Exploring the impact of social relevance on the cortical tracking of speech: viability and temporal response characterisation

Abbreviated title: The neurophysiology of social-speech listening

Emily Y.J. Ip[1,2]*, Asena Akkaya[1,2]*, Martin M. Winchester[1,2], Sonia J. Bishop[2,3], Benjamin R. Cowan[4], Giovanni M. Di Liberto[1,2]**

[1] *School of Computer Science and Statistics, University of Dublin, Trinity College, Dublin 2, Ireland; ADAPT Centre*
[2] *Trinity College Institute of Neuroscience, University of Dublin, Trinity College, Dublin 2, Ireland*
[3] *School of Psychology, University of Dublin, Trinity College, Dublin 2, Ireland*
[4] *School of Information and Communication Studies, University College Dublin, Dublin 4, Ireland*

**Co-first authors*
***Corresponding author*: gdiliber@tcd.ie

Number of pages: 38
Number of figures: 7
Number of tables: 2
Number of words

- Abstract: 234
- Introduction: 632
- Discussion: 1141

Conflict of interest statement: The authors declare no competing interests.

Acknowledgements: This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. This research was conducted with the financial support of Research Ireland at ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology at Trinity College Dublin and University College Dublin [13/RC/2106_P2]. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Abstract

Human speech is inherently social. Yet our understanding of the neural substrates underlying continuous speech perception relies largely on neural responses to monologues, leaving substantial uncertainty about how social interactions shape the neural encoding of speech. Here, we bridge this gap by studying how EEG responses to speech change when the input includes a social element. Specifically, we compared the neural encoding of synthesised undirected monologues, directed monologues, and dialogues in Experiment 1. In Experiment 2, we extended this by using podcasts, addressing the additional challenges of real speech dialogue, such as dysfluencies. Using temporal response function analyses, we show that the presence of a social component strengthens envelope tracking — despite identical acoustic properties — indicating heightened listener engagement. Neural responses to synthesised speech showed a strong correlation with those for real speech podcasts, with a stronger alignment emerging for more socially-relevant speech material. In addition, we demonstrate that robust neural indices of sound and lexical-level processing can be derived using real podcast recordings despite the presence of dysfluencies. Finally, we discuss the importance of dysfluency in social speech experiments, presenting a simulation quantifying the potential impact of dysfluency on lexical level analyses. Together, these findings advance our understanding of continuous speech neurophysiology by highlighting the impact of social elements in shaping auditory neural processing in a controlled manner and providing a framework for future investigation and analysis of social speech listening and speech interaction.

Significance Statement

Human speech is rarely produced or processed in a social vacuum. Yet, our understanding of continuous speech neurophysiology mostly comes from experiments involving speech monologues. This study reveals how social context modulates the neural encoding of speech. We directly contrast neural signals recorded when participants listened to monologues and dialogues, using controlled material from speech synthesis and real podcast recordings. We found that the social element amplifies the neural encoding of speech features, reflecting greater engagement. We also show strong correlation between synthetic and real podcast neural responses, scaling with social relevance. Finally, we demonstrate that lexical processing can be measured robustly even amid natural dysfluencies. These insights advance our understanding of speech neurophysiology, informing future research on social speech.

Introduction

To understand speech, our brains must rapidly process sounds into linguistic meaning (Friederici, 2011; Poeppel et al., 2012). This entails a certain degree of interpretation (Yeshurun et al., 2017). The neural basis of this phenomenon has been studied extensively using controlled paradigms involving targeted manipulations (Osterhout and Holcomb, 1993; Carreiras et al., 2005; Kutas and Federmeier, 2011), and more recently with listening tasks involving more realistic continuous streams of speech (Di Liberto et al., 2021; Gwilliams et al., 2023; Di Liberto and Ip, 2025). That work utilised a variety of neural recording technologies, including electro- and magneto-encephalography (EEG and MEG) (Destoky et al., 2019; Défossez et al., 2023), functional magnetic resonance imaging and functional near-infrared spectroscopy (Formisano et al., 2008; Butler et al., 2020; Chen et al., 2020), and intracranial electrical recordings (e.g., electrocorticography and stereotactic EEG) (Mesgarani et al., 2014; Bernabei et al., 2021; Hamilton et al., 2021). Overall, this research has led to a new understanding of the sound-to-meaning transformation, providing detailed insights into the spatial and temporal dynamics that unfold throughout that process (Uusvuori et al., 2008; Kocagoncu et al., 2017). However, the necessary experimental constraints have led to listening tasks centred around audiobook-style, pre-recorded speech material (Lalor and Foxe, 2010), while ignoring one of the key functions of speech: the social element.

Recent research has begun exploring speech interaction, often using brain-to-brain synchrony and hyperscanning (Kinreich et al., 2017; Schwartz et al., 2024). However, those interactive experiments are far removed from previous speech listening research, as other factors at play (e.g., social anxiety, familiarity, turn-taking) make it difficult to relate findings across the two domains. To achieve this and advance the

study of social speech neurophysiology, we require a mechanistic understanding of how social factors impact the neural processing of speech, which requires careful experimental control. The present study pushes beyond the audiobook listening paradigm by use of novel tasks and analysis methods which allow the social element of speech, termed *social relevance,* to be isolated. In doing so, we investigate how challenges of real speech (e.g., dysfluencies) impact our ability to measure the neural processing of speech and language (Mousavi et al., 2024; Di Liberto and Ip, 2025).

We present results from two EEG experiments (**Figure 1**). The first experiment compares the neural encoding of speech while participants listened to pre-recorded monologues and dialogues (**Experiment 1**). *Dialogue* typically involves turn-taking, mutual feedback, and shared communicative goals (Garrod and Pickering, 2004), while monologues lack immediate reciprocity (excluding sensory modalities other than sound). Nevertheless, monologues can also include a social dimension by being *directed* to a listener*,* as opposed to *undirected monologues*, when they are not aimed at any particular listener. Using the temporal response function analysis (TRF), we quantify and investigate how social relevance impacts the neural processing of speech during listening tasks where participants do not interact. We hypothesised that directed monologues would lead to stronger acoustic and linguistic encoding than undirected monologues, potentially reflecting higher participant engagement. Synthesised text material was used to control for acoustic and lexical differences across conditions, ensuring that differences in the EEG would be due to social relevance as opposed to other factors such as prosody, word rate, or volume level.

Since Experiment 1 used synthesised speech, we designed a separate experiment using natural spontaneous dialogue from popular podcasts (**Experiment 2**). Studying

natural dialogues introduces important challenges compared to synthesised speech, including the presence of dysfluencies that occur in spontaneous, natural speech (Shriberg, 2001). Here, we aimed to determine: (1) if dysfluencies present a barrier for the study of spontaneous dialogue listening, and (2) if social relevance shapes the neural responses to speech in a robust and measurable manner. Specifically, we posited that the correlation between the neural responses to synthesised speech and real speech podcasts would increase with social relevance.

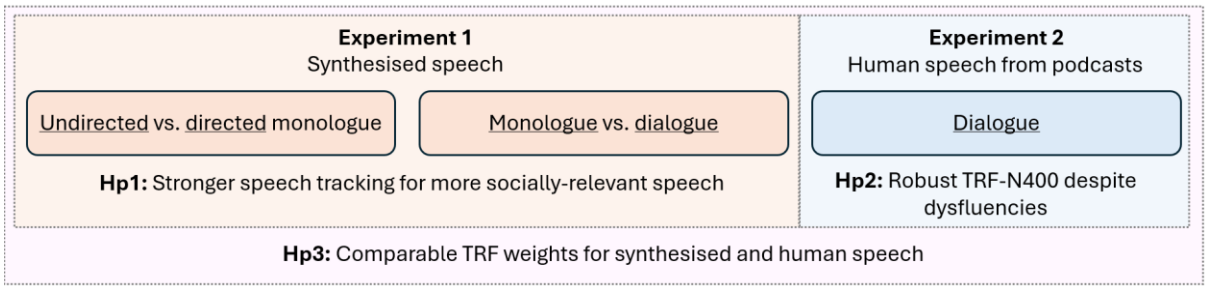

Figure 1. *Study conditions and hypotheses for two speech listening experiments, using synthesised speech or popular podcasts.*

Materials and Methods

EEG Experiment 1

Data acquisition and experimental setup. Twenty healthy individuals (10 male, 10 female; between 18 and 30 years old, median = 22) participated in the first EEG experiment. All participants were native English speakers and reported no history of hearing impairment or neurologic disorder, provided written informed consent, and were paid for participating. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the ethics committee of the School of Computer Science and Statistics at Trinity College Dublin. The experiment was carried out in a single session for each participant. EEG data were recorded from 64 electrode positions, digitised at 512 Hz using a BioSemi ActiveTwo system. Audio stimuli were presented at a sampling rate of 48 kHz using Sennheiser HD 650 headphones and

the Psychtoolbox-3 MATLAB library (Brainard, 1997). Testing was conducted at Trinity College Dublin in a dark room. Participants were instructed to maintain visual fixation on a crosshair centred on the screen and to minimise motor activities during the speech listening task.

