Emotional Semantic Meaning Enhances Neural Tracking of Conversations Through Context-Sensitive Modulation of Auditory – Language Areas

Heikkinen, P.[1]*, Lepistö, S.[1]*, Wikman, P.[1], Koskinen, E.[2], Peräkylä, A.[2], Ravaja, N.[1], & Harjunen, V. J.[1]

[1] *Department of Psychology, Faculty of Medicine, University of Helsinki, Helsinki, Finland*
[2] *Faculty of Social Sciences, University of Helsinki, Helsinki, Finland*

*corresponding authors
pyry.heikkinen@helsinki.fi
santeri.lepisto@helsinki.fi
University of Helsinki, PO Box 21, Haartmaninkatu 3, 00014, Helsinki, Finland.

Abstract

Engagement in interaction arises through the continuous monitoring of conversation content, allowing individuals to prioritize information that is most relevant to them. Emotional salience of semantic content may play a key role in guiding this attentional engagement. By combining audiovisual "Cocktail-party" paradigm with source-localized magnetoencephalography and encoding-decoding models, we examined how emotional salience and narrative coherence shape cortical tracking of concurrent dialogues. Our results showed that emotional salience strengthened neural tracking of attended but not ignored conversations. This enhancement unfolded nonlinearly across the conversational arc, peaking early when the dialogue was coherent and later when incoherent. Cortical temporal response functions revealed latency-specific effects: emotional salience amplified early sensory encoding under uncertainty in auditory and language networks, whereas late semantic integration in these areas was driven by salience alone. These results indicate that emotional meaning dynamically modulates predictive engagement in speech, acting as a high-level relevance signal that sharpens attentional selection and semantic integration moment by moment across conversational turns.

keywords: emotion, neural speech tracking, auditory processing, predictive processing, magnetoencephalography (MEG), semantic coherence, attention, language comprehension

1 Introduction

Attending to relevant information is central in everyday conversations, where complex meanings unfold across multi-turn dialogues. In such contexts, emotional salience of the conversation content stands out as a key candidate for guiding the interactants' attentional engagement (Olano et. al, 2020; Park et al., 2025). Emotional salience signals perceived information's relevance to natural human concerns and goals (e.g., Wharton et al., 2021; Frijda, 2017; Lange et al., 2020), engaging appetitive–defensive motivational systems (Lang et al., 1998; Lang & Bradley, 2013) and prioritizing attentional processing of affectively charged stimuli (e.g., Sander et al., 2003; Vuilleumier, 2005; Yiend, 2010). Yet, much of the existing brain research on processing emotion in speech and language has focused on neural responses to paralinguistic cues (i.e. prosody) or isolated/written emotional words (Berckmoes & Vingerhoets, 2004; Citron, 2012; Hinojosa et al., 2020; Kissler et al., 2006). However, in natural conversation, socio-emotional relevance is distributed across turns and continually reshaped as the evolving conversational context refines and redefines the meaning of unfolding utterances (Xu et al., 2005; Swallow et al., 2011; Hasson et al., 2015). Therefore, studying how emotional semantic content interacts with attentional orientation and predictive mechanisms across the conversational arc is essential for understanding how people sustain engagement in interaction.

In spoken communication, people exchange emotional information through semantic and prosodic channels (e.g., Kotz & Paulmann, 2011). When extracting emotional content from speech, the neural processing stages have been modelled as a three-step sequence (Schirmer & Kotz, 2006; Kotz & Paulmann, 2011; see also Pinheiro, 2025) encompassing early sensory analysis (e.g., N1), derivation of emotional salience based on sensory cues (e.g., P2), and late evaluative processing of affectively charged meaning (e.g., N400). Previous research shows that the modulatory effects of emotion on neural responses extend beyond isolated lexical

items, including studies where target words are embedded in written sentences or narratives (e.g., Bayer et al., 2010; Ding et al., 2015; Ding et al., 2016; Fields & Kuperberg, 2012; Martín-Loeches et al., 2012). For example, emotionally congruent discourse contexts reduce neural response amplitudes, reflected in smaller N400 effects (Zhang et al, 2021). This closely mirrors the more general effect of semantic predictability of target words attenuating N400 response (Nieuwland & Van Berkum, 2006; Van Berkum et al., 2003). Together, these findings suggest that emotional salience and contextual predictability shape neural processing of words under controlled, non-conversational conditions.

When considering natural conversations, emotional salience is rarely conveyed by discrete utterances; rather, it emerges across extended stretches of talk, as meanings often unfold gradually over sustained conversational exchanges (Hoemann et al., 2025; Mar, 2011). In such naturalistic settings, the brain processes the ongoing auditory stream by continuously integrating information across multiple timescales and levels of representation, from rapid fluctuations in auditory features to slower unfolding of semantic structure and conversational context (Brodbeck et al., 2022; Zion Golumbic et al., 2012, 2013). This integration relies on efficient interaction between auditory, language, and higher-order cognitive networks linking sensory input with semantic and contextual representations, alongside social cognition and world-knowledge (Ji et al., 2019, Fedorenko et al., 2024). Neural responses to continuous speech reflect this organization, with early components indexing spectrotemporal encoding and later responses capturing lexical-semantic integration and contextual predictability (Zion Golumbic et al., 2013, Broderick et al., 2018). Recent research, grounded in predictive coding models, proposes that such integration relies on hierarchical exchanges among temporal, frontal, and parietal regions of the language network and auditory areas, where top-down expectations and attentional signals jointly shape sensory encoding and facilitate

comprehension (Broderick et al., 2019, Wikman et al., 2021, Caucheteux et al., 2022; Samara et al., 2025).

To understand how attentional modulation unfolds within more lifelike conditions where people need to continuously maintain attentional engagement to salient conversations, studies have extensively used the Cocktail-party paradigm (Cherry, 1953), where it has shown that attentional engagement is shapes the neural tracking of both attended and ignored speech (Ding & Simon, 2012; Zion Golumbic, et al., 2013; O'Sullivan et al., 2015; Fiedler et al., 2019). Furthermore, attention modulates the neural correlates of both early auditory encoding and later semantic integration, with more pronounced effects in higher-order functionality (Ding et al., 2012, Power et al., 2012; Teoh et al., 2022). Recent empirical work points to selective attention enhancing neural tracking of attended speech streams through recurrent interactions across a distributed network, with key contributions from auditory and language networks (Wikman et al., 2021, 2024). Importantly, this attentional modulation is not static: neural tracking changes non-linearly over the course of presented dialogue. From this viewpoint, emotional salience can be considered as a high-level relevance cue that biases attentional selection, enhancing neural tracking, and facilitating predictive processing of meaningful speech (Schröger et al., 2015; Angeloni & Geffen, 2018; Rutten et al., 2019). Yet, this has not been empirically examined.

Together, the body of evidence suggests that emotional salience is an integral dimension of language processing, interacting with the broader semantic context across multiple timescales to enhance the processing of relevant information (for a view on semantic cohesion, see Dillon et al., 2006). However, it remains unclear how emotional semantic content shapes neural speech encoding in more naturalistic conversational settings.

1.1 Current Study

By examining how emotional salience modulates auditory and semantic processing while listening to conversations, we aim to advance understanding of emotion perception in continuously evolving, naturalistic contexts. Using magnetoencephalography (MEG) with source localization, we recorded brain activity while participants selectively attended to lifelike audiovisual target conversations and were asked to ignore an overlapping distractor conversation in a "cocktail party" paradigm (Wikman et al., 2021, 2024). To examine how emotional salience affects attention-related neural speech processing in different contextual settings, we systematically manipulated both emotional salience (emotional vs. neutral) and dialogue coherence at the sentence level (coherent vs. incoherent) in both target and distractor conversations. All speech was presented with flat prosody and neutral facial expression to isolate semantic effects from lower-order paralinguistic and nonverbal emotional cues. Neural responses were analyzed with a combined encoding-decoding approach, utilizing temporal response functions (TRFs) and speech envelope reconstruction (SER) (Lalor & Foxe, 2010; Crosse, 2016). SER leverages the high-dimensional density of MEG data in a multi-variate manner, providing a high temporal fidelity measure of neural tracking of speech, with increased sensitivity to signal differences when compared to traditional methods. As the weights of the backward-models are not readily anatomically interpretable, TRFs were leveraged to reveal the timing and cortical sources of continuous stimulus-feature encoding.

We hypothesized that emotional salience serves as a high-level relevance signal that modulates neural processing across the speech comprehension hierarchy. Emotionally salient speech was expected to strengthen cortical tracking of the speech envelope, reflecting enhanced coupling between the acoustic envelope and ongoing neural activity. Within the encoding model framework, emotional content was hypothetized to augment early sensory responses and, more prominently, later responses associated with semantic processing. By combining

naturalistic stimuli, controlled semantic manipulations, anatomically grounded source-space analyses, and a combination of decoding and encoding models, we reveal new insights into how emotional meaning dynamically shapes attention and hierarchical language processing in the human brain.

