Neurobiology of Language

an open access journal

RESEARCH ARTICLE

"Um…, It's Really Difficult to… Um… Speak Fluently": Neural Tracking of Spontaneous Speech

Galit Agmon[1,2], Manuela Jaeger[3], Reut Tsarfaty[4], Martin G. Bleichner[3,5], and Elana Zion Golumbic[1]

[1] The Gonda Center for Multidisciplinary Brain Research, Bar-Ilan University, Ramat Gan, Israel
[2] Frontotemporal Degeneration Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3] Neurophysiology of Everyday Life Group, Department of Psychology, University of Oldenburg, Oldenburg, Germany
[4] Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel
[5] Research Center for Neurosensory Science, University of Oldenburg, Oldenburg, Germany

Keywords: disfluencies, speech rate, spontaneous speech, syntactic boundaries, TRF analysis

ABSTRACT

Spontaneous real-life speech is imperfect in many ways. It contains disfluencies and ill-formed utterances and has a highly variable rate. When listening to spontaneous speech, the brain needs to contend with these features in order to extract the speaker's meaning. Here, we studied how the neural response is affected by four specific factors that are prevalent in spontaneous colloquial speech: (1) the presence of fillers, (2) the need to detect syntactic boundaries in disfluent speech, and (3) variability in speech rate. Neural activity was recorded (using electroencephalography) from individuals as they listened to an unscripted, spontaneous narrative, which was analyzed in a time-resolved fashion to identify fillers and detect syntactic boundaries. When considering these factors in a speech-tracking analysis, which estimates a temporal response function (TRF) to describe the relationship between the stimulus and the neural response it generates, we found that the TRF was affected by all of them. This response was observed for lexical words but not for fillers, and it had an earlier onset for opening words vs. closing words of a clause and for clauses with slower speech rates. These findings broaden ongoing efforts to understand neural processing of speech under increasingly realistic conditions. They highlight the importance of considering the imperfect nature of real-life spoken language, linking past research on linguistically well-formed and meticulously controlled speech to the type of speech that the brain actually deals with on a daily basis.

INTRODUCTION

Neural speech tracking has become an increasingly useful tool for studying how the brain encodes and processes continuous speech (Brodbeck & Simon, 2020; Obleser & Kayser, 2019). Importantly, characterizing linguistic attributes of speech on a continuous basis gives a new angle to auditory attention and neurolinguistics, as researchers have been able to dissociate neural responses driven by the acoustics of speech from those capturing higher-order processes in a dynamically changing speech signal, such as phonological identity and semantic expectations (Brodbeck et al., 2018; Gillis et al., 2021; Inbar et al., 2020; Keitel et al., 2018). And yet, the speech stimuli used in these studies are generally highly scripted and edited, taken—for example—from audiobooks or TED talks, which are also extensively rehearsed and are delivered by professionals. These stimuli are in many ways different from colloquial

Citation: Agmon, G., Jaeger, M., Tsarfaty, R., Bleichner, M. G., & Zion Golumbic, E. (2023). "Um…, it's really difficult to… um… speak fluently": Neural tracking of spontaneous speech. *Neurobiology of Language*, *4*(3), 435–454. https://doi.org/10.1162/nol_a_00109

DOI:
https://doi.org/10.1162/nol_a_00109

Supporting Information:
https://doi.org/10.1162/nol_a_00109

Received: 14 July 2022
Accepted: 5 May 2023

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Galit Agmon
galit.agmon@pennmedicine.upenn.edu

Handling Editor:
Sonja A. Kotz

The MIT Press

speech that we hear every day (Blaauw, 1994; Face, 2003; Goldman-Eisler, 1968, 1972; Haselow, 2017; Huber, 2007; Mehta & Cutler, 1988). Spontaneous speech is also markedly less fluent than scripted speech, peppered with fillers and pauses, and often includes partial or grammatically incorrect syntactic structures (Auer, 2009; Haselow, 2017; Linell, 1982; Shriberg, 2001). Spontaneous speech is also more variable than scripted speech in terms of its speech rate (Goldman-Eisler, 1961; Miller et al., 1984). Silent pauses in spontaneous speech occur less consistently on syntactic boundaries compared to reading (Goldman-Eisler, 1972; Wang et al., 2010). These differences may render the syntactic analysis of spoken speech less trivial compared to the planned, well-structured sentences in scripted speech materials. By focusing mostly on scripted speech materials, past research might have overlooked important processes that are essential for understanding naturalistic speech.

Addressing this gap, in the current electroencephalography (EEG) study we assess neural speech tracking of spontaneously generated speech and focus on the specific challenges the brain has to cope with when processing spontaneous speech: the abundant presence of fillers, online segmentation, and detection of syntactic boundaries and variability in speech rate.

Disfluency and Fillers

A prominent feature of spontaneous speech is that it contains frequent pauses, self-corrections and repetitions (Bortfeld et al., 2001; Clark & Wasow, 1998; Fox Tree, 1995). These disfluencies are generally accompanied by fillers, which are nonlexical utterances (or filled pauses) such as “um” or “uh,” or discourse markers such as “you know” and “I mean” (Fox Tree & Schrock, 2002; Tottie, 2014). Fillers can take on different forms, however, their prevalence across a multitude of spoken languages (e.g., Tian et al., 2017; Wieling et al., 2016) suggests it is a core feature of spontaneous speech. Although fillers do not, in and of themselves, contribute specific lexical information, they are also not mere glitches in speech production. Rather, they likely serve several important communicative goals, helping the speaker in transforming their internal thoughts to speech and helping the listener interpret this speech. Specific roles that have been attributed to fillers include signaling hesitation in speech planning (Corley & Stewart, 2008), conveying lack of certainty in the content (Brennan & Williams, 1995; Smith & Clark, 1993), serving as a cue to focus attention on upcoming words or complex syntactic phrases (Clark & Fox Tree, 2002; Fox Tree, 2001; Fraundorf & Watson, 2011; Watanabe et al., 2008), signaling unexpected information to come (Arnold et al., 2004; Barr & Seyfeddinipur, 2010; Corley et al., 2007), and disambiguating syntactic structures (Bailey & Ferreira, 2003). Moreover, studies have shown that the presence of fillers improves accuracy and memory of speech content (Brennan & Schober, 2001; Corley et al., 2007; Fox Tree, 2001; Fraundorf & Watson, 2011) and that surprise-related neural responses (event-related potentials; ERPs) to target words are reduced if they are preceded by a filler (Collard et al., 2008; Corley et al., 2007). And yet, despite their clearly important role in the production and perception of spontaneous speech, fillers and other disfluencies are generally absent in planned speech that is used in most lab speech-tracking experiments. Therefore, how the brain processes fillers has not been studied extensively.

The Unorganized Nature of Spontaneous Sentences

Unlike written text or highly edited spoken scripts, spontaneous speech is constructed “on the fly,” and represents the speaker’s unedited and somewhat unpolished internal train of thought. As a consequence, spontaneous speech does not always contain clear sentence endings, is not always grammatically correct, and sentences can seem extremely long and less concise (e.g., "and, so, then we went to the bus, but it, like, didn't come, the bus back home I mean, so we had to wait, I really don't know for how long"; Auer, 2009; Halliday, 1989; Haselow, 2017; Linell, 1982). This poses a challenge to the listener of how to parse the continuous input stream into meaningful syntactic units correctly.

Syntactic parsing:
The process of analyzing a string of words into a coherent structure. This term can refer either to the cognitive process of analyzing speech input and representing its syntactic structure, or to the computational process performed by natural language processing tools.

Syntactic parsing is the process of analyzing the string of words into a coherent structure and is critical for speech comprehension, even under assumptions of a "good-enough" or noisy parse (Ferreira & Patson, 2007; Traxler, 2014). To detect syntactic boundaries in spoken language, listeners rely on their accumulated syntactic analysis of an utterance as well as on prosodic cues such as pauses and changes in pitch and duration (Ding et al., 2016; Fodor & Bever, 1965; Garrett et al., 1966; Har-Shai Yahav & Zion Golumbic, 2021; Hawthorne & Gerken, 2014; Kaufeld et al., 2020; Langus et al., 2012; Strangert, 1992; Strangert & Strangert, 1993). There are some behavioral and neural indications that words occurring at the final position of syntactic structures have a special status. In production, words in final positions tend to be prosodically marked (Cooper & Paccia-Cooper, 1980; Klatt, 1975). In comprehension, reading studies show that sentence-final words have prolonged reading and fixation times as well as increased ERP responses. These effects are known as *wrap up effects*, which is a general name for the integrative processes triggered by the final word (Just & Carpenter, 1980; Stowe et al., 2018, for review). Independently, there is also evidence for neural responses associated with detecting prosodic breaks (Pannekamp et al., 2005; Peter et al., 2014; Steinhauer et al., 1999), which often coincide with syntactic boundaries. However, the neural correlates of syntactic boundaries have seldom been studied in the context of spoken language, and particularly not for spontaneous speech, where sentence boundaries are not as well-formed as in scripted language materials.

Speech Rate

Another characteristic of spontaneous speech studied here is speech rate across different sentences. Spontaneous speech is produced "on-the-fly," which can yield speech that at times is highly coherent, fast and excited, and at times is prolonged and interspersed with pauses and hesitations (Goldman-Eisler, 1961, 1972; Miller et al., 1984). Generally, a higher speech rate means that information needs to be integrated in a shorter amount of time, which imposes higher cognitive load on the listener and can affect the processing of speech in many ways. For example, processing compressed speech decreases speech comprehension and intelligibility (Ahissar & Ahissar, 2005; Ahissar et al., 2001; Chan & Lee, 2005; Vaughan & Letowski, 1997; Verschueren et al., 2022). Additionally, phonological and lexical decoding are affected by speech rate and dynamically adjusted to local changes in speech rate (Dilley & Pitt, 2010; Dupoux & Green, 1997; Miller et al., 1986). Several studies have shown that neural tracking of continuous speech can be affected by artificially manipulating speech rate (Ahissar et al., 2001; Müller et al., 2019; Verschueren et al., 2022). However, few have looked at the natural variations in speech rate in spontaneous discourse.