Stimuli. Monologue scripts were generated with the state-of-the-art large language model (LLM) GPT-4 (OpenAI, 2024), using prompts such as "Could you create a total of 30 monologues, each approximately 2 minutes long?". Sixteen monologue scripts were selected from a larger set of about 30 scripts to ensure sufficient topic variation between scripts, and to exclude sensitive topics (e.g., scenarios involving extreme sadness and anger). Eight of these monologues were generated to be directed to a listener (appending “Could you turn 8 of them into directed monologues that address the listener” to the prompt); eight of them were undirected (appending “Could you turn 8 of them into undirected monologues in which the listener would passively hear the speaker talking about herself” to the prompt). The resulting monologue scripts (16 monologues (8 undirected, 8 directed)) were converted into dialogues using GPT-4 using the prompt “Could you convert the same monologue texts into dialogues by adding an interlocutor”. We converted monologue scripts to dialogue scripts to equate the content for monologue and dialogue conditions.

Scripts were synthesised using the realistic AI-based voice generation software PlayHT version 2.0[1]. The generated speech sound was manually inspected to maximise voice realism. Generation artifacts and portions sounding unnatural were replaced by re-running the generation for the affected speech chunks (one sentence

[1] https://app.play.ht/

at a time), ensuring a consistent tone and emphasis. Note that this was possible as each generation produces a different result. At the end of these processes 32 generated voice (16 monologues, 16 dialogues) were selected as the final stimuli. Additional details on the stimuli should be noted, which are relevant to a separate analysis beyond the present investigation. Speech stimuli were generated using various emotional tones (happy, sad, fear, anger, neutral), balancing tones across conditions.

Experimental procedure. All thirty-two stimuli were presented in a single experimental session in a randomised order, which was different across participants. Participants were free to take breaks of arbitrary length in-between stimuli. Each stimulus was split into two segments. The first 3 seconds were played first, followed by questions on emotion detection and perception: “What emotion did you detect?” and “What emotion did you feel?”. Participants were asked to indicate one of the following answers: 1: Happiness; 2: Sadness; 3: Fearful; 4: Neutral; 5: Anger. Next, participants were presented with the remaining part of the speech stimulus. The present investigation focussed on the analysis of the second part of each stimulus, ignoring the first 3 seconds and the behavioural task, which targeted a different research question.

EEG Experiment 2

Data acquisition and experimental setup. Twenty healthy individuals (10 male, 10 female; between 20 and 39 years old, median = 26), separate from Experiment 1, participated in the second EEG experiment. All participants were native English speakers and reported no history of hearing impairment or neurologic disorder, provided written informed consent, and were paid for participating. The study was

undertaken in accordance with the Declaration of Helsinki and was approved by the ethics committee of the School of Psychology at Trinity College Dublin. The experiment was carried out in a single session for each participant. EEG data were recorded from 64 electrode positions, digitised at 512 Hz using a BioSemi ActiveTwo system. Audio stimuli were presented at a sampling rate of 44,100 Hz using Sennheiser HD 280 Pro headphones and the PsychoPy Python library version 2022.2.5 (Peirce et al., 2019). Testing was conducted at Trinity College Dublin in a dark room. Participants were instructed to maintain visual fixation on a crosshair centred on the screen and minimise motor activities during the speech listening task.

Stimuli. Podcast recordings were selected from publicly available sources: 1) American podcast shows called *Brains On!* and *Forever Ago;* 2) *WIRED 5 Levels of Difficulty* series on YouTube; and 3) interviews with the same two hosts *Forever Ago,* available online[2]. Dialogues contained general interest topics, such as questions about science and math (e.g., why do we sweat, why memories as a young child are forgotten, etc.). Altogether, forty-two podcast recordings involving 1-to-1 dialogues were selected for this experiment (the same stimuli for all participants), totalling one hour in duration. Sixteen of those recordings involved adult-adult (AA) dialogues, while the remaining twenty-six were adult-child (AC) dialogues. The higher number of recordings for the latter condition ensured that the two conditions had a comparable number of words, as child-speech and child-directed speech are generally slower than adult-speech (Snow, 1972; Ferreira, 2019). To that end, the selected speech registers allowed us to test the robustness of our findings (e.g., the impact of dysfluencies on

[2] https://laist.com/podcasts/servant-of-pod/kids-podcasts-a-true-alternative-to-screen-time; https://creators.spotify.com/pod/profile/martine-severin/episodes/31--How-to-Become-an-Epic-Improv-Comic-with-Joy-Dolo-e1js9c2

the EEG encoding of speech) to variations in acoustic and lexical properties of a speech input.

Experimental procedure. Participants were presented with the forty-two speech stimuli (referred hereafter as trials) in a random order, separated by a short break. After each trials, participants were asked to rate how interesting they found the speech segment on a scale from 1-5 (1 = not at all interesting; 3 = neutral; 5 = very interesting). For a random subset of twenty percent of the trials, the participants were asked a behavioural multiple-choice question to select the set of keywords that best described the trial they just heard. This was to keep the participants engaged and on task throughout the experiment.

EEG data preprocessing

Neural data from both experiments were analysed offline using MATLAB software (MathWorks), using the same preprocessing and analysis pipeline. EEG signals were digitally filtered using Butterworth zero-phase filters with order 2, implemented with the function *filtfilt*. The primary analyses were carried out on EEG data band-pass filtered between 0.5 and 8 Hz, as that range has been shown to be particularly sensitive to the cortical tracking of speech (Doelling et al., 2014; Di Liberto et al., 2015; O'Sullivan et al., 2015; Vanthornhout et al., 2018). Low-frequencies from 0.5 Hz to 1 Hz were included as they were previously shown to be important for measuring the neural encoding of lexical predictions with EEG (Broderick et al., 2018; Carta et al., 2024). Secondary analyses were carried out to pinpoint specific sub-bands driving effects of interest. In Experiment 1, we considered the delta-band (0.5-4 Hz), theta-band (4-8 Hz), alpha-band power (8-12 Hz), and beta-band power (12-30 Hz). In Experiment 2, we considered just the delta-band (0.5-4 Hz) and theta-band (4-8 Hz), as they led to the highest EEG prediction correlation in Experiment 1 (**Figure S1**). The power of

alpha and beta bands was calculated as the instantaneous magnitude of the Hilbert transform in those sub-bands. This selection is based on evidence that alpha and beta rates reflect predictive processing, attention, and working memory (Samaha et al., 2015; Haegens et al., 2017). EEG signals were down sampled to 64 Hz, reducing computational and storage costs. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by a spherical spline interpolation (Perrin et al., 1989; Delorme and Makeig, 2004). Signals were then re-referenced to the average of the two mastoid channels.

Stimulus features

The present study examines the neural encoding of speech and language features during speech listening. The same features and analyses were used across the two experiments.

Sound envelope. The sound envelope was calculated as the instantaneous magnitude of the Hilbert transform of the sound-pressure waveform. The speech envelope captures the slow amplitude modulations of the speech signal, which have been shown to be robustly reflected in the EEG signal (Lalor and Foxe, 2010; Ding and Simon, 2012; Crosse et al., 2016b). The envelope-related features assessed in this study comprised the sound envelope and envelope derivative (calculated from the envelope at the original sampling rate with the MATLAB *diff* function). Matching the processed EEG signal, these acoustic features were then down sampled to 64 Hz.

Word onsets. The forced alignment software DARLA (Reddy and Stanford, 2015) was used to produce a word-by-word time-alignment between each speech segment and its transcript. The onset time of all words in the speech stimuli was then coded into

sparse vectors of zeros and ones with the same length and sampling rate as the other features, where ones represent the onset time of a word, consistent with previous studies (Broderick et al., 2018; Di Liberto et al., 2021; Chalehchaleh et al., 2025).

Lexical surprise and entropy. The second experiment in the present study investigates the possibility of measuring neural signatures of lexical surprise and entropy in the presence of dysfluencies, which are typically part of dialogues (Tree, 1995; Fox Tree, 2001; Clark and Tree, 2002; Bosker et al., 2014). Next-word lexical surprise and entropy values were extracted from Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), an open source state-of-the-art LLM that has been fine-tuned on conversation datasets. For each trial, lexical surprise was calculated as the negative log contextual probability of each word (Heilbron et al., 2022; Slaats and Martin, 2025). Additionally, for each trial lexical entropy was calculated as the weighted sum over the surprise values for all possible next words, and measures the level of uncertainty the model has for its next-word prediction (Goldstein et al., 2022b; Slaats and Martin, 2025). Lexical surprise and entropy feature vectors were built by scaling the word onset vector by its corresponding surprise and entropy value, respectively. Previous work employing similar models (e.g., GPT-2) found lexical expectation values to reliably explain EEG/MEG variance when considering speech listening tasks (Heilbron et al., 2022; Dou et al., 2024). However, that work typically involved scripted speech monologues without dysfluencies.

Dysfluencies. To investigate the impact of dysfluencies on lexical surprise and entropy and their neural signatures, we built two lexical surprise and entropy vectors: (1) one that included dysfluencies and (2) one that excluded dysfluencies in the speech

material (**Figure 4A**). Such dysfluencies were manually transcribed, annotated, and removed (see the dysfluency types considered in **Table S1**). To balance the amount of data in both versions of the surprise and entropy vectors, the annotated dysfluencies were removed from both feature vectors. Therefore, the main difference in the two versions of the surprise and entropy vectors lies primarily in if/how the presence of dysfluencies impacts the predictions of upcoming words. In practice, this led to two surprise and entropy feature vectors with the same word onsets, but different surprise and entropy values.