2 Materials and Methods

2.1 Participants

The sample comprised 32 participants (18 women, 14 men), aged between 20 and 50 years (*M* = 30.22, *SD* = 7.98). The participants were recruited through university mailing lists and campus outreach. Eligibility criteria included the absence of neurological and psychiatric diagnoses, no current use of medications affecting the central nervous system, and normal or corrected-to-normal vision and hearing. Before the experiment, all participants provided written informed consent. As compensation, participants received a reward valued at approximately 30 euros. The research protocol received approval from the University of Helsinki Research Ethics Committee of the Faculty of Medicine, and the study followed the ethical principles outlined in the World Medical Association's Declaration of Helsinki. As one participant's neural signal showed no temporal variation, reflecting technical failure, the data were excluded from further analyses.

2.2 Study design

2.2.1 Procedure Overview

The participants completed an audiovisual "cocktail-party" experiment while their neural activity was recorded (Figure 1a). The experiment, beginning after a 4-minute resting-state recording, consisted of two runs, with 16 trials each. In each trial, two six-line dialogues were presented simultaneously to the participants: an audiovisual target dialogue and an audio-only distractor dialogue, with audio from both dialogues presented from the same speakers

concurrently. The dialogues differed in their conversational content and were performed in male-female pairs. Participants were instructed to focus exclusively on the target dialogue and were informed that they would answer four true–false statements about the dialogue content at the end of each trial. Hence, the task explicitly required attending to semantic information. After responding to the statements, participants received on-screen feedback of their performance. Additionally, participants rated how emotionally charged they perceived each target dialogue.

To manipulate the semantic content in both target and distractor dialogues, the experiment followed a 4 * 4 factorial design, varying two dimensions of semantic content: the level of emotional salience (emotional vs. neutral) and the level of dialogue coherence at sentence level (narratively coherent dialogue vs. narratively incoherent dialogue). Emotional dialogues featured negatively valenced topics (e.g., tragic accident, fear- or disgust-related experience, physical conflict, mental health issue or illness, harassment, loneliness), whereas neutral dialogues described affectively non-salient activities (e.g., grocery shopping, attending a lecture, cooking, working, talking about the weather, having lunch, drinking coffee). Coherent dialogues maintained contextually related conversational turns, while incoherent dialogues were generated by mixing dialogue lines, resulting in turns that were no longer semantically related (for dialogue examples, see Figure 1a; for dialogue material details, see Supplement). Together, the factorial manipulation yielded 16 unique target–distractor combinations (Figure 1b), each trial within a run featuring a different factorial pairing. For example, both of the runs included a trial where an emotionally incoherent target dialogue was presented alongside a neutral coherent distractor dialogue. The presentation order of the different target-distractor factor combinations was randomized within runs and across participants. Additionally, the gender of the active speaker was kept different between the target and distractor conversations.

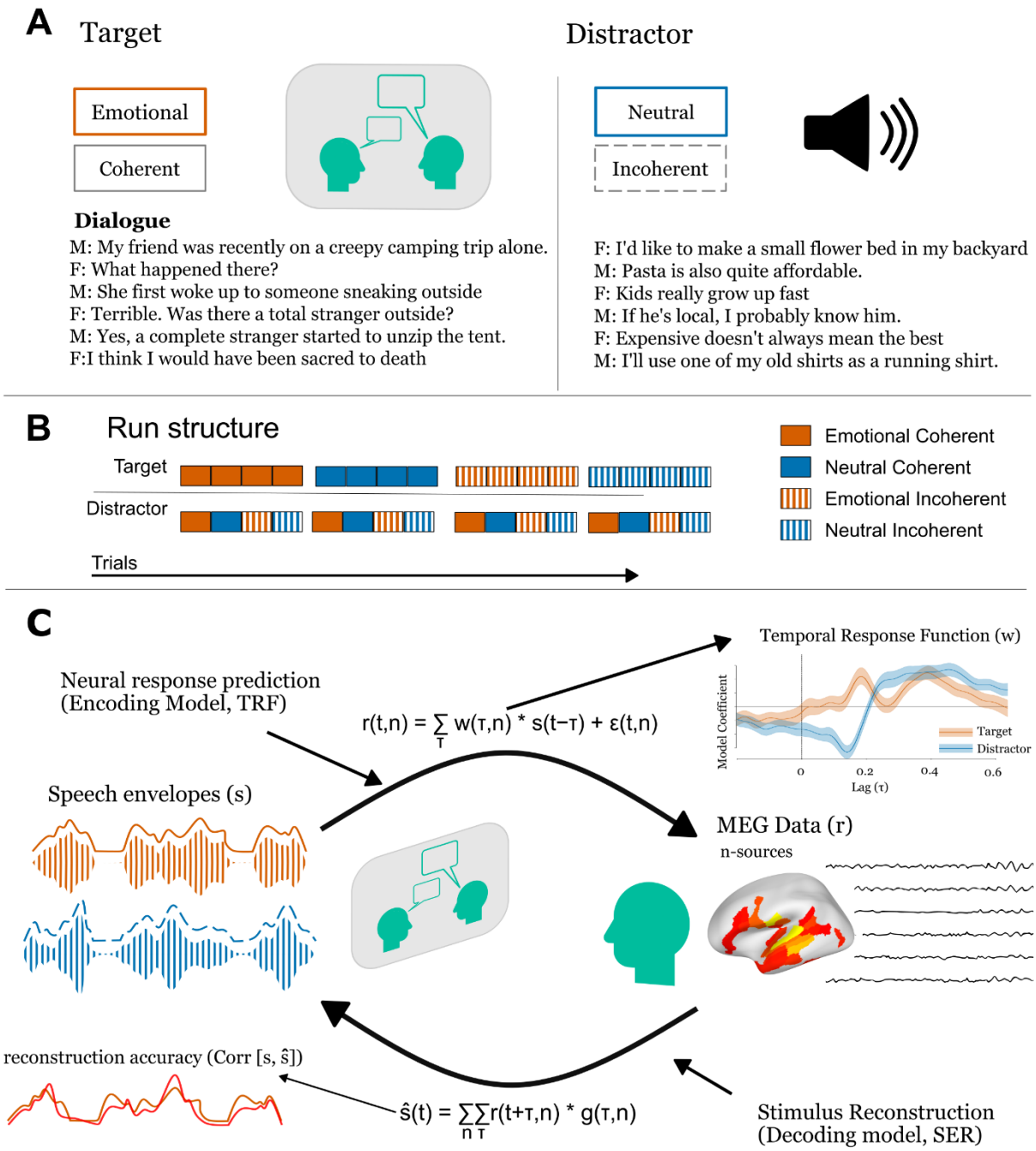

Figure 1. Stimulus Material, Experiment Structure and Modeling Approach. (A) Example of stimulus material used in the experiment, depicting one target-distractor combination, with the target dialogue (video) as emotional-coherent, and distractor (audio) as neutral-incoherent. (B) Structure of one acquisition run within the experiment, showing all the target-distactor combinations used, here shown organized by target conditions, while in practice, the presentation order was pseudo-randomized. (C) Schematic overview of the encoding-decoding modeling approach (adapted from Crosse, et al. 2021), illustrating how the aggregated speech envelopes of the target (orange) and distractor (blue) conversations are used independently to predict neural responses via the encoding model (temporal response functions, TRFs), while in

the decoding model, neural activity is used to reconstruct the envelopes (speech envelope reconstruction, SER). This was done for each condition separately.

2.2.2 Stimulus Presentation

The stimulus was presented with the Presentation® software (Neurobehavioral Systems Inc.). The target dialogues were presented on a projected screen (67 * 38 cm) at a viewing distance of approximately 130 cm (from the participant's eyes). The projector's resolution was 1920 * 1080 (Full HD), with a refresh rate of 120 Hz.

Before each target dialogue, a gray screen with the instruction "Pay attention to the conversation" was shown. Subsequently, target and distractor dialogues were presented line by line. After each line, a gray screen with a fixation cross was presented. For each line, the stimulus volume was set to a comfortable level based on the target, and the paired distractor was attenuated by 10 decibels based on the root mean square of the amplitudes, and each distractor line was set to begin 250 ms before the onset of the corresponding target line.

2.3 Data Acquisition

2.3.1 Magnetoencephalography and Structural Magnetic Resonance Imaging

Brain activity was recorded at 1 kHz with a 306-channel MEGIN Triux system (204 planar gradiometers, 102 magnetometers) at the BioMag Laboratory, Helsinki University Hospital. Horizontal and vertical electro-oculograms (EOG), as well as electrocardiograms (ECG), were recorded to detect eye movements and cardiac artifacts. The head position of the participants were continuously measured using 5 head position indicator (HPI) coils placed around the scalp. To facilitate accurate co-registration between the MEG data and structural magnetic resonance image (MRI), 100 evenly divided scalp points were digitized (Polhemus). For source localization, structural MRI data were acquired using a 1.5 T Siemens Avanto scanner (Erlangen, Germany) with a T1-weighted 3D magnetization-prepared rapid gradient-echo (MP-RAGE) sequence (voxel size = 1 × 1 × 1 mm^3).