To summarize, spontaneous speech differs in many ways from scripted speech. Here we focused on three key characteristics of spontaneous speech and ask how the presence of fillers, the need to detect syntactic boundaries online, and the natural variations in speech rate affect listeners' neural response to speech. To do so, we analyzed EEG-recorded neural responses from individuals listening to a 6 min long monologue rich in those features, recorded from a speaker spontaneously recounting a personal experience. We first analyzed the monologue to identify fillers and syntactic boundaries and estimate speech rate, then used these as data-driven regressors for analyzing the neural activity, using multivariate speech-tracking analysis of the EEG data. In doing so, this study aimed to bridge the gap between the vast literature studying brain responses to meticulously controlled speech and language materials and the type of speech materials that the brain deals with on a regular daily basis.

MATERIALS AND METHODS

Participants

Twenty participants took part in this experiment. All participants were native Hebrew speakers, right-handed (16:4 F:M; age range 20–30, average 23.1 ± 2.65). Prior to their participation, participants signed informed consent approved by the internal review board of Bar-Ilan University and were compensated for their participation with credit or payment.

Stimuli and Procedure

The stimulus was a single 6 min recording of a personal narrative, told in the first person by a female native Hebrew speaker. The narrative was neutral, unscripted, and spontaneously generated, and described the speaker's participation in a Facebook group called (translated from the Hebrew) "Questions With No Point" and a social face-to-face meet-up organized by members of the group. The only instruction given to participants was to listen passively to the story presented without breaks. This 6 min session was used as an interlude between two parts of another experiment by Paz Har-Shai Yahav and colleagues, currently in progress, that focused on changes in low-level auditory responses over time and is orthogonal in its goals to the data reported here.

Linguistic Analysis of the Speech Stimulus

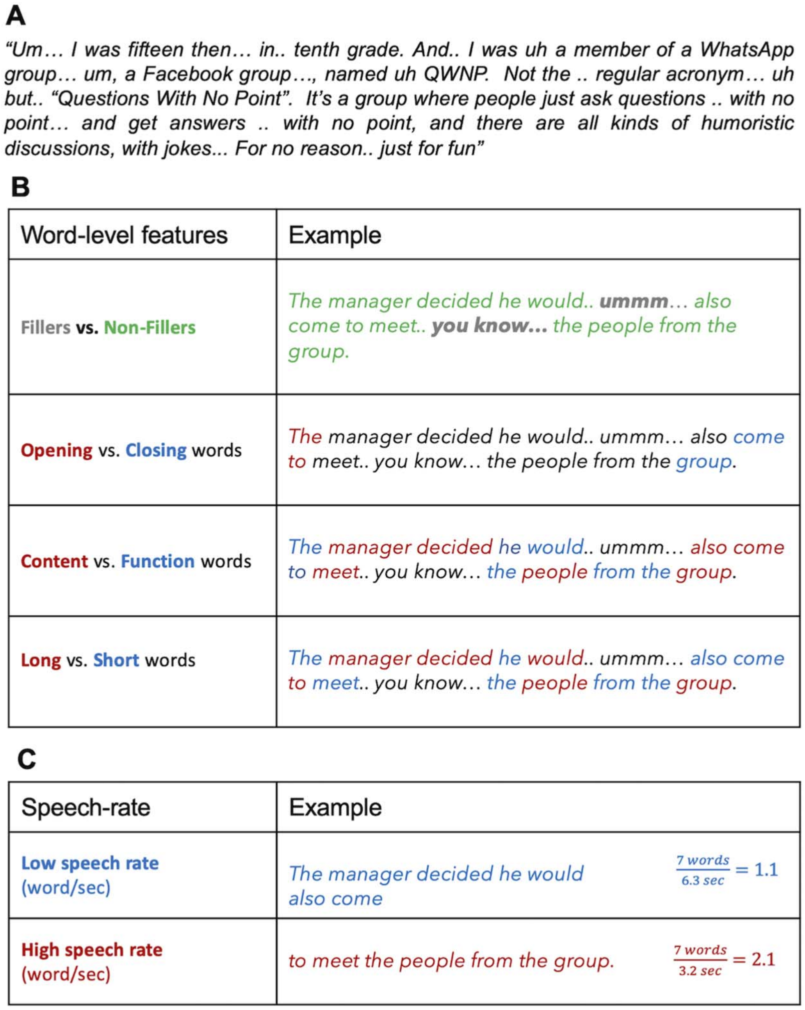
The speech narrative was transcribed manually by two independent annotators who were native speakers of Hebrew, and verified by an expert linguist (GA). The onset and duration of each word and of fillers were identified and time-stamped by the two annotators and confirmed by the linguist, using the software Praat (Boersma & Weenink, 2021).

Based on the transcription, an expert linguist (GA) parsed the speech stimulus into major syntactic units, most notably clauses. We marked boundaries of all main clauses, defined as the minimal unit containing a predicate and all its complements and modifiers. This includes clauses in coordinate constructions (starting with "and" or "but"). In many cases, a clause can contain a subordinate clause, whose boundaries we also marked. These included complement clauses (e.g., "we decided [*that we would meet in one of the gardens*]"), adverbial clauses (e.g., "we met [*because we needed to talk*]"), and relative clauses (e.g., "we met a delegation [*that came from Japan*]"). We also marked the boundaries of heavy phrases such as appositives (e.g., "we decided to go there, [*me and my friends*]") or ellipses (e.g., "I was fifteen years old then, [*in the tenth grade*]").

Although syntactic parsing was done primarily based on the speech transcript, it was double-checked relative to the audio-recording in search of cases where the spoken prosody suggested a different intended parsing than the textual syntactic analysis. For example, the text "I went home with my friends" could be considered a single clause from a purely textual perspective. However, in the audio recording, the speaker inserted a pause after the word "home," and hence, a listener would likely have identified the word "home" as the final word in the clause, before the speaker decided to continue it with a prepositional phrase ("with my friends"). Due to this perceptual consideration, in such cases, we marked both the word "home" and the word "friends" as the final word in the clause.

The time-stamped transcription and syntactic parsing were used to annotate the speech stimulus according to the following four word-level features, which were used to analyze the neural response (described below and summarized in Figure 1):

- *Fillers vs. non-fillers:* Fillers were identified and time-stamped as part of the transcription process. The definition of fillers included both filled pauses (e.g., "um," "uh") and filler discourse markers, which are words that are lexical units, but their use in the utterance is not related to their original lexical meaning (e.g., "like," "well"). A total of 92 fillers were detected in the speech stimulus.
- *Words at syntactic boundaries (opening vs. closing words):* Opening and closing words were identified based on the marking of syntactic clause boundaries described above. There were a total of 166 opening words and 142 closing words (since some clauses were syntactically incomplete and did not have a clear closing word).



A

"Um… I was fifteen then… in.. tenth grade. And.. I was uh a member of a WhatsApp group… um, a Facebook group…, named uh QWNP. Not the .. regular acronym… uh but.. "Questions With No Point". It's a group where people just ask questions .. with no point… and get answers .. with no point, and there are all kinds of humoristic discussions, with jokes... For no reason.. just for fun"

B

Word-level features	Example
Fillers **vs.** **Non-Fillers**	*The manager decided he would.. **ummm**… also come to meet.. **you know…** the people from the group.*
Opening vs. **Closing** words	*The manager decided he would.. ummm… also come to meet.. you know… the people from the group.*
Content vs. **Function** words	*The manager decided he would.. ummm… also come to meet.. you know… the people from the group.*
Long vs. **Short** words	*The manager decided he would.. ummm… also come to meet.. you know… the people from the group.*

C

Speech-rate	Example
Low speech rate (word/sec)	*The manager decided he would also come* $\frac{7\ words}{6.3\ sec} = 1.1$
High speech rate (word/sec)	*to meet the people from the group.* $\frac{7\ words}{3.2\ sec} = 2.1$

Figure 1. Summary of speech stimulus used to analyze the neural response. (A) Excerpt from the speech stimulus used, demonstrating the disfluencies of spontaneous speech. (B) Example of the four word-level features of spontaneous speech analyzed here. (C) Example of the quantification of speech rate at the clause level.

- *Word length* (*short vs. long words):* The length of each word was evaluated based on the time-stamped transcription and the median length was used to distinguish between short and long words (median length: 319 ms).
- *Information content (function vs. content words):* We also differentiated between words that carry the most information (i.e., content words, defined as nouns, verbs, adjectives, and adverbs) and words that mostly play a syntactic role (i.e., function words, defined as pronouns, auxiliary verbs, prepositions, conjunctions, and determiners). In this data set, there were 413 content and 250 function words.

Analysis of Speech Rate

Potential effects of variability in speech rate cannot be assessed at a single-word level, but require assessing the rate of speech over longer periods of times. Here we chose to quantify the mean speech rate within each main clause (including any embedded complement clauses and restrictive relative clauses), following the rationale that a clause is the basic unit over which information needs to be integrated during online listening.