Temporal response function analysis

A system identification approach was employed to determine how specific stimulus features map onto EEG activity. This approach, known as the Temporal Response Function (TRF) (Edmund C. Lalor, 2009; Ding and Simon, 2012), utilises regularised ridge linear regression (Crosse et al., 2016) to estimate a filter, or model, that best captures how the brain converts input stimulus features into the resulting neural responses. One of the two metrics derived from the TRF is the weights of the model over time. The brain's response to a speech stimulus is not instantaneous; rather, it unfolds over time when speech is heard. Therefore, the brain's electrical activity is affected not just at the moment the sound occurs, but also over a subsequent time period. This delayed and extended influence is captured in the time lags of the TRF weights between the stimulus and the EEG response. These time lags can be analysed both spatially (across the scalp) and temporally (in terms of response delays).

As is typical (Crosse et al., 2016a; Crosse et al., 2021), to evaluate the model's ability to generalise and avoid overfitting, we used leave-one-out cross-validation across

trials. In this process, TRF models were trained on all but one trial and then used to predict the EEG signal for the left-out trial. The TRF model performance is then assessed by computing Pearson's correlation between the predicted EEG signal and the actual EEG signal on the left-out fold at each scalp electrode. This constitutes the second metric derived from the TRF: the prediction correlation, or the EEG variance explained by a specific set of speech feature(s). The EEG variance captured by speech features is interpreted as how well the model was able to map the encoding of speech features in the brain signals. Fitting the model with one speech feature results in a *univariate model,* while fitting the model while more than one speech feature results in a *multivariate* model.

Using these objective metrics, univariate forward TRF models were used to assess the corresponding neural response to each speech feature of interest. Three univariate TRF models were fit in Experiments 1 and 2 for the following speech features: the speech envelope, envelope derivative, and the word onset. In Experiment 1, the strength of the neural response to the speech in each of the three conditions was assessed through the TRF models' prediction correlations for any effect of social relevance. In Experiment 2, the temporal dynamics of the EEG response to podcast dialogues were examined from the TRF weights for latencies that would be consistent with previous work on audiobook listening tasks, notably the TRF-N400 response that indexes speech comprehension (Heilbron et al., 2022).

Multivariate TRF (mTRF) analyses were also conducted in Experiment 2 to assess if the cortical encoding of lexical surprise and entropy was more aligned with a model including or disregarding dysfluencies. This enables us to evaluate how much, if at all, dysfluencies impact lexical predictions during speech, informing us on the most appropriate approach for measuring lexical expectations in the human brain in

contexts involving real speech (i.e., with dysfluencies) (Heilbron et al., 2022; Mousavi et al., 2024). Sound envelope and word segmentation information features (e.g., speech, speech derivative, word onset) were concatenated to represent the stimulus, along with one of two versions of surprise: either dysfluent or non-dysfluent surprise.

Statistical analyses

Parallel statistical procedures were applied to the datasets from the two experiments. Results of the univariate TRFs models were analyses using one-tailed *t*-tests ($p < 0.05$) to determine if prediction correlations in each TRF (i.e., speech envelope, envelope derivative, word onset) was different from zero. Secondly, Wilcoxon signed-rank tests were used to compare EEG prediction correlations and TRF weights between conditions, corrected for multiple comparisons for the latter by considering that the same test was repeated at each time lag. To that end, we used the false discovery rate (FDR) of 0.05 according to the procedure set out by (Benjamini and Yekutieli, 2001) for dependence assumptions.

The dysfluency analysis in Experiment 2 included repeated measures ANOVAs. Post-hoc tests using paired *t*-tests with the Holm correction were performed to examine the effect of including/excluding dysfluencies on TRF model predictions, from which we report the adjusted *p*-value. Matlab was used for statistical analysis in Experiment 1. For Experiment 2, statistical analyses were performed using Matlab and JASP software (Love et al., 2019).

Code and data accessibility

The TRF analysis was conducted using the publicly available and open source mTRF-toolbox version 2.6 (Crosse et al., 2016b; Crosse et al., 2021), which can be

downloaded from https://github.com/mickcrosse/mTRF-Toolbox or https://sourceforge.net/projects/ aespa/. Stimulus features and EEG data were stored according to the continuous-event neural data structure CND (Liberto et al., 2024). Analyses were conducted using scripts shared via the GitHub channel (https://github.com/CNSP-Workshop/CNSP-resources/tree/main/CNSP/ exampleScripts) of the open science initiative Cognition and Natural Sensory Processing (CNSP) and Natural Sensory Processing (CNSP-initiative; https://cnspworkshop.net/). Custom scripts were used for data visualisation and statistical analyses. Datasets for Experiments 1 and 2 will be publicly shared via the open science repository OSF (osf.io) after publication and included in the list of datasets of the CNSP-initiative.

Results

Experiment 1

Univariate TRFs were fit to describe the relationship between speech envelope, envelope derivative, and word onset with the broadband EEG signal. Separate TRF models were fit for the three stimulus features. First, we report results for the widely used sound envelope feature as in previous studies (Vanthornhout et al., 2018; Muncke et al., 2022). EEG prediction correlation values calculated with leave-one-out cross validation (see **Materials and Methods**) had magnitudes compatible with envelope TRF analyses in previous research (r~0.2-0.4, after averaging across all participants and EEG channels). Comparing that metric across conditions revealed a statistically significant effect of social relevance when considering envelope TRFs (one-sample ANOVA, $F(2, 57) = 5.98$, $p = 0.006$), with more socially-relevant stimuli producing larger EEG prediction correlation values (post hoc two-tailed Wilcoxon

signed-rank tests; undirected < directed monologue: $p = 0.008$; undirected monologue < dialogue: $p = 0.003$; **Figure 1A**). No statistically significant differences emerged between the directed monologue and dialogue conditions ($p = 1$). No statistically significant effects of social relevance were measured for the other features when considering broadband EEG (envelope derivative: one-sample ANOVA, $F(2, 57) = 2.16$, $p = 0.129$; word onset: one-sample ANOVA, $F(2, 57) = 1.03$, $p = 0.366$).

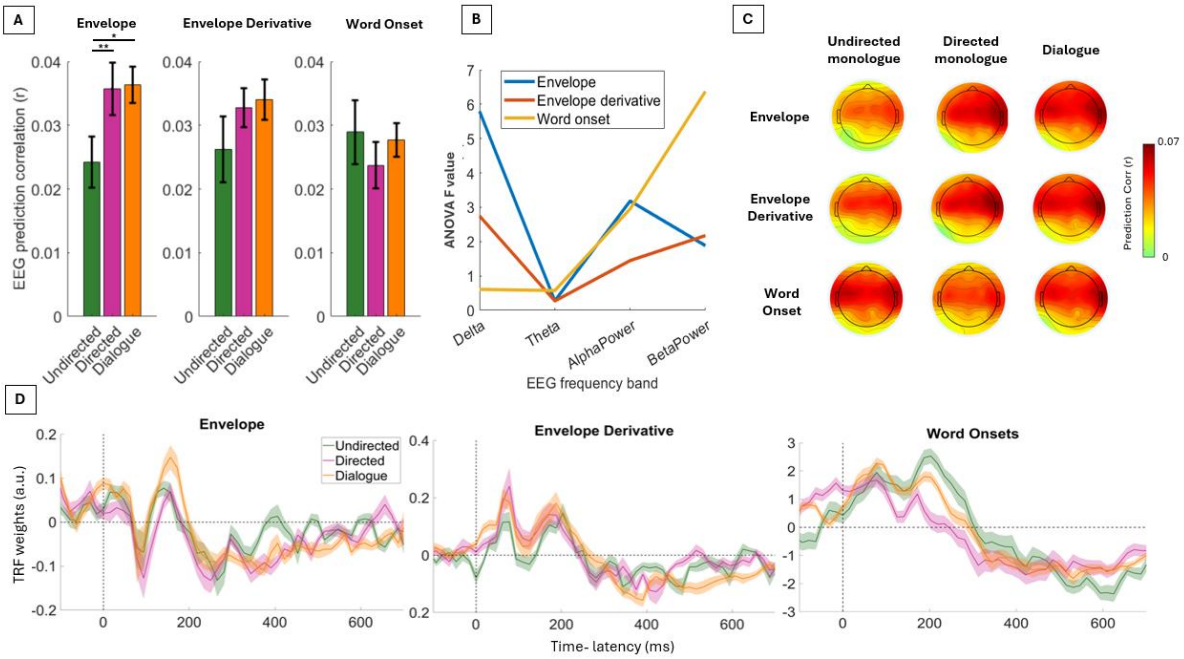

Figure 1. Social relevance impacts the cortical tracking of speech. (A) *Univariate forward TRFs were fit for envelope, envelope derivative, and word onset features to predict the on the broadband EEG in Experiment 1. EEG prediction correlations (Pearson's r), averaged across all EEG channels and participants, are shown (Mean ± SEM) comparing the three experimental conditions i.e., undirected monologue, directed monologue, and dialogue conditions. Envelope TRFs for the socially-relevant conditions exhibited stronger EEG prediction correlations (one-way ANOVA and Wilcoxon signed-rank post hoc tests, *p < 0.05, **p < 0.01).* ***(B)*** *The same analysis was repeated for all EEG sub-bands of interest. F-value of the one-way ANOVA is reported for each sub-band, indicating the effect of social relevance across features and EEG frequency bands. The strongest effects of social relevance for the envelope TRF emerged in the delta-band, and for word-onset in the gamma-band.* ***(C)*** *Scalp topographies of the EEG prediction correlations for each condition and speech feature. Warmer colours indicate higher prediction correlation values. All topographies showed qualitatively similar spatial patterns, compatible with previous studies on continuous speech listening (Di Liberto et al., 2015)* ***(D)*** *TRF weights averaged across all participants at channel Cz are shown for the three conditions and stimulus features. Shaded areas represent the standard error of the mean (SEM) across participants.*