2.3.2 Behavioral Ratings

To evaluate participants' attention to the target dialogues and to prompt attentional engagement across the experiment, participants answered questions regarding the presented target dialogue by using a VPixx response button device. After each trial, participants responded to four true–false statements derived from the content of the target dialogue. For each dialogue, four of the six lines were randomly selected, and a corresponding true–false statement was created for each selected line. The statements were designed to measure participants' detection of central semantic information conveyed in the dialogues, such as recognizing mentioned people, places, objects, or actions (e.g., "The interactant asked whether the other person had an umbrella"; "The interactant is looking for a job"; "The interactant said that their lecturers change often"). The statements were formulated true or false randomly, preventing participants from relying on fixed response patterns. After completing the four statements, feedback was provided on-screen (e.g., "Of your answers, 3 out of 4 were correct"). Based on the participant's correct responses to the true–false statements, a behavioral measure of recall accuracy was computed. Additionally, participants answered the question of how emotionally charged the content of the conversation was. This perceived emotional load was reported through a 3-point scale (1 = "Not at all", 2 = "Somewhat", 3 = "Highly").

2.4 Data Processing

2.4.1 MEG Preprocessing

MEG data were preprocessed using MNE-Python 1.5 (Gramfort et al, 2013). Raw recordings were first denoised using the MNE implementation of MaxFilter (Taulu & Simola, 2006) with temporal signal space separation (tSSS, 20-second sliding window, with the corresponding calibration and crosstalk files for the Elekta Neuromag TRIUX 306 system). The head position of each participant was spatially aligned to the start of run 1, with movement compensation based on the recorded continuous HPI information. Noisy and flat channels were automatically

detected and marked for interpolation at a later stage and the data was resampled at 128 Hz for processing efficiency, while still retaining all the relevant information.

Data were high-pass filtered at 0.5 Hz and low-pass filtered at 15 Hz using zero-phase FIR filters. Bad channel selection was manually confirmed and channels were interpolated with spherical-spline interpolation. Next, ocular and cardiac artefacts were removed via FastICA -algorithm, by flagging components correlated with horizontal or vertical electro-oculogram or with ECG and visually confirming the results before exclusion.

2.4.2 Distributed Source Localisation

For estimating the cortical generators of the measured MEG signal, distributed minimum-norm source reconstruction **(**Hämäläinen & Ilmoniemi, 1994; Gramfort et al., 2014**)** was performed with each participant's Freesurfer reconstructed surfaces. The surfaces were coordinate aligned to MEG digitization points. followed by construction of an individual Boundary Element head Model (BEM). A single-shell BEM (conductivity 0.3 S/m) was computed from the inner skull surface using FreeSurfer reconstructions, and cortical source spaces were defined at oct6 resolution (~5 mm spacing) on the white matter surface, resulting in approx. 10 000 vertices per hemisphere. Noise covariance matrices were estimated from 4 minute long empty-room recordings, with preprocessing steps following the preprocessing steps described above. Minimum-norm inverse operators were built with loose orientation constraint (0.2), depth weighting (0.2), SNR = 1/9, and orientation constrained to the cortical surface normal. The sign of activity was kept to maximize the information available for encoding and decoding models. MEG epochs corresponding to stimulus presentation were projected to source space, separately for target and distractor dialogues, yielding one cortical source current estimates time-locked to each stimulus type, respectively.

2.4.3 Stimulus processing and Data Segmentation

The broadband amplitude envelope of stimulus materials was obtained as the absolute value of the Hilbert transform. Audio envelopes were band-pass filtered (0.5–15 Hz, 4th-order Butterworth filter), and sampling rate was matched to MEG data at 128 Hz. Finally, the envelopes were segmented into speech epochs for each line by detecting silent gaps of ≥1 s where the envelope stayed below 1% of the maximum amplitude. As the Distractor dialogue lines started 250 ms prior to the targets, the start of distractor epochs was cut to align with the onset of target lines.

Next, source-localized MEG data were epoched from each dialogue line onset to offset for both target and distractor dialogues separately. Finally, the data were grouped based on emotional salience and semantic coherence, yielding separate neural data and audio envelope matrices for each Attention-Emotional Salience-Coherence combination.

2.5 Encoding and Decoding Models

Following source localisation and data preparation, decoding and encoding modelling were used to examine how distributed cortical activity tracked the auditory envelopes of the presented dialogues (Figure 1c). Specifically, by using speech envelope reconstruction (SER), it is possible to assess the fidelity at which the cortical activity follows the speech signal, whereas temporal response functions (TRFs) characterize the temporo-spatial dynamics of stimulus-response relationship (Crosse et al., 2016). The temporal lags considered in the models between speech signal and brain activity were restricted to be between -200 – 600 milliseconds. For both approaches, the *receptive field estimation* function in MNE-Python was used (based on the mTRF-Toolbox), utilising the default lagged ridge-regression approach, with 4-fold crossvalidation (regularisation-parameters tested between 1 to 1 * 10^6; see supplementary Figure 1). Separate models were fit for each Attention - Emotional Salience - Dialogue coherence combination, and model fits were assessed based on the fit of each

condition, ultimately providing the best fit was regularization at $1 * 10^4$ for TRFs and chosen for SER to maintain similarity between models.

2.5.1 Speech Envelope Reconstruction (SER)

In SER, the multivariate time series of neural activity serves as input to reconstruct the acoustic speech envelope. As speech-related information is distributed across channels and time lags, the decoding model leverages this multidimensional representation to maximize stimulus recovery (Crosse et al., 2016). To reduce computational strain, the neural input data was first condensed to anatomically-functionally defined cortical areas, based on the human connectome project's multi-modal parcellation (Glasser et al., 2016), by calculating the average time-course within each cortical area (with sign-flip of activations in the non-dominant orientations with > 90 degree deflection). Reconstruction accuracy was quantified as Pearson's product-moment correlation coefficient of the reconstructed and true envelopes across cross-validation folds. To assess the temporal evolution of neural speech representations, the cross-validation test reconstructions from the final chosen model were back-projected to their original temporal locations to build the full reconstructed speech envelope. Then, the mean correlation between SER and the stimulus envelope was calculated for each line.

2.5.2 Temporal Response Functions (TRFs)

TRFs were used to model how changes in the speech envelope drove cortical responses over time. Here, the envelope served as the input and the neural source activity at each vertex as the response. The resulting linear filter describes the timing and strength of stimulus–response relationships across delays within each vertex, independently. Regression coefficients were averaged across folds to yield source-by-time TRFs for each subject and condition. These weights are physiologically interpretable, linear combinations of overlapping responses analogous to ERPs (Crosse et al., 2021; Kaufman & Zion Golumbic, 2023).

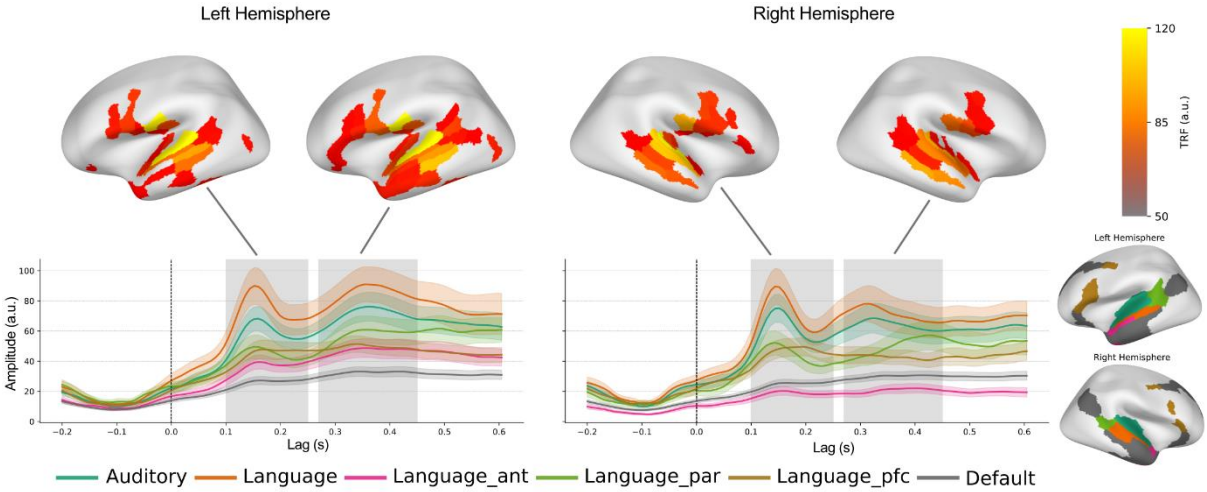

Figure 2. Extracted Source Time-Courses and Windows of Interest. For each hemisphere, grand-averaged source time-courses of analysis regions (with standard error bars), and topographies of mean activations of extracted values within analysis time-windows of 100 – 250 ms and 270 - 450 ms, respectively (depicted in grey).