We tested two metrics for operationalizing speech rate: syllable rate and word rate. These metrics are highly correlated with each other ($r = 0.76$, $p < 10^{-11}$ in the current data set) but emphasize slightly different aspects of information transfer, with syllable rate capturing the rate of acoustic input and word rate capturing the rate of linguistic input and general fluency. The word rate of each clause was quantified as the number of words in a clause (not including fillers) divided by its length. Similarly, the syllable rate was quantified as the number of syllables in a clause (not including fillers) divided by its length.

EEG Recordings

EEG was recorded using a 64 Active-Two system (BioSemi) with Ag-AgCl electrodes, placed according to the 10–20 system, at a sampling rate of 1024 Hz. Additional external electrodes were used to record from the mastoids bilaterally and both vertical and horizontal electrooculography electrodes were used to monitor eye movements. The experiment was conducted in a dimly lit, acoustically and electrically shielded booth. Participants were seated on a comfortable chair and were instructed to keep as still as possible and breathe and blink naturally. Experiments were programmed and presented to participants using PsychoPy (Open Science Tools, 2019; Peirce et al., 2019).

EEG Preprocessing and Speech-Tracking Analysis

EEG preprocessing and analysis were performed using the MATLAB-based FieldTrip toolbox (Oostenveld et al., 2011) as well as custom-written scripts. Raw data were first visually inspected, and time points with gross artifacts exceeding ±50 μV (that were not eye movements) were removed. Independent component analysis was performed to identify and remove components associated with horizontal or vertical eye movements as well as heartbeats (Onton et al., 2006). Any remaining noisy electrodes that exhibited either extreme high-frequency activity (>40 Hz) or low-frequency activity/drifts (<1 Hz), were replaced with the weighted average of their neighbors using an interpolation procedure. The clean EEG data were filtered between 1 and 10 Hz. The broadband envelope of the speech was extracted using an equally spaced filterbank between 100 and 10000 Hz based on Liberman's cochlear frequency map (Liberman, 1982). The narrowband filtered signals were summed across bands after taking the absolute value of the Hilbert transform for each one, resulting in a broadband envelope signal. The speech envelope and EEG data were aligned in time and downsampled to 100 Hz for computational efficiency.

To increase the signal-to-noise ratio of the EEG data and reduce the dimensionality of the data, we applied correlated component analysis (CCA) to the clean EEG data. This procedure identifies spatio-temporal components with high intersubject correlation and is particularly effective for experiments studying neural responses to continuous natural stimuli (Dmochowski et al., 2015; Dmochowski & Poulsen, 2015). We performed speech-tracking analysis on the first three CCA components, but ultimately only the first CCA component showed above-chance speech-tracking responses (see below), therefore subsequent analyses were limited only to this component. Statistical evaluation of effects was restricted only to the CCA data to avoid unnecessary multiple comparisons. Results of identical speech-tracking analysis on the original EEG data from all 64 channels are qualitatively similar to these obtained using the CCA component and are reported in the Supporting Information (Figure S1), available at https://doi.org/10.1162/nol_a_00109.

Speech-tracking analysis was performed using scripts from the STRFpak MATLAB toolbox (University of California Berkeley, 2007), which were adapted for EEG data (Har-Shai Yahav & Zion Golumbic, 2021; Zion Golumbic et al., 2013). In this approach, a normalized reverse correlation is applied in order to estimate a temporal response function (TRF) that best describes the linear relationship between features of the presented speech stimulus S(t) and the neural response R(t). Speech-tracking analysis was performed separately for each of feature of interest, as described below.

Temporal response function (TRF): Describes the linear relationship between the stimulus and the neural response. It can be thought of as the continuous analog of an ERP.

Acoustic-only model

The acoustic-only model was estimated to describe the linear relationship between the broadband envelope of the speech stimulus (S) and the neural response derived from the EEG data (R). For S we used the broadband envelope of the speech, which was estimated by first bandpass filtering the speech stimulus into 10 equally spaced bands between 100 and 10000 Hz based on Liberman's cochlear frequency map, taking the amplitude of each narrow-band using the absolute value of the Hilbert transform, and then summing across bands. For R we used the CCA component (1–20 Hz). This analysis was also repeated for filtered responses in the canonical delta band (0.5–3 Hz) and theta band (4–8 Hz; zero-phase FIR filter); responses are reported in Figure S2. S and R were segmented into 12 equal-length non-overlapping epochs (~30 s each). To minimize the effects of overfitting, TRFs were estimated using a jackknife leave-one-out cross-validation procedure, whereby TRFs are estimated using a five-fold train–test procedure. In each iteration, the model was trained on 10 epochs and tested on the remaining 2 epochs that were not included in the training. To select the optimal regularization tolerance value, in each fold of the training, we performed a jackknife leave-one-out cross-validation procedure in which TRFs were estimated for a range of regularization values (0.005–0.01) using all but one of the training trials (i.e., 9 trials), and the Pearson's correlation (*r*) was calculated between the estimated and actual EEG signal for the remaining validation trial. After this procedure was repeated for all possible jackknifes (10 times), the tolerance level that yielding the highest *r* value across all jackknifes was selected, and the estimated TRFs for that tolerance level were averaged across jackknifes. The predictive power of the resulting TRF was then assessed using the test data (the 2 trials that were not included in the training), by comparing the EEG signals of the test data with the predicted response using the TRF estimated from the training data and calculating the Pearson's correlation (*r*) between them. This entire procedure was repeated five times, using different partitioning of the data into independent training–testing data sets, and the mean predictive power and TRFs across these five folds are reported. The optimized tolerance values estimated for the acoustic-only model were then used in all subsequent analyses.

Statistical significance of the speech-tracking response using the acoustic-only model was evaluated using a permutation test as follows: We repeated the TRF estimation procedure on 100 different permutations of mismatched S–R pairs (S from one segment with R from another). This yielded a null distribution of predictive power values that could be obtained by chance. The predictive power of the correct S–R pairing was compared to this distribution and was considered significant if it fell in the top 5% of the null distribution (one-sided).

Word-level models

For the word-level models, the full acoustic envelope was separate into independent regressors containing only the envelope of words/utterances corresponding to a particular feature. We then performed a series of multivariate TRF analyses to test different word-level research questions.

- *Fillers vs. non-filler words model:* comparing a regressor representing all the fillers vs. a regressor with non-filler words (see Figure 3A in the Results section). Since non-filler words were extremely more prevalent than filler words, a subset of non-filler words were randomly selected for this analysis to match the number of fillers. To avoid selection bias, this analysis was conducted 10 times using different subselections of non-filler words.
- *Syntactic boundary model:* comparing a regressor with opening vs. closing words (see Figure 4A in Results section). Note that although the research question addressed in this analysis was the difference in opening vs. closing words, mid-sentence words were also included in this model as a separate regressor, so as to maintain a full representation of the entire speech envelope for the purpose of cross-validation. However, the TRF for mid-sentence words was not compared statistically to the others, since they were found to be too variable acoustically.

We also tested two control models, to test potential alternative explanations for differences between fillers and non-filler words or opening and closing words.

- *Information content model:* comparing a regressor with content vs. function words (see Figure 4B in Results).
- *Word length model:* comparing a regressor with short vs. long words (median split; Figure 4C in Results). The fillers regressor was included in all word-level models, to allow us to compare TRFs to fillers vs. words with specific features.

Each model was tested using a multivariate TRF analysis that estimates a separate TRF for each regressor, using the five-fold train–test regime described above. We then compared the amplitude of the TRFs estimated for the different regressors in each model, using a paired *t* test at each time point, and corrected for multiple comparisons over time using a global-statistic permutation test (1,000 permutations).

Speech rate models

Speech rate was estimated for each clause (as described above) and clauses were divided into two groups corresponding to “low”/”high” speech rate. This was initially done using a median split of speech rates, but given that clauses with faster rates may be shorter than clauses with lower speech rates, the cutoff for separating the two was shifted slightly, to ensure that each group represented roughly half of the full stimulus (cutoff values: syllable rate: 5.23 syllable/s; word rate 2.14 words/s; see solid line shown in Figure 6A in the Results section). Using this criterion, clauses with low speech rate had a mean syllable rate of 4.16 ± 0.67 syllable/s and a mean word rate of 1.72 ± 0.25 words/s; and clauses with high speech rate had a mean syllable rate of 7.36 ± 1.42 syllable/s and a mean word rate of 3.23 ± 0.81 words/s. Two separate regressors were created, representing the envelope of the clauses with low vs. high speech rate (Figure 6B). A multivariate TRF analysis was performed using the two regressors for each feature, using the five-fold train–test regime described above. We then tested for significant differences between the TRFs estimated for low vs. high values of each feature, using a paired *t* test at each time point, and corrected for multiple comparisons over time using a global-statistic permutation test (1,000 permutations).

RESULTS

Acoustic-Only Model

The topography of the first CCA component corresponded to the traditional EEG topography for auditory responses (Figure 2 and Figure S1), and this was the only CCA component that showed a significant speech-tracking response ($r = 0.006$, $p < 0.01$ relative to permutations; Figure 2C). Therefore, all subsequent analyses were restricted only to this component. The acoustic TRF showed a positive deflection between 0 and 200 ms (with several local peaks riding on), and a negative deflection between 300 and 400 ms. This temporal profile is in line with TRF typically obtained for speech stimuli, although the precise timing and modulation depth of different peaks can be highly dependent on the specific regularization parameters selected and the characteristics of the regressor used.