TRF and ANOVA analyses were repeated for different EEG sub-bands to determine if the effect was specific to certain cortical rhythms. The envelope TRF analyses showed strongest effects in low-frequency EEG (**Figure 1B**), with a statistically significant effect in the delta-band (one-sample ANOVA, $F(2, 57) = 5.80$, $p = 0.006$); a trending effect in the alpha-band ($F(2, 57) = 3.18$, $p = 0.053$); and no statistically significant effects elsewhere (theta-band: $F(2, 57) = 0.27$, $p = 0.763$; beta-band: $F(2, 57) = 1.89$, $p = 0.166$). Conversely, the word onset TRF showed strongest effects in higher EEG frequencies, with a statistically significant effect in the beta-band (one-sample ANOVA, $F(2, 57) = 6.38$, $p = 0.004$); a trending effect in the alpha-band ($F(2, 57) = 2.95$, $p = 0.064$); and no statistically significant effects elsewhere (delta-band: $F(2, 57) = 0.61$, $p = 0.551$; theta-band: $F(2, 57) = 0.57$, $p = 0.559$). No statistically significant effects were measured for the envelope derivative TRF (delta-band: $F(2, 57) = 2.75$, $p = 0.077$; theta-band: $F(2, 57) = 0.27$, $p = 0.765$; alpha-band: $F(2, 57) = 1.45$, $p = 0.248$; gamma-band: $F(2, 57) = 2.17$, $p = 0.128$).

We also report TRF weights for the models in the EEG broadband. Based on previous research, we expected a P1-N1-P2 pattern with latencies around 40, 80, and 140ms respectively (Lalor and Foxe, 2010; Di Liberto et al., 2015; Crosse et al., 2016b), where longer latencies have been linked with factors such as selective attention (Power et al., 2012). The literature also informs us on the neural dynamics corresponding with higher order features like word onsets, such as a large slow negative deflection at a latency of approximately 400ms (Broderick, 2018). Here, we found TRF patterns that were qualitatively similar to that literature in all those aspects.

Experiment 2

EEG signals were recorded as participants listened to spontaneous two-party dialogues from popular podcasts, each involving a podcast host and a guest. Forward univariate TRFs were derived for the two speech features showing an effect of social relevance: the sound envelope and word onset features (**Figure 2**; we also report the envelope derivative results in **Figure S2**). As in Experiment 1, EEG prediction correlations were calculated with leave-one-out cross validation for each stimulus feature separately. EEG prediction correlations were above chance for both sound envelope (**Figure 2A-left**; *t*-tests, $p = 3.64 \times 10^{-11}$) and word onsets (**Figure 2B-left**; *t*-tests, $p = 2.77 \times 10^{-10}$), reflecting a statistically significant EEG encoding of the two features.

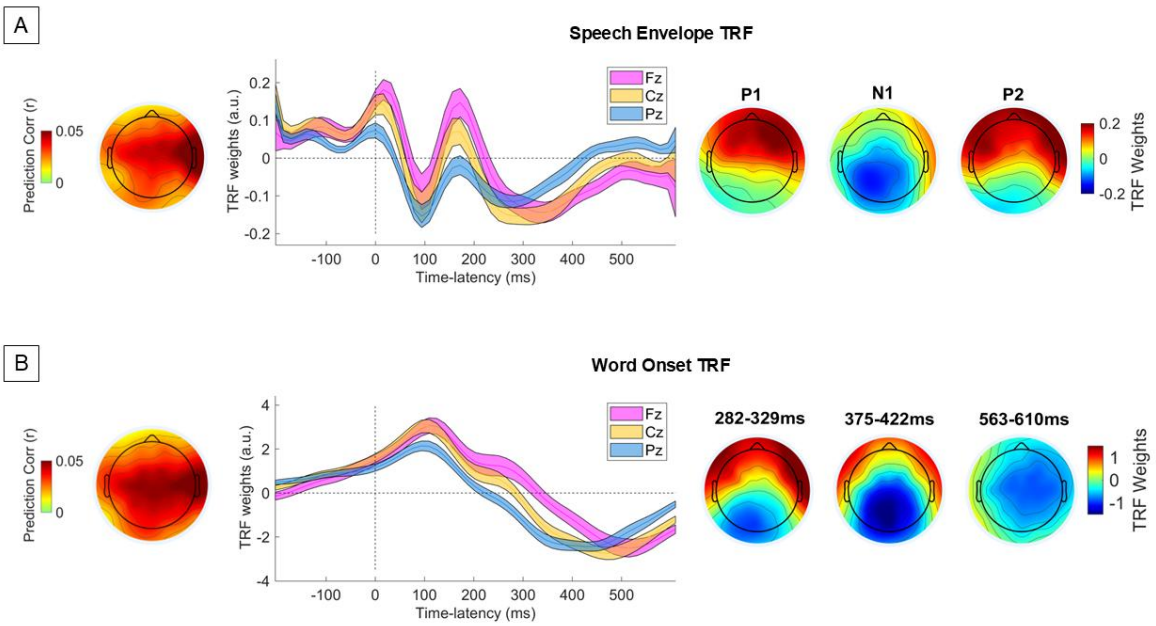

Figure 2. Characterising the cortical tracking of speech using real podcast dialogues. *Univariate forward TRF models in Experiment 2 for envelope* ***(A)*** *and word onset* ***(B)*** *features. These analyses aggregated data from both listening conditions (AA and AC). Scalp distributions of the average EEG prediction correlations are reported (left panels). TRF weights at channels Fz, Cz and Pz (middle-panels) and the TRF weights at three key time-lags exhibit patterns compatible with the literature (Lalor, 2009; Di Liberto et al., 2015; Chalehchaleh et al., 2025).*

In Experiment 2, the podcast host was always an adult speaker, while the podcast guest was either an adult or a child. This led to two conditions (adult-adult conversations - AA; adult-child conversations - AC) that we compared to determine if their multi-faceted differences (e.g., prosody, lexis, word-rate) would impact the envelope and word onset TRF model fits. While there may be other differences that exist elsewhere between AA and AC speech, for the purposes of the current analyses, we focused on just these two speech features. EEG prediction correlations averaged across scalp locations did not differ significantly between AA and AC stimuli (Wilcoxon paired signed-rank test (*p <0.05): envelope: p = 0.17; word onset: p = 0.16). The effect of stimulus type (AA versus AC) upon TRF feature weights at channel Cz was also examined for each time lag. No statistically significant differences emerged (FDR-corrected Wilcoxon paired signed-rank test, *p > 0.05 across all time lags). On this basis, the analyses that follow are carried out by combining AA and AC stimuli. While no effects of speech register emerged in this investigation, further work could investigate possible effects on other linguistic features (Piazza et al., 2025).

The TRF weights were qualitatively comparable with the previous literature (Lalor et al., 2009; Di Liberto et al., 2015) and with the results from Experiment 1, with the envelope TRF presenting the typical P1-N1-P2 characteristics, and the word onset TRF exhibiting the typical dominant negative deflection at about 400 ms (**Figure 2-middle**). We measured Pearson's correlations between the TRF weights across the two experiments.

