NEUROSCIENCE

A speech envelope landmark for syllable encoding in human superior temporal gyrus

Yulia Oganian and Edward F. Chang*

The most salient acoustic features in speech are the modulations in its intensity, captured by the amplitude envelope. Perceptually, the envelope is necessary for speech comprehension. Yet, the neural computations that represent the envelope and their linguistic implications are heavily debated. We used high-density intracranial recordings, while participants listened to speech, to determine how the envelope is represented in human speech cortical areas on the superior temporal gyrus (STG). We found that a well-defined zone in middle STG detects acoustic onset edges (local maxima in the envelope rate of change). Acoustic analyses demonstrated that timing of acoustic onset edges cues syllabic nucleus onsets, while their slope cues syllabic stress. Synthesized amplitude-modulated tone stimuli showed that steeper slopes elicited greater responses, confirming cortical encoding of amplitude change, not absolute amplitude. Overall, STG encoding of the timing and magnitude of acoustic onset edges underlies the perception of speech temporal structure.

INTRODUCTION

The most basic representation of the speech signal is the acoustic waveform (Fig. 1A). It is prominently defined by the undulating sequence of peaks and valleys in its intensity profile over time. These modulations in speech intensity are captured by the low-frequency amplitude envelope of speech and are critical for intelligibility (*1*–*4*). It is well established that neural activity in auditory areas reflects these fluctuations in the speech envelope (*5*), but the neural computations that underlie this representation are under active debate.

One prevailing model is that cortex contains an analog representation of the moment-by-moment fluctuations of the amplitude envelope, based on the well-documented neurophysiological correlation between cortical activity and the speech amplitude envelope (*6*–*8*). Alternatively, it has been suggested that cortex detects discrete acoustic landmarks. The most prominent candidate landmarks are the peaks in the envelope (*9*, *10*) and rapid increases in amplitude (also called auditory onset edges) (*11*–*13*). Thus, a fundamental question is whether the cortical representation of the speech envelope is analog or discrete; and if discrete, which landmark is represented.

It is not clear why the amplitude envelope is necessary for intelligibility. Alone, it is not sufficient for comprehension (*14*), and it does not contain spectral cues from the phonetic units of consonants and vowels. Because envelope modulations correlate with the syllable rate, a common interpretation is that the envelope underlies the detection of syllable boundaries in continuous speech. However, direct evidence for neural extraction of syllabic boundaries from the envelope is lacking. Understanding what features in the envelope are encoded, and how they relate to linguistic information, will advance our understanding of what aspects of the speech signal are most critical for comprehension.

A challenge in understanding the neural encoding of the speech envelope is that amplitude changes are highly correlated with concurrent changes in phonetic content. One major reason is that vowels have more acoustic energy (sonority) than consonants. Therefore, it is difficult to ascertain whether encoding is specific to amplitude modulations alone or to the concurrent spectral content associated with phonetic transitions.

Our goal was to determine the critical envelope features that are encoded in the nonprimary auditory cortex in the human superior temporal gyrus (STG), which has been strongly implicated in phonological processing of speech. The human STG is a likely locus of cortical broadband envelope representation due to its complex spectral selectivity (*15*), unlike the narrow frequency tuning in primary auditory cortex (*16*, *17*). To address this, we used direct, high-density intracranial recordings from the cortical surface [electrocorticography (ECoG)], whose high temporal and spatial resolution allowed us to distinguish between model alternatives. The high spatial resolution of ECoG allowed us to localize specific envelope encoding neural populations on STG and to distinguish them from neural populations encoding other temporal features, such as onsets, or acoustic-phonetic features (*18*). The high temporal and spatial resolution of ECoG is particularly advantageous for the study of online speech processing. Signals recorded with noninvasive techniques, such as magnetoencephalography/electroencephalogragy (M/EEG), likely reflect a mix of neural responses to different input features due to the spatial proximity of their cortical representations (e.g., the envelope, onset, and spectral phonetic structure). Determining how the speech envelope is neurally encoded may redefine the neurolinguistic understanding of how we perceive the temporal structure of speech.