Word-Level Models

When comparing the TRF for fillers vs. non-filler words (Figure 3) we found that the TRFs for fillers was extremely flat and was significantly lower than TRF for non-filler words in the two

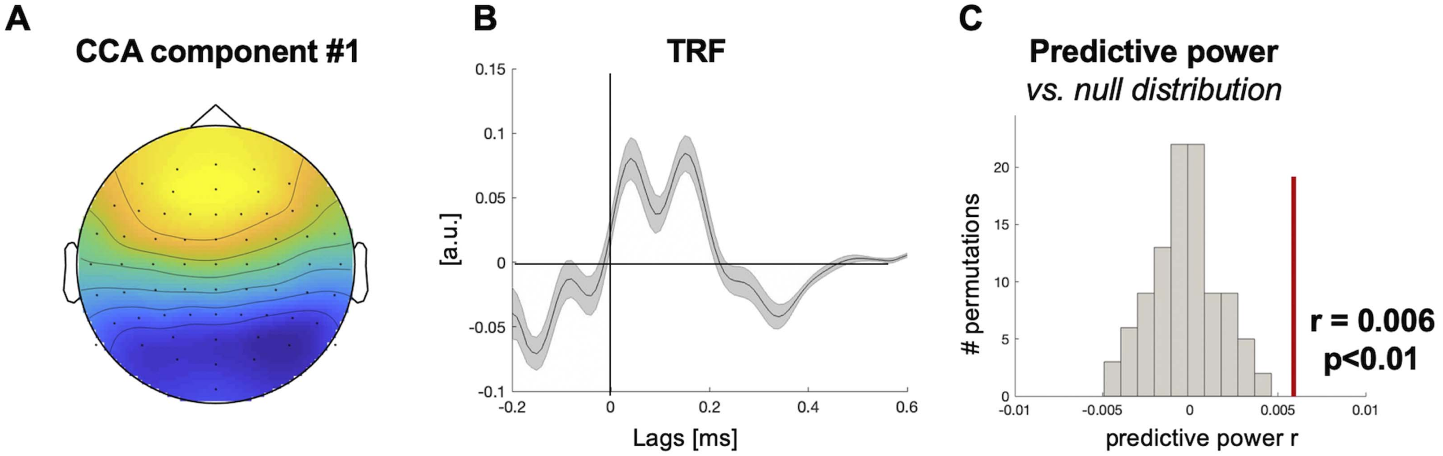


Figure 2. The neural response of the first component of the correlated correspondence analysis (CCA). (A) The scalp distribution of CCA component #1, which corresponds to the expected auditory scalp topography. (B) The temporal response function (TRF) estimated for CCA component #1. (C) The predictive power of the TRF shown in B was significantly larger than could be obtained by chance, shown here relative to the null distribution of predictive power from 100 randomly shuffled S–R combinations.

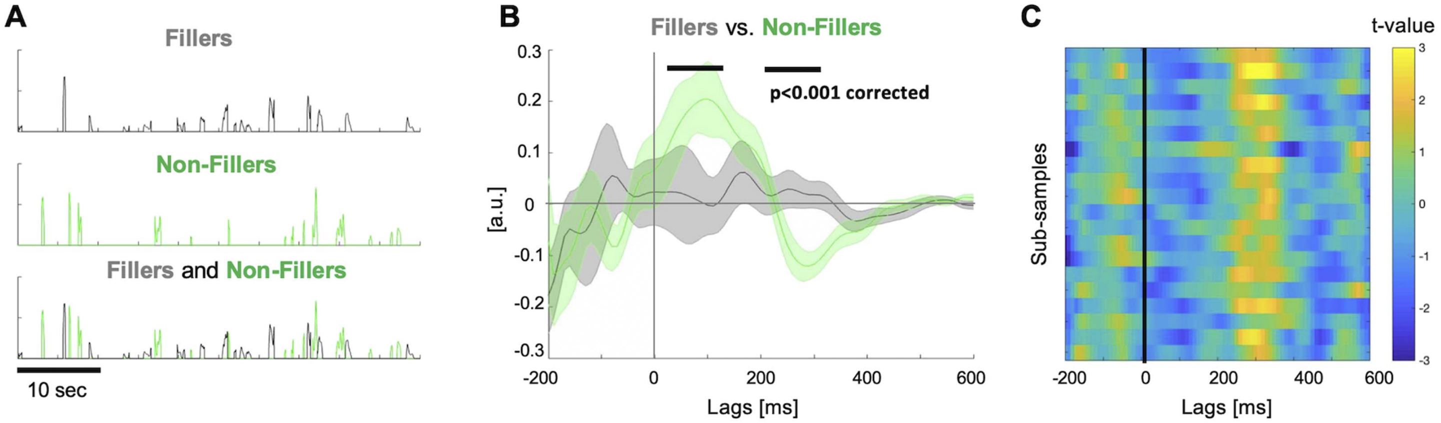


Figure 3. Comparison of the temporal response function (TRF) for fillers vs. non-filler words. (A) Illustration of the regressors use to assess TRFs for fillers vs. non-filler words. Top two rows: Each regressor contains only the envelope of the relevant portion of the stimulus. Bottom row: the two regressors overlaid. For the non-filler regressor, 92 words were randomly sampled, to equate with the number of fillers. (B.) TRFs for fillers vs. non-filler words, which differed significantly in both an early (20–120 ms) and late time window (220–360 ms; $p < 0.001$). (C) Results from 20 different repetitions of this analysis using different subsamples of non-filler words. Figure shows the t value at each time lag reflecting the difference between TRFs for the filler vs. the non-filler regressors.

time windows surrounding peaks in the speech-tracking response (20–120 ms and 220–360 ms; $p < 0.001$ corrected). This analysis was repeated for 20 different subsamples of nonlexical words (proportion matched to the fillers), which yielded similar results, indicating the robustness of these results (Figure 3C and see also Figure 5C).

The other word-level regressors tested included comparison of the TRFs derived for opening vs. closing words, content vs. function words, and long vs. short words (Figure 4 and Figure 5). The TRFs to each word type were also compared to TRFs for fillers.

Alongside the TRF analysis comparing the neural responses to different word-level features, we also evaluated the acoustic differences between the different type of words, by calculating the average envelope amplitude across words and calculating their duration, for each feature

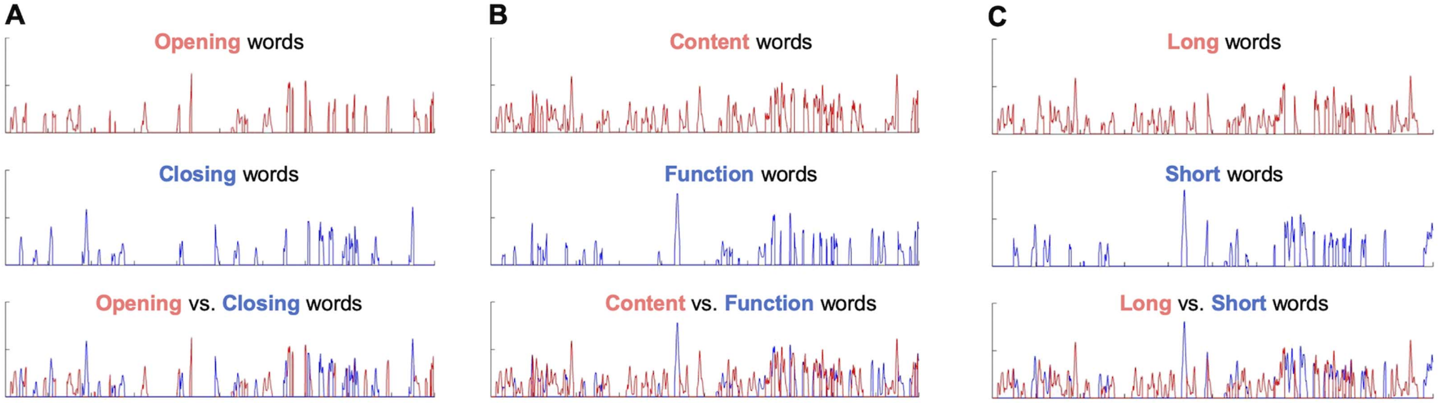


Figure 4. Regressors used to assess temporal response functions. (A) Opening vs. closing words. (B) Content vs. function words. (C) Long vs. short words (median split of stimulus). Top two rows: Each panel shows the individual regressors and contains only the envelope of the relevant portion of the stimulus. Bottom row: The two regressors overlaid. For B and C this constitutes the full stimulus.