A

		Correlations with real dialogue TRF (Exp 2)		
		Envelope (ENV)	**Envelope Derivative (ENVDER)**	**Word Onset (WO)**
Exp 1 Condition	Synthesised Dialogue	0.795	0.656	0.922
	Directed Monologue	0.793	0.649	0.883
	Undirected Monologue	0.699	0.475	0.755

B

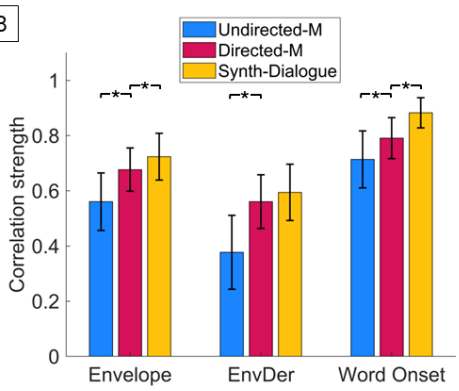

Figure 3. Speech podcast TRFs are more correlated with synthesised speech TRFs involving socially-relevant speech than undirected monologues. (A) *Correlations between average TRF weights of the 64 EEG channels from the two experiments are reported. TRF weights for the real podcast dialogues in Experiment 2 shows high correlations with the three conditions in Experiment 1 in the following descending order: synthesised dialogue, directed monologue, undirected monologue. This pattern emerges for all speech features in the analysis.* ***(B)*** *A bootstrap resampling procedure confirmed the robustness of this pattern, with statistically significant differences consistently observed for the envelope and word onset features across all conditions.*

The two datasets differed in both stimuli material and participants. To this end, to ensure the robustness of this cross-experiment comparison, we performed a bootstrap resampling procedure on the group-averaged TRF weights and correlated them for the following stimulus features: envelope, envelope derivative, and word onset. The resulting correlations after averaging the TRF weights of the 64 EEG channels across participants are shown in **Figure 3A.** The general assumption in averaging TRFs (as for ERPs) is that TRFs for individual participants may vary, but that participants share certain consistent dynamics, which would emerge by averaging. As such, the assumption is that averaging TRFs from two distinct samples of participants (from the same population) would lead to the same TRF (when enough participants are available). Accordingly, we tested the robustness of the result via a bootstrap resampling procedure that derived 1,000 average TRFs for each of the conditions in Experiment 1 and for Experiment 2, generated using a random sample with replacement on each participant cohort. This led to 1,000 correlations between

bootstrap average TRF weights, revealing a statistically significant effect of condition (two-way repeated measures ANOVA: $F(2,1998) = 2436.51$, $p = 0.00$, $\eta^2 = 0.22$). *Post hoc* comparisons indicated significantly higher correlations between conditions (**Figure 3B**), whereby the highest correlations occurred in the real dialogues with the conditions in Experiment 1 in the order of synthesised dialogues > directed monologue > undirected monologue for the envelope (*t*-tests on envelope feature; $r_{\text{directed-M vs. real-Dialog}} > r_{\text{undirected-M vs. real-Dialog}}$, $t(999) = -31.60$, $p_{holm} = 3.51 \times 10^{-200}$, Cohen's $d = -1.22$; $r_{\text{synth-Dialog vs. real-Dialog}} > r_{\text{directed-M vs. real-Dialog}}$, $t(999) = -9.73$, $p_{holm} = 2.08 \times 10^{-21}$, Cohen's $d = -0.38$) and word onset feature (*t*-tests on word onset feature; $r_{\text{directed-M vs. real-Dialog}} > r_{\text{undirected-M vs. real-Dialog}}$, $t(999) = -23.59$, $p_{holm} = 3.13 \times 10^{-116}$, Cohen's $d = -0.91$; $r_{\text{synth-Dialog vs. real-Dialog}} > r_{\text{directed-M vs. real-Dialog}}$, $t(999) = -23.13$, $p_{holm} = 5.19 \times 10^{-112}$, Cohen's $d = -0.90$). This result also emerged with the envelope derivative feature, but only in the correlations between the real dialogues and two conditions of Experiment 1: directed monologue and undirected monologue ($r_{\text{directed-M vs. real-Dialog}} > r_{\text{undirected-M vs. real-Dialog}}$, $t(999) = -70.19$, $p_{holm} = 0.00$, Cohen's $d = -3.00$).

Our expectation in this analysis was that social relevance would exert similar impacts on all socially relevant conditions, thus making the real dialogue TRF from Experiment 2 more strongly correlated with the socially-relevant TRFs in Experiment 1 (directed monologue and dialogue) than the socially-irrelevant undirected monologues. Our results revealed patterns compatible with this premise for each of the speech features in the analysis (**Figure 3**). Remarkably, although the two datasets differed in both stimuli material and participants, the synthesised dialogue TRF and real podcast TRF exhibited very high correlation scores, especially for the word onset TRF ($r = 0.92$).

Investigating the impact of dysfluencies on lexical expectation measurements

Experiment 2 involved real podcast recordings, which include dysfluencies by nature. Indeed, the dysfluency types and frequencies can vary due to a variety of factors (e.g., speaker, social context, social anxiety, fatigue, familiarity with interlocutor, knowledge of topic, speaking a native or non-native language). To investigate the impact of dysfluencies on the lexical surprise and entropy feature and its corresponding brain response, two lexical surprise and entropy vectors were constructed, including and excluding dysfluencies in the speech material, respectively (**Figure 4A**). As such, we carried out a characterisation of the dysfluency type and frequency for this specific dataset split across conditions AA and AC (**Figure 4B**), finding a majority of filler-type dysfluency and an overall frequency of dysfluent words of 3.36%, whereby ~5% of all words were filler-type dysfluencies (e.g., “uh”/”like”/”you know”), ~1% were repetitions, with less than 1% consisting of replacements and restarts (see **Materials and Methods** and **Table S1**). The frequency distribution of these dysfluencies types was consistent across podcast episodes, with a statistically significant main effect of dysfluency type (repeated measures ANOVA: $F(15,4) = 28.49$, $p = 2.79 \times 10^{-13}$, $\eta^2 = 0.42$; *post hoc* tests in **Table S2**). We found no statistically significant interaction of dysfluency type and experimental condition (adult-adult podcast AA; adult-child podcast AC) (repeated measures ANOVA: $F(15,4) = 1.04$, $p = 0.39$, $\eta^2 = 0.02$).

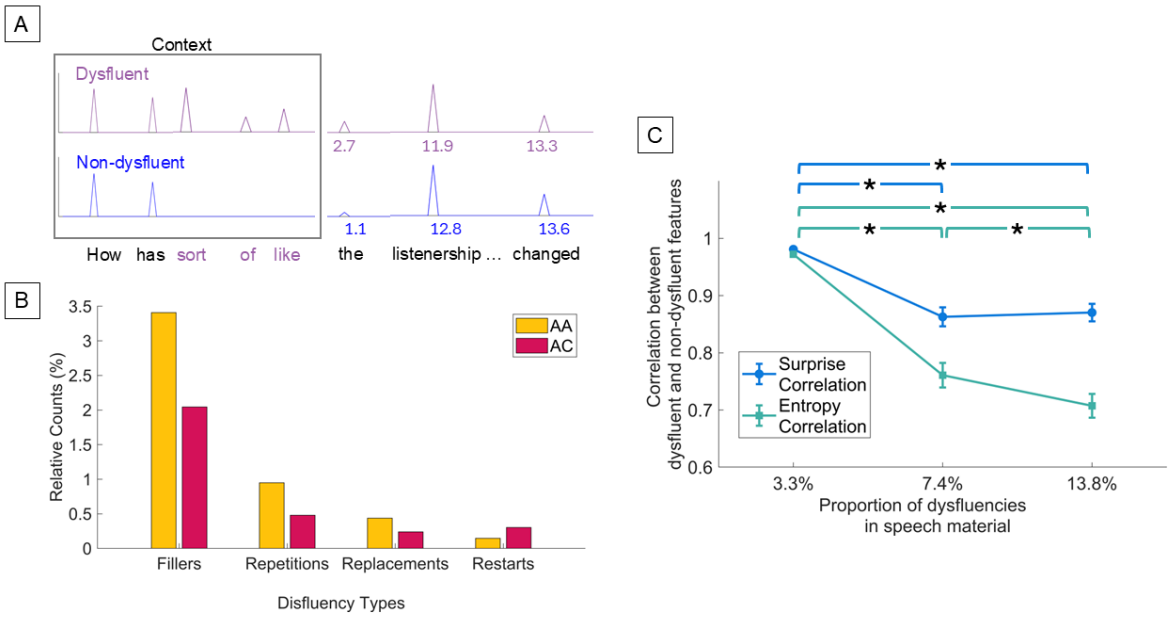

Figure 4. Investigating the impact of dysfluencies on LLMs. (A) *Audio transcripts were analysed using two different lexical surprise and entropy vectors. While both vectors only included non-dysfluent words, for sake of comparison, lexical surprise and entropy were calculated using Mistral by including or excluding dysfluent words from the text input, producing the dysfluent and non-dysfluent feature vectors.* ***(B)*** *Dysfluencies were identified manually in the text material. We report the dysfluency frequency organised according to dysfluency type and experimental condition (adult-adult podcast AA; adult-child podcast AC).* ***(C)*** *The correlations between Mistral's expectations when including or excluding dysfluent words in the calculation of lexical surprise and entropy are shown, quantifying the potential impact of dysfluency on the measurement of lexical expectations. Doubling the number of dysfluencies led to a statistically significant effect of dysfluency. On this basis, future research should measure dysfluencies in the speech material, since differences across participant cohorts or conditions may reflect differences in the prevalence of dysfluent words in the heard or produced speech.*

Here, we investigated if the presence of dysfluent words in the speech material (and its transcript) significantly impacts our ability to model the neural encoding of lexical expectations. To that end, we introduced two additional stimulus features in Experiment 2. These comprised word onset vectors modulated by lexical surprise and entropy values calculated using Mistral (see **Materials and Methods**). The first vector involved surprise values calculated when including all words, of which 3.3% of the were dysfluent. The second vector only included non-dysfluent words (dysfluencies were identified manually). However, dysfluent and non-dysfluent surprise values did

not differ in this dataset, with a Pearson's correlation between the two surprise vectors of 0.98. Due to this high correlation, we also did not expect differences to report in the coupling between these vectors and the corresponding EEG signals. While we could measure a statistically significant neural encoding of lexical surprise (using either dysfluent or non-dysfluent vectors) when compared to a baseline acoustic model (**Figure S3**), this is not guaranteed to be the case in speech material with more dysfluencies.