First, we asked whether STG neural populations encode instantaneous envelope values or detect a discrete landmark (*5*). Results from two experiments, one with continuous speech at normal speed and one with slowed speech, showed that STG responses encode the amplitude envelope via evoked responses to acoustic onset edges. We then analyzed the linguistic structure of speech around acoustic onset edge and found that they co-occur with vowel onsets, thus representing the temporal structure of speech at the syllabic level. Furthermore, we asked whether the encoding of the amplitude envelope is distinct from the processing of complex spectral patterns that define consonants and vowels and are encoded in the STG (*18*, *19*). Last, to unequivocally establish whether the amplitude envelope is encoded independently of spectral changes, we isolated neural responses to amplitude modulations in an additional experiment with amplitude-modulated nonspeech tones, which provided converging evidence for the encoding of peakRate in middle STG.

Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA.
*Corresponding author. Email: edward.chang@ucsf.edu

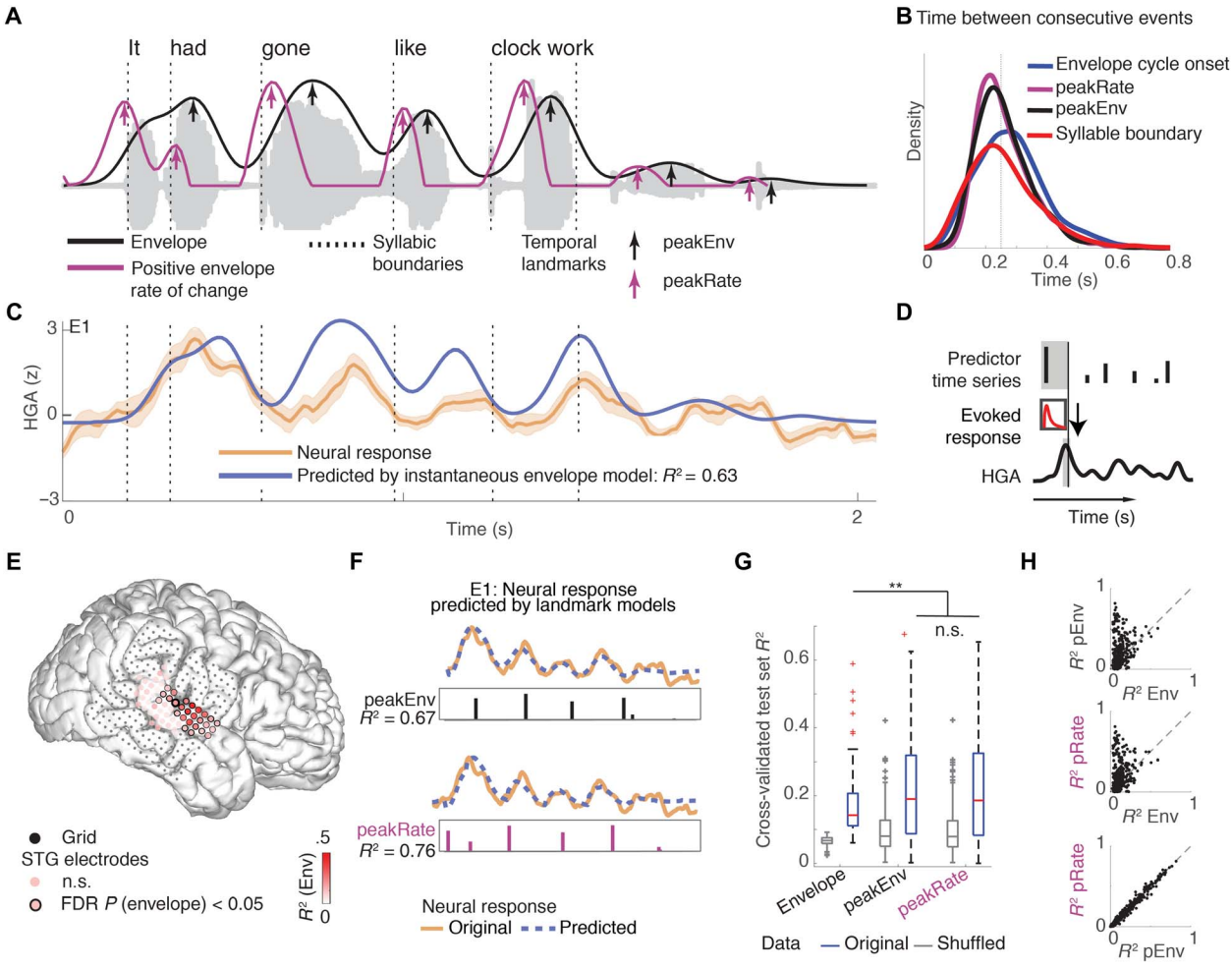

Fig. 1. STG responses to speech amplitude envelope reflect encoding of discrete events. (**A**) Acoustic waveform of example sentence, its amplitude envelope (black) and half-rectified rate of amplitude change (purple). Arrows mark local peaks in envelope (peakEnv) and rate of change of the envelope (peakRate), respectively. (**B**) Rate of occurrence of syllabic boundaries, envelope cycles, peaks in the envelope, and peaks in the rate of change of the envelope in continuous speech across all sentences in stimulus set. All events occur on average every 200 ms, corresponding to a rate of 5 Hz. (**C**) Average HGA response to the sentence in (A) for electrode E1 (yellow). The predicted response based on a time-lagged representation of the envelope (blue) is highly correlated with the neural response for this electrode E1 and the example sentence (R^2 = 0.58). (**D**) Schematic of temporal receptive field (TRF) model. The neural response is modeled as convolution of a linear filter and stimulus time series in a prior time window. (**E**) Variance in neural response explained by representation of instantaneous amplitude envelope in an example participant's superior temporal gyrus (STG) electrodes. Neural activity in a cluster of electrodes in middle STG follows the speech envelope. n.s., not significant. (**F**) Predicted neural response to the example sentence, based on discrete time series of peakEnv events (top) and