuracy, and TRFs, such that data for three neighboring SNR levels were averaged. Averaging across three stories was calculated to reduce noise in the data and match the averaging of three stories for the clear condition. For TRFs, analyses focused on the P1-N1 and the P2-N1 amplitude differences. For the statistical analyses, the clear condition was compared to each SNR level (resulting from the sliding average) using a paired samples t-test. False discovery rate (FDR) was used to account for multiple comparisons (Benjamini and Hochberg, 1995; Genovese et al., 2002). In cases where the data indicated a breaking point in behavior or brain response as a function of SNR, an explorative piece-wise regression (broken-stick analysis) with two linear pieces was calculated (McZgee and Carleton, 1970; Vieth, 1989; Toms and Lesperance, 2003). Identification of the breaking point and two pieces was calculated on the across-participant average as the minimum root mean squared error. A linear function was then fit

to each participant's data as a function of SNR and the estimated slope was tested against zero using a one-sample t-test, separately for each of the two pieces.

## Results and discussion

Comprehension accuracy and gist ratings did not differ significantly between clear speech and any of the SNR levels ($p_{FDR} > 0.05$; Figure 1A). Because gist ratings appeared to change somewhat with SNR (Figure 1A, right), an explorative piece-wise regression was calculated. The piece-wise regression revealed a breaking point at +15.6 dB SNR, such gist ratings decreased for +15.6 dB SNR and lower ($t_{21} = 3.008$, $p = 0.007$, $d = 0.641$; right to left in Figure 1A), whereas gist ratings did not linearly change for +15.6 dB SNR and above ($t_{21} = 0.214$, $p = 0.832$, $d = 0.046$).

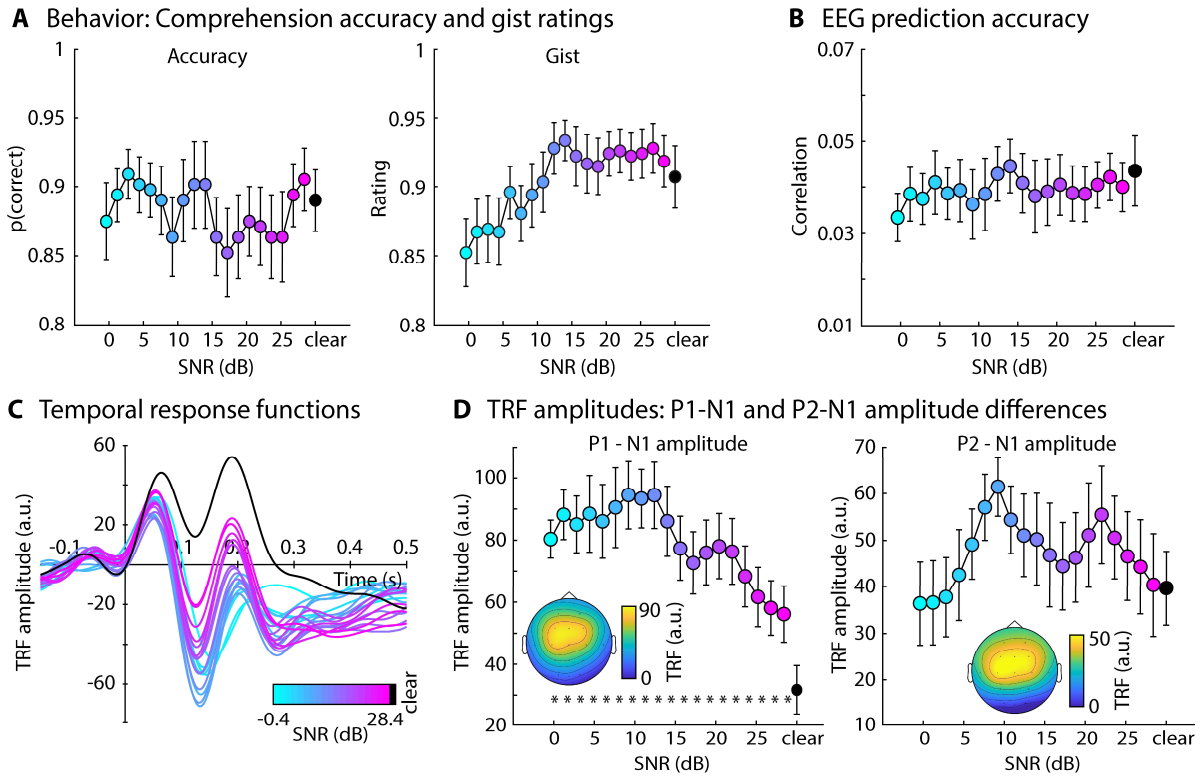


**Figure 1: Results for Experiment 1. A:** Accuracy of story comprehension (left) and gist ratings (right). **B:** EEG prediction accuracy. **C:** Temporal response functions (TRFs). **D:** P1-N1 and P2-N1 amplitude difference for different speech-clarity conditions. Topographical distributions reflect the average across all speech-clarity conditions. The black asterisk close to the x-axis indicates a significant difference relative to the clear condition (FDR-thresholded). The absence of an asterisk indicates that there was no significant difference. Error bars reflect the standard error of the mean.

EEG prediction accuracy did not differ between clear speech and any SNR level ($p_{FDR} > 0.05$; Figure 1B). In contrast, the P1-N1 amplitude of the TRF was significantly greater for all SNR

levels relative to clear speech ($p_{FDR} \leq 0.05$; Figure 1C, D). Explorative piece-wise regression revealed a breaking point at +9.2 dB SNR, showing a significant linear increase in P1-N1 amplitude from +28.4 dB to +9.2 dB SNR (right to left in Figure 1D; $t_{21} = -5.131$, $p = 4.4 \cdot 10^{-5}$, $d = 1.094$), whereas no significant trend was observed for SNRs from about +9.2 dB to -0.4 dB SNR ($t_{21} = 1.001$, $p = 0.328$, $d = 0.214$). No differences were found for the P2-N1 amplitude ($p_{FDR} > 0.05$; Figure 1C, D).

Experiment 1 replicates the previously observed enhancement in the neural tracking of the speech onset-envelope for speech presented in moderate background babble (Yasmin et al., 2023; Panela et al., 2024). Critically, Experiment 1 expands the previous work by showing that background babble also leads to enhanced neural speech tracking for very minimal background masking levels (~30 dB SNR). The noise-related enhancement at moderate babble levels has previously been interpreted to result from increased attention or effort to understand speech (Yasmin et al., 2023; Panela et al., 2024). However, speech for the very high SNR levels (>15 dB SNR) used in Experiment 1 is highly intelligible and thus should require little or no effort to understand. It is therefore unlikely that increased attention or effort drive the increase in neural speech tracking in background babble. Nonetheless, participants attended to the speech in Experiment 1, and it can thus not be fully excluded that attention investment played a role in the noise-related enhancement. To investigate the role of attention further, participants in Experiment 2 listened to stories under the same speech-clarity conditions as for Experiment 1 while performing a visual task and ignoring the stories.

## Experiment 2: Noise-related enhancement in speech tracking is unrelated to attention

### Methods and materials

#### *Participants*

Twenty-two adults participated in Experiment 2 (median: 23 years; range: 18–31 years; 11 male, 10 female, 1 transgender). Data from 5 additional participants were recorded but excluded from the analysis because they performed below 60% in the visual n-back task that was used to distract participants from listening to the concurrently presented speech. A low performance could mean that the participants did not fully attend to the visual task and instead attended to the spoken speech. To avoid this possibility, data from low performers were excluded.

### *Stimuli and procedure*

A new set of 24 stories was generated and participants were presented with 4 stories in each of 6 blocks. Speech-clarity levels were the same as in Experiment 1 (i.e., clear speech and SNRs ranging from +30 to –2 dB SNR). In Experiment 2, participants were instructed to ignore the stories and instead perform a visual 1-back task. In no part of the experiment were participants instructed to attend to the speech.

For the visual 1-back task, images of white digits (0 to 9) on black background were taken from the MNIST Handwritten Digit Classification Dataset (Deng, 2012). The digit images were selected, because different images of the same digit differ visually and thus make it challenging to use a simple feature-matching strategy to solve the 1-back task. A new digit image was presented every 0.5 seconds throughout the time over which a story was played (1:30 to 2:30 min). A digit image was presented for 0.25 s followed by a 0.25 s black screen before the next image was presented. The continuous stream of digits contained a digit repetition (albeit a different image sample) every 6 to 12 digits. Participants were tasked with pressing a button on a keyboard as soon as they detected a repetition. We did not include comprehension questions or gist ratings in Experiment 2 to avoid that participants feel they should pay attention to the speech materials. Hit rate and response time were used as behavioral measures.

### *Analyses*

Analyses examining the effects of SNR in Experiment 2 were similar to the analyses in Experiment 1. Behavioral data (hit rate, response time in 1-back task) and EEG data (prediction accuracy, TRFs) for the three clear stories were averaged and a sliding average procedure across SNRs (three neighbors) was used for stories in babble. Statistical tests compared SNR conditions to the clear condition, including FDR-thresholding. An explorative piece-wise regression with two linear pieces was calculated in cases where the data indicated a breaking point in behavior or brain response as a function of SNR.

## Results

The behavioral results showed no significant differences for hit rate or response time in the visual 1-back task between the clear condition and any of the SNR levels ($p_{FDR} > 0.05$; Figure 2A).