These virtually identical regressors suggest that at this low volume of dysfluencies, using speech material that includes or excludes dysfluencies would render little to no difference in the neural processing and encoding of speech. However, social speech can widely vary in the number of dysfluencies depending on numerous factors (e.g., speaker, social context, anxiety, tiredness, familiarity with interlocutor, knowledge of topic, speaking a native or non-native language). As a consequence, the neural encoding of speech may be impacted when more dysfluencies are present. An excessive number of dysfluencies (e.g., more dysfluent than non-dysfluent words) may disrupt lexical prediction in both the LLM and the human brain. Hence, we then assessed what level of dysfluency frequencies constitute an issue for lexical prediction in LLMs and to what extent this sensitivity persists by manually increasing the number of dysfluencies in the speech transcript. Specifically, we doubled and quadrupled the existing dysfluencies in Experiment 2, calculating the correlation between the lexical surprise and entropy values between the dysfluent and non-dysfluent vectors (**Figure 4C**). Lower correlations indicate a stronger impact of dysfluencies. We found a significant impact of dysfluency frequency (FDR-corrected Wilcoxon paired sign-ranked test: $p_{originalSur\ vs.\ doubledSur} = 3.57 \times 10^{-8}$, $p_{originalSur\ vs.\ quadSur} = 3.57 \times 10^{-8}$, $p_{originalEntro}$

$_{vs.\ doubledEntro}$ = 3.57 x 10^{-8}, $p_{doubledEntro\ vs.\ quadEntro}$ = 1.60 x 10^{-6}, $p_{originalEntro\ vs.\ quadEntro}$ = 3.57 x 10^{-8}), whereby there was practically no sensitivity to dysfluencies for the frequency in the original material (dysfluent-nondysfluent correlation ~ 1) and significant sensitivity to dysfluencies once they were doubled, though this effect seems to level out. Additionally, the strongest sensitivity overall emerged for entropy vectors (correlation down to ~0.7).

Discussion

Speech communication is an ideal testbed for investigating human cognition, as it involves a multitude of processes involved in the transformation of sound to meaning. Such processes encompass mechanisms such as statistical prediction, attention and cognitive effort. The possibility of extending that work to social speech opens up opportunities for probing a whole range of processes that are typically excluded from monologue listening research, such as social role and speaker familiarity, communication accommodation, social anxiety, and many others (Di Liberto and Ip, 2025). However, this opportunity comes with the important challenge of having to isolate the phenomena of interests from a much richer and harder-to-control set of behaviours and neural processes. One key for disentangling those processes and phenomena is to pinpoint how exactly these socially-relevant factors impact the spatial and temporal patterns reflecting the neural processing of speech and language. Here, we take a first step in that direction by means of a principled step-by-step approach, determining the impact of including a social element in speech listening tasks (excluding interaction in our scope of investigation).

The results from Experiment 1 indicate that social relevance substantially enhances the neural tracking of speech. That effect was directly related to social relevance and was not due to acoustic differences in the input. The increased encoding emerged specifically for the speech envelope, but not for word-level processing (EEG prediction correlation metric; **Figure 1A**). Previous work on selective attention has also shown similar increases in envelope tracking. Interestingly, however, the increased speech tracking in selective attention tasks also corresponded with changes in the TRF patterns, especially with a modulation of the TRF-P2 component (Power et al., 2012; Vanthornhout et al., 2018; Carta et al., 2024), while the increased envelope tracking here does not correspond to changes in the temporal or spatial dynamics of the TRF (**Figure 1D**). As such, our results are compatible with a general increased engagement with the task and stimulus, but with important differences from the widely studied selective attention phenomenon. It should be noted that we also detected sensitivity to the social relevance of speech in the beta band, specifically for word onset encoding. This result supports the potential and relevance of measuring cortical tracking at higher EEG rates (even though stronger speech tracking typically emerges in the delta and theta bands). While we did not have specific hypotheses pertaining to the EEG beta band in this study, we note that studying speech tracking at different frequency rates may provide an additional dimension for more clearly disentangling the cortical tracking of distinct speech features, and should be considered (without forgetting about the risks that come with filtering operations) (De Cheveigné and Nelken, 2019).

While Experiment 1 isolated the impact of social relevance on speech tracking, our design in Experiment 2 takes us a step further by measuring neural dynamics in more

naturalistic settings, wherein spontaneous speech introduces additional complexities. The results in Experiment 2 (**Figure 2**) demonstrated temporal response characterisations that are not only consistent with previous monologue-listening literature (Broderick et al., 2018; Di Liberto et al., 2021; Heilbron et al., 2022), but also with the neural speech encoding results obtained in Experiment 1 (**Figure 3**). Our results indicated similar spatial and temporal patterns for synthesised and real speech recordings; the EEG encoding of speech features, including the speech envelope and word onset, was highly correlated across synthesised and real speech recordings, achieving correlations of up to 0.92 for word onset TRFs. This is quite remarkable, considering that the two experiments involved completely different speech material and distinct participant cohorts. The fact that we observed a consistent increase in that similarity with social relevance is also a strong indication that the impact of social relevance on the spatial and temporal patterns of speech is consistent across material and participants. Further work is needed to better characterise and test the boundaries of consistency in neural tracking of speech features across synthesised and real speech and to characterise the sensitivity of TRFs to variations in synthesiser model and configuration. Nonetheless, our findings suggest that the similarity between synthesised and real speech TRFs and their modulation by the social relevance of speech are remarkably high.

Experiment 2 demonstrates the feasibility of employing the TRF using naturalistic and socially-relevant speech material. Our dysfluency analysis indicated that the presence of dysfluencies observed in the podcast stimuli utilised does not impact the ability to derive indices of surprise and entropy, which can, in turn, be used to improve TRF models. This is promising for future research, as it suggests the possibility of studying

speech tracking even in presence of dysfluencies. However, note here that the LLM dysfluency simulation analysis highlights that when a larger portion of the speech material includes dysfluencies, they can exert a notable impact on lexical expectation indices.

From this analysis, two key conclusions emerge. First, dysfluencies do not pose a problem up to a certain threshold, which may eliminate the need for manual transcript cleaning when using LLMs. Second, it is essential to quantify and report the dysfluency profile — both type and frequency — used in a given experiment, particularly when generating new stimuli. This task is relatively straightforward in listening experiments, where all participants are exposed to identical stimuli. In contrast, speech interaction paradigms, such as hyperscanning studies, introduce greater methodological complexity. A central challenge is that individual participants often exhibit unique dysfluency profiles (Befi-Lopes et al., 2014; Pistono et al., 2024), which can differentially affect measures of lexical expectations — both in terms of LLM-based estimation and neural encoding (Goldstein et al., 2022a). This variability reinforces the need for systematic dysfluency measurement and individualised reporting to ensure robust and interpretable findings. Finally, it is worth noting that the observed sensitivity of LLMs to dysfluency opens new avenues for research focused on its impact on the neural encoding of speech. Future investigations could specifically examine this effect in participant cohorts with speech production challenges, such as individuals who stutter or those experiencing social anxiety. Such studies would not only deepen our understanding of speech processing under naturalistic conditions but also offer insights into the neural mechanisms underlying communicative difficulties.

In summary, this study demonstrates that incorporating social relevance into speech listening tasks enhances the neural tracking of speech and reveals remarkable consistent spatial-temporal patterns across varied speech materials and participant cohorts. By extending analysis to spontaneous speech and dysfluency, we demonstrate the possibility of modelling the neural tracking of speech, as well as lexical prediction mechanisms even in presence of dysfluencies. Based on our simulation, we provide evidence for the importance of considering the issue of dysfluency when studying social speech, when listening to speech interactions and especially in experiments where participants engage in dialogue. Altogether, these findings lay the groundwork for future investigations into the neural encoding of socially complex and dysfluent speech, offering a principled path toward understanding communication in real-world contexts.

