JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Phase alignment of low-frequency neural activity to the amplitude envelope of speech reflects evoked responses to acoustic edges, not oscillatory entrainment

https://doi.org/10.1523/JNEUROSCI.1663-22.2023

Cite as: J. Neurosci 2023; 10.1523/JNEUROSCI.1663-22.2023

Received: 23 August 2022
Revised: 27 February 2023
Accepted: 2 March 2023

Phase alignment of low-frequency neural activity to the amplitude envelope of speech reflects evoked responses to acoustic edges, not oscillatory entrainment

Abbreviated title: Evoked responses underlie speech envelope tracking

Yulia Oganian*[1,2], Katsuaki Kojima*[1,3,4], Assaf Breska[5], Chang Cai[3], Anne Findlay[3], Edward Chang[1]‡, Srikantan Nagarajan[3]‡

1. Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA

2. Center for Integrative Neuroscience, University Medical Center Tuebingen, Ottfried-Mueller-Str. 25, 72076 Tuebingen, Germany

3. Department of Radiology, University of California, San Francisco, 513 Parnassus Avenue, S362, San Francisco, CA 94143-0628

4. Neurodevelomental Disorders Prevention Center, Perinatal Institute, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229-3039

5. Max-Planck-Institute for biological Cybernetics, Max-Planck-Ring 8-14, 72076 Tuebingen, Germany

*, ‡ authors contributed equally

Corresponding authors:

Yulia Oganian: yulia.oganian@uni-tuebingen.de

Edward Chang: Edward.chang@ucsf.edu

Srikantan Nagarajan: Srikantan.Nagarajan@ucsf.edu

Number of figures: 6

Number of tables: 0

Number of extended multimedia files: 2

Number of extended data tables: 1

Number of pages: 39

Abstract word count: 245

Introduction word count: 650

Discussion word count: 1500

Abstract

The amplitude envelope of speech is crucial for accurate comprehension. Considered a key stage in speech processing, the phase of neural activity in the theta-delta bands (1 - 10 Hz) tracks the phase of the speech amplitude envelope during listening. However, the mechanisms underlying this envelope representation have been heavily debated. A dominant model posits that envelope tracking reflects entrainment of endogenous low-frequency oscillations to the speech envelope. Alternatively, envelope tracking reflects a series of evoked responses to acoustic landmarks within the envelope. It has proven challenging to distinguish these two mechanisms. To address this, we recorded magnetoencephalography while participants (n=12, 6 female) listened to natural speech, and compared the neural phase patterns to the predictions of two computational models: An oscillatory entrainment model and a model of evoked responses to peaks in the rate of envelope change. Critically, we also presented speech at slowed rates, where the spectro-temporal predictions of the two models diverge. Our analyses revealed transient theta phase-locking in regular speech, as predicted by both models. However, for slow speech we found transient theta and delta phase-locking, a pattern that was fully compatible with the evoked response model but could not be explained by the oscillatory entrainment model. Furthermore, encoding of acoustic edge magnitudes was invariant to contextual speech rate, demonstrating speech rate normalization of acoustic edge representations. Taken together, our results suggest that neural phase locking to the speech envelope is more likely to reflect discrete representation of transient information rather than oscillatory entrainment.

Significance statement

Oganian and colleagues probe a highly debated topic in speech perception – the neural mechanisms underlying the cortical representation of the temporal envelope of speech. It is well established that the slow intensity profile of the speech signal, its envelope, elicits a robust brain response that "tracks" these envelope fluctuations. The oscillatory entrainment model posits that envelope tracking reflects phase alignment of endogenous neural oscillations. Here the authors provide evidence for a distinct mechanism. They show that neural speech envelope tracking arises from transient evoked neural responses to rapid increases in the speech envelope. Explicit computational modeling provides direct and compelling evidence that evoked responses are the primary mechanism underlying cortical speech envelope representations, with no evidence for oscillatory entrainment.

Introduction

Speech comprehension is essential to human communication. A major computational step in neural processing of speech is the extraction of its amplitude envelope, the overall intensity of speech across spectral bands. The speech envelope is dominated by fluctuations in the range of ~1–10Hz, which are temporally correlated with the syllabic structure of speech, and the removal of which from speech severely impairs intelligibility (Drullman et al., 1994a, 1994b). Many studies have shown a consistent relationship between the phase of band-limited low-frequency neural activity measured in M/EEG over auditory cortical areas and the phase of the amplitude envelope of speech, a phenomenon widely known as envelope tracking (Ahissar et al., 2001; Luo & Poeppel, 2007). The strength of envelope tracking is correlated with speech intelligibility, suggesting that it could constitute an essential stage in speech comprehension (Abrams et al., 2008; Peelle et al., 2013). However, the neural computations underlying speech envelope tracking are controversial (Gwilliams, 2019; Obleser & Kayser, 2019; Zoefel, ten Oever, et al., 2018).

A dominant theory of speech envelope tracking posits that it reflects the entrainment (i.e., phase alignment) of endogenous neural oscillations to envelope fluctuations. According to this, phase correction is driven by discrete acoustic landmarks in the speech signal and occurs primarily for oscillators in the delta-theta range (1-10 Hz), matching the syllabic rate of the speech signal (Ding et al., 2015; Zoefel, 2018; Giraud & Poeppel, 2012). Functionally, oscillatory entrainment is thought to benefit speech processing via the self-sustaining property of oscillating dynamical systems, resulting in automatically-driven temporal prediction of upcoming information (Haegens & Zion Golumbic, 2018; Helfrich et al., 2019).

However, recent work has demonstrated that phase alignment of low-frequency neural activity can be the outcome of transient neural responses rather than oscillatory dynamics (Breska & Deouell, 2017; Capilla et al., 2011). This becomes pertinent in the case of speech, as it has been suggested that the speech envelope is encoded in evoked responses to the same acoustic landmarks that supposedly drive the entrainment process. Recent electrophysiology recordings suggest that these events are peaks in the rate

of amplitude envelope change, marking the perceived onset of vowels. To date it remains unclear, which of these processes drive phase adjustments in speech envelope tracking. The two competing models have drastically disparate functional and mechanistic implications (Bree et al., 2021; Doelling & Assaneo, 2021; Ruhnau et al., 2020; Zoefel et al., 2019).

To address this, we combined a model-based computational approach with neurophysiological (MEG) recordings of neural responses in an ecologically valid context, using natural continuous speech. We implemented an oscillatory entrainment model and an evoked responses model, quantified the spectral content and temporal dynamics of neural activity predicted by each model in response to speech, identified diverging model predictions, and tested them against MEG data.

Our modeling approach had two critical features. First, we analyzed phase patterns as event-locked to acoustic landmarks. This allowed us to have an extremely high number of events (2106 within-participant), and to probe phase alignment in a time-resolved manner. Particularly, it enabled us to quantify reverberation following a phase-reset, a hallmark of oscillatory processes. Second, we additionally presented continuous speech at, equally intelligible, 1/3 of its original rate. In natural speech, the speech rate, and hence the expected frequency of an entrained oscillator, overlaps with the spectral content of evoked responses. Moreover, the duration of an evoked response is longer than the time between phase-resetting events, where oscillatory reverberation is expected to occur. We hypothesized that slowing speech would solve both.

This manipulation also allowed us to address the neural mechanisms of speech rate normalization, listeners' ability to adjust perceptual processes to differences in speech rate. It has previously been proposed that speech rate normalization relies on shifts in the frequency of the phase-locked oscillator towards the speech rate (Kösem et al., 2018; Nourski et al., 2009; Pefkou et al., 2017). Here we examined this hypothesis in naturalistic speech.

Methods

Participants

Twelve healthy, right-handed volunteers (six females; age range 22-44 years, median 25 years) participated in the study. All participants were native speakers of English. All participants provided informed written consent and received monetary compensation for their participation. The study was approved by the University of California, San Francisco Committee on Human Research.

Speech stimulus

Participants listened to two stories (one male, one female speaker) from the Boston University Radio Speech Corpus (BURSC, Table S1 for full stimulus transcripts) (Ostendorf et al., 1995), each once at regular speech rate and once slowed to 1/3 speech rate. Overall, the stimuli contained 26 paragraphs (each containing 1 - 4 sentences) of 10 – 60 s duration, with silent periods of 500 – 1100 ms inserted between paragraphs to allow measuring onset responses in the MEG without distortion from preceding speech. Boundaries between paragraphs corresponded to breaks between phrases, such that silences were perceived as natural. Speech stimuli were slowed using the Pitch Synchronous Overlap and Add (PSOLA) algorithm, as implemented in the software Praat (Boersma & Weenik, 2019), which slows down the temporal structure of the speech signal while keeping its spectral structure constant (Moulines & Charpentier, 1990). Overall, the regular speech stimulus was 6.5 min long and the slowed stimulus was 19.5 min long. An example excerpt of the stimulus at slow and regular speech rate is provided in the extended data section for download.