2.6 Statistical Analyses

All statistical analyses were conducted in R (version 4.5.1). Linear mixed-effects models (LMMs) were fitted using the lmerTest package, and cumulative link mixed models (CLMMs) using the ordinal package. Significance for fixed effects in LMMs was assessed using Type-III ANOVA with Satterthwaite's approximation of degrees of freedom. Post hoc comparisons were based on estimated marginal means (emmeans package) with Holm correction for multiple comparisons in SER, and FDR-correction for TRF-analysis.

In behavioral analyses, recall accuracy (true–false statements) was quantified as the proportion of correct answers (number of correct responses divided by four) and analyzed using a repeated-measures LMM. The model included Emotional Salience (emotional vs. neutral) and Coherence (coherent vs. incoherent) as fixed within-subject factors, and a random intercept for participant, to account for repeated observations. Additionally, mean-centered SER accuracy was included as a continuous predictor to explore the influence of speech tracking on recall performance. Ratings of perceived emotional load were analyzed using a CLMM (nAGQ

= 7) with Emotional Salience and Coherence as within-subject factors and a random intercept for participants.

To examine how emotional salience, semantic coherence, and attention modulated cortical tracking of continuous speech, a LMM was built with speech-envelope reconstruction (SER) accuracy values Fisher's *z*-transformed prior to analysis. The model included Attention (attended vs. distractor), Emotional Salience (emotional vs. neutral), and Coherence (coherent vs. incoherent) as fixed factors, with all interactions and random intercepts for participants.

To further characterize how emotional salience and dialogue coherence influenced the temporal evolution of speech tracking, we modeled line-by-line SER accuracy for attended speech only. The model included Emotional Salience, Coherence, and dialogue turns with all interactions as fixed effects, and participants as random intercepts. To capture nonlinear temporal effects, a polynomial model including linear, quadratic, and cubic components was compared to a linear-only model using likelihood-ratio testing. The polynomial model provided a significantly better fit and was retained for inference (χ^2 (8) = 49.28, $p < .001$).

To assess how emotional salience and coherence modulated neural speech encoding, a LMM was fitted to assess two separate time windows of interest, reflecting the prominent peaks observed in the grand average TRFs. The parcellation-level TRFs were grouped within functionally relevant areas derived from Cole-Anticevic Brain-wide Network Partition (Ji et al., 2019). From this partition, the language areas were further subdivided to posterior language, anterior language, prefrontal language and parietal language areas. These, in addition to the auditory and default mode networks were included in the analysis (Figure 2). To analyze the earlier peak, local positive peak within 100–200 ms was identified for each network and condition separately. The mean amplitude within a 50 ms window around this peak was computed, capturing network-specific early encoding strength while accommodating latency

differences across cortical regions. The later activation was quantified as the overall mean across 270-500 ms time window, due to less pronounced curvature in the source waveforms.

The resulting amplitudes were subjected into a LMM with Emotional Salience (emotional vs. neutral), Dialogue coherence (coherent vs. incoherent), area, and Hemisphere with all interactions as fixed effects, and random intercepts for participants. Significance was assessed via Type-III ANOVA (Satterthwaite approximation), and post hoc pairwise contrasts were FDR-corrected.

3 Results

3.1 Behavioral Ratings

Descriptive results indicated that participants' recall accuracy (scores ranging between 0 and 1) was generally high across all conditions, from Emotional–Coherent (M = 0.89, SD = 0.18) and Neutral–Coherent (M = 0.88, SD = 0.17) to Emotional–Incoherent (M = 0.81, SD = 0.19) and Neutral–Incoherent (M = 0.80, SD = 0.23) dialogues. In the LMM, results showed a significant main effect of Coherence, b = 0.041, SE = 0.006, $t(958) = 6.95$, $p < .001$, indicating that participants recalled more accurately narrative content of coherent dialogues compared to incoherent dialogues. The main effect of Emotional Salience was not significant, b = 0.005, SE = 0.006, $t(958) = 0.85$, $p = .40$, and there was no significant Emotional Salience * Coherence interaction, b = –0.002, SE = 0.006, $t(958) = –0.25$, $p = .80$. Additionally, the inclusion of trial-based SER correlation as a continuous predictor revealed no significant relationship with recall accuracy (b = –0.036, SE = 0.07, $t(985) = -0.50$, $p = 0.62$).

Similarly, descriptive results for perceived emotional load were computed across Emotional–Coherent (M = 2.43, SD = 0.68), Neutral–Coherent (M = 1.06, SD = 0.25), Emotional–Incoherent (M = 2.38, SD = 0.70), and Neutral–Incoherent (M = 1.06, SD = 0.23) dialogues. In the cumulative link mixed model, the analysis revealed a significant main effect of Emotional Salience, indicating that emotional dialogues were rated as more emotionally

loaded than neutral ones, β = 3.50, *SE* = 0.18, *z* = 19.1, *p* < .001. No significant effects were found for Coherence (β = 0.067, *SE* = 0.11, *z* = 0.60, *p* = .55) or for the Emotional Salience * Coherence interaction (β = 0.029, *SE* = 0.11, *z* = 0.26, *p* = .79).

3.2 Speech Envelope Reconstruction Results

Higher SER accuracy was observed for attended, compared to distractor speech ($F(1, 210) = 339.89$, $p < .001$). SER accuracy was also higher for emotional than for neutral dialogues, $F(1, 210) = 16.66$, $p < .001$, whereas no main effect of coherence was observed, $F(1, 210) = 0.04$, $p = .85$. Critically, these effects were qualified by a significant Attention * Emotional Salience, $F(1, 210) = 13.09$, $p < .001$, and Attention * Emotional Salience * Coherence, $F(1, 210) = 5.27$, $p = .023$, interactions (Figure 3a; see supplementary Table 1 for full fixed effects). The model explained a substantial portion of the variance ($R^2_m = .58$, R^2□ = .62).

Follow-up comparisons of marginal means (see supplementary Table 2) showed that emotional salience enhanced SER only for attended speech in both coherent (β = 0.045, *SE* = 0.011, $t(210) = 4.16$, $p < .001$), and incoherent dialogues (β = 0.038, *SE* = 0.011, $t(210) = 3.55$, $p = .002$) with no difference for distractors ($p = .29$, and p = .16, respectively; see supplementary Table 3 for full pairwise tests). Examination of marginal effects indicated that the three-way interaction was primarily driven by small, opposite-signed emotionality effects across coherence conditions within the distractor stream—negative for coherent (β = -0.019, *SE* = 0.011, $t(210) = -1.76$, $p = .291$) and positive for incoherent speech (β = 0.024, *SE* = 0.011, $t(210) = 2.22$, $p = .164$). Thus, although statistically significant, the omnibus interaction likely reflects minor distractor-related variability rather than a meaningful coherence-dependent modulation of emotional enhancement.