peakRate events (bottom), in electrode E1. Both discrete event models outperform the continuous envelope model shown in (C). (**G**) Boxplot of R^2 distributions for the instantaneous envelope, peakEnv, and peakRate models and shuffled null distributions. Bars represent the 0.25 and 0.75 quantiles, respectively, across electrodes. Both discrete event models are significantly better than the continuous envelope model, but they do not significantly differ from each other, $^{**}P < 0.05$. (**H**) Portion of variance explained by the continuous envelope (Env), peakEnv (pEnv), and peakRate (pRate) models in single speech-responsive electrodes that tracked the envelope. Each dot represents one speech-responsive electrode.

RESULTS

Continuous speech: Discrete events are extracted from the speech envelope in bilateral STG

We asked whether neural populations in human STG represent the instantaneous moment-by-moment values of the amplitude envelope or whether they detect temporally discrete acoustic landmarks in the speech envelope and encode their occurrence and magnitude. We refer to an instantaneous representation as one that reflects the amplitude of the speech signal at each time point. We compared this to two independent models of encoding of prominent temporal landmarks: Peaks in the speech envelope (peakEnv; Fig. 1A, black arrows) and acoustic onset edges, defined as peaks in the first derivative of the envelope (peakRate; Fig. 1A, purple arrows). Figure 1A shows the timing of each of these landmarks in a sample sentence, with peakRate preceding peakEnv landmarks within each cycle of the envelope (between two consecutive envelope troughs). Both landmarks appear within each envelope cycle (i.e., envelope between two consecutive troughs), such that envelope cycle onset, peakRate, and peakEnv events are equally frequent in speech (Fig. 1B). Note also that all three events occur as frequently as single syllables, a prerequisite for one of these events to serve as a marker of syllables.

We used high-density ECoG recordings from the lateral temporal lobe of 11 participants (four left hemispheres; see table S1 for patient details), who were undergoing clinical monitoring for intractable epilepsy and volunteered to participate in the research study. Participants passively listened to 499 sentences from the TIMIT acoustic-phonetic corpus (*20*)

(see Fig. 1A for example sentence). We extracted the analytic amplitude of neural responses in the high gamma range (HGA; 70 to 150 Hz), which is closely related to local neuronal firing and can track neural activity at the fast rate of natural speech (*21*).