Supplementary

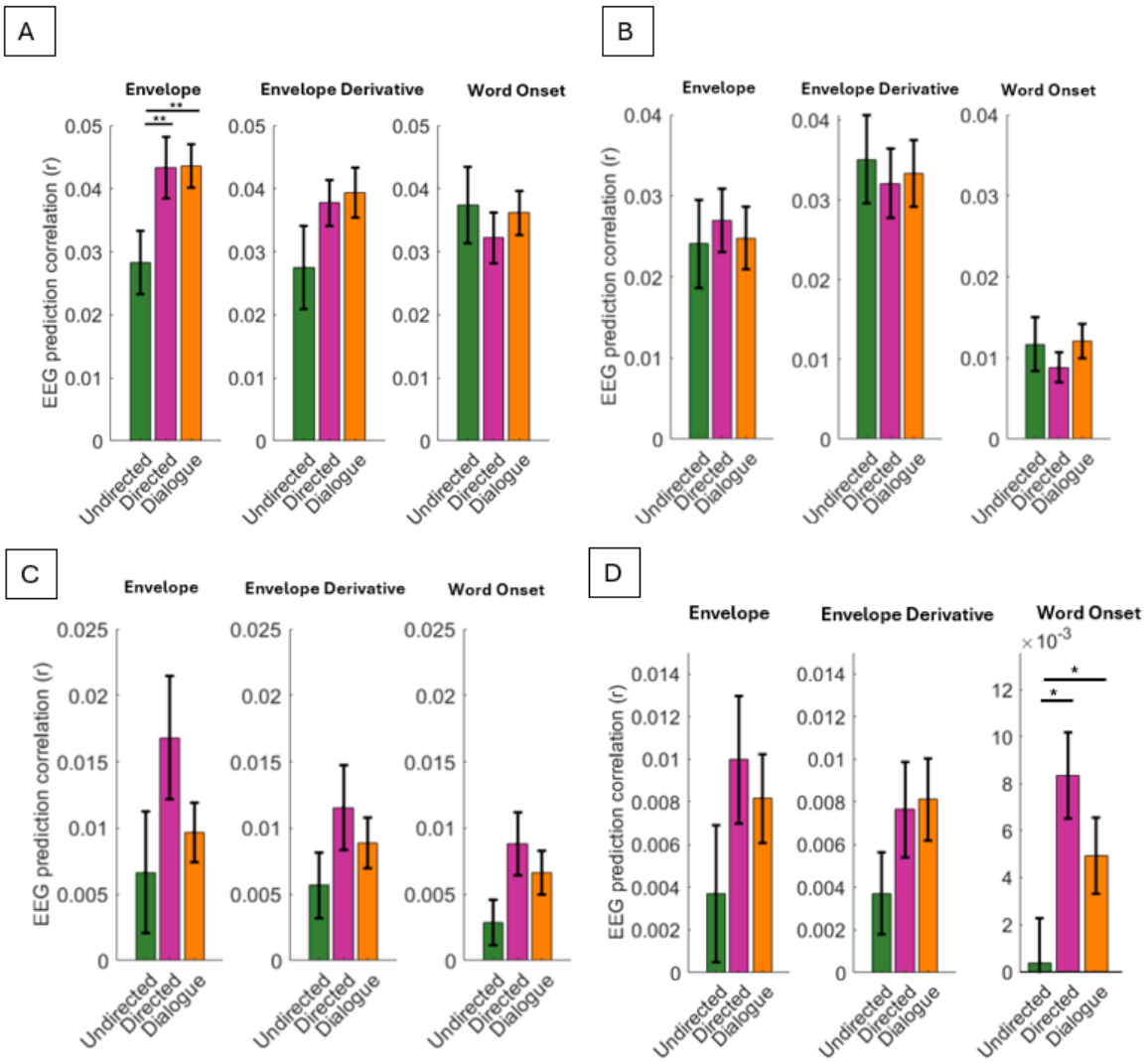

Figure S1. EEG prediction correlations for different frequency bands (Experiment 1). *Univariate forward TRFs were fitted for envelope, envelope derivative, and word onset features to predict the EEG signal at frequency bands delta (A), theta (B), alpha (C), and beta (D). EEG prediction correlations (Pearson's r), averaged over all channels and participants, are presented as mean ± SEM, comparing the three experimental conditions: undirected monologue, directed monologue, and dialogue. One-way ANOVA and Wilcoxon signed-rank post hoc tests were conducted (*p < 0.05, **p < 0.01).This data extends the results in Figure 1.*

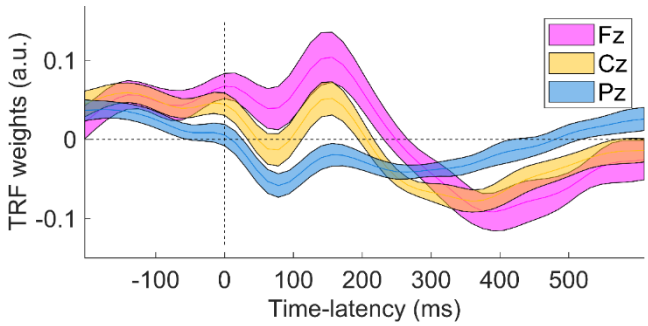

Figure S2. *TRF weights for the envelope derivative feature from Experiment 2 at three selected EEG channels. This data extends the results in Figure 2.*

Dysfluency Type	**Definition**
Fillers	uh / um, you know, sort of, like, I'll say, I don't know
Repetitions	I-I-I
Restarts (also coined *false start* in (Shriberg, 1999))	*The speaker abandons completely the initial utterance and restarts it. This type of dysfluency does not actually involve a repair as the speaker begins a completely new utterance(Passali et al., 2022).*
Replacements (also coined *repair* in (Shriberg, 1999))	*The speaker replaces the dysfluent word(s) or phrase with the fluent one. In replacements, the reparandum is replaced by the repair (Passali et al., 2022)*

Table S1. Dysfluency types. *The dysfluency analyses in Figures 4 and 5 constitute a preliminary approach for handling dysfluencies that are a natural part of social speech. This table indicates the four dysfluency types that were identified – fillers, repetitions, restarts (also known as false starts), and replacements (also known as repairs)* (Shriberg, 1999)*. See Figure 4B for the distribution of these dysfluency types in the speech material for Experiment 2. Note that Experiment 1 did not include any dysfluencies.*

		Mean Difference	SE	t	Cohen's d	p_{holm}
Fillers	**Repetitions**	5.812500	0.801358	7.253313	1.671350	7.394361×10^{-9}
	Replacements	6.718750	0.801358	8.384206	1.931937	9.777329×10^{-11}
	Restart	6.968750	0.801358	8.696177	2.003823	3.213573×10^{-11}

Table S2. Post Hoc Comparisons of dysfluency types*. A statistically significant main effect of dysfluency type emerged when comparing the frequency distribution of dysfluencies across trials in both AA and AC conditions with a repeated measures ANOVA. Post hoc tests show how fillers significantly outnumber any other dysfluency produced in real podcast dialogues. This data extends the results in Figure 4B.*

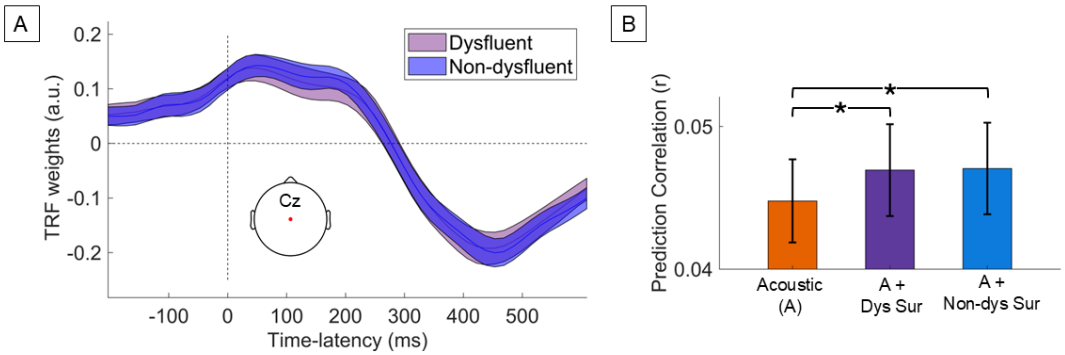

Figure S3. Investigating the impact of dysfluencies on speech TRFs. (C) *Multivariate TRFs were computed to predict the EEG signal using the dysfluent or non-dysfluent vectors, concatenated with nuisance regressors capturing speech acoustics and word segmentation (envelope, envelope derivative, and word onset). We report the TRF weights corresponding to the dysfluent and non-dysfluent surprise features at electrode Cz.* ***(D)*** *Both versions of the models that included acoustics and dysfluent surprises returned a significant gain in the average EEG prediction correlations in comparison to the acoustics-only mTRF, reflecting a statistically significant neural encoding of lexical surprise. These data extend the results in Figure 4.*

Forward multivariate TRFs were fit for each of the two vectors and including the three features used in the earlier analysis as nuisance regressors (envelope, envelope derivative, and word onsets), leading to a *non-dysfluent surprise TRF* and a *dysfluent surprise TRF*. Note that, for sake of comparison, dysfluent and non-dysfluent vectors

were built to be identical in everything but the surprise values i.e., they had the same word onsets. Finally, a third TRF was fit for the nuisance regressors without any lexical surprise information (*baseline* TRF). When comparing the model weights between the non-dysfluent and dysfluent surprise TRFs (Wilcoxon paired signed-rank test: p = 0.951; **Figure S3a**), no statistically significant differences emerged. In addition, EEG prediction correlations were compared across the three TRF models, showing a statistically significant differences across the three TRF models (repeated measures ANOVA: $F(1, 19) = 7.96$, $p = 0.001$, $\eta^2 = 0.30$; **Figure S3b**). *Post hoc* comparisons indicated larger EEG prediction correlations when including either surprise vector (*t*-tests; $r_{\text{non-dysfluent surprise}} > r_{\text{no-surprise}}$, $t(19) = -3.54$, $p_{holm} = 0.003$, Cohen's $d = -0.16$; $r_{\text{dysfluent surprise}} > r_{\text{no-surprise}}$, $t(19) = -3.36$, $p_{holm} = 0.004$, Cohen's $d = -0.16$). As expected, no statistically significant differences were measured between non-dysfluent and dysfluent surprise TRF prediction correlations (*t*-test: $t(19) = -0.18$, $p_{holm} = 0.861$, Cohen's $d = -0.01$).