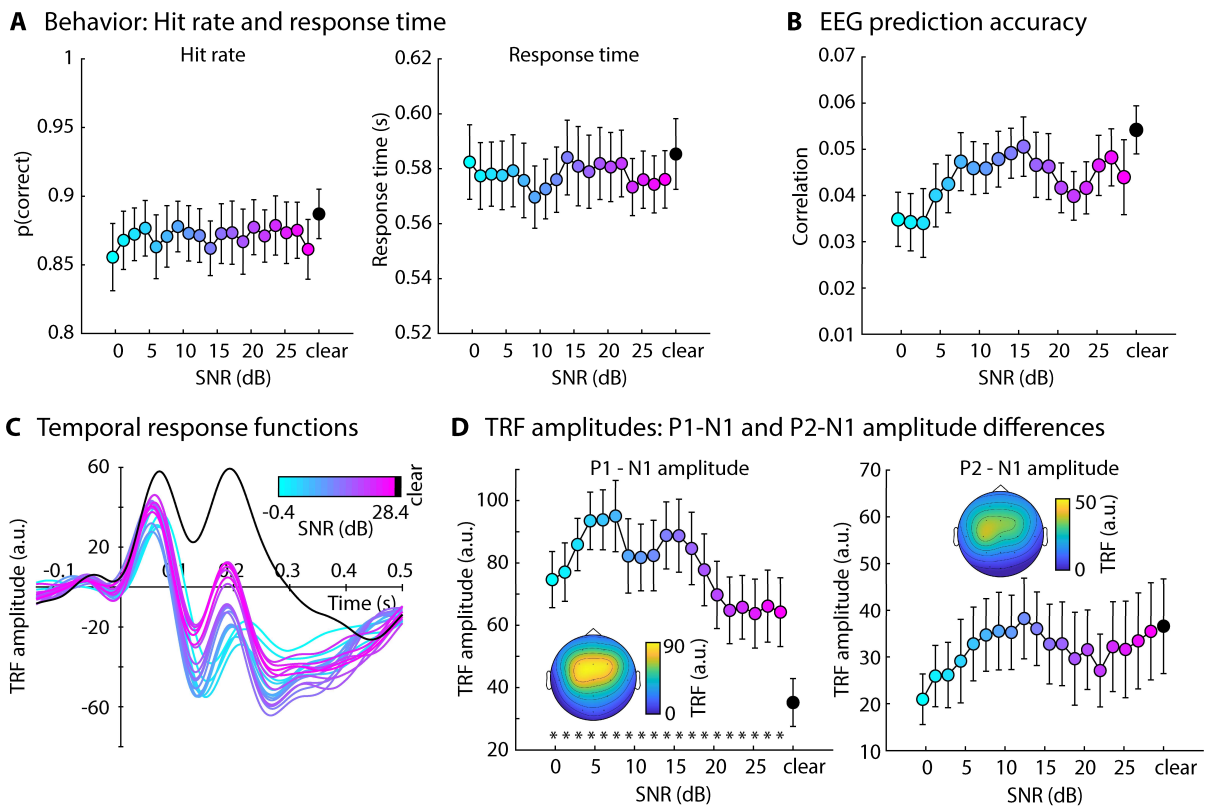


**Figure 2: Results for Experiment 2. A:** Hit rate (left) and response times (right) for the visual n-back task. **B:** EEG prediction accuracy. **C:** Temporal response functions (TRFs). **D:** P1-N1 and P2-N1 amplitude difference for different speech-clarity conditions. Topographical distributions reflect the average across all speech-clarity conditions. The black asterisk close to the x-axis indicates a significant difference relative to the clear condition (FDR-thresholded). The absence of an asterisk indicates that there was no significant difference. Error bars reflect the standard error of the mean.

EEG prediction accuracy did not differ between clear speech and any SNR level ($p_{FDR}$ > 0.05; Figure 3B). In contrast, and similar to Experiment 1, the P1-N1 amplitude of the TRF was significantly greater for all SNR levels relative to clear speech ($p_{FDR} \leq 0.05$; Figure 3C, D). An explorative piece-wise regression revealed a breaking point at +4.4 dB SNR, showing a significant linear increase in P1-N1 amplitude from +28.4 dB to +4.4 dB SNR ($t_{21}$ = -3.506, p = 0.002, d = 0.747; right to left in Figure 3D), whereas the P1-N1 amplitude decreased from +4.4 dB to -0.4 dB SNR ($t_{21}$ = 2.416, p = 0.025, d = 0.515). No differences were found for the P2-N1 amplitude ($p_{FDR}$ > 0.05; Figure 3C, D)

The results of Experiment 2 show that, under diverted attention, neural speech tracking is enhanced for speech presented in background babble at very high SNR levels (~30 dB SNR) relative to clear speech. Because participants did not pay attention to the speech in Experiment 2, the results indicate that attention is unlikely to drive the masker-related enhancement in neural tracking

observed here and previously (Yasmin et al., 2023; Panela et al., 2024). The enhancement may thus be exogenously rather endogenously driven. Previous work suggests the type of masker may play an important role in whether speech tracking is enhanced by background sound. A masker-related enhancement was reported for a 12-talker babble masker (Yasmin et al., 2023; Panela et al., 2024), whereas the effect was absent, or relatively small, for a stationary noise that spectrally matched the speech signal (Ding and Simon, 2013; Zou et al., 2019; Synigal et al., 2023). Sound normalization also differed. The work observing the enhancement normalized all SNR conditions to the same overall amplitude (Yasmin et al., 2023; Panela et al., 2024). This leads to a reduction in the speech-signal amplitude as SNR decreases, which, in fact, works against the observation that neural speech tracking is enhanced for speech in babble. In the other studies, the level of the speech signal was kept the same across SNRs and background noise was added at across SNRs (Ding and Simon, 2013; Synigal et al., 2023). Using this normalization approach, the overall amplitude of the sound mixture increases with decreasing SNR. Experiment 3 was conducted to disentangle these different potential contributions to the masker-related enhancement in neural speech tracking.

## Experiment 3: Masker-related enhancement of neural tracking is greater for babble than speech-matched noise

### Methods and materials

#### *Participants*

Twenty-three people (median: 25 years; range: 19–33 years; 7 male, 15 female, 1 transgender) participated in Experiment 3. Data from one additional person were recorded, but due to technical issues no triggers were recorded. The data could thus not be analyzed and were excluded.

#### *Stimuli and procedures*

A new set of 24 stories, comprehension questions, and multiple-choice options were generated. Participants listened to 4 stories in each of 6 blocks. Four of the 24 stories were presented under clear conditions. Ten stories were masked by 12-talker babble at 5 different SNRs (two stories each: +5, +10, +15, +20, +25 dB SNR), whereas the other 10 stories were masked by a stationary noise that spectrally matched the speech signal at 5 different SNRs (two stories each: +5, +10, +15, +20, +25 dB SNR). To obtain the spectrally matched noise, the long-term spectrum of the clear speech

signal was calculated using a fast Fourier transform (FFT). The inverted FFT was calculated using the frequency-specific amplitudes estimated from the speech signal jointly with randomly selected phases for each frequency. Figure 4 shows a time course snippet, a power spectrum, and a short snippet of the spectrogram for clear speech, speech masked by babble, and speech masked by the spectrally matched noise. Twelve-talker babble and speech-matched noise spectrally overlap extensively (Figure 4), but 12-talker babble recognizably contains speech (although individual elements cannot be identified), whereas the speech-matched noise does not contain recognizable speech elements.

For five of the 10 stories per masker type (babble, noise), the speech-plus-masker sound signal (mixed at a specific SNR) was normalized to the same root-mean-square (RMS) amplitude as the clear speech stories. This normalization results in a decreasing level of the speech signal in the speech-plus-masker mixture as SNR decreases. For the other five stories, the RMS amplitude of the speech signal was kept the same for all stories and a masker was added at the specific SNRs. This normalization results in an increasing sound level (RMS) of the sound mixture as SNR decreases. In other words, the speech signal in the sound mixture is played at a slightly lower intensity for the former than for the latter normalization type. We thus refer to this manipulation as Speech Level with lower and higher levels for the two normalization types, respectively. Please note that these differences in speech level are very minor due to the high SNRs used here. Stories were presented in randomized order and the assignment of stories to conditions was randomized across participants.