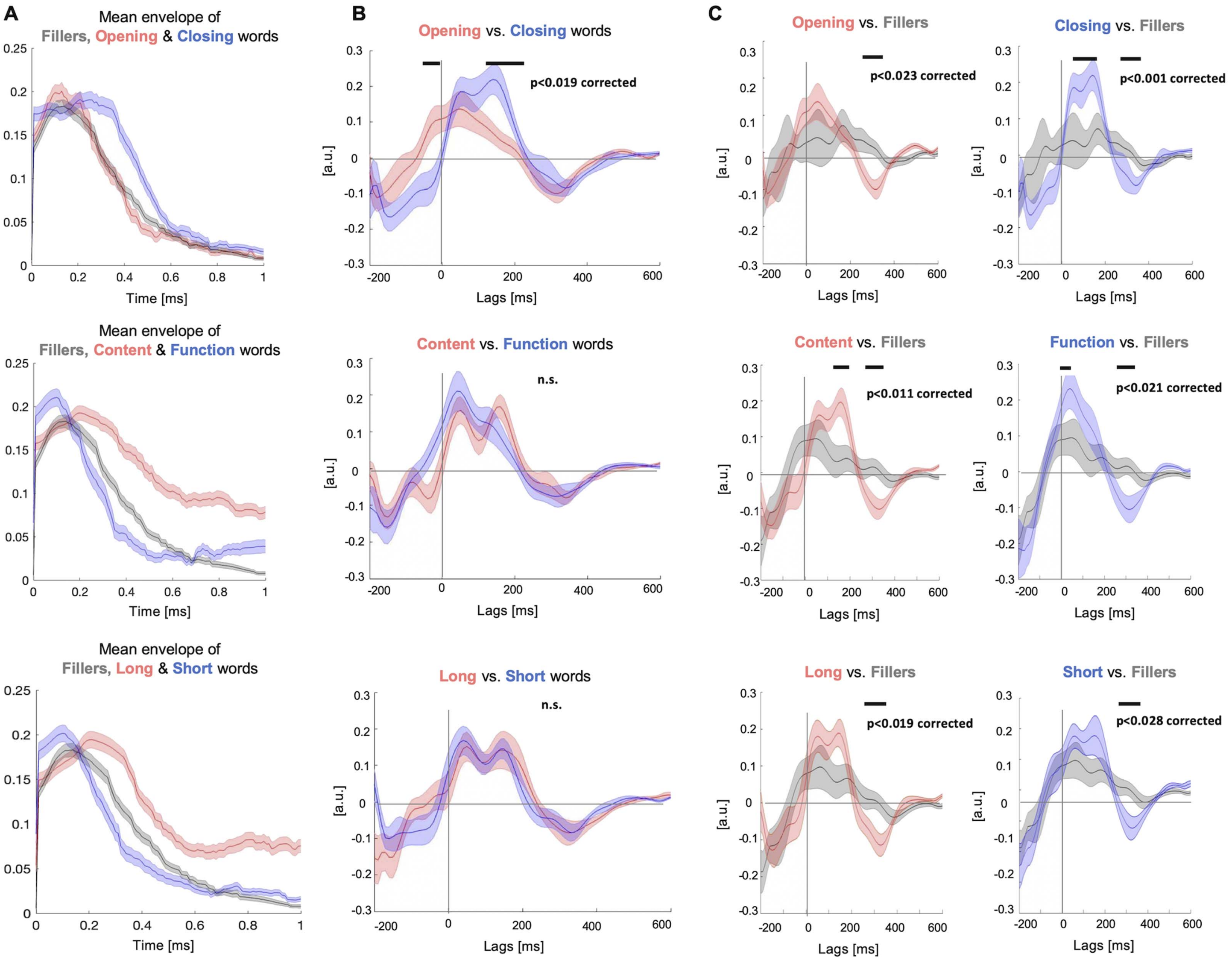


Figure 5. Results for word-level regressors. (A) The average envelope across words\fillers, for the features tested in each word-level model. This comparison indicates that the mean amplitude (volume) is relatively similar across word-level features; however, closing words, content words and long words are more prolonged relative to opening words, function words, and short words, respectively. (B) The TRFs estimated in each word-level model. These results reveal significant differences in the neural response to opening vs. closing words ($p < 0.019$, corrected), but not differences between content vs. function words or between long vs. short words. (C) Comparison of the TRFs to different types of words vs. fillers (estimated as part of each multivariate word-level model). In all cases, significant differences were observed for the late negative TRF response (~250–400 ms), and sometimes for the early positive TRF responses as well (~50–180 ms).

(Figure 5A). This is important in order to determine whether differences in neural responses might be due to systematic differences in the acoustic properties of different word types. This revealed that closing words had a similar mean amplitude relative to opening words, but were more prolonged (median duration 485 ± 164 ms vs. 300 ± 208 ms; $p < 10^{10}$). This lengthening for closing words has been previously documented in final positions of major syntactic constituents and shown to be an important cue for segmentation (Klatt, 1975; Langus et al., 2012; Vaissière, 1983). As expected, we also found that content words were more prolonged relative to function words (median duration 400 ± 188 ms vs. 207 ± 136 ms, $p < 10^{10}$); however, these did not differ significantly in their mean peak amplitude, as was also the case for our separation between long words (median duration 460 ± 160 ms) vs. short words (208 ± 68 ms). Fillers had a similar overall amplitude as all word types, and their duration was comparable to those of the longer words (median duration 481 ± 310 ms).

Interestingly, these differences in acoustics were not reflected in the comparison between the TRFs estimated for each word-level feature. The only significant difference observed was between opening vs. closing words, with the response to closing words peaking later and with a higher amplitude, relative to opening words (Figure 5B top; $p < 0.019$, corrected). However, no differences were found between the neural response to content vs. function words (Figure 5B middle) or between long vs. short words (Figure 5B bottom), suggesting that word duration alone probably did not drive the difference between opening vs. closing words. In line with the results showing reduced TRFs for fillers vs. non-filler words for both the early and late peaks of the TRF (Figure 3), we find that this effect is replicated in all comparisons of TRFs to specific types of words (opening words, closing words, short words, long words, content words, and function words) vs. fillers, further supporting the robustness of this result.

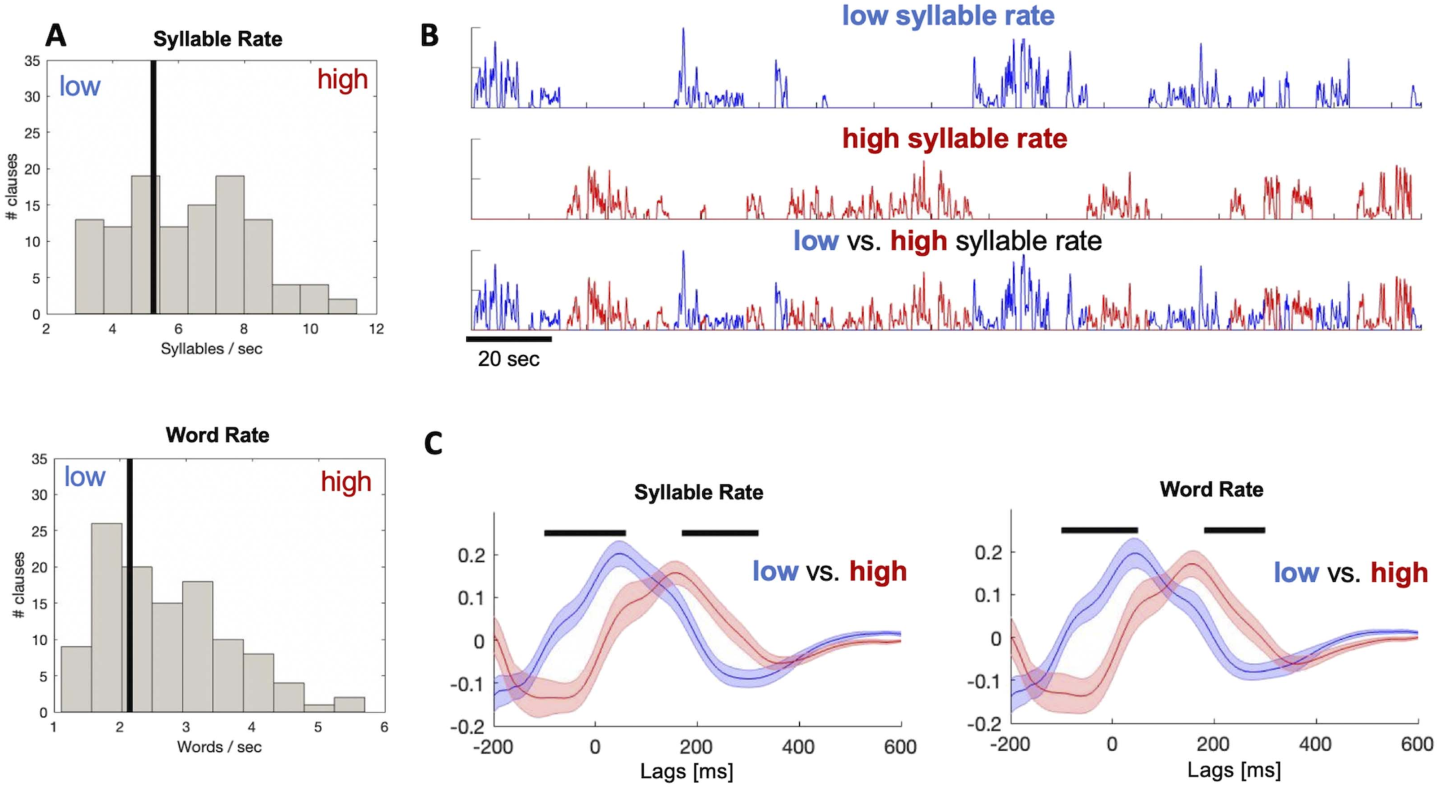


Figure 6. Results for clauses with slow vs. fast speech. (A) The distribution of syllabe-rate (top) and word-rate (bottom) values across all clauses. The vertical line in each panel indicates the cutoff value used to split the clauses into low/high speech rate (median split of the stimulus). (B) Excerpts demonstrating the time course of the regressors representing clauses with low/high syllable rate. The top two rows show the individual regressors, and the bottom row shows both regressors overlaid, which together constitute the full stimulus. (C) The TRFs estimated for clauses with low (blue) vs. high (red) speech rate, shown separately for the two operationalizations of this construct (syllable rate and word rate). Horizonal black lines indicate the time windows where significant differences were observed between the TRFs ($p < 0.001$, corrected).

Speech Rate Models

As expected, results are highly similar for the two operationalizations of speech rate—syllable rate and word rate—as they are highly correlated with each other. For both measures we found significant differences between the TRFs for clauses with fast vs. slow speech, which manifested primarily as a latency shift, with earlier response for clauses with slow vs. fast speech (Figure 6). The differences between TRFs were significant in an early time window starting even prior to time lag 0 (–90–50 ms) and in a later time window (170–300 ms; $p < 0.001$, corrected for both; reported time windows are the overlap between the two measures).