Procedure and stimulus presentation

All stimuli were presented binaurally at a comfortable ambient loudness (~ 70 dB) through MEG compatible headphones using custom-written MATLAB R2012b scripts (Mathworks, https://www.mathworks.com). Speech stimuli were sampled at 16 kHz. Participants were asked to listen to the stimuli attentively and to keep their eyes closed throughout.

Participants listened to the radio stories once at regular and once at slowed rate in separate but interleaved blocks, such that each participant heard one story first at regular speech rate and the other at slowed

speech rate. Comprehension was assessed with 3-4 multiple choice comprehension questions posed after each story (Table S2 for list of comprehension questions.). For each participant, a different randomly selected subset of questions was used for each block. Percentage correct was compared between regular and slow blocks using a two-sided paired t-test.

Neural data acquisition and preprocessing

MEG recordings were obtained with a 275-axial gradiometers whole-head MEG system (CTF, Coquitlam, British Columbia, Canada) at a sampling rate of 1,200 Hz. Three fiducial coils were placed on the nasion and left and right pre-auricular points to triangulate the position of the head relative to the MEG sensor array. The position of the patient's head in the device relative to the MEG sensors was determined using indicator coils before and after each recording interval to verify an adequate sampling of the entire field. The fiducial markers were later co-registered onto a structural magnetic resonance imaging scan to generate head shape (Teichmann et al., 2013).

Data analysis and modeling

All analyses were conducted in MATLAB R2019a - MATLAB R2021b (Mathworks, https://www.mathworks.com) using custom-written scripts and the FieldTrip toolbox (Oostenveld et al., 2011).

Acoustic feature extraction

We extracted the broad amplitude envelope of speech stimuli by applying rectification, low-pass filtering at 10 Hz, and down-sampling to 100 Hz, to the original stimulus waveform (in this order). We then calculated the derivative of the resulting envelopes as a measure of its rate of change. Finally, we extracted the sparse time series of local peaks in the amplitude envelope (peakEnv) and its derivative (peakRate). All features are depicted in Figure 1A, for an example stimulus excerpt. Overall, the stimulus set contained 2106 peakRate and 2106 peakEnv events per speech rate condition.

Evoked response and oscillatory entrainment models for IEPC simulation

We implemented two computational models that predict neural activity in response to continuous speech, one based on oscillatory entrainment and another based on evoked responses. We then submitted their output to the same phase analysis as for MEG data. We assumed that both processes were driven by peakRate events, based on our analysis of responses to acoustic landmarks and previous work (Oganian & Chang, 2019). As input, each model received a time series that contained peakRate values, scaled within speech rate between 0.5 and 1, at times of peakRate events, and zeros otherwise. We scaled to this range as our analyses revealed that neural phase alignment to speech is normalized within each speech rate, and that its magnitude for the bottom quantile is ~50% of the top quantile (see Results, Figure 5). To capture the variable latency of the neural response to non-transient sensory events such as acoustic landmarks, we added random temporal jitter (gaussian distribution, SD = 10 and 30 ms in regular and slow speech, respectively) to the timestamp of each peakRate event. Subsequent phase analyses were conducted using the original, non-jittered time stamps. To account for the non-uniform spectral impact of the 1/f noise that is typical to neurophysiological measurement, we added noise with this spectral content to the predicted neural response output by each model, with a signal-to-noise ratio of 1/10. To create the noise, we filtered gaussian white noise to the 1/f shape with the Matlab function firls.m. The temporal and amplitude jitter parameters were fitted to maximize the similarity between the predicted and observed spectrotemporal patterns of phase alignment. Importantly, to not favor one model, this was done across both models and speech rates. To ensure that results would not be biased by the introduction of simulated random noise, we repeated the randomization procedure 2560 times for each model and each speech rate (64 iterations of temporal noise X 40 iterations of amplitude noise), calculated the phase analyses (below) on the predicted neural signal from each randomization, and then averaged across randomizations.

For the oscillator model, peakRate events induce phase corrections of a fixed-frequency oscillator whose frequency is centered on the speech rate (5.7 and 1.9 Hz for regular and slow speech, respectively), as is assumed by oscillatory entrainment models and confirmed in previous work (Breska & Deouell, 2017;

Large & Snyder, 2009). Following Large & Snyder (Large & Snyder, 2009), this process was modeled using a coupled oscillator dynamical system:

$$\frac{d\theta}{dt} = 2\pi F - c \cdot \frac{s(t)}{r} \cdot sin\ sin\ \theta$$

$$\frac{dr}{dt} = r(1 - r^2) + c \cdot s(t) \cdot cos\ cos\ \theta$$

The system produces periodic limit cycle behavior at a radius of $r = 1$ (attractor point) and a frequency F in the absence of input ($s(t) = 0$) and follows phase correction towards an angle of $\theta = 0$ when presented with input ($s(t) > 0$). The magnitude of phase correction depends on the strength of the input, the current angle, and the coupling parameter c. At low values of c, no oscillator was able to entrain to speech, whereas at high values, entrainment spread across all oscillator frequencies. Crucially, as predicted, at intermediate values, only the oscillator with the correct frequency was entraining to our speech stimulus (Fig. 2B). We thus focused on an oscillator model with intermediate entrainment strength and oscillator frequency corresponding to the speech rate in each task condition for further analyses. Specifically, the value of c was set such that the maximal phase correction possible (when $s(t) = 1$ and $\theta = \frac{\pi}{2}$ or $-\frac{\pi}{2}$) would be 70% of the maximal phase shift. We reconstructed the predicted response as: $PredResp_i = cos\ cos\ \theta_i \cdot r_i$.

peakRate events trigger a prototypical evoked response with its amplitude proportional to the strength of the input. For the evoked response model, this process was modeled using a linear convolution of the time series of peakRate events with the waveform of an evoked response to peakRate events. The latter was estimated directly from the MEG data, using a time-delayed linear encoding model (Temporal Receptive Field, TRF (Holdgraf et al., 2017; Oganian & Chang, 2019)), with a time window of -150 to 450 ms relative to peakRate events. While we found no effect of speech slowing on the shape of the neural response to peakRate events in our previous intracranial work (Oganian & Chang, 2019), we assumed that neural responses recorded with MEG will be additionally shaped by other speech features that occur in temporal proximity to peakRate events (e.g., vowel onsets), even though our dataset did not allow us to

explicitly model such additional features. Rather, we estimated the evoked response separately within each speech rate. We used the TRF approach instead of simple averaging due to the high rate of peakRate events (average interval ~170 ms), which would have distorted the averaging-based estimate due to overlap between evoked responses.

MEG data preprocessing

Offline data preprocessing included (in this order) artifact rejection with dual signal subspace projection (DSSP) and down-sampling to 400 Hz. DSSP is a MEG interference rejection algorithm based on spatial and temporal subspace definition (Sekihara et al., 2016). Its performance has been recently validated using clinical data (Cai et al., 2019). In all subsequent analyses of segmented data, segments containing single sensor data above 1.5pT and visually identified artifacts (including muscle, eye blink, and motion) were flagged as bad events and removed from further processing (0.2 % of segments).

Sensor selection

To focus analyses on responses originating in temporal auditory areas, we selected sensors based on the magnitude of the group-averaged M100 response to the onset of utterances (independent of responses to acoustic features within the utterance, which were the focus of subsequent analyses). For this purpose, we segmented the broadband signal around utterance onsets (- 200 to 500 ms), averaged these epochs across utterances and participants, applied baseline correction (-200 ms to 0 ms relative to utterance onset), and extracted the M100 amplitude as the average activity between 60-100 ms after utterance onset. We then selected the ten sensors with maximal M100 responses from each hemisphere. All subsequent analyses were conducted on these 20 sensors.

Event related analysis and sensor selection

For broadband evoked response analysis, we first extracted the broadband signal by band-pass filtering the data between 1 and 40 Hz (second-order Butterworth filter).

To identify which landmark in the speech envelope drives evoked responses, we analyzed evoked responses to peakRate and peakEnv events. We reasoned that with alignment to an incorrect landmark, evoked responses would have reduced magnitude due to smearing, and latency that is shifted away from the acoustic event. For this purpose, we segmented the broadband signal around acoustic landmark events (-100 to 300 ms), averaged these epochs across events within each participant separately for peakRate and peakEnv events, and applied baseline correction (-100 ms to 0 ms relative to event onset). Based on our previous work (Oganian & Chang, 2019), we hypothesized that peakRate events would be the driving acoustic landmark. We compared evoked responses to peakRate and peakEnv using timepoint by timepoint t-tests.