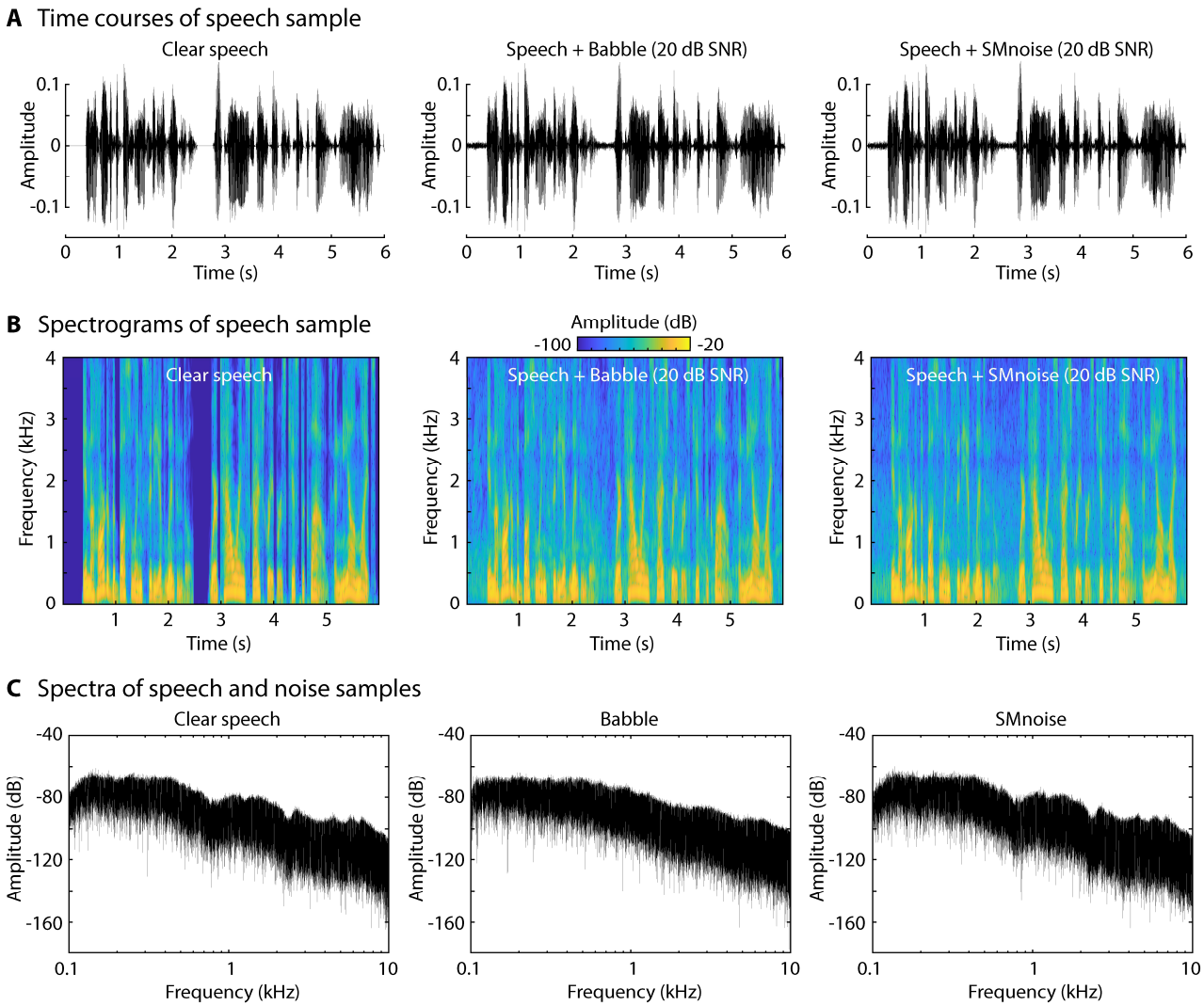


**Figure 3: Depiction of stimulus samples. A:** Time courses for clear speech and speech to which background babble of speech-matched noise was added at 20 dB SNR (all sound mixtures were normalized to the same root-mean-square amplitude). The first 6 s of a story are shown. **B:** Spectrograms of the samples in Panel A. **C:** Power spectra for clear speech, babble, and speech-matched noise. In C, only background babble/noise is displayed, without added speech.

### *Analysis*

Behavioral data (comprehension accuracy, gist ratings), EEG prediction accuracy, and TRFs for the four clear stories were averaged. For the stories in babble and speech-matched noise, a sliding average across SNR levels was calculated for behavioral data, EEG prediction accuracy, and TRFs, such that data for four neighboring SNR levels were averaged, separately for the two masker types (babble, noise) and normalization types (adjusted speech level, non-adjusted speech level). Averaging across four stories was calculated to reduce noise in the data and match the number of

stories included in the average for the clear condition. For TRFs, analyses focused on the P1-N1 and the P2-N1 amplitude differences. For the statistical analyses, the clear condition was compared to each SNR level (resulting from the sliding average) using a paired samples t-test. False discovery rate (FDR) was used to account for multiple comparisons (Benjamini and Hochberg, 1995; Genovese et al., 2002). Differences between masker types and normalization types were examined using a repeated measures analysis of variance (rmANOVA) with the within-participant factors SNR (10, 15, 20 dB SNR), Masker Type (babble, speech-match noise), and Normalization Type (lower vs higher speech levels). Post hoc tests were calculated for significant main effects and interactions. Holm's methods was used to correct for multiple comparisons (Holm, 1979). Statistical analyses were carried out in JASP software (v0.18.3; JASP, 2023). Note that JASP uses pooled error terms and degrees of freedom from an rmANOVA for the corresponding post hoc effects. The reported degrees of freedom are thus higher than for direct contrasts had they been calculated independently from the rmANOVA.

## Results

The analysis of behavioral performance revealed significantly lower comprehension performance and gist ratings, compared to clear speech, for babble-masked speech that was normalized such that the speech level was slightly lower than for clear speech (Figure 4A). The rmANOVA for the proportion of correctly answered comprehension questions did not reveal any effects or interactions (for all $p > 0.1$). The rmANOVA for gist ratings revealed an effect of Normalization Type ($F_{1,22} = 4.300$, $p = 0.05$, $\omega^2 = 0.008$), showing higher gist ratings when the speech signal had a 'higher' compared to a 'lower' intensity, whereas the other effects and interactions were not significant (for all $p > 0.05$).

The analysis of EEG prediction accuracy revealed no differences between clear speech and any of the masked speech (for all $p_{FDR} > 0.05$; Figure 4B). The rmANOVA did not reveal any significant effects nor interactions (for all $p > 0.25$).

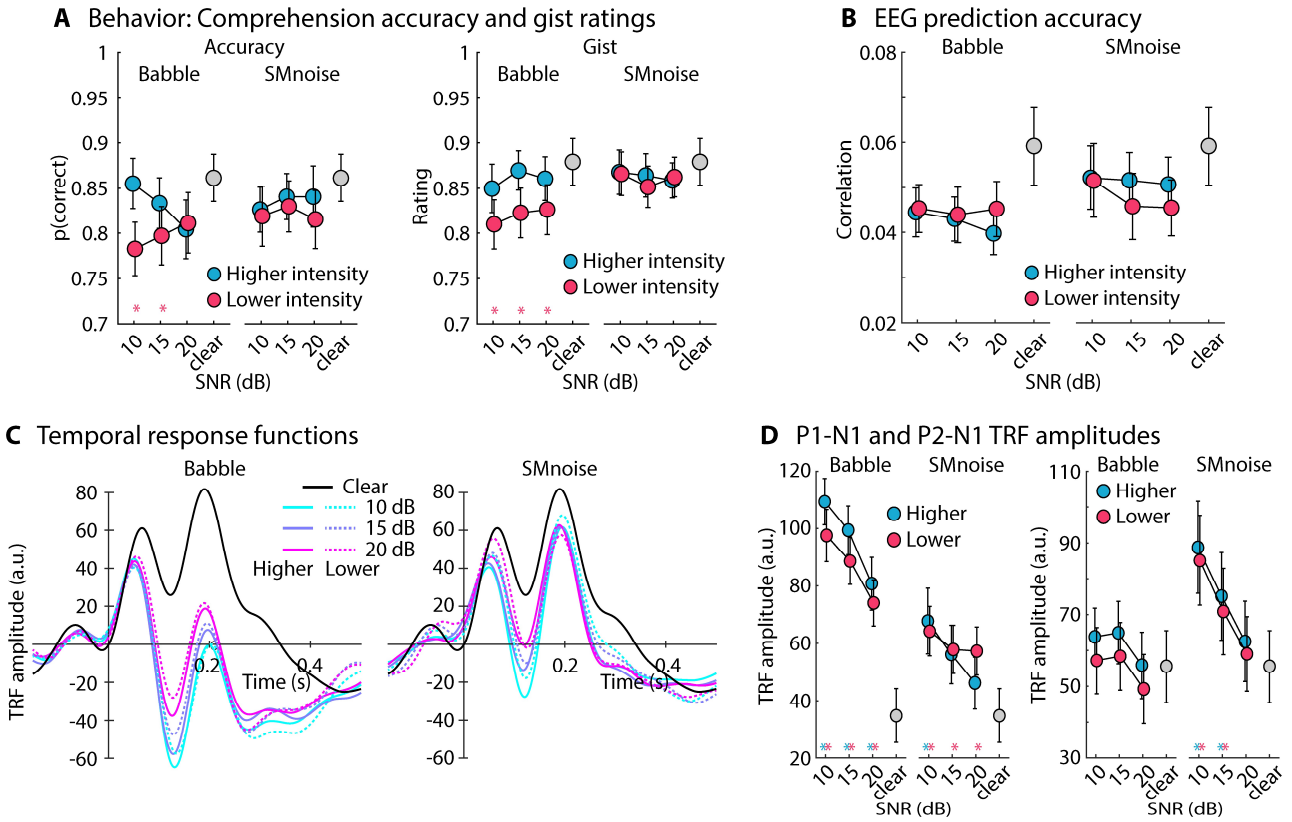


**Figure 4: Results for Experiment 3. A:** Accuracy of story comprehension (left) and gist ratings (right). Higher vs. lower intensity refers to the two sound-level normalization types, one resulting a slightly lower intensity of the speech signal in the sound mixture than the other. **B:** EEG prediction accuracy. **C:** Temporal response functions (TRFs). **D:** P1-N1 and P2-N1 amplitude difference for clear speech and different speech-masking and sound normalization conditions. In panels A, B, and D, a colored asterisk close to the x-axis indicates a significant difference relative to the clear condition (FDR-thresholded). The specific color of the asterisk – blue vs red – indicates the normalization type (higher vs lower speech level, respectively). The absence of an asterisk indicates that there was no significant difference relative to clear speech. Error bars reflect the standard error of the mean.