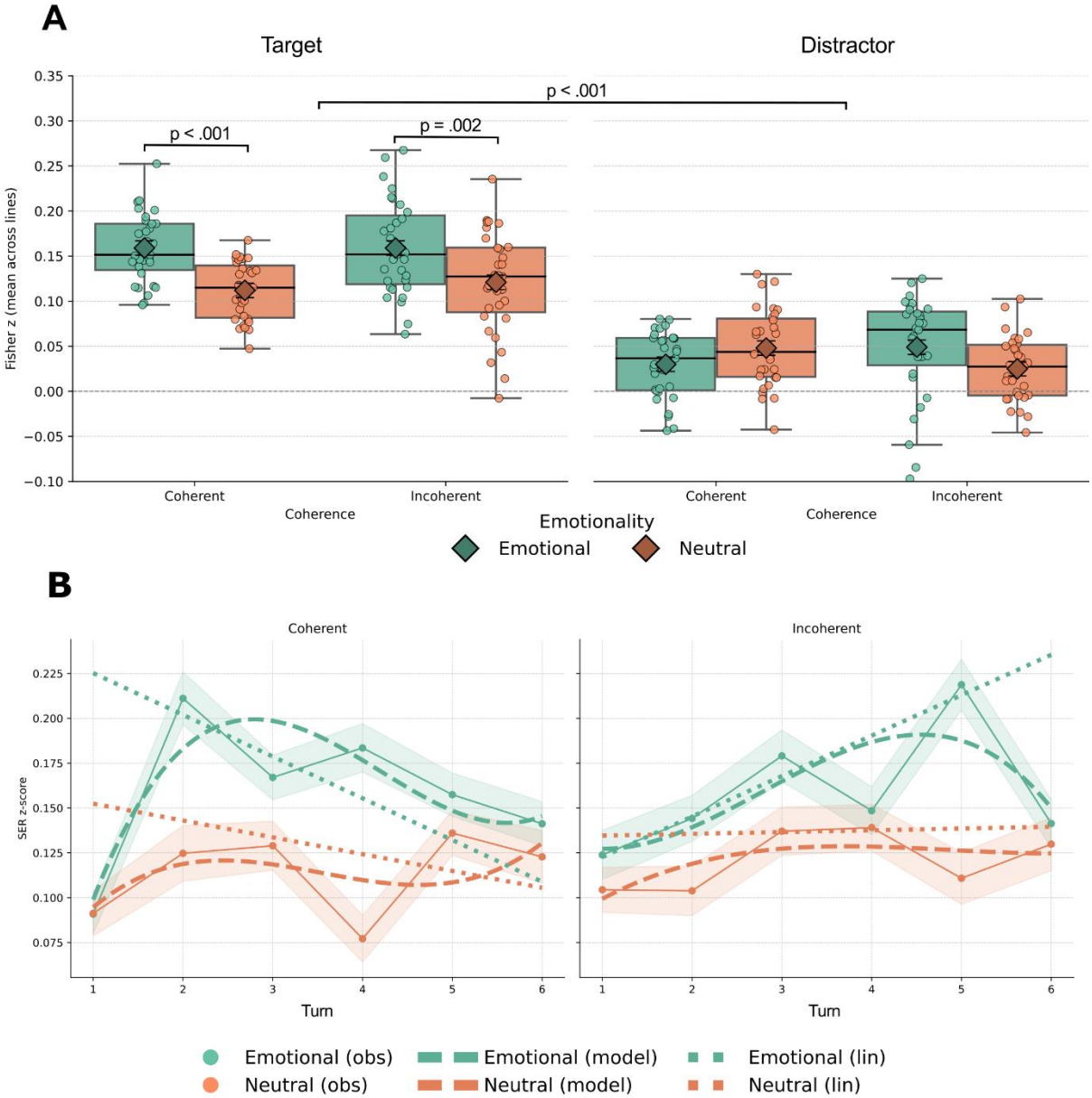

Figure 3. Speech Envelope Reconstruction Results. (A) Mean reconstructions accuracy (Fisher's Z-transformed) for each condition combination in both target and distractor stimuli. Boxplots and individual observations are based on empirical data. Diamonds depict lmm-based marginal means, with error bars showing standard errors. (B) The temporal evolution of SER accuracy in target dialogues as a function of turn index. Circle and solid line shows the observed means (with standard error band), while the full modeled response (linear-quadratic-cubic) and linear-only (mean offset for visualisation purposes) components are shown in dashed and dotted lines, respectively.

3.2.1 Temporal Modulation of SER Accuracy in Target Dialogues

In the turn-by-turn model for target dialogue SER, neural tracking varied systematically across the dialogue, showing significant linear (β = 0.049, *SE* = 0.013, *t* = 3.84 (5,921), p < .001) and quadratic (β = -0.016, *SE* = 0.006, *t* = -2.45 (5,921), *p* = .014) temporal components. Coherence significantly modulated the temporal effect in linear (β = 0.036, *SE* = 0.013, *t* = 2.83 (5,921), p = .005), quadratic (β = -0.019, *SE* = 0.006, *t* = 3.03 (5,921), *p* = .002), and cubic components (β = 0.003, *SE* = 0.001, *t* = 3.03 (5,921), *p* < .025). Emotionality further modulated these temporal effects, yielding significant Emotional Salience * Coherence * Turn (β = 0.029, *SE* = 0.013, *p* = .025) and Emotional Salience * Coherence * Turn² (β = -0.013, *SE* = 0.006, *p* = .038) interactions (Figure 3b; see supplementary Table 4 for full fixed effects).

In the linear component, follow-up analysis (see supplementary Table 5) show emotional coherent dialogues having a negative trajectory compared to emotional incoherent dialogues, reflecting stronger reductions in SER accuracy across time when the dialogue was coherent (β = -0.046, *SE* = 0.013, *z* = –3.59, *p* = .002). For the quadratic component, emotional coherent stories also exhibited a more pronounced curvature than neutral coherent stories (β = -0.011, *SE* = 0.003, *z* = –3.43, *p* = .004). Finally, in the cubic component, emotional coherent stories showed a significantly higher coefficient, reflecting an earlier inflection point followed by a steeper decline, when compared to emotional incoherent stories (β = 0.008, *SE* = 0.002, *z* = 3.432, *p* = .003). Overall model performance was modest (R^2_m = .019, R^2 = .029).

Together, these results indicate that emotional salience enhances neural tracking during active attention and that this effect evolves over time based on the conversations narrative coherence, showing stronger early coupling followed by adaptation or decline as the story unfolds.

3.3 Temporal Response Functions (TRFs) Results

3.3.1 Early TRF Window (100-250 ms)

In the early response range, amplitudes varied significantly with emotionality and coherence. Emotionally salient speech elicited stronger early responses than neutral speech, $F(1,1410) = 37.51$, $p < .001$, and incoherent dialogues evoked larger amplitudes than coherent ones, $F(1,1410) = 9.85$, $p = .002$ (Figure 4). Importantly, modulation by emotional semantic content varied with both dialogue coherence and cortical regions seen in the Emotional Salience * Coherence, $F(1,1410) = 11.39$, $p < .001$, and emotional salience * Region interactions $F(5,1410) = 2.62$, $p = .023$. See supplementary Table 6 for the full ANOVA table.

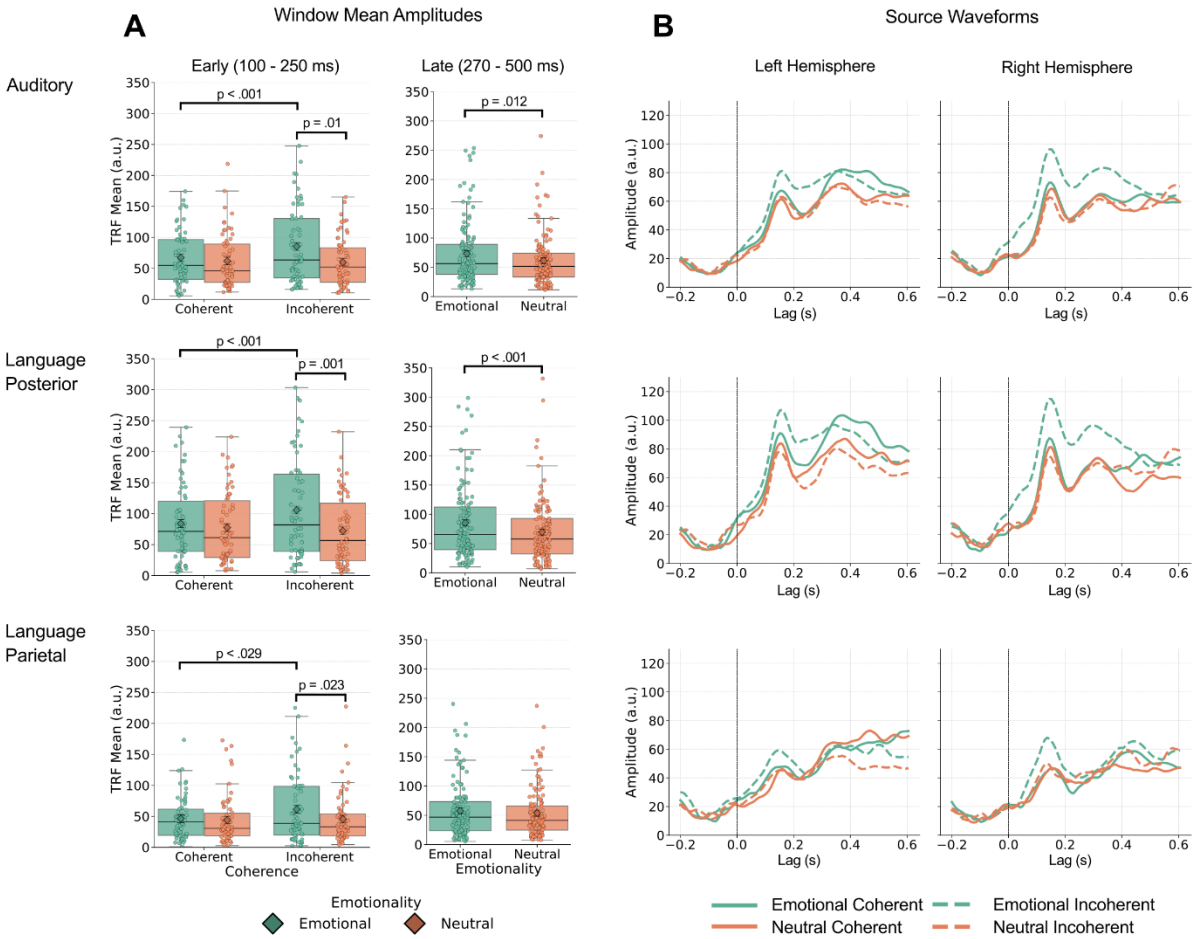

Figure 4. Temporal Response Functions (TRFs) Results Across Time Windows for Cortical Regions. Only regions with significant findings are reported here to maintain clarity, see supplementary fig. 2, for the non-significant activation patterns. (A) Boxplots showing TRF

model performance for target speech in the early (100–250 ms) and late (270–450 ms) time windows across cortical areas with significant effects. The early window revealed a coherence * emotion interaction, whereas the late window showed a main effect of emotion only. (B) Diamonds depict LMM-based marginal means, with error bars showing standard errors. Grand averaged TRF source waveforms for the left and right hemispheres in each significant area, illustrating temporal and hemispheric patterns of neural responses across the modelled lags.