Time-Frequency decomposition

Identical time-frequency (TF) analyses were performed on the continuous MEG data and on the continuous simulated signal from the Evoked Response and Oscillatory Entrainment models. To evaluate the instantaneous phase of the signal at individual frequency bands (logarithmically spaced between 0.67 and 9 Hz, 0.1 octave steps), we applied non-causal band-pass Butterworth filters around each frequency of interest, performed the Hilbert transform, and obtained the amplitude and phase as the absolute value and phase angle, respectively, of the Hilbert signal. Filter order was chosen to achieve maximal 3 dB of passband ripple and at least 24 dB of stopband attenuation. We conducted this TF analysis with a narrow filter width (±0.1 octave of the frequency of interest) for analyses of spectral patterns to increase frequency resolution, and again with a wider filter (±0.5 octave) for analyses of temporal dynamics to increase temporal resolution.

Cerebro-acoustic phase coherence (CAC)

To assess cerebro-acoustic phase coherence between the speech envelope and MEG responses, the speech envelope was processed using the same procedure that was applied to the MEG responses: Down-sampling and TF analysis using the wide filter settings. Phase locking between the speech envelope and

MEG response was calculated across the entire duration of every utterance within each frequency band, using the Cerebro-acoustic phase coherence (CAC):

$$CAC(\varphi) = \frac{1}{N}\left|\sum_{t=1}^{T} exp(i * (ph(\varphi, t) - phs(\varphi, t)))\right|$$

where φ is the center frequency of a frequency band, T is the number of time samples in an utterance, ph is the phase of the neural signal, and phs is the phase of the speech envelope in band φ at time t. To equate the number of time points entering the analysis for slow and regular speech, slow speech utterances were split into three equal parts before CAC calculation, and resultant CAC values were averaged. CAC was averaged across sensors for each hemisphere.

A priori, we hypothesized that CAC would differ between conditions in the frequency bands corresponding to the average frequency of peakRate events in each rate condition (regular: 5.7 Hz; slow: 1.9 Hz, Figure 1B). We tested this hypothesis using a 3-way repeated-measures ANOVA with factors frequency band (high/low), factor speech rate (slow/regular), and hemisphere (left/right). To test for further differences in each frequency band, we assessed the effect of speech rate and hemisphere onto CAC using a two-way repeated-measures ANOVA with factor speech rate (slow/regular) and hemisphere (left/right). Significance in this analysis was Bonferroni-corrected for multiple comparisons across bands.

Inter-event phase coherence (IEPC)

Both IEPC analyses were conducted on the actual MEG data and the neural responses predicted by the evoked response and oscillatory entrainment models. To assess neural phase locking around peakRate events, we segmented the continuous phase data around peakRate events (see below), and obtained a time-resolved inter-event phase coherence (IEPC) (Lachaux, Rodriguez, Martinerie, & Varela, 1999). For each timepoint, IEPC was calculated using the following formula:

$$IEPC(\varphi, t) = \frac{1}{N}\left|\sum_{k=1}^{N} exp(i * ph_k(\varphi, t))\right|$$

where N is the number of events, ph is the phase of the neural signal in trial k, for the frequency band φ and timepoint t. IEPC were first calculated within each of the selected sensors, then averaged across sensors.

Spectral patterns of IEPC

To assess the spectral distribution of phase-locking following peakRate events with increased frequency resolution, we segmented the phase data outputted by the narrow filter TF analysis around peakRate events (-500 to 500 ms) and calculated the IEPC. To prevent distortion of the estimated phase by subsequent peakRate events, we only used ones that were not followed by another peakRate event within the 0-500 ms window (n=813 within each participant). To identify whether in this time window and frequency range there was a significant increase in IEPC in the MEG data, the resulting time x frequency IEPC was compared with the pre-event baseline using 2-D cluster-based permutation t-tests (Maris & Oostenveld, 2007) with 3000 permutations, a peak t threshold of $p < 0.01$, and a cluster threshold of $p < 0.01$. Baseline IEPC was calculated as the average IEPC between -400 ms to -100 ms relative to event onset in each frequency band.

To compare between model predictions and data, IEPC spectral profiles were calculated, separately for each speech rate condition, by averaging IEPC TF images following peakRate event onset across a time window that conforms to one cycle of an oscillator whose frequency matches the speech rate, i.e. 0-170 ms at regular speech rate and 0 – 500 ms at slowed speech rate.

Temporal extent of IEPC

To assess the temporal extent of IEPC between peakRate events, we focused on the slowed speech condition, where phase-locking originating from the evoked response and from putative oscillatory entrainment occupy distinct spectral bands. We segmented the phase data outputted by the broad filter TF analysis around peakRate events (-500 to 1000 ms). with a temporal interval of more than two oscillatory cycles for half an octave around the frequency of peakRate events (1.9 Hz) - that is at least 1040 ms to the

next peakRate (n = 114 peakRate events per participant). As this analysis was focused on the temporal dynamics of IEPC, we examined IEPC dynamics as a function of time, averaged across single frequency bands in this range. For the MEG data, this time course was tested against a theoretical chance level, defined as the expected IEPC value for randomly sampling a matched number of angles from a uniform Von-Miese distribution.

Effect of peakRate magnitude on IEPC

In each rate condition, peakRate events were split into five quantiles, and IEPC was separately calculated within each quantile. Then, we extracted the average IEPC in the theta band (4 – 8 Hz) across all the time points for one cycle of the given frequency band after the event. IEPC in each quantile was compared using 2-way ANOVA with factors quantile and speech rate (regular speech, slow speech).

Effect sizes and power

With over 1000 events (trials) per participant, our data set is well-powered beyond what is typically discussed in psycholinguistic studies, where the number of trials is mostly limited by stimulus selection (e.g., (Brysbaert, 2019)). For all comparisons we report post-hoc power analyses with effect sizes (dz) and beta power, calculated with the software G*power ((Faul et al., 2009)).

Results

Speech Envelope Tracking for regular and slow speech as seen in MEG

[Figure 1 about here]

We recorded MEG while participants (n = 12) listened to continuous speech containing 2106 instances of each envelope landmark, at the original rate (Regular speech condition 6.5 minutes duration), and once slowed to 1/3 of the original speech rate (Slow speech condition, 19.5 minutes duration, Figure 1A). With

this high number of events per condition, we were able to see clear and robust effects based on data from 12 participants (Stefanics et al. 2010, see Methods page 10 for details on power calculation). Stimuli were split into 26 utterances of 10-69 seconds duration (30 – 210 s in Slow speech condition), with additional silence periods inserted between them. This allowed us to estimate an auditory evoked response to speech onset from the data, without altering the original temporal dynamics of the stimulus within sentences.

In a first step, we characterized the temporal dynamics of acoustic landmark events in our speech stimulus, focusing on peaks in the rate of envelope change (peakRate, n = 2106 per condition, Figure 1A) and on peaks in the envelope (peakEnv, n = 2106 per condition, black in Figure 1A). In the regular speech condition, the average frequency of landmarks (similar for peakRate and peakEnv) was 5.7 Hz (SD = 2.9 Hz, Figure 1B), as is typical in natural speech (Ding et al., 2017). In the slow speech condition, the average frequency of landmarks was 1.9 Hz (SD = 1 Hz, similar for peakRate and peakEnv), shifting the peak of the envelope power spectrum to the delta band. Slowing did not impair participants' comprehension, as probed by multiple choice comprehension questions after each story (3-4 questions per story, chance-level per question: 50 %; accuracy in regular speech: mean = 83%, SD = 13%; accuracy in slow speech: mean = 90%, SD = 9.5%; $t(11) = -1.85$, $p = 0.09$; Figure 1C).