For the analysis of the TRF revealed the following results. For the babble masker, the P1-N1 amplitudes were larger for all SNR levels compared to clear speech, for both sound-level normalization types (for all $p_{FDR} \leq 0.05$). For the speech-matched noise, P1-N1 amplitudes were larger for all SNR levels compared to clear speech for the normalization resulting in lower speech intensity (for all $p_{FDR} \leq 0.05$), but only for 10 dB SNR for the normalization resulting in higher speech intensity (15 dB SNR was significant for an uncorrected t-tests). The rmANOVA revealed larger P1-N1 amplitudes for the 12-talker babble compared to the speech-matched noise masker (effect of Masker Type: $F_{1,22} = 32.849$, $p = 9.1 \cdot 10^{-6}$, $\omega^2 = 0.162$) and larger amplitudes for lower

SNRs (effect of SNR: $F_{2,44} = 22.608$, $p = 1.8 \cdot 10^{-7}$, $\omega^2 = 0.049$). None of the interactions nor the effect of Normalization Type were significant (for all $p > 0.05$).

Analysis of the P2-N1 revealed larger amplitudes for speech masked by speech-matched noise at 10 dB and 15 dB SNR, for both normalization types (for all $p_{FDR} \leq 0.05$). None of the other masked speech conditions differed from clear speech. The rmANOVA revealed an effect of SNR ($F_{2,44} = 26.851$, $p = 2.7 \cdot 10^{-8}$, $\omega^2 = 0.024$), Masker Type ($F_{1,22} = 5.859$, $p = 0.024$, $\omega^2 = 0.023$), and a SNR × Masker Type interaction ($F_{2,44} = 7.684$, $p = 0.001$, $\omega^2 = 0.007$). The interaction was due to an increase in P2-N1 amplitude with decreasing SNR for the speech-matched noise (all $p_{Holm} \leq 0.05$), whereas P2-N1 amplitudes for babble did not differ significantly between SNR conditions (all $p_{Holm} > 0.05$).

The results of Experiment 3 show that neural speech tracking increases for babble and speech-matched noise maskers compared to clear speech, but that the 12-talker babble masker leads to a greater enhancement compared to the speech-matched noise. Slight variations in the level of the speech signal in the sound mixture (resulting from different sound-level normalization procedures) do not seem to overly impact the results. Because Experiment 3 indicates that the type of background noise may affect the degree of masker-related enhancement, we conducted Experiment 4 to investigate whether different types of commonly used noises lead to similar enhancements in neural speech tracking.

## Experiment 4: Neural-tracking enhancements generalize across different masker types

### Methods and materials

#### *Participants*

Twenty individuals participated in Experiment 4 (median: 25.5 years; range: 19–34 years; 4 male, 15 female, 1 transgender). Data from one additional person were recorded, but due to technical issues no triggers were recorded. The data could thus not be analyzed and were excluded.

#### *Stimuli and procedures*

A new set of 24 stories, corresponding comprehension questions, and multiple-choice options were generated. Participants listened to 4 stories in each of 6 blocks. After each story, they answered 5 comprehension questions and rated gist understanding. Three stories were presented in each of 8

conditions: clear speech (no masker), speech with added white noise, pink noise, stationary noise that spectrally matched the speech signal (Experiment 3), 12-talker babble (Experiments 1-3), and three additional 12-talker babbles. The additional babble maskers were created to ensure there is nothing specific about the babble masker used in our previous work (Yasmin et al., 2023; Panela et al., 2024) and in Experiments 1-3 that could lead to an enhanced tracking response and to vary spectral properties of the babble associated with different voice genders (male, female). For the three additional babble maskers, 24 text excerpts of about 800 words each were taken from Wikipedia (e.g., about flowers, forests, insurance, etc.). The 24 text excerpts were fed into Google's AI speech synthesizer to generate 24 continuous speech materials (~5 min) of which 12 were from male voices and 12 from female voices. The three 12-talker babble maskers were created by adding speech from 6 male and 6 female voices (mixed gender 12-talker babble), speech from the 12 male voices (male gender 12-talker babble), and speech from the 12 female voices (female gender 12-talker babble). Maskers were added to the speech signal at 20 dB SNR and all acoustic stimuli were normalized to the same root-mean-square amplitude. A power spectrum for each of the 7 masker types is displayed in Figure 5. Stories were presented in randomized order and assignment of stories to the 8 different conditions were randomized across participants.

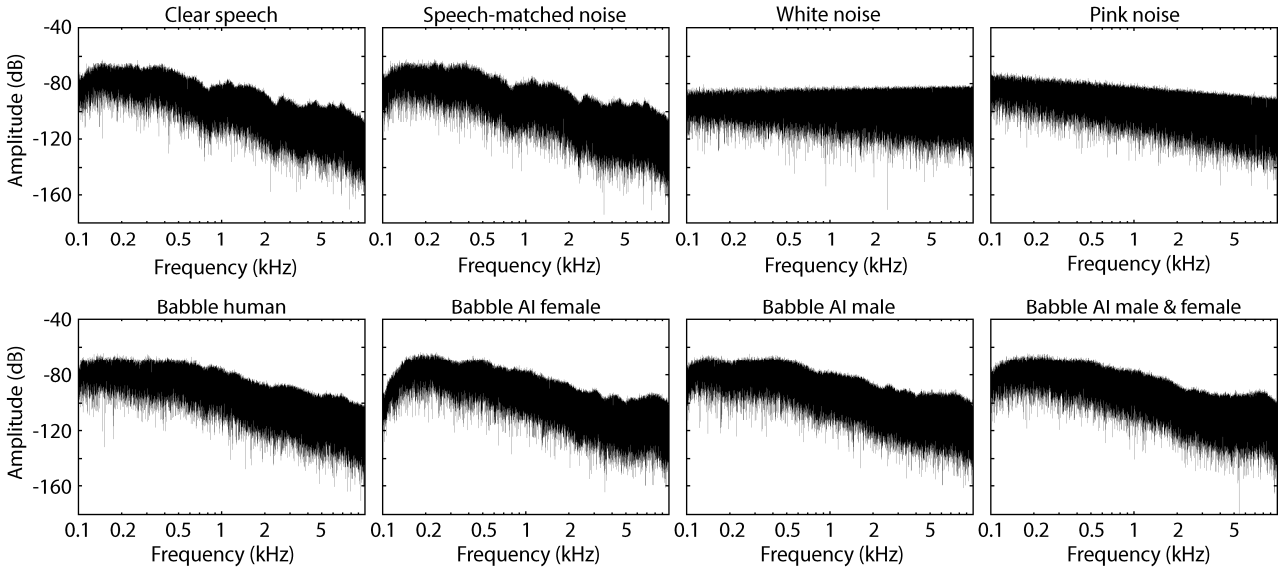


**Figure 5: Spectra for clear speech and different background noises.**

### *Analysis*

Behavioral data (comprehension accuracy, gist ratings), EEG prediction accuracy, and TRFs were averaged across the three stories for each condition. For TRFs, analyses focused on the P1-N1 and

the P2-N1 amplitude differences. For the statistical analyses, the clear condition was compared to the masker conditions using a paired samples t-test. FDR was used to account for multiple comparisons (Benjamini and Hochberg, 1995; Genovese et al., 2002). Differences between the four different babble maskers, between babble and noise maskers, and between the three noise maskers were also investigated using paired samples t-tests.

## Results and discussion

Comprehension accuracy and gist ratings for the clear story did not significantly differ from the data for masked speech (for all $p_{FDR} > 0.05$; Figure 6A).