DISCUSSION

In the current study, we looked at how the neural speech-tracking response (i.e., TRF) is affected by different features of spontaneous speech. We tested effects of a word's lexicality (whether it was a filler or not), its role in the structure (whether it opened or closed a syntactic clause) and in the message (whether it was a function or content word), as well as potential differences due to word duration and speech rate. We found that, indeed, the estimated TRFs were modulated by many of these features, suggesting that this response does not merely follow the acoustic of the stimulus but adjusts flexibly to accommodate the linguistic complexity of spontaneous speech. Specifically, we found that fillers (e.g., "um," "uh," "like") elicited a weaker speech-tracking response than words that were part of the core/semantic meaning of the utterance. We also found differences in the TRF latencies for opening vs. closing words in a clause, possibly providing a neural marker for online syntactic segmentation. TRFs estimated for entire clauses were also affected by the rate of speech, with earlier latencies found for clauses with slower speech rate. While many findings are in line with previous research, some of the results are more difficult to interpret mechanistically at this point and require additional follow-up research. However, taken together, these findings highlight the importance of considering the complexities and imperfections of speech when building theoretical models of neural processing of speech, particularly if we strive to understand how the brain contents with the type of speech encountered in actual real life.

The Speech-tracking Response

Before discussing the specific effects found here, it is worth considering what the TRF represents and how it relates to more traditional ERP measures capturing neural responses to language material. The ERP world has a long tradition of associating different deflections in the signal (*components*) to specific perceptual and/or cognitive operations, assigning meaning to both the amplitude and latency of these time-locked responses (Hagoort & Brown, 2000). The recent introduction of speech-tracking methods and estimation of TRFs is often thought of as an equivalent to ERPs for continuous stimuli (Brodbeck & Simon, 2020; Dikker et al., 2020). Supporting this notion, TRFs to speech share many commonalities with auditory ERPs, which display similar time courses of peaks and troughs, and similar scalp topographies and source estimations (Lalor et al., 2009). Early TRF components between 50–150 ms have been shown to be localized to auditory cortex and are primarily driven by the acoustics of the speech (Di Liberto et al., 2015; Ding et al., 2014), whereas linguistic and semantic features, such as predictability, surprise, lexical frequency, and semantic expectations, have been shown to affect TRF responses in later time windows between 200–400 ms (Brodbeck et al., 2018, 2022; Broderick et al., 2018, 2019; Donhauser & Baillet, 2020; Gillis et al., 2021), in ways reminiscent of the P2 and N400 ERP responses to written words (Kutas & Federmeier, 2011). Hence, it is appealing to interpret the observed TRP responses as an extension of the vast ERP literature to the domain of continuous speech (Gwilliams & Davis, 2022).

At the same time, we must be careful not to overinterpret these similarities for several reasons. For one, the TRF is an impulse function that estimates the linear relationship between the acoustic signal and the neural response, and thus by definition it is affected by the shape and intensity of the regressor it is trained on. Moreover, TRFs represent a linear estimation of the neural response to the entire regressor they are trained on, and therefore are highly affected by the variability over time across different features of the stimulus. To this point, here we find that the acoustic TRF, which was trained on the entire speech stimulus, exhibited a double early positive peak between 0 and 200 ms. However, when separating the analysis according to specific features (e.g., speech rate, opening/closing words), we no longer find a double early peak but rather a single early response peaking at different latencies for different features of the speech. Hence, it is possible that the early double peak in the acoustic model stems from averaging together portions of the speech with different temporal characteristics, and should not be interpreted as a sequence of independent components. This also highlights the importance of taking this variability into account when modeling the neural response to spontaneous speech. Interestingly, the negative peak in the TRF, seen around 200–400 ms, was more stable in latency across the different analyses, suggesting that perhaps it can more reliably be associated with known ERP responses, such as the N400 component, in line with previous speech-tracking studies (Brodbeck et al., 2018, 2022; Broderick et al., 2018, 2019; Donhauser & Baillet, 2020; Gillis et al., 2021). Bearing these caveats in mind, and with the goal of gaining a unified perspective on how findings from more controlled experiments generalize to continuous real-life speech, we now turn to discuss the neural responses to features of spontaneous speech observed here.

Neural Tracking of Fillers

The comparison between the TRFs estimated for lexical words vs. fillers indicates a sharp difference in their encoding in the brain. The flat response to fillers was replicated even when parcellating the words according to their length or function. Although fillers clearly serve a communicative function, and some have even evolved from words with lexical meaning (e.g., pragmatic markers that express reduced commitment, such as “like” and “you know”; Ziv, 1988), they do not contribute to the direct comprehension of the narrative and construction of semantic meaning. The substantially flat TRF response observed here for fillers suggests that they are quickly identified as less important, and their encoding is suppressed, perhaps as a means for “cleaning-up” the input from irrelevant insertions before applying syntactic and semantic integrative analysis at the sentence level.

This finding contributes to an ongoing debate about whether the insertion of fillers into spontaneous speech is a conscious choice of the speaker, just as it is for meaningful words, or if their production is more similar to the insertion of pauses, which are often simply a by-product of production delays (Clark & Fox Tree, 2002; Silber-Varod et al., 2021; Tian et al., 2017). To our knowledge, this is the first study to look directly at the neural response to fillers. Our results suggest that, at least from the listener’s perspective, fillers are treated by the brain more like pauses than words. It is still an open question whether this filtering mechanism is sensitive to the lexicality of the input or rather to the extent to which it contributes to meaning. The fact that no difference was found between function words and content words, which differ in their levels of contribution to meaning, suggests that the lexicality of the input is the important feature. However, since informativeness can be operationalized in various ways that are possibly more accurate than considering only parts of speech, we leave this question open for future research.

Syntactic Boundaries in Spontaneous Speech

Speech comprehension relies on listeners ability to segment continuous speech into syntactic clauses that serve as basic processing units (Fodor & Bever, 1965; Garrett et al., 1966; Jarvella, 1971). Previous studies have shown that, for highly structured speech stimuli, neural tracking is sensitive to the underlying syntactic structure of the utterances (Ding et al., 2016; Har-shai Yahav & Zion Golumbic, 2021; Kaufeld et al., 2020). However, how this extends to speech that is spontaneous and less well-formed remains unknown. By comparing the neural response to opening vs. closing words of syntactic clauses we sought to find neural markers of detecting the boundaries of syntactic structures "on the fly." Supporting this, here we found that TRFs to closing words had a later latency and larger peak compared to opening words. Importantly, although closing words are longer on average than opening words, this did not seem to explain the observed effects, since no differences were found between TRFs for short vs. long words. Rather, we infer that slightly different neural processing is applied to words that open a new clause vs. the closing word of a clause.

We offer two possible explanations for this result. One explanation pertains to the special status of words at the end of a clause or sentence. As a sentence unfolds, listeners gradually form and update its syntactic representation (Brennan et al., 2016; Nelson et al., 2017; Pallier et al., 2011). However, only once the sentence ends is it possible to fully integrate across all words to form a complete structure and compute its final syntactic and semantic interpretation. These processes are often generally referred as *wrap up effects*, and were studied mostly through reading times or eye-tracking experiments (Hirotani et al., 2006; Just & Carpenter, 1980; see review by Stowe et al., 2018; Warren et al., 2009). In the ERP literature, final words in written sentences have been shown to trigger a late positive response, which may be related to the later response seen here for closing words (Friedman et al., 1975; Kutas & Hillyard, 1980). However, since researchers were often advised to avoid analyzing responses to final words (due to wrap up effects), there is very little literature to date directly comparing neural responses to opening vs. closing words, particularly for continuous speech (see Stowe et al., 2018, for a review).

Another possible explanation for the difference in responses to opening vs. closing words relates to the prosodic features of closing words. In speech, syntactic boundaries often align with prosodic breaks (Cooper & Paccia-Cooper, 1980; Klatt, 1975; Strangert, 1992; Strangert & Strangert, 1993) and prosodic cues such as pauses, longer word duration or lower intensity are more prominent on words that close a syntactic unit (Hawthorne & Gerken, 2014; Langus et al., 2012; Vaissière, 1983). Although here opening vs. closing words had similar average intensity and we found that duration alone did not account for the differences, our study was not designed to rule out all prosodic explanations, and it is likely that closing words carry other prosodic cues such as pitch deflection and prolonged pauses. Hence, it is possible that the later and larger response seen here for closing words is related to the closure positive shift, a centroparietal ERP response observed often at prosodic breaks (Pannekamp et al., 2005; Peter et al., 2014; Steinhauer et al., 1999). Interestingly, the closure positive shift is evoked even without clear prosodic cues, suggesting that it interacts with syntactic boundaries and not simply with acoustic features (Itzhak et al., 2010; Kerkhofs et al., 2007; Steinhauer & Friederici, 2001).

We do not attempt to dissociate the independent contributions of prosodic processing and syntactic processing to online segmentation, and it is highly likely that syntactic and prosodic interpretations are not mutually exclusive, particularly when processing spontaneous speech. Moreover, as noted above, making direct comparisons between ERP components and the TRF responses measured here, is not straightforward. Given the exploratory nature of our study and the lack of extensive literature pertaining to spontaneous speech processing, we offer these results as the basis for data-driven hypothesis and hope that future research will better characterize the underlying mechanism of the latency shift found here for opening vs. closing words in spontaneous speech.