Acoustic edges drive MEG evoked responses

We first asked which landmark in the speech envelope drives evoked responses and phase locking to the envelope in regular speech. To focus our analyses on sensors that capture auditory sensory processing, we selected ten sensors with the largest M100 response to speech onsets after silence periods from each hemisphere for all further analyses (Figure 1D). The M100 response showed the typical dipole pattern in each hemisphere (Chait et al., 2004). First, we examined the characteristics of evoked responses (band-pass filtered 1-40 Hz and averaged in the time domain) locked to peakRate and peakEnv landmark events. While peakEnv closely follows on peakRate in regular speech, the interval between them varies. Thus, aligning to the incorrect landmark should lead to (1) a reduced magnitude of the averaged evoked neural signal due to smearing, and (2) shifts in response onset times away from the acoustic event. We found

transient evoked responses with both alignments (Figure 1E). Crucially, the evoked response was of larger magnitude when aligned to peakRate than to peakEnv (peak magnitude: $t(11) = 5.9$, $p < 0.001$). Moreover, this response started after peakRate events, but before peakEnv events (response latency relative to the event for peakEnv: -12.5 ms; peakRate: +50 ms, determined as the first significant time point in a cluster-based permutation test against 0). Together, these analyses indicated that peakRate events, that is, acoustic edges, rather than peakEnv events, that is, envelope peaks, triggered the evoked response in MEG, in line with previous results (Brodbeck et al., 2018; Doelling et al., 2014; Gross et al., 2013; Oganian & Chang, 2019).

Cerebro-acoustic phase coherence between speech envelope and MEG

To confirm that cortical speech envelope tracking was present in our data (Peelle & Davis, 2012), we calculated the cerebro-acoustic phase coherence (CAC) between neural responses and the speech envelope in frequency bands below 10 Hz. CAC is typically increased at the frequency corresponding to the speech rate (Pefkou et al., 2017), which in our data corresponds to the frequency of peakRate in each rate condition (regular: 5.7 Hz, slow: 1.9 Hz). Indeed, speech rate had opposite effects on CAC in these two frequency bands (repeated-measures ANOVA, interaction $F(1, 11) = 31.20$, $p < 0.001$, $\eta^2 = 0.30$, Figure 1F). At 5.7 Hz, CAC was higher for regular speech ($t(11) = 5.6$, $p < 0.001$, $\eta^2 = 0.42$), while at 1.9 Hz it was higher for slow speech ($t(11) = 3.4$, $p = 0.006$, $\eta^2 = 0.29$). Moreover, CAC was overall higher at lower frequencies ($F(1, 11) = 16.44$, $p < 0.001$, $\eta^2 = 0.39$), as is typical for this measure (Cohen, 2014). No other frequency band showed a significant effect of speech rate on CAC (all Bonferroni-corrected $p > 0.05$). Overall, this result replicates previous findings of cortical speech envelope tracking in frequency bands corresponding to the speech rate of the stimulus. However, as this measure is calculated across the entire stimulus time course, it cannot capture local temporal dynamics in the neural phase, driven by phase resets at acoustic edges. To evaluate local temporal and spectral patterns of neural phase-locking following peakRate events, we calculated inter-event phase coherence (IEPC) across peakRate events in

394 the speech stimulus. In contrast to prior studies of CAC, which quantified phase consistency across time,
395 IEPC is calculated across single event occurrences (i.e., single trials) for each time point. IEPC thus
396 enables tracking of the temporal dynamics of phase locking (Gross et al., 2013).
397
398 **Oscillator and evoked response models predict distinct patterns of phase alignment to slowed**
399 **natural speech**
400
401 [Figure 2 about here]
402
403 To obtain a quantitative estimate of neural phase patterns predicted by oscillatory entrainment and evoked
404 response mechanisms, we implemented computational models of neural envelope tracking as predicted by
405 both processes (see methods for a full description of both models). The input to both models was the
406 acoustic stimulus reduced to peakRate events: a continuous time-series down-sampled to match the MEG
407 sampling frequency and containing non-zero values corresponding to peakRate magnitudes at times of
408 peakRate events, and 0 otherwise. The oscillator model was implemented as a coupled oscillator
409 dynamical system with a non-decaying amplitude attractor point, that followed phase resetting whenever
410 the input was different from 0 (at peakRate events), at a magnitude determined by an entrainment
411 parameter (Breska & Deouell, 2017). A preliminary analysis verified that indeed an oscillator whose
412 endogenous frequency corresponds to the average rate of the speech stimulus would be best suited to
413 entrain to the speech stimulus. The evoked response model was designed as a linear convolution of the
414 peakRate event time series with a stereotypical evoked response, which was extracted from the actual
415 MEG data using a time-lagged linear encoding model (rather than simulated to have an ideal shape)
416 (Holdgraf et al., 2017; Oganian & Chang, 2019). To both models, we added 1/f shaped noise, as is
417 observed in neurophysiological data, and a temporal jitter around peakRate event occurrence to each
418 model. See methods for a full description of both models. Both models output a predicted neural response
419 time series (Fig. 2A), from which we extracted predicted spectral and temporal patterns of inter-event

phase coherence (IEPC) in the theta-delta frequency ranges following peakRate events for each condition (Fig. 2B).

To identify distinct predictions of the two models, we focused on two aspects of the overall predicted pattern of IEPC. First, we quantified the spectral shape of predicted responses, by examining the average IEPC pattern in the first oscillatory cycle after peakRate events. We found that in regular speech, both the evoked response model and the oscillatory model predicted a transient increase in theta IEPC following peakRate events (Figure 2B+C, left). However, their predictions for the slow speech condition diverged significantly (Figure 2B+C, middle). The oscillator model predicted a single peak in IEPC around the oscillator frequency in IEPC (Figure 2B, right). In contrast, the evoked response model predicted two IEPC peaks, around 5.7 Hz and around 1.9 Hz, reflective of the shape of the evoked response (the higher frequency peak) and its frequency of occurrence (i.e., the frequency of peakRate events, the lower frequency peak), respectively (Figure 2C, right). We verified this by manually morphing the shape of the evoked response and the frequency of evoked responses, which shifted the location of the upper and lower IEPC peaks, respectively.

Second, we examined the temporal extent of IEPC predicted by each model. A key feature of an oscillatory entrainment mechanism, that is central to the cognitive functions ascribed to oscillatory models, is that the endogenous oscillator will continue to reverberate after phase reset beyond the duration of a single oscillatory cycle, resulting in increased phase alignment for a prolonged time window (Haegens & Zion Golumbic, 2018; Helfrich et al., 2019; Meyer et al., 2019). In our data, this should be expressed as an increase in IEPC extending beyond a single oscillatory cycle after peakRate events. In contrast, if phase locking is the result of evoked responses to peakRate events, the increase in IEPC should be limited to the duration of an evoked response. To quantify this, we focused our analysis on the first two cycles after peakRate events. To prevent interference from subsequent phase-resetting events, we only included peakRate events that were not followed by another peakRate event in this interval (n=114). Importantly, such events were distributed throughout the speech stimulus and not limited to sentence or phrase ends. As in regular speech rate the duration of the evoked response (~350 ms, Figure 1E) extends

446 across two putative cycles at the speech rate frequency (~350 ms at 5.7 Hz), which would not allow to
447 dissociate the two models, we focused this analysis on the slow speech condition. We then examined the
448 time course of IEPC in a range of frequencies surrounding 1.9 Hz, the frequency of the putative oscillator
449 that best entrains to the slow speech rate. As expected, we found divergent predictions: the oscillator
450 model predicts that IEPC remains increased for multiple oscillatory cycles (Figure 2D). In contrast, the
451 evoked response model predicts that the increase in IEPC is temporally limited to the duration of a single
452 evoked response (Figure 2E). Taken together, this model comparison identified two divergent predictions
453 for IEPC patterns in slow speech: The spectral distribution of IEPC and its temporal extent. Next, we
454 performed these identical analyses on our neural data and compared the patterns in the data with the
455 models' predictions.
456
457 **Spectral pattern of Delta-Theta phase-locking to acoustic edges is best described by the evoked**
458 **response model**
459
460 [Figure 3 about here]
461
462 We next turned to testing the two divergent predictions of the two models against MEG data, starting with
463 predictions for spectral distribution. Based on the models' predictions (Fig. 2 and Fig. 3A), we first took a
464 hypothesis-based approach, testing whether average IEPC values in predefined time-frequency ROIs
465 increased: within a single oscillatory cycle post peakRate event in the theta (4-8 Hz) and delta (1 – 3Hz)
466 ranges (Fig. 3B). In regular speech, we found significant IEPC increase (from theoretical baseline based
467 on Von-Mises distribution) in the theta band ($t(11) = 6.9$, $p < .001$, $d=2.1$), but not the delta band ($p > .5$),
468 consistent with both models (Fig. 3A). We then turned to the slow speech condition, where the
469 predictions of the two models diverge. We found two spectral peaks in IEPC to peakRate events in slow
470 speech, with a significant increase from baseline in the theta band ($t(11) = 8.5$, $p < .001$, $d=3.1$) and in the
471 delta band ($t(11) = 5.2$, $p < .001$, $d = 1.9$). This pattern is in line with the predictions of the evoked

response model but not of the oscillator entrainment model (Fig 3A), as the latter cannot explain the increased theta IEPC. To verify that these findings did not reflect the specific predefined time-frequency ROIs, we complemented the ROI analysis with a data-driven 2D cluster-based permutation test. This analysis found one cluster in the theta band in the regular speech condition and a large cluster encompassing both theta and delta bands in the slowed speech condition ($p < 0.001$; Fig. 3C, white borders).