Follow-up comparisons of marginal means (FDR-adjusted; see supplementary tables 7 & 8) revealed that emotional speech produced higher early TRF amplitudes than neutral speech under incoherent conditions in the Auditory (t(1, 410) = 3.16, p = .010), posterior Language (t(1, 410) = 3.82, p = .001), and parietal Language (t(1, 410) = 2.65, p = .032) areas (Figure 4). Within these same networks, emotional-incoherent responses were significantly greater than emotional-coherent responses (Auditory: t(1, 410) = –4.44, p < .001; Language posterior: t(1, 410) = –5.89, p < .001; Language parietal: t(1, 410) = –2.75, p = .029).

3.3.2 Late TRF Window (270–450 ms)

In the late window, TRF amplitudes remained sensitive to emotionality but were less dependent on dialogue coherence. Emotionally salient speech again produced stronger responses than neutral speech, $F(1, 1410) = 18.15$, $p < .001$, whereas coherence did not show an effect $F(1, 1410) = 0.51$, $p = .475$. Robust main effects of cortical region, $F(5, 1410) = 80.80$, $p < .001$, and hemisphere, $F(1, 1410) = 43.59$, $p < .001$, as well as the region * hemisphere, $F(5, 1410) = 5.81$, $p < .001$, interaction shows that the overall TRF responses differed between network areas and hemispheres. There was also a significant coherence * hemisphere interaction, $F(1, 1410) = 5.00$, $p = .025$ (for the full ANOVA-table, see supplementary Table 9). Follow-up analysis of the marginal mean contrasts (FDR-adjusted; see supplementary Tables 10 & 11) showed significantly greater late TRF amplitudes for emotional compared with neutral speech in the Auditory ($t(1,410) = –2.89$, $p = .012$) and posterior language regions ($t(1,410) = –3.81$, $p < .001$).

Taken together, the observed effects point to temporally distinct roles of emotional salience in speech processing—an early, context-sensitive amplification of sensory encoding when comprehension is uncertain, and a later enhancement of higher-order engagement with the semantic dimensions of speech with emotionally salient semantic content.

4 Discussion

While prior body of evidence demonstrates that emotional content modulates neural responses at the level of discrete target words (e.g., Bayer et al., 2010; Ding et al., 2015; Ding et al., 2016; Fields & Kuperberg, 2012; Kissler et al., 2006; Martín-Loeches et al., 2012; Zhang et al, 2021), the present magnetoencephalography study extends this work via examining how emotionally salient content and dialogue coherence shape the neural processing in naturalistic spoken conversations. By using a combined encoding-decoding approach to quantify how the brain responds to dynamic changes of the speech envelope in a cocktail-party setting, we revealed enhanced neural tracking for attended emotionally salient conversations, with distinct nonlinear temporal dynamics depending on the contextual coherence of the dialogue. Further analysis via temporal response functions revealed both context-dependent and independent modulations in attentional and semantic processing of the speech stream. Together, these findings show that emotional semantic content of conversation modulates neural tracking by incorporating information from the larger conversational context, and that emotional semantic information can modulate attentional mechanisms within the auditory and language areas. These are consistent with our theory that higher-order construing of emotional relevance can affect language processing via top-down modulatory mechanisms.

4.1 Emotional Semantic Content and Coherence Affect Perception and Recall

In the preliminary behavioral analyses, participants rated emotionally salient dialogues as more affectively intense and recalled incoherent dialogues less accurately than coherent ones.

Sensitivity to the emotional manipulation aligns with prior work on detecting emotional meaning in spoken language (e.g., Berckmoes & Vingerhoets, 2004; Cicero et al., 1999; Vingerhoets et al., 2003; see also Citron, 2012; Hinojosa et al., 2020; Kissler et al., 2006). Similarly, the effect of coherence accords with evidence that contextual continuity facilitates memory integration and recall (e.g., Becker, 1979; Cohn-Sheehy et al., 2021; Delarazan et al., 2025; Zwaan & Radvansky, 1998). Notably, although coherent and incoherent dialogues were constructed from the same dialogue lines and followed by identical true–false statements, semantic relatedness between adjacent utterances in coherent dialogues appeared to provide a scaffold supporting recall. Emotionality did not enhance recall performance, possibly because the task's memory demands were relatively modest and recall did not benefit further from emotionally charged content (see Park et al., 2025 for evidence of emotion-driven enhancement in recall). Together, these outcomes demonstrate that emotional conversation content and dialogue coherence shaped perception and memory processing, confirming successful experimental manipulation and establishing a basis for the neural-level analyses.

4.2 Emotional Content Enhances Neural Tracking of Attended Speech

Attentional orientation to target conversation was evident in the reconstruction accuracy of the speech envelope of said dialogue. This attentional enhancement of neural tracking has been reliably demonstrated in prior research (Ding & Simon., 2012; Zion Golumbic, et al., 2013; O'Sullivan et al., 2015; Fiedler et al., 2019). Interestingly, we show that emotional salience of the dialogue content systematically enhanced the cortical tracking of the attended speech envelope, irrespective of narrative coherence of the dialogues**.** As emotionally salient semantic content strengthens the coupling between neural activity and the acoustic speech envelope only when the speech is subjected to full linguistic evaluation, we theorize the observed emotional salience effect reflects top-down modulation of auditory-lexical processing by higher order language and associative areas. Here, social-cognitive understanding and accumulated world-

knowledge are leveraged for the comprehending meaning, including detection of emotional saliency within the attended dialogues (Fedorenko et al., 2024, Mar, 2011).

Although our incoherent dialogues exhibited less overall semantic cohesion, we did not observe clear modulations in SER accuracy, as has been observed in previous research (Broderick, 2019, Wikman, 2021). However, the effect of reduced coherence was reflected in a wider distribution of SER-accuracy scores within the semantically incoherent dialogues. This added variability could reflect less stable neural representations of the speech objects, possibly due to decreases in predictive confidence of the system leading to modulations of neural representations of the speech streams (Winkler et al. 2009; Schröger et al., 2015). We found no general association between dialogue-level SER strength and later recall accuracy performance, possibly reflecting temporal fluctuations in the relation between SER accuracy and recollection task-demands, observed in previous research (Wikman et al., 2024).

4.3 Dialogue Coherence Affects Temporal Modulation of Speech Tracking

As neural tracking of lifelike dialogues has been shown to display variability in the time domain, we extended the analysis to include the temporal changes across the conversational arc (Wikman et al., 2021, 2024). The turn-to-turn analysis corroborated that the influence of emotional salience is not static but temporally structured, with dialogue coherence shaping SER trajectories. In emotional conversations, coherent narratives showed a sharp early rise and subsequent decline in neural tracking, indicating that emotional semantic content strongly engages the auditory-language network at the beginning of a coherent narrative.

This pattern can be interpreted through a predictive-coding framework, where attention and emotional relevance heighten the gain of precision errors, and as the model error stabilizes, less resources are utilised within the earlier stages of the processing hierarchy (Winkler et al., 2009, Schröger et al., 2015). In contrast, emotional incoherent dialogues exhibited a flatter nonlinear trajectory peaking in the later stages of the dialogue, with the linear component

showing tracking to increase gradually across the narrative. Thus, when contextual coherence is disrupted, emotional salience does not elicit early enhancement but instead sustains a slower, ongoing increase in neural tracking until an inflection point is reached. Such a pattern may reflect continued attempts by the predictive system to construct stable expectations from disorganized but salient input. As the predictions fail to converge into a reliable model, increased energy is allocated to the close tracking of the speech stream.

The temporal evolution of emotional salience induced enhancement can alternatively be interpreted thought the expansion – renormalization model of cortical plasticity (Kilgard, 2012), where attentional engagement transiently heightens neural responsiveness though recruitment of maximal resources to adapt to task-specific goals, followed by renormalizing to efficient configurations favoring sparsity. Emotionally salient speech may thus induce an initial expansion phase of heightened gain and cross-network engagement, followed by renormalization, enhancing the brain's ability to encode behaviorally significant cues via increased capacity for adaptation (Clark, 2013, Diederen and Fletcher, 2021). Accordingly, emotional salience transiently increases large-scale cortical integration, reconfiguring connectivity to enhance readiness for adaptive learning and memory (Park et al., 2025). This reorganization links affective relevance detection to preparatory states for action and communication, facilitating flexible adaptation to socially and motivationally significant contexts (Song et al., 2021; Nau et al., 2021).