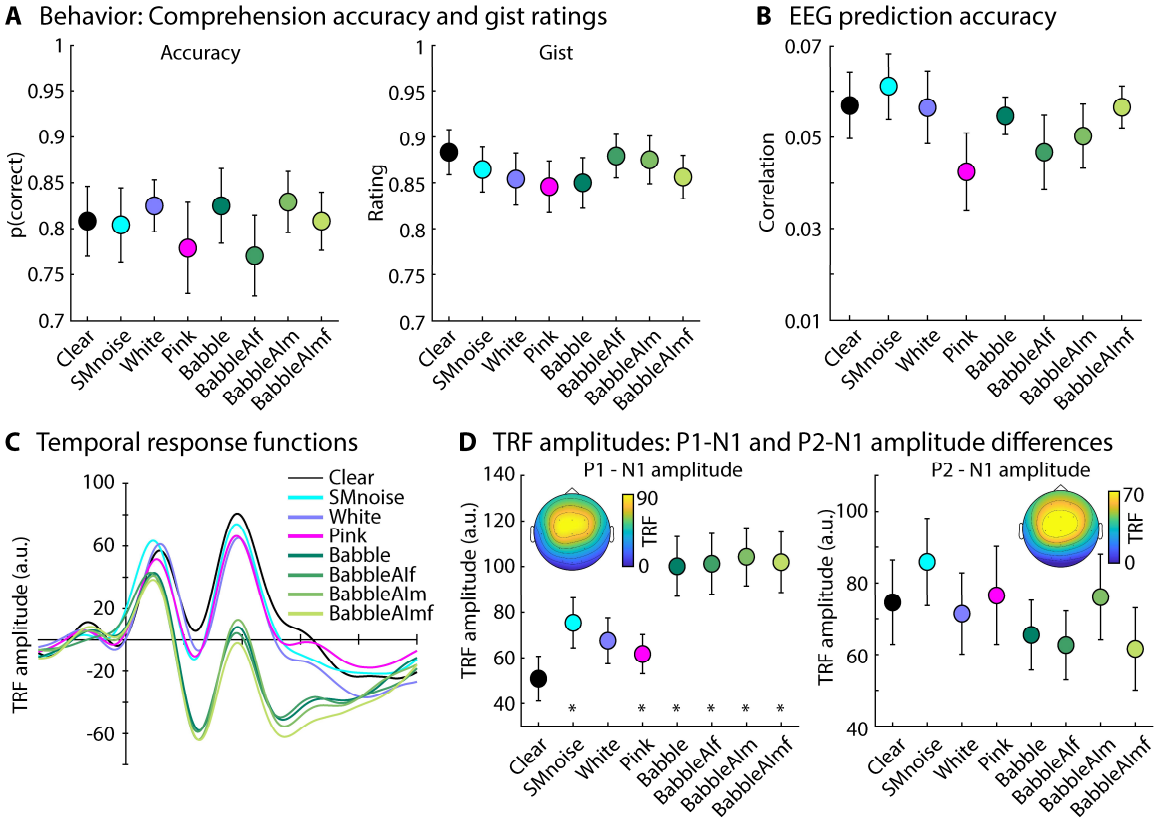


**Figure 6: Results for Experiment 4. A:** Accuracy of story comprehension (left) and gist ratings (right). **B:** EEG prediction accuracy. **C:** Temporal response functions (TRFs). **D:** P1-N1 and P2-N1 amplitude difference for clear speech and different speech-masking conditions. Topographical distributions reflect the average across all conditions. In panels A, B, and D, the black asterisk close to the x-axis indicates a significant difference relative to the clear condition (FDR-thresholded). The absence of an asterisk indicates that there was no significant difference relative to clear speech. Error bars reflect the standard error of the mean.

EEG prediction accuracy did not significantly differ between clear speech and any of the masker types (for all $p_{FDR} > 0.05$; Figure 6B). In contrast, the TRF P1-N1 amplitude was larger for

all masker types, expect for white noise, compared to clear speech (for all $p_{FDR} \leq 0.05$; Figure 6D; the difference between clear speech and speech masked by white noise was significant when uncorrected; $p = 0.05$). There were no differences among the four different babble maskers (for all $p > 0.6$), indicating that different voice genders of the 12-talker babble do not differentially affect the masker-related enhancement in neural speech tracking. However, the P1-N1 amplitude was larger for speech masked by the babble maskers (collapsed across the four babble maskers) compared to speech masked white noise ($t_{19} = 4.133$, $p = 5.7 \cdot 10^{-4}$, $d = 0.68$), pink noise ($t_{19} = 5.355$, $p = 3.6 \cdot 10^{-5}$, $d = 1.197$), and the noise spectrally matched to speech ($t_{19} = 3.039$, $p = 0.007$, $d = 0.68$). There were no differences among the three noise maskers (for all $p > 0.05$). The P2-N1 amplitude for clear speech did not different from the P2-N1 amplitude for masked speech (for all $p_{FDR} > 0.05$; Figure 6D, right).

The results of Experiment 4 replicate the results from Experiments 1-3 by showing that babble noise at a high SNR (20 dB) increases neural speech tracking. Experiment 4 further shows that the neural-tracking enhancement generalizes across different noises, albeit a bit less for white noise (significant only when uncorrected for multiple comparisons). Results from Experiment 4 also replicate the larger tracking enhancement for speech in babble noise compared to speech in speech-matched noise observed in Experiment 3. Sounds in Experiments 1-4 were presented via headphones, which is comparable to previous work using headphones or in-ear phones (Alain et al., 2009; Alain et al., 2012; Ding and Simon, 2013; Alain et al., 2014; Broderick et al., 2018; Decruy et al., 2019; Tune et al., 2021; Synigal et al., 2023; Yasmin et al., 2023; Panela et al., 2024). However, headphones or in-ear phones attenuate external sound sources such that clear speech is arguably presented in 'unnatural' quiet. In everyday life, speech typically reaches our ears in free-field space. Experiment 5 examines whether the noise-related enhancement in neural speech tracking also generalizes to free-field listening.

## Experiment 5: Neural-tracking enhancement generalizes to free-field listening

### Methods and materials

#### *Participants*

Twenty-two adults participated in Experiment 5 (median: 26 years; range: 19–34 years; 10 male, 11 female, 1 transgender).

514 ***Stimuli and procedures***

515 A new set of 24 stories, comprehension questions, and multiple-choice options were generated.
516 Participants listened to 4 stories in each of 6 blocks. After each story, they answered 5
517 comprehension questions and rated gist understanding. For 3 of the 6 blocks, participants listened
518 to the stories through headphones, as for Experiments 1-4, whereas for the other 3 blocks,
519 participants listened to the stories via computer loudspeakers placed in front of them. Blocks for
520 different sound-delivery conditions (headphones, loudspeakers) alternated within each participant's
521 session, and the starting condition was counter-balanced across participants. For each sound-
522 delivery condition, participants listened to three stories each under clear conditions, +10 dB, +15
523 dB, and +20 dB SNR (12 talker babble, generated using Google's AI voices as described for
524 Experiment 4). Stories were distributed such that the four speech-clarity conditions were presented
525 in each block in randomized order. All acoustic stimuli were normalized to the same root-mean-
526 square amplitude. The sound level of the headphones and the sound level of the loudspeakers (at
527 the location of a participant's head) were matched.

528 ***Analysis***

529 Behavioral data (comprehension accuracy, gist ratings), EEG prediction accuracy, and TRFs were
530 averaged across the three stories for each speech-clarity and sound-delivery condition. For TRFs,
531 analyses focused on the P1-N1 and the P2-N1 amplitude differences. For the statistical analyses,
532 the clear condition was compared to the masker conditions using a paired samples t-test. FDR was
533 used to account for multiple comparisons (Benjamini and Hochberg, 1995; Genovese et al., 2002).
534 To test for differences between sound-delivery types, a rmANOVA was calculated, using the within-
535 participant factors SNR (10, 15, 20 dB SNR; clear speech was not included, because a difference to
536 clear speech was tested directly as just described) and Sound Delivery (headphones, loudspeakers).
537 Post hoc tests were calculated using dependent samples t-tests, and Holm's methods was used to
538 correct for multiple comparisons (Holm, 1979).

539 **Results and discussion**

540 Comprehension accuracy and gist ratings are shown in Figure 7A. There were no differences
541 between clear speech and speech masked by background babble for any of the conditions, with the
542 exception of a lower gist rating for the 20 dB SNR loudspeaker condition (Figure 7A, right).

For the EEG prediction accuracy, there were no differences between the clear conditions and any of the masked speech conditions (Figure 7B). The rmANOVA did not reveal effects or interactions involving SNR or Sound Delivery (ps > 0.15).

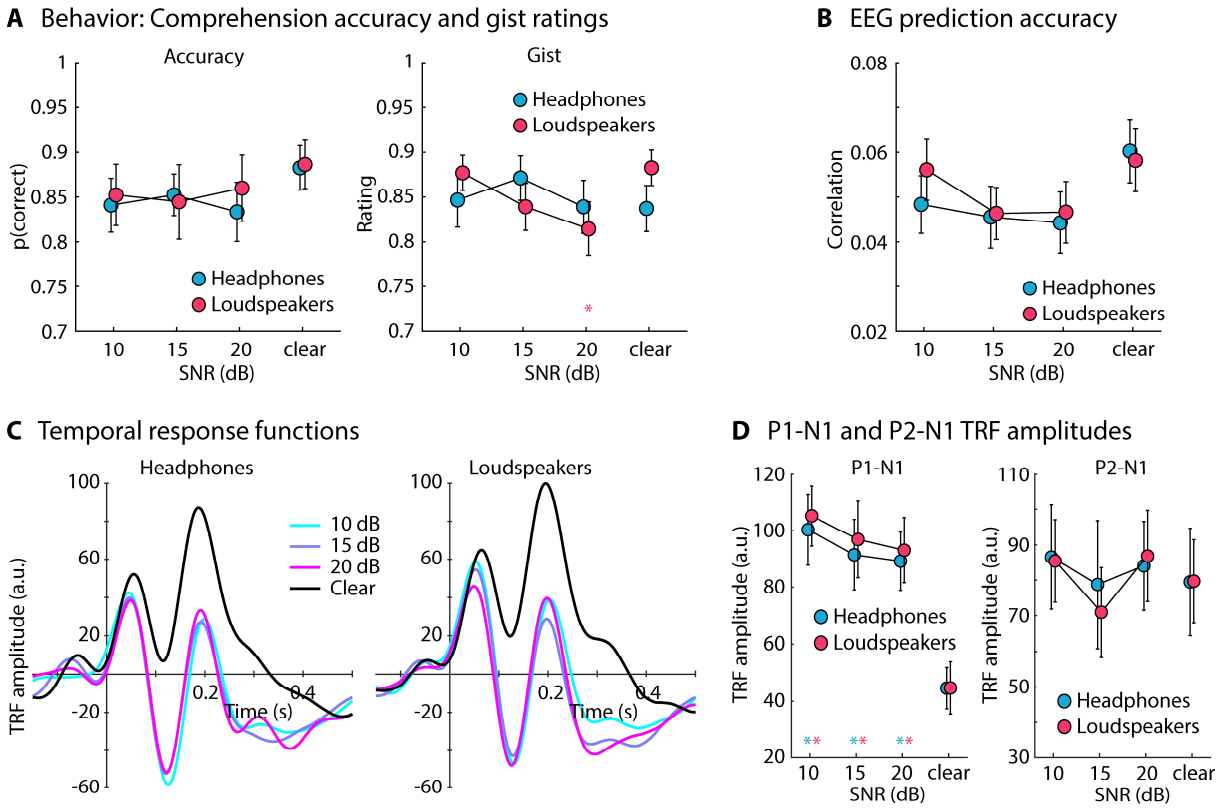


**Figure 7: Results for Experiment 5. A:** Accuracy of story comprehension (left) and gist ratings (right). **B:** EEG prediction accuracy. **C:** Temporal response functions (TRFs). **D:** P1-N1 and P2-N1 amplitude difference for clear speech and different speech-masking and sound-delivery conditions. In panels A, B, and D, a colored asterisk close to the x-axis indicates a significant difference relative to the clear condition (FDR-thresholded). The specific color of the asterisk – blue vs red – indicates the sound-delivery type. The absence of an asterisk indicates that there was no significant difference relative to clear speech. Error bars reflect the standard error of the mean.