Finally, we directly compared how the predictions of both models fit with the spectral IEPC pattern in the data (Fig. 3D for spectral patterns and Fig. 3E for model comparisons). As expected, the difference between models was not significant in the regular speech condition (oscillatory model: mean r = 0.86, evoked response model mean r = 0.81, t(11) = 1.9, p = 0.06). Crucially, in the slowed speech condition, the evoked response model captured the IEPC dynamics significantly better than the oscillatory model (model comparison t (11) = 3.8, p = 0.002), with a large effect size (d = 1.1, post-hoc beta = 0.93). This was because while both models captured the delta-band peak in IEPC, only the evoked response model captured the IEPC dynamics in higher frequencies (oscillatory model: mean r = 0.46, evoked response model mean r = 0.7). Overall, the results of this analysis favor the evoked response model over the oscillatory model.

Temporal extent of Delta phase locking is limited to a single cycle after peakRate events.

[Figure 4 about here]

We then examined the temporal extent of increased IEPC following peakRate events in the slowed speech condition. The oscillator model predicted that neural IEPC would remain elevated for at least oscillatory cycle, whereas the evoked response model predicted a transient increase in IEPC and return to baseline within 500 ms after the phase reset (Fig. 4A). We calculated IEPC for the MEG data on the same peakRate events as for the model simulations (duration of at least two cycles to subsequent peakRate

events), which allowed us to test for continuous entrainment without interference by a subsequent event. We found that IEPC was elevated above baseline for a single cycle following peakRate events, but returned to baseline immediately after (Fig. 4B, cluster-based permutation test against theoretical baseline based on Von-Mises distribution). Notably, this pattern, including the latency of peak IEPC, closely followed the predictions of the evoked response model. Indeed, direct test of the fit of the models' predictions to the MEG data revealed strong significant correlation with the evoked response model (mean r = 0.59), but not with the oscillator model (mean r = -0.18). This was also reflected in a large significant effect in the direct comparison between models ($t(11) = 3.11$, $p = 0.009$, effect size $d = 0.9$, post-hoc power beta = 0.8).

Finally, we explicitly tested in a hierarchical multiple regression model (data ~ OSC-model + ER-model) whether the oscillatory model would explain variance in the data beyond the variance explained by the evoked response model. Second level analyses on betas across participants showed a significant effect for the ER-model ($t(11) = 3.34$, $p = .003$), but no significant addition to the explained variance by the oscillatory entrainment model ($t(11) = -0.8$, $p = .2$). Note, that this is in line with the negative correlation between data and the oscillatory model, which is due to the reduction in IEPC in the MEG data in the second oscillatory cycle, whereas IEPC remains high in the oscillatory model.

This analysis thus illustrates the transient nature of neural phase locking to peakRate events, which is more consistent with an evoked response mechanism of speech envelope tracking, rather than with an oscillatory entrainment model. Collectively, our findings disagree with an oscillatory entrainment account, which postulates an oscillatory phase-reset after an event, followed by continuous oscillatory reverberation. A more parsimonious account of our results is that the low-frequency phase locking to the speech envelope in MEG is driven by evoked responses to peaks in the envelope rate of change (peakRate). Furthermore, our analysis shows that IEPC to peakRate events reflects the superposition of two different sources: (1) local responses to individual peakRate events and (2) the rate of occurrence of responses to peakRate events. Our analyses also demonstrate that the shift in IEPC frequency bands with changes in speech rate may be the product of a time-frequency decomposition of a series of evoked

524 responses, rather than a shift in the frequency of an entrained oscillator. This finding is a powerful
525 illustration of the importance of explicit computational modeling of alternative neural mechanisms.
526 In the past, it has been suggested that evoked responses are reduced at slower speech rate, where
527 peakRate magnitudes are smaller, limiting the usability of the evoked response model. In a final analysis
528 we thus tested whether IEPC to peakRate is normalized to account for changes in speech envelope
529 dynamics induced by changes in speech rate.
530
531 **Speech rate normalization of peakRate IEPC**
532
533 [Figure 5 about here]
534
535 The perceptual ability to adapt to variation in the speech signal resulting from changes in the speech rate,
536 i.e., the number of syllables produced per second, is referred to as speech rate normalization. Changes in
537 speech rate results in acoustic changes in the speech signal, including slower amplitude increases at
538 acoustic edges, that is lower peakRate magnitudes (Figure 5A, B). We had previously found that
539 responses to peakRate monotonically scale with peakRate magnitude, being larger for faster changes in
540 the speech amplitude (Oganian & Chang, 2019). Efficient envelope tracking across speech rates would
541 thus require remapping of neural responses to peakRate magnitude, to account for this overall reduction.
542 Here, we assessed the effect of speech rate on the magnitude of theta IEPC to peakRate events. In the
543 slowed speech, stimuli peakRate magnitudes were 1/3 of those in regular speech (Figure 5C). If no
544 normalization occurs, IEPC magnitudes in slow speech should reflect absolute peakRate values, resulting
545 in an overall reduction in IEPC (Figure 5F, dark dots). In contrast, if theta IEPC to peakRate is invariant
546 to speech rate, it should reflect peakRate values relative to the contextual speech rate, resulting in similar
547 IEPC magnitudes in both speech rate conditions (Figure 5F, light dots).
548 An evaluation of IEPC after peakRate events, split by peakRate magnitude quantiles, showed comparable
549 theta IEPC in both speech rate conditions (Figure 5D-E), such that average theta IEPC was more robust

for larger peakRate magnitudes across both rate conditions (the main effect of peakRate quantile: b = 0.01, SD = 0.001, t = 1.4, χ^2= 55.0, $p = 10^{-13}$). Crucially, they did not differ between regular and slow speech (Interaction effect: b = 0.003, SD = 0.005, t = 0.6, n.s., Figure 5G), as expected in case of speech rate normalization (Figure 5F, dark dots). The same pattern was observed for the magnitude of peak evoked responses (Fig. 5H). Thus, the magnitude of phase reset induced by peakRate depended on its magnitude relative to the local speech rate context, allowing for the flexible encoding of peakRate information at different speech rates.

Evoked low-frequency power following peakRate events

[Figure 6 about here]

Evoked increase in power is a marker of evoked neural responses and is used to distinguish between evoked responses and oscillatory activity. In addition to calculating the ERP to peakRate events, we thus also tested whether band-passed power would increase after peakRate events. However, we found no significant effects of peakRate on evoked power in theta or delta bands ($p > 0.05$, cluster-based permutation test, data not shown). Our hypothesis that this was due to higher susceptibility of power measures to noise was confirmed in a simulation of the evoked response model (see below).

We hypothesized that this lack of increase in power in theta or delta bands following peakRate events might reflect the high susceptibility of power increases to noise. To assess the effect of noise onto power and phase measures, we tested the evoked response model at noise levels of 1 to 10 relative to response magnitude. We evaluated the effect of noise onto power and IEPC in the theta band (4-8Hz) in the window of a single cycle for a given frequency band after event onset. The effects of noise on power and IEPC were compared using two-sided paired t-tests at each noise level (n = 20 simulated responses), with Bonferroni correction for the number of comparisons. As predicted, we found continuously large effect

sizes for IEPC even at high levels of noise, whereas the effect size for power deteriorated rapidly with the addition of noise.

Discussion

We evaluated local temporal dynamics in MEG neural representation of the continuous speech envelope against the predictions of oscillatory entrainment and evoked response models, derived from explicit computational models of both processes. In line with previous work, we found that acoustic edges (peakRate events) drove evoked responses and phase locking over auditory cortical areas (Brodbeck et al., 2018; Hertrich et al., 2012; Oganian & Chang, 2019). Critically however, only the evoked response model captured the spectral and temporal extent of phase-locking to acoustic edges: a transient local component in the theta range, reflective of the evoked response, and – spectrally distinct in slow speech - a separate global component, which captured the frequency of acoustic edges in the stimulus. An analysis of temporally sparse acoustic events further supported the evoked response model: phase locking was transient and limited to the duration of the evoked response. This contradicts the pattern predicted by entrainment models, namely sustained oscillatory phase locking at the speech rate (Helfrich et al., 2019; Peelle & Davis, 2012). Finally, we found that the magnitude of the evoked phase reset to acoustic edges reflected the speech-rate-normalized amplitude slope at the acoustic edge, offering novel evidence for speech rate normalization. Our results establish acoustic edges as the basis for the representation of the speech envelope across methodologies and provide additional support against the representation of envelope peaks in the human speech cortex. Overall, our findings suggest that neural phase locking induced by evoked responses to acoustic edges is the primary source of speech envelope tracking in the theta-delta band.