The Effect of Speech Rate

When estimating the speech-tracking response at the clause level we found that TRFs were significantly earlier for clauses with slow rates vs. those with fast rates. Previous studies have shown that the neural response to speech is highly affected by speech rate, and that speech tracking and comprehension are poorer for faster speech (Ahissar & Ahissar, 2005; Doelling et al., 2014; Nourski & Brugge, 2011; Verschueren et al., 2022). It is important to note, however, that past research has not studied the natural variability in speech rate within a given stimulus, but typically artificially manipulates speech rate of the entire stimulus (e.g., by compression).

One interpretation for the earlier response found here for slower speech could be acoustic, since the envelope of slower speech can also have a shallower slope at acoustic edges (Oganian et al., 2023). These have been shown in a previous study to affect the amplitude of the TRF response, which was larger for steeper slopes (Oganian & Chang, 2019). Another recent study found that faster speech induced larger TRF amplitude as well as an increased TRF latency, interpreting this finding as stemming from the more time that is needed to process fast speech (Verschueren et al., 2022). However, another interpretation may relate not to durational features of words, which could be longer in slow speech, but to the frequent disfluencies that are found in slow speech. Pauses and fillers allow more time for the listener to adjust their expectations and predict upcoming words in the utterance (Clark & Fox Tree, 2002; Fox Tree, 2001; Fraundorf & Watson, 2011; Goldman-Eisler, 1958a, 1958b; Watanabe et al., 2008). Hence, the earlier latency of the TRF may reflect the effects of predictive processing that may be easier to apply to slower speech (Zion Golumbic et al., 2012).

Summary

As speech processing research moves toward understanding how the brain encodes speech under increasingly realistic conditions, it is important to recognize that the type of speech we hear daily is linguistically imperfect. The current study offers insight into some of the ways that these imperfections are dealt with in the neural response and demonstrates the robustness of the speech processing system for understanding spontaneous speech despite its disfluent and highly complex nature. The mechanistic and linguistic insights from this proof-of-concept study provide a basis for forming specific hypotheses for the continued investigation of how the brain deals with the natural imperfections of spontaneous speech.

ACKNOWLEDGMENTS

We would like to thank Paz Har-shai Yahav, Rotem Budagzad-Jacobson, Orel Levi, and Shiraz Levi for their assistance in speech annotation and data collection.

FUNDING INFORMATION

Elana Zion Golumbic, Ministry of Science and Technology, Israel (https://dx.doi.org/10.13039/501100006245), Award ID: 88962. Elana Zion Golumbic, Deutsche Forschungsgemeinschaft (https://dx.doi.org/10.13039/501100001659), Award ID: 490839860. Martin G. Bleichner, Deutsche Forschungsgemeinschaft (https://dx.doi.org/10.13039/501100001659), Award ID: 490839860. Martin G. Bleichner, Deutsche Forschungsgemeinschaft (https://dx.doi.org/10.13039/501100001659), Award ID: 411333557. Manuela Jaeger, Deutsche Forschungsgemeinschaft (https://dx.doi.org/10.13039/501100001659), Award ID: 490839860.

AUTHOR CONTRIBUTIONS

Galit Agmon: Conceptualization; Data curation; Formal analysis; Writing – original draft; Writing – review & editing. **Manuela Jaeger:** Conceptualization; Funding acquisition; Writing – review & editing. **Reut Tsarfaty:** Resources; Writing – review & editing. **Martin G. Bleichner:** Conceptualization; Funding acquisition; Methodology; Writing – review & editing. **Elana Zion Golumbic:** Conceptualization; Formal analysis; Funding acquisition; Methodology; Resources; Visualization; Writing – original draft; Writing – review & editing.

DATA AVAILABILITY STATEMENT

The auditory stimulus used in this study and the regressors extracted from it are available on OSF: https://osf.io/hbzpx/. For inquiries about the codes used for data analysis, please contact the authors.

REFERENCES

Ahissar, E., & Ahissar, M. (2005). Processing of the temporal envelope of speech. In R. König, P. Heil, E. Budinger, & H. Scheich (Eds.), *The auditory cortex: A synthesis of human and animal research* (pp. 295–313). Erlbaum.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367–13372. https://doi.org/10.1073/pnas.201400998, PubMed: 11698688

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582. https://doi.org/10.1111/j.0956-7976.2004.00723.x, PubMed: 15327627

Auer, P. (2009). On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences*, *31*(1), 1–13. https://doi.org/10.1016/j.langsci.2007.10.004

Bailey, K. G. D., & Ferreira, F. (2003). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, *49*(2), 183–200. https://doi.org/10.1016/S0749-596X(03)00027-5

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*(4), 441–455. https://doi.org/10.1080/01690960903047122

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, *14*(4), 359–375. https://doi.org/10.1016/0167-6393(94)90028-0

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* (Version 6.1.37). https://www.praat.org

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*(2), 123–147. https://doi.org/10.1177/00238309010440020101, PubMed: 11575901

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008, PubMed: 27208858

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*(2), 274–296. https://doi.org/10.1006/jmla.2000.2753

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383–398. https://doi.org/10.1006/jmla.1995.1017

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A. L., 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, PubMed: 35060904

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, *172*, 162–174. https://doi.org/10.1016/j.neuroimage.2018.01.042, PubMed: 29366698

Brodbeck, C., & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*, *18*, 25–31. https://doi.org/10.1016/j.cophys.2020.07.014, PubMed: 33225119

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. https://doi.org/10.1016/j.cub.2018.01.080, PubMed: 29478856

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, PubMed: 31371424

Chan, A. H. S., & Lee, P. S. K. (2005). Intelligibility and preferred rate of Chinese speaking. *International Journal of Industrial Ergonomics*, *35*(3), 217–228. https://doi.org/10.1016/j.ergon.2004.09.001

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3, PubMed: 12062148

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242. https://doi.org/10.1006/cogp.1998.0693, PubMed: 9892548

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 696–702. https://doi.org/10.1037/0278-7393.34.3.696, PubMed: 18444766

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and speech*. Harvard University Press. https://doi.org/10.4159/harvard.9780674283947

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010, PubMed: 17173887

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589–602. https://doi.org/10.1111/J.1749-818X.2008.00068.X

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030, PubMed: 26412129

Dikker, S., Assaneo, M. F., Gwilliams, L., Wang, L., & Kösem, A. (2020). Magnetoencephalography and language. *Neuroimaging Clinics of North America*, *30*(2), 229–238. https://doi.org/10.1016/J.NIC.2020.01.004, PubMed: 32336409

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670. https://doi.org/10.1177/0956797610384743, PubMed: 20876883

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054, PubMed: 24188816

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186, PubMed: 26642090

Dmochowski, J. P., Greaves, A. S., & Norcia, A. M. (2015). Maximally reliable spatial filtering of steady state visual evoked potentials. *NeuroImage*, *109*, 63–72. https://doi.org/10.1016/j.neuroimage.2014.12.078, PubMed: 25579449

Dmochowski, J. P., & Poulsen, A. T. (2015). *rca* [Software]. https://github.com/dmochow/rca

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*(Pt. 2), 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035, PubMed: 23791839

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385–393. https://doi.org/10.1016/J.NEURON.2019.10.019, PubMed: 31806493

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 914–927. https://doi.org/10.1037/0096-1523.23.3.914, PubMed: 9180050

Face, T. L. (2003). Intonation in Spanish declaratives: Differences between lab speech and spontaneous speech. *Catalan Journal of Linguistics*, *2*, 115–131. https://doi.org/10.5565/rev/catjl.46

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. https://doi.org/10.1111/J.1749-818X.2007.00007.X

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, *4*(5), 414–420. https://doi.org/10.1016/S0022-5371(65)80081-0

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*(6), 709–738. https://doi.org/10.1006/JMLA.1995.1032

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and Cognition*, *29*(2), 320–326. https://doi.org/10.3758/BF03194926, PubMed: 11352215

Fox Tree, J. E., & Schrock, J. C. (2002). Basic meanings of *you know* and *I mean*. *Journal of Pragmatics*, *34*(6), 727–747. https://doi.org/10.1016/S0378-2166(02)00027-9

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175. https://doi.org/10.1016/j.jml.2011.03.004, PubMed: 21765590

Friedman, D., Simson, R., Ritter, W., & Rapin, I. (1975). The late positive component (P300) and information processing in sentences. *Electroencephalography and Clinical Neurophysiology*, *38*(3), 255–262. https://doi.org/10.1016/0013-4694(75)90246-1, PubMed: 46803

Garrett, M., Bever, T., & Fodor, J. (1966). The active use of grammar in speech perception. *Perception & Psychophysics*, *1*(1), 30–32. https://doi.org/10.3758/BF03207817

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, *41*(50), 10316–10329. https://doi.org/10.1523/JNEUROSCI.0812-21.2021, PubMed: 34732519

Goldman-Eisler, F. (1958a). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96–106, https://doi.org/10.1080/17470215808416261

Goldman-Eisler, F. (1958b). The predictability of words in context and the length of pauses in speech. *Language and Speech*, *1*(3), 226–231. https://doi.org/10.1177/002383095800100308

Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language and Speech*, *4*(3), 171–174. https://doi.org/10.1177/002383096100400305

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech*, *15*(2), 103–113. https://doi.org/10.1177/002383097201500201, PubMed: 4653677

Gwilliams, L., & Davis, M. H. (2022). Extracting language content from speech sounds: The information theoretic approach. In L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), *Speech perception* (Vol. 74, pp. 113–139). Springer. https://doi.org/10.1007/978-3-030-81542-4_5

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia*, *38*(11), 1518–1530.

https://doi.org/10.1016/S0028-3932(00)00052-X, PubMed: 10906377

Halliday, M. A. K. (1989). Spoken language: Grammatical intricacy. In *Spoken and written language* (pp. 76–91). Oxford University Press.