The analysis of TRF amplitudes revealed significantly larger P1-N1 amplitudes for all masked speech conditions compared to clear speech (for all $p_{FDR} \leq 0.05$; Figure 7D, left). A rmANOVA did not reveal any effects or interaction involving SNR or Sound Delivery (ps > 0.1). For the P2-N1 amplitude, there were no differences between clear speech and any of the masked speech conditions (for all $p_{FDR} > 0.05$; Figure 7D, right). The rmANOVA revealed an effect of SNR ($F_{2,42} = 3.953$, $p = 0.027$, $\omega^2 = 0.005$), caused by the lower P2-N1 amplitude for the +15 dB SNR conditions compared to the +10 dB SNR ($t_{42} = 2.493$, $p_{Holm} = 0.05$, $d = 0.171$) and the +20 dB SNR

condition ($t_{42}$ = 2.372, $p_{Holm}$ = 0.05, d = 0.162). There was no effect of Sound Delivery nor a SNR × Sound Delivery interaction (ps > 0.5).

The results of Experiment 5 show that the enhanced neural tracking of speech associated with minor background babble is unspecific to delivering sounds via headphones (which typically attenuate sounds in the environment). Instead, minor background babble at +20 dB SNR also increased the neural tracking of speech under free-field (loudspeaker) conditions, thus pointing to the generalizable nature of the phenomenon to conditions more akin to naturalistic listening scenarios.

## Discussion

In 5 EEG experiments, the current study investigated the degree to which background masking sounds at high SNRs, for which speech is highly intelligible, affect neural speech tracking. The results show that 12-talker babble enhances neural tracking at very high SNRs (~30 dB) relative to clear speech (Experiment 1) and that this enhancement is present even when participants carry out an unrelated visual task (Experiment 2), suggesting that attention or effort do not cause the noise-related neural-tracking enhancement. The results further show that the enhancement of neural speech tracking is greater for speech in the presence of 12-talker babble compared to a stationary noise that spectrally matched the speech (Experiments 3 and 4), although both masker types spectrally overlap. The enhancement was also greater for 12-talker babble compared to pink noise and white noise (Experiment 4). Finally, the enhanced neural speech tracking generalizes from headphone to free-field listening (Experiment 5), pointing to the real-world nature of the tracking enhancement. Overall, the current study paints a clear picture of a highly generalized enhancement of neural speech tracking in the presence of minimal background noise, making links to speech intelligibility and listening challenges in noise challenging.

### Enhanced neural tracking of speech under minimal background noise

Across all five experiments of the current study, speech masked by background noise at high SNRs (up to 30 dB SNR) led to an enhanced neural tracking of the amplitude-onset envelope of speech. The enhancement was present for different background maskers, but most prominently for 12-talker babble. Previous work on neural speech tracking also observed enhanced neural tracking for speech masked by 12-talker babble at moderate SNRs (~12 dB; Yasmin et al., 2023; Panela et al., 2024), consistent with the current study. The current results are also consistent with studies showing a

noise-related enhancement to tone bursts (Alain et al., 2009; Alain et al., 2012; Alain et al., 2014), syllable onsets (Parbery-Clark et al., 2011), and high-frequency temporal modulations in sounds (Ward et al., 2010; Shukla and Bidelman, 2021). Other work, using a noise masker that spectrally matches the target speech, have not reported tracking enhancements (Ding and Simon, 2013; Zou et al., 2019; Synigal et al., 2023). However, in these works, SNRs have been lower (<10 dB) to investigate neural tracking under challenging listening conditions. At low SNRs, neural speech tracking decreases (Ding and Simon, 2013; Zou et al., 2019; Yasmin et al., 2023; Figures 1 and 2), thus resulting in an inverted u-shape in relation to SNR for attentive and passive listening (Experiments 1 and 2). Moreover, the speech-tracking enhancement was smaller for speech-matched noise compared to babble noise (Figures 4 and 6), potentially explaining the absence of the enhancement for speech-matched noise at low SNRs in previous work (Ding and Simon, 2013; Zou et al., 2019; Synigal et al., 2023).

The noise-related enhancement in the neural tracking of the speech envelope was greatest for 12-talker babble, but it was also present for speech-matched noise, pink noise, and, to some extent, white noise. The latter three noises bare no perceptional relation to speech, but resemble stationary, background buzzing from industrial noise, heavy rain, waterfalls, wind, or ventilation. Twelve-talker babble – which is also a stationary masker – is clearly recognizable as overlapping speech, but words or phonemes cannot be identified (Bilger, 1984; Bilger et al., 1984; Wilson, 2003; Wilson et al., 2012b). There may thus be something about the naturalistic, speech nature of the background babble that facilitates neural speech tracking.

The spectral power for both the 12-talker babble and the speech-matched noise overlaps strongly with the spectral properties of the speech signal, although the speech-matched noise most closely resembles the spectrum of speech. Spectral overlap of the background sound and the speech signal could cause energetic masking in the auditory periphery and degrade accurate neural speech representations (Brungart et al., 2006; Mattys et al., 2012; Wilson et al., 2012a; Kidd et al., 2019). Although peripheral masking would not explain why neural speech tracking is enhanced in the first place, more peripheral masking for the speech-matched noise compared to the 12-talker babble would be consistent with a reduced enhancement for the former compared to the latter masker. However, pink noise and white noise overlap spectrally much less with speech than the other two background maskers, but the neural tracking enhancement did not differ between the speech-matched noise, pink noise, and white noise maskers. This again suggests that there may be something about the speech-nature of the babble masker that drives the larger neural tracking enhancement.

Critically, the current results have implications for research and clinical applications. The neural tracking of the speech envelope has been linked to speech intelligibility (Ding et al., 2014; Vanthornhout et al., 2018; Lesenfants et al., 2019) and has been proposed to be a useful clinical biomarker for speech encoding in the brain (Dial et al., 2021; Gillis et al., 2022; Schmitt et al., 2022; Kries et al., 2024; Panela et al., 2024). However, speech intelligibility, assessed here via gist ratings (Davis and Johnsrude, 2003; Ritz et al., 2022), only declines for SNRs below 15 dB SNR (consistent with intelligibility scores; Irsik et al., 2022), whereas the neural tracking enhancement is already present for ~30 dB SNR. This result questions the link between neural envelope tracking and speech intelligibility, or at least makes the relationship non-linear (cf. Yasmin et al., 2023). Research on the neural tracking of speech using background noise must thus consider that the noise itself may enhance the tracking.

**Potential mechanisms associated with enhanced neural speech tracking**

Enhanced neural tracking associated with a stationary background masker or noise-vocoded speech has been interpreted to reflect an attentional gain when listeners must invest cognitively to understand speech (Hauswald et al., 2022; Yasmin et al., 2023; Panela et al., 2024). However, these works used moderate to low SNRs or moderate speech degradations, making it challenging to distinguish between attentional mechanisms and mechanisms driven by background noise per se. The current study demonstrates that attention unlikely causes the enhanced neural tracking of the speech envelope. First, the tracking enhancement was observed for very high SNRs (~30 dB) at which speech is highly intelligible (Holder et al., 2018; Spyridakou et al., 2020; Irsik et al., 2022). Arguably, little or no effort is needed to understand speech at ~30 dB SNR, making attentional gain an unlikely explanation. Importantly, neural speech tracking was enhanced even when participants performed a visual task and passively listened to speech (Figure 2). Participants are unlikely to invest effort to understand speech when performing an attention-demanding visual task. Taken together, the current study provides little evidence that noise-related enhancements of neural speech tracking are due to attention or effort investment.

Another possibility put forward in the context of enhanced neural responses to tone bursts in noise (Alain et al., 2009; Alain et al., 2012; Alain et al., 2014) is that background noise increases arousal, which, in turn, amplifies the neural response to sound. However, a few pieces of evidence are inconsistent with this hypothesis. Arousal to minimal background noise habituates quickly (Alvar and Francis, 2024) and arousal does not appear to affect early sensory responses but rather later, non-sensory responses in EEG (>150 ms; Han et al., 2013). Moreover, pupil dilation – a

measure of arousal (Mathôt, 2018; Joshi and Gold, 2020; Burlingham et al., 2022) – is similar for speech in noise at SNRs ranging from +16 dB to +4 dB SNR (Ohlenforst et al., 2017; Ohlenforst et al., 2018), for which neural tracking increases (Figure 1). Hence, arousal is unlikely to explain the noise-related enhancement in neural speech tracking, but more direct research is needed to clarify this further.