Neural phase resetting may be fully explained by the superposition of evoked responses or additionally also contain the entrainment of endogenous oscillatory activity. To distinguish between neural responses reflective of each, we derived the spectral and temporal patterns of phase locking to acoustic edges using

simulations of both mechanisms. Model predictions diverged in the slowed speech condition: Spectrally, the evoked response model predicted two spectral peaks in phase reset, in both theta and delta ranges, whereas oscillatory models predicted delta phase locking only. Temporally, the evoked response model predicted only transient phase locking at the speech rate, whereas oscillatory entrainment predicted reverberation: a persisting oscillation for at least 2 cycles after phase-reset (Helfrich et al., 2019). Note, that the precise temporal extent of IEPC in the oscillator model depends on the decay parameter. However, the hallmark prediction of oscillatory models is that phase-locking will continue after phase-reset beyond a single oscillatory cycle, which is the minimal temporal extent that allows for the model's proposed functional benefits. It was thus not necessary to include a decay parameter in our models.

In our data, both spectral and temporal patterns of phase locking favored the evoked response model: two spectral peaks and temporally transient phase locking. Notably, both models generated the low frequency phase-locking component in the slow speech condition, corresponding to the frequency of acoustic edge events. While previous work interpreted this component in favor of oscillatory entrainment, our results show that only its temporal extent distinguishes between the two models (van Bree et al., 2022). Overall, our analyses show that a linear convolution of evoked responses to discrete acoustic edge events in speech is sufficient to account for the pattern of neural phase locking to continuous speech. This finding has major implications for theories of speech perception. For instance, instead of oscillatory resonance, predictive processing of speech could rely on non-oscillatory temporal prediction mechanisms guided by statistical learning (Friston et al., 2020; Sohoglu & Davis, 2016).

Speech rate normalization is a central behavioral (Reinisch, 2016; Wade & Holt, 2005) and neural phenomenon in speech perception. Shifting of the entrained oscillatory frequency to match the input speech rate was previously proposed as its neural mechanism (Alexandrou et al., 2018b; Kösem et al., 2018). Here, however, we find that the shift of neural phase locking to lower frequencies with speech slowing is an epiphenomenon of spectral analysis of a series of evoked responses. Instead, the magnitude of phase locking to acoustic edges was normalized relative to the distribution of peakRate magnitudes at each rate. Namely, phase locking was comparable across speech rates, despite flatter acoustic edges in

627 slow speech. This suggests that the cortical representations of acoustic edges reflect the magnitude of an
628 edge relative to the contextual speech rate. Such shifting of the dynamic range for acoustic edge
629 magnitudes constitutes a flexible mechanism that maximizes the sensitivity to speech temporal dynamics
630 (Diehl et al., 1980; Hirataa & Lambacher, 2004) and might not be limited to speech sounds.
631 Our approach represents a methodological departure from previous investigations of speech envelope
632 tracking. Namely, previous studies focused on cerebro-acoustic coherence (CAC), which reflects the
633 consistency of phase differences between the neural signal and the acoustic stimulus across time (Peelle et
634 al., 2013). CAC is primarily sensitive to regularities across time, such as the rate of phase resets. In
635 contrast, we used inter-event phase coherence (IEPC), which focuses on assessing temporally local
636 similarities in neural phase across repeated occurrences of the same acoustic event (see (Gross et al.,
637 2013) for IEPC to speech onsets). Our approach revealed that both local phase resets and their rate of
638 occurrence are reflected in IEPC to acoustic edges. In regular speech, both components overlapped,
639 whereas slowing of the speech signal revealed their distinct sources.
640 Speech rate manipulations are frequently used to study speech envelope tracking (Ahissar et al., 2001;
641 Ghitza & Greenberg, 2009; Nourski et al., 2009; Pefkou et al., 2017). Most previous studies used
642 compressed speech to study temporal boundaries on envelope tracking and intelligibility. In contrast, here
643 we used slowed speech to spread distinct acoustic envelope features out in time. Notably, our approach
644 required us to slow the speech signal by a factor of 3, which is rarely encountered in natural speech,
645 except in clinical populations (e.g. subcortical degeneration), where speech can get very slow (Volkmann
646 et al. 1992). Crucially as our participants adapted to the slow speech immediately, it is likely that our
647 stimulus relies on the same perceptual mechanisms that are at play in the regular speech condition. This is
648 also supported by our intracranial work, where responses to acoustic edges in slow (up to slowing factor
649 of 4) and regular speech were qualitatively identical (Oganian & Chang 2019). It is essential to
650 reconsider previous findings under the evoked response framework. For example, while envelope
651 tracking and intelligibility deteriorate for speech rates higher than 8 Hz, insertion of brief silence periods
652 in compressed speech, which returns the effective speech rate to below 8 Hz, improves intelligibility

(Ghitza & Greenberg, 2009). While this result is typically interpreted as evidence for oscillatory envelope tracking in the theta range, within an evoked response framework it might be reflective of the minimal refractory period of neural populations that encode acoustic edges in speech.

Natural speech does not have a robust temporal rhythmicity (Alexandrou et al., 2018a). Our focus on envelope tracking for natural speech indicates that in this case, neural signatures of envelope tracking are well explained by an evoked response model without the need for an oscillatory component. These results seemingly contradict recent findings of predictive entrainment to music (Doelling et al., 2019). However, our study employed natural speech with considerable variability in inter-edge intervals, unlike in rhythmic musical stimuli. Critically, recent neuropsychological work dissociated neural mechanisms for prediction based on rhythmic streams from predictions in non-rhythmic streams (Breska & Ivry, 2018). This adds an important caveat to the current debate, suggesting that previous results may perhaps not extend to natural speech with inherent temporal variability and reduced rhythmicity. The present study thus calls to reevaluate the role of oscillatory entrainment in natural speech comprehension. However, it does not preclude the possibility that the introduction of additional rhythmicity to speech, e.g., in poetry or song, or occasionally more temporally regular everyday speech, particularly in longer utterances, recruits additional neural processes associated with the processing of rhythms.

Such additional processes might support speech comprehension and could underlie some of the recent findings obtained with a rhythmic speech stimulus (Ding et al., 2015; ten Oever & Sack, 2015; Zoefel et al., 2019). On the other hand, while intelligibility and phase patterns are affected by increased speech rhythmicity or concurrent rhythmic brain stimulation, such findings indicate that oscillations may enhance speech processing, but not that they are necessary for the representation of the significantly less periodic natural speech. Therefore, caution needs to be exercised when extending findings from rhythmic stimuli (e.g., (Ding et al., 2015; Doelling et al., 2019; Zoefel, Archer-Boyd, et al., 2018)) to natural speech.

Overall, our results show that an evoked response model accounts for the main neural signatures of speech envelope tracking in MEG. This neural representation of acoustic edges informs about speech rate via inter-event intervals. Moreover, the speech rate normalization of these responses renders this

679 mechanism flexibly adaptable to changes in speech rate. Thus, evoked responses to acoustic edges track
680 the syllabic rate in speech and provide a flexible framework for temporal analysis and prediction during
681 speech perception.
682
683 **Data and code availability**
684 All custom-written analysis code will be publicly available upon publication on github
685 (https://github.com/ChangLabUcsf/MEG-SlowSpeech). Data will be made available upon request from
686 the corresponding authors.
687
688 **Author contributions:** Y.O and E.F.C conceived the study; Y.O, K.K and S.N. designed the experiments
689 and analyzed the data; A.B, Y.O and S.N developed and implemented model simulations; K.K., C.C., and
690 A.F collected and preprocessed the data; K.K. and Y.O. wrote the manuscript; K.K., Y.O, A.B., E.F.C,
691 and S.N revised the manuscript.
692
693 **References**
694 Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-Hemisphere Auditory Cortex Is
695 Dominant for Coding Syllable Patterns in Speech. *J Neurosci*, *28*(15), 3958–3965.
696 https://doi.org/10.1523/JNEUROSCI.0187-08.2008
697 Ahissar, E., Nagarajan, S. S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001).
698 Speech comprehension is correlated with temporal response patterns recorded from auditory cortex.
699 *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–
700 13372. https://doi.org/10.1073/pnas.201400998
701 Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2018a). Cortical entrainment: What we can
702 learn from studying naturalistic speech perception. *Language, Cognition and Neuroscience*, *0*(0),
703 1–13. https://doi.org/10.1080/23273798.2018.1518534