4.4 Neural Processing Underlying the Effects of Emotional Salience

While the Decoding/backward-modelling approach captures the macro-level slow fluctuations in neural engagement and its modulations by emotional salience and conversation coherence, we complement this analysis by using temporal-response functions, thus extending the investigation into the temporal evolution and topographical organisation of continuous speech processing.

In the early window (100-250 ms), emotional speech elicited enhanced responses in auditory, as well as posterior and parietal language areas when compared to neutral speech, particularly under incoherent conditions. According to the three-stage model of emotion speech processing (Schirmer & Kotz, 2006; Kotz & Paulmann, 2011), this latency range corresponds to the early sensory encoding stages, including the initial detection of emotional salience based on the auditory cues (e.g., Paulmann & Kotz, 2008). Although such detection of emotional significance is traditionally associated with prosody (Kotz & Paulmann, 2011; Pinheiro, 2025), evidence for early semantic processing has also been reported (e.g., Kissler et al., 2006; Kissler et al., 2009). In the present study, where emotionality was manipulated in the semantic domain, this early enhancement may reflect an initial evaluation of emotionally salient dialogue content preceding more elaborated semantic analysis. Interestingly, however, the effect is tapered when listening to the emotional dialogues with high narrative coherence. Thus, we interpret this P200-like enhancement as reflecting increased sensory gain under conditions of conversational unpredictability (incoherence), with emotional salience acting as a higher-order relevance cue that modulates predictive processing.

In the later window (270–500 ms), emotional dialogues again elicited stronger neural responses across both coherence conditions, encompassing auditory and higher-order language regions. This temporal range aligns with semantic integration and contextual updating processes traditionally reflected in the N400 and late positive components (e.g., see Citron, 2012; Brodbeck et al., 2022; Nieuwland & Van Berkum, 2006; Van Berkum et al., 2003; Zhang et al, 2021). The persistence of these effects suggests that emotional meaning acts as a relevance signal, strengthening the construction and maintenance of semantic representations even under varying coherence. Within the multistep model of emotion speech processing (Schirmer & Kotz, 2006; Kotz & Paulmann, 2011), this stage corresponds to the evaluation

phase of emotional content, where affectively charged semantic information is integrated into the broader contextual model and interpreted in relation to prior discourse.

Notably, the neural engagement of classical language areas is most evident in our data, as the responses were specifically modelled in relation to the speech envelope, which acted as a gestalt feature of the speech input. We observed no differences in the extended semantic networks, possibly due to them being coupled to the more general language and comprehension networks via nonlinear dynamics and latent representations of semantic structure (Chauteaux, 2023; Fedorenko et al., 2024; Samara et al., 2025).

To summarize, the dissociation between earlier, context-dependent modulations and later, context-general semantic effects supports a distributed hierarchical model of context modulated semantic processing (Xu et al., 2005). Emotional meaning appears to sharpen lexical access and increase processing precision when coherence is low, while also facilitating later conceptual integration. This positions emotion as a modulator of predictive hierarchies in language comprehension, adjusting the depth of semantic inference according to relevance and uncertainty. Hence, rather than a dual prosodic–semantic system, our results reveal a purely semantic pathway of emotional influence: emotional content biases predictive modelling and meaning integration within the core auditory-language network itself, independent of acoustic or prosodic cues.

Limitations and future directions

There were some limitations in the present study that should be considered when interpreting the findings. Although the emotional and neutral dialogue contents were carefully matched for length and structure, subtle differences in linguistic complexity or speech delivery might have influenced neural tracking responses (see Supplement for semantic similarity analysis). More elaborated examination linking linguistic features with neural dynamics could help clarify the specific effects of emotional dialogue content. Future research should also integrate emotion

semantics and affective prosody analyses to capture a more comprehensive picture of attentional enhancement in naturalistic emotional communication. It is also of note that the TRFs did not reliably capture early auditory processing, namely M50, M100, possibly due to high regularizations and multimodal nature of our stimuli.

Conclusions

Together, these findings show that emotional salience dynamically shapes predictive engagement during natural speech comprehension. Emotional semantic content enhances neural tracking and modulates the effect of contextual uncertainty on cortical gain, linking affective relevance to the updating and stabilization of meaning. This interaction between emotion and prediction supports the adaptive tuning of auditory–language networks to socially and motivationally significant cues. Rather than acting as a simple amplifier of sensory encoding, emotional salience emerges as an organizing principle, shaping predictive language processing and balancing flexibility and stability to optimize participating in human communication.

Acknowledgements

The authors would like to thank Mariel Wuolio for her valuable help in preparing materials, participant recruitment and assistance in MEG measurements. We are also grateful to Markus Söderman for his assistance with the preparation of the stimulus material and study design, and Jan Lindström for his role in the conceptualization and planning of the project. Finally, we thank students Leevi Lintunen, Jenna Jakonen, Aino Innanen and Aarni Arvilommi for their valuable contributions in producing the study's stimulus material.

Funding

The research has been funded by The Society of Swedish Literature in Finland [Project title: EnTiTy: Understanding Engagement in interaction through language, emotions, personality,

and Technology] and Research Council of Finland [Project title: EngIne: Engagement in the era of remote social Interactions: Manifestations, variation, and consequences, Project number: 356592].

References

Angeloni, C., & Geffen, M. (2018). Contextual modulation of sound processing in the auditory cortex. *Current Opinion in Neurobiology*, *49*, 8–15. https://doi.org/10.1016/j.conb.2017.10.012

Bayer, M., Sommer, W., & Schacht, A. (2010). Reading emotional words within sentences: The impact of arousal and valence on event-related potentials. *International Journal of Psychophysiology*, *78*(3), 299–307. https://doi.org/10.1016/j.ijpsycho.2010.09.004

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(2), 252–259. https://doi.org/10.1037/0096-1523.5.2.252

Berckmoes, C., & Vingerhoets, G. (2004). Neural Foundations of Emotional Speech Processing. *Current Directions in Psychological Science*, *13*(5), 182–185. https://doi.org/10.1111/j.0963-7214.2004.00303.x

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, *11*, e72056. https://doi.org/10.7554/eLife.72056

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.e3. https://doi.org/10.1016/j.cub.2018.01.080

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

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), 430–441. https://doi.org/10.1038/s41562-022-01516-2

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Cicero, B. A., Borod, J. C., Santschi, C., Erhan, H. M., Obler, L. K., Agosti, R. M., Welkowitz, J., & Grunwald, I. S. (1999). Emotional versus nonemotional lexical perception in patients with right and left brain damage. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *12*(4), 255–264.

Citron, F. M. M. (2012). Neural correlates of written emotion word processing: A review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and Language*, *122*(3), 211–226. https://doi.org/10.1016/j.bandl.2011.12.007

Clark, A. (2013). The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology*, *4*, 270. https://doi.org/10.3389/fpsyg.2013.00270

Cohn-Sheehy, B. I., Delarazan, A. I., Reagh, Z. M., Crivelli-Decker, J. E., Kim, K., Barnett, A. J., Zacks, J. M., & Ranganath, C. (2021). The hippocampus constructs narrative memories across distant events. *Current Biology*, *31*(22), 4935-4945.e7. https://doi.org/10.1016/j.cub.2021.09.013

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00604

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*. https://doi.org/10.3389/fnins.2021.705621

Delarazan, A. I., Bosak, E., Cohn-Sheehy, B. I., Foureaux-Lee, V., Zacks, J. M., & Reagh, Z. M. (2025). Narrative coherence warps the timeline of recalled naturalistic events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001516

Diederen, K. M. J., & Fletcher, P. C. (2021). Dopamine, Prediction Error and Beyond. *The Neuroscientist*, *27*(1), 30–46. https://doi.org/10.1177/1073858420907591

Dillon, D. G., Cooper, J. J., Grent-'t-Jong, T., Woldorff, M. G., & LaBar, K. S. (2006). Dissociation of event-related potentials indexing arousal and semantic cohesion during emotional word encoding. *Brain and Cognition*, *62*(1), 43–57. https://doi.org/10.1016/j.bandc.2006.03.008

Ding, J., Wang, L., & Yang, Y. (2015). The dynamic influence of emotional words on sentence processing. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(1), 55–68. https://doi.org/10.3758/s13415-014-0315-6

Ding, J., Wang, L., & Yang, Y. (2016). The dynamic influence of emotional words on sentence comprehension: An ERP study. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(3), 433–446. https://doi.org/10.3758/s13415-016-0403-x

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. https://doi.org/10.1073/pnas.1205381109

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

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. https://doi.org/10.1016/j.neuroimage.2018.10.057

Fields, E. C., & Kuperberg, G. R. (2012). It's All About You: An ERP Study of Emotion and Self-Relevance in Discourse. *NeuroImage*, *62*(1), 562–574. https://doi.org/10.1016/j.neuroimage.2012.05.003