A third potential explanation of enhanced neural tracking is stochastic resonance, reflecting an automatic mechanism in neural circuits (Stein et al., 2005; McDonnell and Abbott, 2009; McDonnell and Ward, 2011). Stochastic resonance has been described as the facilitation of a near-threshold input through background noise. That is, a near-threshold stimulus or neuronal input, that alone may not be sufficient to drive a neuron beyond its firing threshold, can lead to neuronal firing if noise is added, because the noise increases the stimulus or neuronal input for brief periods, causing a response in downstream neurons (Ward et al., 2002; Moss et al., 2004). However, the term is now used more broadly to describe any phenomenon where the presence of noise in a nonlinear system improves the quality of the output signal than when noise is absent (McDonnell and Abbott, 2009). Stochastic resonance has been observed in humans in several domains, such as in tactile, visual, and auditory perception (Kitajo et al., 2003; Wells et al., 2005; Tabarelli et al., 2009, but see Rufener et al., 2020). In the current study, speech was presented at suprathreshold levels, but stochastic resonance may still play a role at the neuronal level (Stocks, 2000; McDonnell and Abbott, 2009). EEG signals reflect the synchronized activity of more than 10,000 neurons (Niedermeyer and da Silva, 2005). Some neurons may not receive sufficiently strong input to elicit a response when a person listens to clear speech but may be pushed beyond their firing threshold by the additional, acoustically elicited noise in the neural system. Twelve-talker babble was associated with the greatest noise-related enhancement in neural tracking, possibly because the 12-talker babble facilitated neuronal activity in speech-relevant auditory regions, where the other, non-speech noises were less effective.

## Conclusions

The current study provides a comprehensive account of a generalized increase in the neural tracking of the amplitude-onset envelope of speech due to minimal background noise. The results show that a) neural speech tracking is enhanced for speech masked by background noise at very high SNRs (~30 dB), b) this enhancement is independent of attention, c) it generalizes across different stationary background maskers, although being strongest for 12-talker babble, and d) it is present

for headphone and free-field listening, suggesting the neural-tracking enhancement generalizes to real-life situations. The work paints a clear picture that minimal background noise enhances the neural representation of the speech envelope. The work further highlights the non-linearities of neural speech tracking as a function of background noise, challenging the feasibility of neural speech tracking as a biological marker for speech processing.

## Acknowledgements

We thank Priya Pandey, Tiffany Lao, and Saba Junaid for their help with data collection. The research was supported by the Canada Research Chair program (CRC-2019-00156) and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant: RGPIN-2021-02602).

## Authors contributions

**Björn Herrmann:** Conceptualization, methodology, formal analysis, investigation, data curation, writing - original draft, writing - review and editing, visualization, supervision, project administration, funding acquisition.

## Declaration of conflict of interest

None.

## References

Alain C, McDonald K, Van Roon P (2012) Effects of age and background noise on processing a mistuned harmonic in an otherwise periodic complex sound. Hearing Research 283:126-135.

Alain C, Roye A, Salloum C (2014) Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex. Frontiers in Systems Neuroscience 8:Article 8.

Alain C, Quan J, McDonald K, Van Roon P (2009) Noise-induced increase in human auditory evoked neuromagnetic fields. European Journal of Neuroscience 30:132-142.

Alvar A, Francis AL (2024) Effects of background noise on autonomic arousal (skin conductance level). JASA Express Letters 4:013601.

Auerbach BD, Rodrigues PV, Salvi RJ (2014) Central gain control in tinnitus and hyperacusis. Frontiers in Neurology 5:Article 206.

Bell AJ, Sejnowski TJ (1995) An information maximization approach to blind separation and blind deconvolution. Neural Computation 7:1129-1159.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57:289-300.

Biesmans W, Das N, Francart T, Bertrand A (2017) 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 25:402-412.

Bilger RC (1984) Manual for the clinical use of the revised SPIN Test. Champaign, IL, USA: The University of Illinois.

Bilger RC, Nuetzel JM, Rabinowitz WM, Rzeczkowski C (1984) Standardization of a Test of Speech Perception in Noise. Journal of Speech, Language, and Hearing Research 27:32-48.

Brodbeck C, Presacco A, Anderson S, Simon JZ (2018) Over-representation of speech in older adults originates from early response in higher order auditory cortex. Acta Acust United Acust 104:774-777.

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Current Biology 28:803-809.

Broderick MP, Di Liberto GM, Anderson AJ, Rofes A, Lalor EC (2021) Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. Scientific Reports 11:4963.

Brungart DS, Chang PS, Simpson BD, Wang D (2006) Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. The Journal of the Acoustical Society of America 120:4007-4018.

Burlingham CS, Mirbagheri S, Heeger DJ (2022) A unified model of the task-evoked pupil response. Science Advances 8:eabi9979.

Cohen SS, Parra LC (2016) Memorable Audiovisual Narratives Synchronize Sensory and Supramodal Neural Responses. eNeuro 3:e0203.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (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 MJ, Zuk NJ, Di Liberto GM, Nidiffer AR, Molholm S, Lalor EC (2021) Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research. Frontiers in Neuroscience 15.

Davis MH, Johnsrude IS (2003) Hierarchical Processing in Spoken Language Comprehension. The Journal of Neuroscience 23:3423-3431.

Decruy L, Vanthornhout J, Francart T (2019) Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. Journal of Neurophysiology 122:601-615.

Decruy L, Vanthornhout J, Francart T (2020) Hearing impairment is associated with enhanced neural tracking of the speech envelope. Hearing Research 393:107961.

Deng L (2012) The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine 29:141-142.

Di Liberto Giovanni M, O'Sullivan James A, Lalor Edmund C (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Current Biology 25:2457-2465.

Dial HR, Gnanateja GN, Tessmer RS, Gorno-Tempini ML, Chandrasekaran B, Henry ML (2021) Cortical Tracking of the Speech Envelope in Logopenic Variant Primary Progressive Aphasia. Frontiers in Human Neuroscience 14.

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. The Journal of Neuroscience 33:5728-5735.

Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage 88:41-46.

Dmochowski JP, Sajda P, Dias J, Parra LC (2012) Correlated components of ongoing EEG point to emotionally laden attention – a possible marker of engagement? Frontiers in Human Neuroscience 6:Article 112.

Dmochowski JP, Bezdek MA, Abelson BP, Johnson JS, Schumacher EH, Parra LC (2014) Audience preferences are predicted by temporal reliability of neural processing. Nature Communications 29:4567.

Fiedler L, Wöstmann M, Herbst SK, Obleser J (2019) Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. Neuroimage 186:33-42.

Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J (2017) Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. Journal of Neural Engineering 14:036020.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15:870–878.

Gillis M, Van Canneyt J, Francart T, Vanthornhout J (2022) Neural tracking as a diagnostic tool to assess the auditory pathway. Hearing Research 426:108607.

Han L, Liu Y, Zhang D, Jin Y, Luo Y (2013) Low-Arousal Speech Noise Improves Performance in N-Back Task: An ERP Study. PLOS ONE 8:e76261.

Hauswald A, Keitel A, Chen Y-P, Rösch S, Weisz N (2022) Degradation levels of continuous speech affect neural speech tracking and alpha power differently. European Journal of Neuroscience 55:3288-3302.

Heffernan E, Coulson NS, Henshaw H, Barry JG, Ferguson MA (2016) Understanding the psychosocial experiences of adults with mild-moderate hearing loss: An application of Leventhal's self-regulatory model. International Journal of Audiology 55:S3-S12.

Herrmann B (2023) The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. International Journal of Speech Technology 26:395-415.

Herrmann B, Johnsrude IS (2020) A Model of Listening Engagement (MoLE). Hearing Research 397:108016.

Herrmann B, Butler BE (2021) Hearing Loss and Brain Plasticity: The Hyperactivity Phenomenon. Brain Structure & Function 226:2019-2039.

Herrmann B, Henry MJ, Obleser J (2013) Frequency-specific adaptation in human auditory cortex depends on the spectral variance in the acoustic stimulation. Journal of Neurophysiology 109:2086-2096.

Herrmann B, Maess B, Johnsrude IS (2018) Aging Affects Adaptation to Sound-Level Statistics in Human Auditory Cortex. The Journal of Neuroscience 38:1989-1999.

Hertrich I, Dietrich S, Trouvain J, Moos A, Ackermann H (2012) Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. Psychophysiology 49:322-334.

Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12:55-67.

Holder JT, Levin LM, Gifford RH (2018) Speech Recognition in Noise for Adults With Normal Hearing: Age-Normative Performance for AzBio, BKB-SIN, and QuickSIN. Otology & Neurotology 39:e972-e978.

Holm S (1979) A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6:65-70.

Irsik VC, Johnsrude IS, Herrmann B (2022) Neural activity during story listening is synchronized across individuals despite acoustic masking. Journal of Cognitive Neuroscience 34:933-950.

Irsik VC, Almanaseer A, Johnsrude IS, Herrmann B (2021) Cortical Responses to the Amplitude Envelopes of Sounds Change with Age. The Journal of Neuroscience 41:5045-5055.

JASP (2023) JASP [Computer software]. In: https://jasp-stats.org/.

Joshi S, Gold JI (2020) Pupil Size as a Window on Neural Substrates of Cognition. Trends in Cognitive Sciences 24:466-480.