704 Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2018b). Cortical Tracking of Global and
705 Local Variations of Speech Rhythm during Connected Natural Speech Perception. *Journal of*
706 *Cognitive Neuroscience*, *26*(3), 1–16. https://doi.org/10.1162/jocn_a_01295
707 Boersma, P., & Weenik, D. (2019). Praat: Doing phonetics by computer [Computer program]. Version
708 6.1.08, retrieved 5 December 2019 from http://www.praat.org/. *Amsterdam: University of*
709 *Amsterdam*.
710 Bree, S. van, Sohoglu, E., Davis, M. H., & Zoefel, B. (2021). Sustained neural rhythms reveal
711 endogenous oscillations supporting speech perception. *PLOS Biology*, *19*(2), e3001142.
712 https://doi.org/10.1371/journal.pbio.3001142
713 Breska, A., & Deouell, L. Y. (2017). Neural mechanisms of rhythm-based temporal prediction: Delta
714 phase-locking reflects temporal predictability but not rhythmic entrainment. *PLOS Biology*, *15*(2),
715 e2001665. https://doi.org/10.1371/journal.pbio.2001665
716 Breska, A., & Ivry, R. B. (2018). Double dissociation of single-interval and rhythmic temporal prediction
717 in cerebellar degeneration and Parkinson's disease. *Proceedings of the National Academy of*
718 *Sciences*, *115*(48), 12283–12288. https://doi.org/10.1073/pnas.1810596115
719 Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic
720 Representations of Continuous Speech. *Current Biology*, *28*(24), 3976-3983.e5.
721 https://doi.org/10.1016/j.cub.2018.10.042
722 Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory
723 Encoding of Natural Speech. *The Journal of Neuroscience*, 0584–19.
724 https://doi.org/10.1523/JNEUROSCI.0584-19.2019
725 Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A
726 tutorial of power analysis with reference tables. *Journal of Cognition*, *2*.
727 https://doi.org/10.5334/joc.72
728 Cai, C., Xu, J., Velmurugan, J., Knowlton, R., Sekihara, K., Nagarajan, S. S., & Kirsch, H. (2019).
729 Evaluation of a dual signal subspace projection algorithm in magnetoencephalographic recordings

730 from patients with intractable epilepsy and vagus nerve stimulators. *Neuroimage*, *188*, 161–170.
731 https://doi.org/10.1016/j.neuroimage.2018.11.025
732 Capilla, A., Pazo-Alvarez, P., Darriba, A., Campo, P., & Gross, J. (2011). Steady-State Visual Evoked
733 Potentials Can Be Explained by Temporal Superposition of Transient Event-Related Responses.
734 *PLoS ONE*, *6*(1), e14543. https://doi.org/10.1371/journal.pone.0014543
735 Chait, M., Simon, J. Z., & Poeppel, D. (2004). Auditory M50 and M100 responses to broadband noise:
736 Functional implications: *NeuroReport*, *15*(16), 2455–2458. https://doi.org/10.1097/00001756-
737 200411150-00004
738 Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press.
739 Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech
740 reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.
741 https://doi.org/10.1016/j.cub.2015.08.030
742 Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech
743 perception. *Percept Psychophys*, *27*(5), 435–443. https://doi.org/10.3758/bf03204461
744 Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical
745 linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164.
746 https://doi.org/10.1038/nn.4186
747 Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta
748 oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*(0
749 2), 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035
750 Doelling, K. B., & Assaneo, M. F. (2021). Neural oscillations are a start toward understanding brain
751 activity rather than the end. *PLOS Biology*, *19*(5), e3001234.
752 https://doi.org/10.1371/journal.pbio.3001234
753 Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model
754 better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*,
755 *116*(20), 201816414. https://doi.org/10.1073/pnas.1816414116

Drullman, R., Festen, J. M., & Plomp, R. (1994a). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, *95*(2), 1053–1064. https://doi.org/10.1121/1.408467

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, *95*(5), 2670–2680. https://doi.org/10.1121/1.409836

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2020). Active listening. *Hearing Research*, 107998. https://doi.org/10.1016/j.heares.2020.107998

Ghitza, O., & Greenberg, S. (2009). On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence. *Phonetica*, *66*(1–2), 113–126. https://doi.org/10.1159/000208934

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12), e1001752. https://doi.org/10.1371/journal.pbio.1001752

Gwilliams, L. (2019). Hierarchical oscillators in speech comprehension: A commentary on Meyer Sun and Martin 2019. *Language, Cognition and Neuroscience*. https://doi.org/10.1080/23273798.2020.1740749

Haegens, S., & Zion Golumbic, E. (2018). Rhythmic facilitation of sensory processing: A critical review. *Neuroscience & Biobehavioral Reviews*, *86*, 150–165. https://doi.org/10.1016/j.neubiorev.2017.12.002

782 Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to
783 Speech in the Human Superior Temporal Gyrus. *Current Biology*, *28*(12), 1860-1871.e4.
784 https://doi.org/10.1016/j.cub.2018.04.033
785 Helfrich, R. F., Breska, A., & Knight, R. T. (2019). Neural entrainment and network resonance in support
786 of top-down guided attention. *Current Opinion in Psychology*, *29*, 82–89.
787 https://doi.org/10.1016/j.copsyc.2018.12.016
788 Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., & Ackermann, H. (2012). Magnetic brain activity phase-
789 locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech
790 signal. *Psychophysiology*, *49*(3), 322–334. https://doi.org/10.1111/j.1469-8986.2011.01314.x
791 Hirataa, Y., & Lambacher, S. G. (2004). Role of word-external contexts in native speakers' identification
792 of vowel length in Japanese. *Phonetica*, *61*(4), 177–200. https://doi.org/10.1159/000084157
793 Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017).
794 Encoding and Decoding Models in Cognitive Electrophysiology. *Frontiers in Systems*
795 *Neuroscience*, *11*. https://doi.org/10.3389/fnsys.2017.00061
796 Hovsepyan, S., Olasagasti, I., & Giraud, A.-L. (2020). Combining predictive coding and neural
797 oscillations enables online syllable recognition in natural speech. *Nature Communications*, *11*(1).
798 https://doi.org/10.1038/s41467-020-16956-5
799 Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural
800 Entrainment Determines the Words We Hear. *Current Biology*, *28*(18), 2867-2875.e3.
801 https://doi.org/10.1016/j.cub.2018.07.023
802 Large, E. W., & Snyder, J. S. (2009). Pulse and Meter as Neural Resonance. *Annals of the New York*
803 *Academy of Sciences*, *1169*(1), 46–57. https://doi.org/10.1111/j.1749-6632.2009.04550.x
804 Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in
805 Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004
806 Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of*
807 *Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

808 Meyer, L., Sun, Y., & Martin, A. E. (2019). Synchronous, but not entrained: Exogenous and endogenous
809 cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, 1–11.
810 https://doi.org/10.1080/23273798.2019.1693050
811 Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-
812 speech synthesis using diphones. *Speech Communication*, *9*(5–6), 453–467.
813 https://doi.org/10.1016/0167-6393(90)90021-Z
814 Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., & Brugge,
815 J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory
816 Cortex. *Journal of Neuroscience*, *29*(49), 15564–15574.
817 https://doi.org/10.1523/JNEUROSCI.3065-09.2009
818 Obleser, J., & Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain.
819 *Trends Cogn Sci*, *23*(11), 913–926. https://doi.org/10.1016/j.tics.2019.08.004
820 Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human
821 superior temporal gyrus. *Science Advances*. https://doi.org/10.1101/388280
822 Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for
823 advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci*,
824 *2011*, 156869. https://doi.org/10.1155/2011/156869
825 Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus.
826 *Linguistic Data Consortium*, 1–19.
827 O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney,
828 M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment
829 Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.
830 https://doi.org/10.1093/cercor/bht355
831 Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension.
832 *Frontiers in Psychology*, *3*(SEP), 1–17. https://doi.org/10.3389/fpsyg.2012.00320