Frijda, N. H. (2017). *The Laws of Emotion* (1st ed.). Psychology Press. https://doi.org/10.4324/9781315086071

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42. https://doi.org/10.1007/BF02512476

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313. https://doi.org/10.1016/j.tics.2015.04.006

Hinojosa, J. A., Moreno, E. M., & Ferré, P. (2020). Affective neurolinguistics: Towards a framework for reconciling language and emotion. *Language, Cognition and Neuroscience*, *35*(7), 813–839. https://doi.org/10.1080/23273798.2019.1620957

Hoemann, K., Lee, Y., Dussault, È., Devylder, S., Ungar, L. H., Geeraerts, D., & Mesquita, B. (2025). The construction of emotional meaning in language. *Communications Psychology*, *3*(1), 99. https://doi.org/10.1038/s44271-025-00255-0

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., & Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57. https://doi.org/10.1016/j.neuroimage.2018.10.006

Kaufman, M., & Zion Golumbic, E. (2023). Listening to two speakers: Capacity and tradeoffs in neural speech tracking during Selective and Distributed Attention. *NeuroImage*, *270*, 119984. https://doi.org/10.1016/j.neuroimage.2023.119984

Kilgard, M. P. (2012). Harnessing plasticity to understand learning and treat disease. *Trends in Neurosciences*, *35*(12), 715–722. https://doi.org/10.1016/j.tins.2012.09.002

Kissler, J., Assadollahi, R., & Herbert, C. (2006). Emotional and semantic networks in visual word processing: Insights from ERP studies. In *Progress in Brain Research* (Vol. 156, pp. 147–183). Elsevier. https://doi.org/10.1016/S0079-6123(06)56008-X

Kissler, J., Herbert, C., Winkler, I., & Junghofer, M. (2009). Emotion and attention in visual word processing—An ERP study. *Biological Psychology*, *80*(1), 75–83. https://doi.org/10.1016/j.biopsycho.2008.03.004

Kotz, S. A., & Paulmann, S. (2011). Emotion, Language, and the Brain. *Language and Linguistics Compass*, *5*(3), 108–125. https://doi.org/10.1111/j.1749-818X.2010.00267.x

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *The European Journal of Neuroscience*, *31*(1), 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Lang, P. J., & Bradley, M. M. (2013). Appetitive and Defensive Motivation: Goal-Directed or Goal-Determined? *Emotion Review*, *5*(3), 230–234. https://doi.org/10.1177/1754073913477511

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological Psychiatry*, *44*(12), 1248–1263. https://doi.org/10.1016/S0006-3223(98)00275-3

Lange, J., Dalege, J., Borsboom, D., Van Kleef, G. A., & Fischer, A. H. (2020). Toward an Integrative Psychometric Model of Emotions. *Perspectives on Psychological Science*, *15*(2), 444–468. https://doi.org/10.1177/1745691619895057

Mar, R. A. (2011). The Neural Bases of Social Cognition and Story Comprehension. *Annual Review of Psychology*, *62*(Volume 62, 2011), 103–134. https://doi.org/10.1146/annurev-psych-120709-145406

Martín-Loeches, M., Fernández, A., Schacht, A., Sommer, W., Casado, P., Jiménez-Ortega, L., & Fondevila, S. (2012). The influence of emotional words on sentence processing: Electrophysiological and behavioral evidence. *Neuropsychologia*, *50*(14), 3262–3272. https://doi.org/10.1016/j.neuropsychologia.2012.09.010

Nau, M., Schmid, A. C., Kaplan, S. M., Baker, C. I., & Kravitz, D. J. (2024). Centering cognitive neuroscience on task demands and generalization. *Nature Neuroscience*, *27*(9), 1656–1667. https://doi.org/10.1038/s41593-024-01711-6

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. https://doi.org/10.1162/jocn.2006.18.7.1098

Olano, M. A., Elizalde Acevedo, B., Chambeaud, N., Acuña, A., Marcó, M., Kochen, S., & Alba-Ferrara, L. (2020). Emotional salience enhances intelligibility in adverse acoustic conditions. *Neuropsychologia*, *147*, 107580. https://doi.org/10.1016/j.neuropsychologia.2020.107580

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2019). Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron*, *104*(6), 1195-1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007

Park, J. S., Gollapudi, K., Ke, J., Nau, M., Pappas, I., & Leong, Y. C. (2025). Emotional arousal enhances narrative memories through functional integration of large-scale brain networks. *Nature Human Behaviour*, 1–20. https://doi.org/10.1038/s41562-025-02315-1

Paulmann, S., & Kotz, S. A. (2008). Early emotional prosody perception based on different speaker voices. *NeuroReport*, *19*(2), 209–213. https://doi.org/10.1097/WNR.0b013e3282f454db

Pinheiro, A. P. (2025). Behind a Voice There is a Speaker: Why Vocal Emotion Research Needs to Become 'Personal.' *Affective Science*, *6*(3), 562–574. https://doi.org/10.1007/s42761-025-00317-w

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503. https://doi.org/10.1111/j.1460-9568.2012.08060.x

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., & Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour*, *3*(9), 974–987. https://doi.org/10.1038/s41562-019-0648-9

Samara, A., Zada, Z., Vanderwal, T., Hasson, U., & Nastase, S. (2025). Cortical language areas are coupled via a soft hierarchy of model-based linguistic features. In *bioRxiv: The preprint server for biology*. https://doi.org/10.1101/2025.06.02.657491

Sander, D., Grafman, J., & Zalla, T. (2003). The Human Amygdala: An Evolved System for Relevance Detection. *Reviews in the Neurosciences*, *14*(4). https://doi.org/10.1515/REVNEURO.2003.14.4.303

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24–30. https://doi.org/10.1016/j.tics.2005.11.009

Schröger, E., Marzecová, A., & SanMiguel, I. (2015). Attention and prediction in human audition: A lesson from cognitive psychophysiology. *European Journal of Neuroscience*, *41*(5), 641–664. https://doi.org/10.1111/ejn.12816

Song, H., Park, B., Park, H., & Shim, W. M. (2021). Cognitive and Neural State Dynamics of Narrative Comprehension. *Journal of Neuroscience*, *41*(43), 8972–8990. https://doi.org/10.1523/JNEUROSCI.0037-21.2021

Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., & Zacks, J. M. (2011). Changes in Events Alter How People Remember Recent Information. *Journal of Cognitive Neuroscience*, *23*(5), 1052–1064. https://doi.org/10.1162/jocn.2010.21524

Taulu, S., & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine & Biology*, *51*(7), 1759. https://doi.org/10.1088/0031-9155/51/7/008

Teoh, E. S., Ahmed, F., & Lalor, E. C. (2022). Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment. *Journal of Neuroscience*, *42*(4), 682–691. https://doi.org/10.1523/JNEUROSCI.1455-20.2021

van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research*, *17*(3), 701–718. https://doi.org/10.1016/S0926-6410(03)00196-4

Vingerhoets, G., Berckmoes, C., & Stroobant, N. (2003). Cerebral hemodynamics during discrimination of prosodic and semantic emotion in speech studied by transcranial doppler ultrasonography. *Neuropsychology*, *17*(1), 93–99.

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585–594. https://doi.org/10.1016/j.tics.2005.10.011

Wharton, T., Bonard, C., Dukes, D., & Oswald, S. (2021). Relevance and emotion. *Journal of Pragmatics*, *181*, 259–269. https://doi.org/10.1016/j.pragma.2021.06.001

Wikman, P., Sahari, E., Salmela, V., Leminen, A., Leminen, M., Laine, M., & Alho, K. (2021). Breaking down the cocktail party: Attentional modulation of cerebral audiovisual speech processing. *NeuroImage*, *224*, 117365. https://doi.org/10.1016/j.neuroimage.2020.117365

Wikman, P., Salmela, V., Sjöblom, E., Leminen, M., Laine, M., & Alho, K. (2024). Attention to audiovisual speech shapes neural processing through feedback-feedforward loops between different nodes of the speech network. *PLOS Biology*, *22*(3), e3002534. https://doi.org/10.1371/journal.pbio.3002534

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *NeuroImage*, *25*(3), 1002–1015. https://doi.org/10.1016/j.neuroimage.2004.12.013

Yiend, J. (2010). The effects of emotion on attention: A review of attentional processing of emotional information. *Cognition and Emotion*, *24*(1), 3–47. https://doi.org/10.1080/02699930903205698

Zhang, Q., Ding, J., Zhang, Z., Yang, X., & Yang, Y. (2021). The effect of congruent emotional context in emotional word processing during discourse comprehension. *Journal of Neurolinguistics*, *59*, 100989. https://doi.org/10.1016/j.jneuroling.2021.100989

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and Language*, *122*(3), 151–161. https://doi.org/10.1016/j.bandl.2011.12.010

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

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162