Kidd G, Jr., Mason CR, Best V, Roverud E, Swaminathan J, Jennings T, Clayton K, Steven Colburn H (2019) Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss. The Journal of the Acoustical Society of America 145:440-457.

Kitajo K, Nozaki D, Ward LM, Yamamoto Y (2003) Behavioral Stochastic Resonance within the Human Brain. Physical Review Letters 90:218103.

Kitajo K, Doesburg SM, Yamanaka K, Nazaki D, Ward LM, Yamamoto Y (2007) Noise-induced large-scale phase synchronization of human-brain activity associated with behavioural stochastic resonance. Europhysics Letters 80:40009.

Krauss P, Tziridis K, Metzner C, Schilling A, Hoppe U, Schulze H (2016) Stochastic Resonance Controlled Upregulation of Internal Noise after Hearing Loss as a Putative Cause of Tinnitus-Related Neuronal Hyperactivity. Frontiers in Neuroscience 10:Article 597.

Kries J, De Clercq P, Gillis M, Vanthornhout J, Lemmens R, Francart T, Vandermosten M (2024) Exploring neural tracking of acoustic and linguistic speech representations in individuals with post-stroke aphasia. Human Brain Mapping 45:e26676.

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.

Lin FR, Albert M (2014) Hearing loss and dementia – who is listening? Aging & Mental Health 18:671-673.

Makeig S, Bell AJ, Jung T-P, Sejnowski TJ (1995) Independent component analysis of electroencephalographic data. In: Advances in Neural Information Processing Systems (Touretzky D, Mozer M, Hasselmo M, eds), pp 145-151. Cambridge, MA, USA: MIT Press.

Mathiesen SL, Van Hedger SC, Irsik VC, Bain MM, Johnsrude IS, Herrmann B (2023) Exploring age differences in absorption and enjoyment during story listening. PsyArXiv.

Mathôt S (2018) Pupillometry: Psychology, physiology, and function. Journal of Cognition 1.

Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: A review. Language and Cognitive Processes 27:953-978.

McDermott Josh H, Simoncelli Eero P (2011) Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. Neuron 71:926-940.

McDonnell MD, Abbott D (2009) What Is Stochastic Resonance? Definitions, Misconceptions, Debates, and Its Relevance to Biology. PLOS Computational Biology 5:e1000348.

McDonnell MD, Ward LM (2011) The benefits of noise in neural systems: bridging theory and experiment. Nature Reviews Neuroscience 12:415-425.

McZgee VE, Carleton WT (1970) Piecewise Regression. Journal of the American Statistical Association 65:1109-1124.

Moss F, Ward LM, Sannita WG (2004) Stochastic resonance and sensory information processing: a tutorial and review of application. Clinical Neurophysiology 115:267-281.

Näätänen R, Picton TW (1987) The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24:375-425.

Nachtegaal J, Smit JH, Smits C, Bezemer PD, van Beek JH, Festen JM, Kramer SE (2009) The association between hearing status and psychosocial health before the age of 70 years: results from an internet-based national survey on hearing. Ear & Hearing 30:302-312.

Niedermeyer E, da Silva FHL (2005) Electroencephalography: Basic Principles, Clinical Applications, and Related Fields: Lippincott Williams & Wilkins.

Ohlenforst B, Wendt D, Kramer SE, Naylor G, Zekveld AA, Lunner T (2018) Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. Hearing Research.

Ohlenforst B, Zekveld AA, Lunner T, Wendt D, Naylor G, Wang Y, Versfeld NJ, Kramer SE (2017) Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. Hearing Research 351:68-79.

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Computational Intelligence and Neuroscience 2011:Article ID 156869.

OpenAI et al. (2023) GPT-4 Technical Report. arXiv.

Palana J, Schwartz S, Tager-Flusberg H (2022) Evaluating the use of cortical entrainment to measure atypical speech processing: A systematic review. Neuroscience & Biobehavioral Reviews 133:104506.

Panela RA, Copelli F, Herrmann B (2024) Reliability and generalizability of neural speech tracking in younger and older adults. Neurobiology of Aging 134:165-180.

Panza F, Lozupone M, Sardone R, Battista P, Piccininni M, Dibello V, La Montagna M, Stallone R, Venezia P, Liguori A, Giannelli G, Bellomo A, Greco A, Daniele A, Seripa D, Nicola Q, Logroscino G (2019) Sensorial frailty: age-related hearing loss and the risk of cognitive impairment and dementia in later life. Therapeutic Advances in Chronic Disease 10:1-17.

Parbery-Clark A, Marmel F, Bair J, Kraus N (2011) What subcortical–cortical relationships tell us about processing speech in noise. European Journal of Neuroscience 33:549-557.

Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BWY, Humes LE, Lemke U, Lunner T, Matthen M, Mackersie CL, Naylor G, Phillips NA, Richter M, Rudner M, Sommers MS, Tremblay KL, Wingfield A (2016) Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). Ear & Hearing 37 Suppl 1:5S-27S.

Picton TW, John SM, Dimitrijevic A, Purcell DW (2003) Human auditory steady-state responses. International Journal of Audiology 42:177-219.

Presacco A, Simon JZ, Anderson S (2016) Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. Journal of Neurophysiology 116:2346-2355.

Presacco A, Simon JZ, Anderson S (2019) Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss. PLoS ONE 14:e0213899.

Ritz H, Wild CJ, Johnsrude IS (2022) Parametric cognitive load reveals hidden costs in the neural processing of perfectly intelligible degraded speech. The Journal of Neuroscience 42:4619-4628.

Rufener KS, Kauk J, Ruhnau P, Repplinger S, Heil P, Zaehle T (2020) Inconsistent effects of stochastic resonance on human auditory processing. Scientific Reports 10:6419.

Ruhnau P, Herrmann B, Schröger E (2012) Finding the right control: The mismatch negativity under investigation. Clinical Neurophysiology 123:507-512.

Schmitt R, Meyer M, Giroud N (2022) Better speech-in-noise comprehension is associated with enhanced neural speech tracking in older adults with hearing impairment. Cortex 151:133-146.

Shukla B, Bidelman GM (2021) Enhanced brainstem phase-locking in low-level noise reveals stochastic resonance in the frequency-following response (FFR). Brain Research 1771:147643.

Spyridakou C, Rosen S, Dritsakis G, Bamiou D-E (2020) Adult normative data for the speech in babble (SiB) test. International Journal of Audiology 59:33-38.

Stein RB, Gossen ER, Jones KE (2005) Neuronal variability: noise or part of the signal? Nature Reviews Neuroscience 6:389-397.

Stocks NG (2000) Suprathreshold Stochastic Resonance in Multilevel Threshold Systems. Physical Review Letters 84:2310-2313.

Synigal SR, Anderson AJ, Lalor EC (2023) Electrophysiological indices of hierarchical speech processing differentially reflect the comprehension of speech in noise. BioRxiv.

Tabarelli D, Vilardi A, Begliomini C, Pavani F, Turatto M, Ricci L (2009) Statistically robust evidence of stochastic resonance in human auditory perceptual system. The European Physical Journal B 69:155-159.

Toms JD, Lesperance ML (2003) Piecewise regression: A tool for identifying ecological thresholds. Ecology 84:2034-2041.

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

Van Hirtum T, Somers B, Dieudonné B, Verschueren E, Wouters J, Francart T (2023) Neural envelope tracking predicts speech intelligibility and hearing aid benefit in children with hearing loss. Hearing Research 439:108893.

Vanthornhout J, Decruy L, Wouters J, Simon JZ, Francart T (2018) Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. Journal of the Association for Research in Otolaryngology 19:181-191.

Vieth E (1989) Fitting piecewise linear regression functions to biological responses. Journal of Applied Physiology 67:390-396.

Ward LM, Neiman A, Moss F (2002) Stochastic resonance in psychophysics and in animal behavior. Biological Cybernetics 87:91-101.

Ward LM, MacLean SE, Kirschner A (2010) Stochastic Resonance Modulates Neural Synchronization within and between Cortical Sources. PLoS ONE 5:e14371.

Wells C, Ward LM, Chua R, Inglis JT (2005) Touch Noise Increases Vibrotactile Sensitivity in Old and Young. Psychological Science 16:313-320.

Wilson RH (2003) Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. Journal of the American Academy of Audiology 14:453-470.

Wilson RH, Trivette CP, Williams DA, Watts KL (2012a) The Effects of Energetic and Informational Masking on the Words-in-Noise Test (WIN). J Am Acad Audiol 23:522-533.

Wilson RH, McArdle RA, Watts KL, Smith SL (2012b) The Revised Speech Perception in Noise Test (R-SPIN)in a Multiple Signal-to-Noise Ratio Paradigm. Journal of the American Academy of Audiology 23:590-605.

Yasmin S, Irsik VC, Johnsrude IS, Herrmann B (2023) The effects of speech masking on neural tracking of acoustic and semantic features of natural speech. Neuropsychologia 186:108584.

Zeng F-G (2013) An active loudness model suggesting tinnitus as increased central noise and hyperacusis as increased nonlinear gain. Hearing Research 295:172-179.

Zeng F-G (2020) Tinnitus and hyperacusis: central noise, gain and variance. Current Opinion in Physiology 18:123-129.

Zou J, Feng J, Xu T, Jin P, Luo C, Zhang J, Pan X, Chen F, Zheng J, Ding N (2019) Auditory and language contributions to neural encoding of speech features in noisy environments. NeuroImage 192:66-75.

Zuk NJ, Murphy JW, Reilly RB, Lalor EC (2021) Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. PLOS Computational Biology 17:e1009358.