833 Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory
834 cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387.
835 https://doi.org/10.1093/cercor/bhs118
836 Pefkou, M., Arnal, L. H., Fontolan, L., & Giraud, A.-L. (2017). θ-Band and β-Band Neural Activity
837 Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech. *The*
838 *Journal of Neuroscience*, *37*(33), 7930–7938. https://doi.org/10.1523/JNEUROSCI.2882-16.2017
839 Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate.
840 *Applied Psycholinguistics*, *37*(6), 1397–1415. https://doi.org/10.1017/S0142716415000612
841 Ruhnau, P., Rufener, K. S., Heinze, H.-J., & Zaehle, T. (2020). Pulsed transcranial electric brain
842 stimulation enhances speech comprehension. *Brain Stimulation*.
843 https://doi.org/10.1016/j.brs.2020.07.011
844 Sekihara, K., Kawabata, Y., Ushio, S., Sumiya, S., Kawabata, S., Adachi, Y., & Nagarajan, S. S. (2016).
845 Dual signal subspace projection (DSSP): A novel algorithm for removing large interference in
846 biomagnetic measurements. *J Neural Eng*, *13*(3), 036007. https://doi.org/10.1088/1741-
847 2560/13/3/036007
848 Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction
849 error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747–E1756.
850 https://doi.org/10.1073/pnas.1523266113
851 Teichmann, M., Kas, A., Boutet, C., Ferrieux, S., Nogues, M., Samri, D., Rogan, C., Dormont, D.,
852 Dubois, B., & Migliaccio, R. (2013). Deciphering logopenic primary progressive aphasia: A
853 clinical, imaging and biomarker investigation. *Brain*, *136*(Pt 11), 3474–3488.
854 https://doi.org/10.1093/brain/awt266
855 ten Oever, S., & Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the*
856 *National Academy of Sciences*, *112*(52), 15833–15837. https://doi.org/10.1073/pnas.1517519112

van Bree, S., Alamia, A., & Zoefel, B. (2022). Oscillation or not—Why we can and need to know (commentary on Doelling and Assaneo, 2021). *European Journal of Neuroscience*, *55*(1), 201–204. https://doi.org/10.1111/ejn.15542

Volkmann, J., Hefter, H., Lange, H. W., & Freund, H.-J. (1992). Impairment of temporal organization of speech in basal ganglia diseases. Brain and Language, 43(3), 386–399. https://doi.org/10.1016/0093-934X(92)90108-Q

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950. https://doi.org/10.3758/BF03193621

Zoefel, B. (2018). Speech Entrainment: Rhythmic Predictions Carried by Neural Oscillations. *Current Biology*, *28*(18), R1102–R1104. https://doi.org/10.1016/j.cub.2018.07.048

Zoefel, B., Allard, I., Anil, M., & Davis, M. H. (2019). Perception of Rhythmic Speech Is Modulated by Focal Bilateral Transcranial Alternating Current Stimulation. *Journal of Cognitive Neuroscience*, *32*(2), 226–240. https://doi.org/10.1162/jocn_a_01490

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. *Current Biology*, *28*(3), 401-408.e5. https://doi.org/10.1016/j.cub.2017.11.071

Zoefel, B., ten Oever, S., & Sack, A. T. (2018). The Involvement of Endogenous Neural Oscillations in the Processing of Rhythmic Input: More Than a Regular Repetition of Evoked Neural Responses. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00095

878 **Figure Legends**
879
880

881 **Figure 1. Task design and envelope tracking in neural data A.** The acoustic waveform of an example
882 utterance ('Tarantino says…"), with syllable boundaries, amplitude envelope, and first temporal
883 derivative of the envelope superimposed on it. The same utterance is shown at a regular rate (left) and
884 slowed (right) speech rate. Arrows mark candidate temporal landmark that might induce phase locking
885 (Black: local peaks in the envelope, peakEnv; Purple: acoustic edges, defined as local peaks in the first
886 temporal derivative (rate of change) of the envelope, peakRate). See Table 1-1 for transcripts of the entire
887 speech stimulus. See Sound 1-1 and annotation 1-2 for example stimulus excerpts at two different speech
888 rates. **B.** Frequency of occurrence for peakRate/peakEnv events. Dashed vertical lines mark the average
889 frequency of peakRate events in slow (blue, 1.9 Hz) and regular speech (green, 5.7 Hz). **C.** Single-
890 subject (black) and group-average (red) comprehension performance. See Table 1-2 for a list of all
891 comprehension questions. **D.** Sensor selection was based on M100 response to utterance onsets. Top:
892 Group-averaged evoked response across all 20 sensors included in the analysis. Error bars are ± 1 SEM
893 across subjects. Bottom: Topographic map of a group-averaged M100 response with selected sensors
894 marked in red. **E.** Group-averaged evoked response aligned to peakRate and peakEnv events. Dotted lines
895 mark clusters with $p < 0.05$ with a cluster-based permutation test against 0. Error bars are ± 1 SEM across
896 subjects. **F**. Cerebro-acoustic phase coherence (CAC) between MEG responses and speech envelope
897 (upper panel), and the difference between slow and regular speech (ΔCAC, lower panel). Data were
898 filtered in semi-logarithmically spaced bands between 0.3 and 10 Hz for this analysis. Dashed vertical
899 lines mark the average frequency of peakRate events in each condition, as shown in D. * $p < 0.01$ in post-
900 hoc t-tests with interaction $p < 0.01$. Error bars are ± 1 SEM across subjects.
901

902 **Figure 2. Spectral and temporal signatures of inter-event phase coherence (IEPC) in oscillatory**
903 **entrainment and evoked response models. A.** Schematic illustrations of the predicted neural response to
904 the utterance in Figure 1A using three different models. Top: speech signal. Middle: oscillatory

entrainment model; Bottom: Evoked response model. **B.** IEPC patterns predicted by oscillatory entrainment model for regular and slow speech with a focus on spectral precision. Dashed lines indicate the frequency of peakRate events in each condition. **C**. As B for evoked response model. **D.** Temporal dynamics of delta-IEPC predicted by oscillatory entrainment model, based on peakRate events that are at least 1000ms apart from following events (n = 113 events) in the Slow speech condition. **E.** Same as D for the evoked response model.

Figure 3. Spectral patterns of IEPC in MEG data. A. Predictions of oscillatory and evoked response models for spectral distribution of phase locking to peakRate events. **B.** Average IEPC magnitudes observed in regular and slowed speech conditions within time-frequency ROIs in theta and delta bands one oscillatory cycle post peakRate event. **C.** IEPC patterns observed in MEG responses to speech at regular (left) and slowed (middle) rates. **D.** Spectral IEPC profile averaged across time corresponds to predictions of the evoked response models (A, bottom panel). Significance contours in C,D based on 2D cluster-based permutation testing against pre-event baseline, p<.001. **E.** Correlation between IEPC time courses predicted by the models and observed in the neural data. * $p < 0.05$.

Figure 4. Delta phase locking is limited to a single oscillatory cycle after peakRate events. A. Delta IEPC across selected peakRate events that were at least 200 ms away from preceding, and 1000 ms away from subsequent events. **B**. Delta IEPC time course. Bottom panel shows the IEPC average across the delta range. Red horizontal line marks baseline, red dots mark timepoints of significant deviance from baseline. **C**. Correlation between IEPC time courses predicted by the models and observed in the neural data. * $p < 0.05$.

Figure 5. Normalization of peakRate IEPC for contextual speech rate. A. Histogram of peakRate magnitudes in regular speech, with quantile boundaries marked in red. **B.** Same as A for slow speech **C**. Quantile-Quantile plot of peakRate magnitudes in regular and slowed speech stimulus. peakRate values in

slowed speech stimulus are 1/3 of peakRate values in regular speech stimulus. **D.** IEPC in 1st, 3rd, 5th peakRate magnitude quantile. Horizontal lines mark the theta frequency range (4-8Hz). **E**. Same as D for slow speech. **F.** Predicted quantile-quantile plots of theta IEPC in regular and slowed speech with (dark) or without (light) normalization. **G.** Quantile-quantile plot of theta-band IEPC (mean, error bars mark ± 1 SEM across subjects) in regular and slow speech. Theta IEPC quantile-quantile values are close to the diagonal, indicating similar magnitudes of theta IEPC in regular and slowed speech conditions. H. Quantile-quantile plot of broadband evoked response peak magnitudes (mean, error bars mark ± 1 SEM across subjects) in regular and slow speech. Quantile-quantile values are close to the diagonal, indicating similar magnitudes of the broadband evoked response to peakRate events in regular and slowed speech conditions.

Figure 6. Effect of noise level on IEPC (black) and power (red) after peakRate events in theta band (4-8Hz) for regular speech. * $p < 0.01$.

Extended data legends.

Table 1-1. Speech stimulus transcription.

Table 1-2. Comprehension questions.

Sound 1 -1. Sound files for an example stimulus at regular and slowed speech rates.

Annotation 1-1. Annotation of content of Sound 1-1 in a praat textgrid format.