Har-Shai Yahav, P., & Zion Golumbic, E. (2021). Linguistic processing of task-irrelevant speech at a cocktail party. *eLife*, *10*, e65096. https://doi.org/10.7554/elife.65096, PubMed: 33942722

Haselow, A. (2017). *Spontaneous spoken English: An integrated approach to the emergent grammar of speech*. Cambridge University Press. https://doi.org/10.1017/9781108265089

Hawthorne, K., & Gerken, L. A. (2014). From pauses to clauses: Prosody facilitates learning of syntactic constituency. *Cognition*, *133*(2), 420–428. https://doi.org/10.1016/j.cognition.2014.07.013, PubMed: 25151251

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, *54*(3), 425–443. https://doi.org/10.1016/J.JML.2005.12.001

Huber, J. E. (2007). Effect of cues to increase sound pressure level on respiratory kinematic patterns during connected speech. *Journal of Speech, Language, and Hearing Research*, *50*(3), 621–634. https://doi.org/10.1044/1092-4388, PubMed: 17538105

Inbar, M., Grossman, E., & Landau, A. N. (2020). Sequences of intonation units form a ~ 1 Hz rhythm. *Scientific Reports*, *10*(1), Article 15846. https://doi.org/10.1038/s41598-020-72739-4, PubMed: 32985572

Itzhak, I., Pauker, E., Drury, J. E., Baum, S. R., & Steinhauer, K. (2010). Event-related potentials show online influence of lexical biases on prosodic processing. *NeuroReport*, *21*(1), 8–13. https://doi.org/10.1097/WNR.0B013E328330251D, PubMed: 19884867

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 409–416. https://doi.org/10.1016/S0022-5371(71)80040-3

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329, PubMed: 7413885

Kaufeld, G., Bosker, H. R., Oever, S. T., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *Journal of Neuroscience*, *40*(49), 9467–9475. https://doi.org/10.1523/JNEUROSCI.0302-20.2020, PubMed: 33097640

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*(3), Article e2004473. https://doi.org/10.1371/journal.pbio.2004473, PubMed: 29529019

Kerkhofs, R., Vonk, W., Schriefers, H., & Chwilla, D. J. (2007). Discourse, syntax, and prosody: The brain reveals an immediate interaction. *Journal of Cognitive Neuroscience*, *19*(9), 1421–1434. https://doi.org/10.1162/jocn.2007.19.9.1421, PubMed: 17714005

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*(3), 129–140. https://doi.org/10.1016/S0095-4470(19)31360-9

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123, PubMed: 20809790

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/SCIENCE.7350657, PubMed: 7350657

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. https://doi.org/10.1152/JN.90896.2008, PubMed: 19439675

Langus, A., Marchetto, E., Bion, R. A. H., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language*, *66*(1), 285–306. https://doi.org/10.1016/j.jml.2011.09.004

Liberman, M. C. (1982). The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency. *The Journal of the Acoustical Society of America*, *72*(5), 1441–1449. https://doi.org/10.1121/1.388677, PubMed: 7175031

Linell, P. (1982). *The written language bias*. Department of Communication Studies, University of Linköping.

Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, *31*(2), 135–156. https://doi.org/10.1177/002383098803100203, PubMed: 3256770

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1–3), 106–115. https://doi.org/10.1159/000261764

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*(4), 215–225. https://doi.org/10.1159/000261728, PubMed: 6535162

Müller, J. A., Wendt, D., Kollmeier, B., Debener, S., & Brand, T. (2019). Effect of speech rate on neural tracking of speech. *Frontiers in Psychology*, *10*, 1–15. https://doi.org/10.3389/fpsyg.2019.00449, PubMed: 30906273

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114, PubMed: 28416691

Nourski, K. V., & Brugge, J. F. (2011). Representation of temporal sound features in the human auditory cortex. *Reviews in the Neurosciences*, *22*(2), 187–203. https://doi.org/10.1515/rns.2011.016, PubMed: 21476940

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences*, *23*(11), 913–926. https://doi.org/10.1016/j.tics.2019.08.004, PubMed: 31606386

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advance*, *5*(11), Article eaay6279. https://doi.org/10.1126/sciadv.aay6279, PubMed: 31976369

Oganian, Y., Kojima, K., Breska, A., Cai, C., Findlay, A., Chang, E., & Nagarajan, S. (2023). Phase alignment of low-frequency neural activity to the amplitude envelope of speech reflects evoked responses to acoustic edges, not oscillatory entrainment. *Journal of Neuroscience*, *43*(21), 3909–3921. https://doi.org/10.1523/JNEUROSCI.1663-22.2023, PubMed: 37185238

Onton, J., Westerfield, M., Townsend, J., & Makeig, S. (2006). Imaging human EEG dynamics using independent component analysis. *Neuroscience & Biobehavioral Reviews*, *30*(6), 808–822. https://doi.org/10.1016/j.neubiorev.2006.06.007, PubMed: 16904745

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, Article 156869. https://doi.org/10.1155/2011/156869, PubMed: 21253357

Open Science Tools. (2019). *PsychoPy* [Software]. https://psychopy.org/about/overview.html

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/PNAS.1018711108, PubMed: 21224415

Pannekamp, A., Toepel, U., Alter, K., Hahne, A., & Friederici, A. D. (2005). Prosody-driven sentence processing: An event-related brain potential study. *Journal of Cognitive Neuroscience*, *17*(3), 407–421. https://doi.org/10.1162/0898929053279450, PubMed: 15814001

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/S13428-018-01193-Y, PubMed: 30734206

Peter, V., McArthur, G., & Crain, S. (2014). Using event-related potentials to measure phrase boundary perception in English. *BMC Neuroscience*, *15*(1), Article 129. https://doi.org/10.1186/S12868-014-0129-z, PubMed: 25424987

Shriberg, E. (2001). To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153–169. https://doi.org/10.1017/S0025100301001128

Silber-Varod, V., Gósy, M., & Lerner, A. (2021). Is it a filler or a pause? A quantitative analysis of filled pauses in Hebrew. In A. Karpov & R. Potapova (Eds.), *Speech and computer* (pp. 638–648). Springer. https://doi.org/10.1007/978-3-030-87802-3_57

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, *2*(2), 191–196. https://doi.org/10.1038/5757, PubMed: 10195205

Steinhauer, K., & Friederici, A. D. (2001). Prosodic boundaries, comma rules, and brain responses: The closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Research*, *30*(3), 267–295. https://doi.org/10.1023/A:1010443001646, PubMed: 11523275

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition*, *176*, 232–247. https://doi.org/10.1016/j.cognition.2018.03.011, PubMed: 29609098

Strangert, E. (1992). Prosodic cues to the perception of syntactic boundaries. In *Proceedings 2nd International Conference on Spoken Language Processing (ICSLP 92)* (pp. 1283–1285). https://doi.org/10.21437/ICSLP.1992-345

Strangert, E., & Strangert, B. (1993). Prosody in the perception of syntactic boundaries. In *EUROSPEECH '93: Third European conference on speech communication and technology* (pp. 1209–1210). ISCA Archive. https://isca-speech.org/archive_v0/archive_papers/eurospeech_1993/e93_1209.pdf

Tian, Y., Maruyama, T., & Ginzburg, J. (2017). Self addressed questions and filled pauses: A cross-linguistic investigation. *Journal of Psycholinguistic Research*, *46*(4), 905–922. https://doi.org/10.1007/S10936-016-9468-5, PubMed: 28028662

Tottie, G. (2014). On the use of *uh* and *um* in American English. *Functions of Language*, *21*(1), 6–29. https://doi.org/10.1075/fol.21.1.02tot

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, *18*(11), 605–611. https://doi.org/10.1016/j.tics.2014.08.001, PubMed: 25200381

University of California Berkeley. (2007). *STRFpak* [Software]. https://strfpak.berkeley.edu/index.html

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurement* (pp. 53–66). Springer. https://doi.org/10.1007/978-3-642-69103-4_5

Vaughan, N. E., & Letowski, T. (1997). Effects of age, speech rate, and type of test on temporal auditory processing. *Journal of Speech, Language, and Hearing Research*, *40*(5), 1192–1200. https://doi.org/10.1044/jslhr.4005.1192, PubMed: 9328889

Verschueren, E., Gillis, M., Decruy, L., Vanthornhout, J., & Francart, T. (2022). Speech understanding oppositely affects acoustic and linguistic neural tracking in a speech rate manipulation paradigm. *Journal of Neuroscience*, *42*(39), 7442–7453. https://doi.org/10.1523/JNEUROSCI.0259-22.2022, PubMed: 36041851

Wang, Y.-T., Green, J. R., Nip, I. S. B., Kent, R. D., & Kent, J. F. (2010). Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatrica et Logopaedica*, *62*(6), 297–302. https://doi.org/10.1159/000316976, PubMed: 20588052

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E–Z Reader. *Cognition*, *111*(1), 132–137. https://doi.org/10.1016/j.cognition.2008.12.011, PubMed: 19215911

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81–94. https://doi.org/10.1016/j.specom.2007.06.002

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *6*(2), 199–234. https://doi.org/10.1163/22105832-00602001

Zion Golumbic, E. M., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *Journal of Neuroscience*, *33*(4), 1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013, PubMed: 23345218

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, PubMed: 22285024

Ziv, Y. (1988). Lexical hedges and non-committal terms. *Acta Linguistica Hungarica*, *38*(1–4), 261–274.