References

Befi-Lopes DM, Cáceres-Assenço AM, Marques SF, Vieira M (2014) School-age children with specific language impairment produce more speech disfluencies than their peers. In: CoDAS, pp 439-443: SciELO Brasil.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29:1165-1188.

Bernabei JM, Arnold TC, Shah P, Revell A, Ong IZ, Kini LG, Stein JM, Shinohara RT, Lucas TH, Davis KA, Bassett DS, Litt B (2021) Electrocorticography and stereo EEG provide distinct measures of brain connectivity: implications for network models. Brain Commun 3:fcab156.

Bosker HR, Quené H, Sanders T, de Jong NH (2014) What makes speech sound fluent? The contributions of pauses, speed and repairs. Language Testing 31:159-175.

Brainard DH (1997) The Psychophysics Toolbox. Spatial Vis 10:433-436.

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.e803.

Broderick MP, 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.

Butler LK, Kiran S, Tager-Flusberg H (2020) Functional Near-Infrared Spectroscopy in the Study of Speech and Language Impairment Across the Life Span: A Systematic Review. American Journal of Speech-Language Pathology 29:1674-1701.

Carreiras M, Vergara M, Barber H (2005) Early event-related potential effects of syllabic processing during visual word recognition. Journal of Cognitive Neuroscience 17:1803-1817.

Carta S, Aličković E, Zaar J, Valdés AL, Di Liberto GM (2024) Cortical encoding of phonetic onsets of both attended and ignored speech in hearing impaired individuals. PloS One 19:e0308554.

Chalehchaleh A, Winchester MM, Di Liberto GM (2025) Robust assessment of the cortical encoding of word-level expectations using the temporal response function. Journal of Neural Engineering 22:016004.

Chen W-L, Wagner J, Heugel N, Sugar J, Lee Y-W, Conant L, Malloy M, Heffernan J, Quirk B, Zinos A, Beardsley SA, Prost R, Whelan HT (2020) Functional Near-Infrared Spectroscopy and Its Clinical Application in the Field of Neuroscience: Advances and Future Directions. Frontiers in Neuroscience 14.

Clark HH, Tree JEF (2002) Using uh and um in spontaneous speaking. Cognition 84:73-111.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016a) The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. Frontiers in Human Neuroscience 10:219245-219245.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016b) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Frontiers in Human Neuroscience.

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.

De Cheveigné A, Nelken I (2019) Filters: when, why, and how (not) to use them. Neuron 102:280-293.

Défossez A, Caucheteux C, Rapin J, Kabeli O, King J-R (2023) Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence 5:1097-1107.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134:9-21.

Destoky F, Philippe M, Bertels J, Verhasselt M, Coquelet N, Vander Ghinst M, Wens V, De Tiège X, Bourguignon M (2019) Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. NeuroImage 184:201-213.

Di Liberto GM, Ip EYJ (2025) Speech Neurophysiology in Realistic Contexts: Big Hype or Big Leap? arXiv preprint arXiv:250605494.

Di Liberto GM, Nie J, Yeaton J, Khalighinejad B, Shamma SA, Mesgarani N (2021) Neural representation of linguistic feature hierarchy reflects second-language proficiency. NeuroImage 227:117586-117586.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Current Biology 25:2457-2465.

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences 109:11854-11859.

Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage 85:761-768.

Dou J, Anderson AJ, White AS, Norman-Haignere SV, Lalor EC (2024) Dynamic modeling of EEG responses to natural speech reveals earlier processing of predictable words. bioRxiv:2024.2008.2026.609779.

Edmund C. Lalor AJP, Richard B. Reilly and John J. Foxe (2009) Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. Journal of Neurophysiology 102.

Ferreira VS (2019) A Mechanistic Framework for Explaining Audience Design in Language Production. Annu Rev Psychol 70:29-51.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. Science 322:970-973.

Fox Tree JE (2001) Listeners' uses of um and uh in speech comprehension. Memory & cognition 29:320-326.

Friederici AD (2011) The Brain Basis of Language Processing: From Structure to Function. Physiological Reviews 91:1357-1392.

Garrod S, Pickering MJ (2004) Why is conversation so easy? Trends in Cognitive Sciences 8:8-11.

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A (2022a) Shared computational principles for language processing in humans and deep language models. Nat Neurosci 25:369-380.

Goldstein A et al. (2022b) Shared computational principles for language processing in humans and deep language models. Nature Neuroscience 2022 25:3 25.

Gwilliams L, Marantz A, Poeppel D, King J-R (2023) Top-down information shapes lexical processing when listening to continuous speech. Language, Cognition and Neuroscience 39:1-14.

Haegens S, Vergara J, Rossi-Pool R, Lemus L, Romo R (2017) Beta oscillations reflect supramodal information during perceptual judgment. Proceedings of the National Academy of Sciences 114:13810-13815.

Hamilton LS, Oganian Y, Hall J, Chang EF (2021) Parallel and distributed encoding of speech across human auditory cortex. Cell 184:4626-4639.e4613.

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, De Lange FP (2022) A hierarchy of linguistic predictions during natural language comprehension. Proceedings of the National Academy of Sciences of the United States of America 119:e2201968119-e2201968119.

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D, Bressand F, Lengyel G, Lample G, Saulnier L (2023) Mistral 7b. arxiv. arXiv preprint arXiv:231006825 10.

Kinreich S, Djalovski A, Kraus L, Louzoun Y, Feldman R (2017) Brain-to-Brain Synchrony during Naturalistic Social Interactions. Scientific Reports 7:17060.

Kocagoncu E, Clarke A, Devereux BJ, Tyler LK (2017) Decoding the Cortical Dynamics of Sound-Meaning Mapping. J Neurosci 37:1312-1319.

Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology 62:621-647.

Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European Journal of Neuroscience 31:189-193.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. Journal of neurophysiology 102:349-359.

Lalor ECP, Alan J.; Reilly, Richard B.; Foxe, John J. (2009) Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. Journal of Neurophysiology 102.

Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, Ly A, Gronau QF, Šmíra M, Epskamp S (2019) JASP: Graphical statistical software for common statistical designs. Journal of Statistical Software 88:1-17.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006-1010.

Mousavi SM, Roccabruna G, Alghisi S, Rizzoli M, Ravanelli M, Riccardi G (2024) Are LLMs Robust for Spoken Dialogues? arXiv preprint arXiv:240102297.

Muncke J, Kuruvila I, Hoppe U (2022) Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise. Frontiers in Neuroscience 16:876421.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb Cortex 25:1697-1706.

Osterhout L, Holcomb P (1993) Event-Related Potentials and Syntactic Anomaly: Evidence of Anomaly Detection during the Perception of Continuous Speech. Language and Cognitive Processes 8:413-437.

Passali T, Mavropoulos T, Tsoumakas G, Meditskos G, Vrochidis S (2022) LARD: Large-scale artificial disfluency generation. arXiv preprint arXiv:220105041.

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK (2019) PsychoPy2: Experiments in behavior made easy. Behavior research methods 51:195-203.

Perrin F, Pernier J, Bertrand O, Echallier JF (1989) Spherical splines for scalp potential and current density mapping. Electroencephalography and Clinical Neurophysiology 72:184-187.

Piazza G, Carta S, Ip EY, Pérez-Navarro J, Kalashnikova M, Martin CD, Di Liberto GM (2025) Are you talking to me? How the choice of speech register impacts listeners' hierarchical encoding of speech. Imaging Neuroscience.

Pistono A, Maziero S, Chaix Y, Jucla M (2024) Speech disfluencies in children with developmental dyslexia: How do they differ from typical development? International Journal of Language & Communication Disorders 59:1032-1042.

Poeppel D, Emmorey K, Hickok G, Pylkkänen L (2012) Towards a New Neurobiology of Language. Journal of Neuroscience 32:14125-14131.

Power AJ, Foxe JJ, Forde E-J, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. European Journal of Neuroscience 35.

Reddy S, Stanford J (2015) A web application for automated dialect analysis. In, pp 71-75.

Samaha J, Bauer P, Cimaroli S, Postle BR (2015) Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. Proceedings of the National Academy of Sciences 112:8439-8444.

Schwartz L, Levy J, Hayut O, Netzer O, Endevelt-Shapira Y, Feldman R (2024) Generation WhatsApp: inter-brain synchrony during face-to-face and texting communication. Scientific Reports 14:2672.

Shriberg E (2001) To 'errrr'is human: ecology and acoustics of speech disfluencies. Journal of the international phonetic association 31:153-169.

Shriberg EE (1999) Phonetic consequences of speech disfluency.

Slaats S, Martin AE (2025) What's surprising about surprisal. Computational Brain & Behavior 8:233-248.

Snow CE (1972) Mothers' Speech to Children Learning Language. Child Development 43:549-565.

Tree JEF (1995) The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. Journal of Memory and Language 34:709-738.

Uusvuori J, Parviainen T, Inkinen M, Salmelin R (2008) Spatiotemporal Interaction between Sound Form and Meaning during Spoken Word Perception. Cerebral Cortex 18:456-466.

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