To compare the three models of envelope encoding, we first tested how well neural responses could be predicted from each model. For the instantaneous envelope model, we used the standard approach of cross-correlating neural activity and the speech envelope to determine the optimal lag at which the neural response resembles the speech envelope most closely. To model the neural data as a series of evoked responses to peakEnv or peakRate events, we used time-delayed multiple regression [also known as temporal receptive field (TRF) estimation; Fig. 1D] (*22*). This model estimates time-dependent linear filters that describe the neural responses to single predictor events. All models were trained on 80% of the data and subsequently tested on the remaining 20% that were held out from training, repeated five times for full cross-validation. Model comparisons were based on held-out test set R^2 values. For the comparison between models, we excluded sentence onsets because they induce strong transient responses in posterior STG after periods of silence typically found at the onset of a sentence or phrase but do not account for variance related to the ongoing envelope throughout an utterance (*18*).

In a representative electrode E1, HGA was well correlated with the speech amplitude envelope [across test set sentences: $R^2_{mean} = 0.19$, false discovery rate (FDR)-corrected $P < 0.001$, $R^2_{max} = 0.59$, mean lag = 60 ms; Fig. 1C], but prediction accuracies were significantly higher for the landmark models (peakEnv model: $R^2_{mean} = 0.63$, FDR-corrected $P < 0.001$, $R^2_{max} = 0.89$; peakRate model: $R^2_{mean} = 0.61$, FDR-corrected $P < .001$, $R^2_{max} = 0.85$; Fig. 1F). This pattern held across all speech-responsive STG electrodes (see Fig. 1E for electrode grid of a representative patient). Namely, HGA in up to 80% of speech-responsive electrodes was correlated with the speech envelope (n = 220 electrodes with FDR-corrected permutation $P < 0.05$, 6 to 42 per patient, average optimal lag: +86 ms, SD = 70 ms, $R^2_{mean} = 0.17$, $R^2_{max} = 0.59$; see fig. S1A for example traces from all speech-responsive electrodes). However, across these electrodes, landmark models outperformed the instantaneous envelope model (peakEnv model: $R^2_{mean} = 0.22$, $R^2_{max} = 0.65$; peakRate model: $R^2_{mean} = 0.22$, $R^2_{max} = 0.68$; signed-rank tests for comparison to continuous envelope model across electrodes, $P < 0.05$), whereas both landmark models predicted the neural data equally well (signed-rank test, $P > 0.5$; Fig. 1G). Notably, at the single-electrode level, the sparse landmark models robustly outperformed the envelope model (Fig. 1H). These results demonstrate that the STG neural responses to the speech envelope primarily reflect discrete peakEnv or peakRate landmarks, not instantaneous envelope values.

Slowed speech: Selective encoding of peakRate landmark

Next, we wanted to understand which of the two landmarks was driving neural responses to the speech envelope in STG. However, at natural speech rate, peakEnv and peakRate events occur on average within 60 ms of each other (Fig. 2B), which is why the encoding model approach used above could not disambiguate between them. To solve this, we created samples of slow speech that had longer envelope cycles (Fig. 2, A and C) and thus also longer time windows between peakRate and peakEnv events (Fig. 2B). These sentences were still fully intelligible (*23*) (see Supplementary Materials for example sentences and methods for technical details on speech slowing) and had the same spectral composition as the original speech samples (Fig. 2E). For example, in speech slowed to $^1/_4$ of normal speed, the average time between consecutive peakRate and peakEnv events was 230 ms, sufficient for a neural response evoked by peakRate to return to baseline before occurrence of the peakEnv landmark. Four participants listened to a set of four sentences that were slowed to $^1/_2$, $^1/_3$, and $^1/_4$ of the original speech speed (Fig. 2A). We predicted that, in the context of the slowed sentences, evoked responses would be more clearly attributable to one of the envelope features (Fig. 2F).

Figure 2D shows neural responses to a single sentence at different speech rates for an example electrode, alongside predicted responses based on the peakEnv and peakRate models. Neural responses at this electrode had the same number of peaks across speech rates, corresponding to single envelope cycles. At $^1/_2$ rate, predictions from both models were almost identical, whereas at $^1/_4$ rate a distinct lag between the predictions was readily apparent. Specifically, the predicted responses based on the peakEnv model lagged behind both the predictions from the peakRate model and the neural response.

Across all speech-responsive electrodes (n = 55, 5 to 20 per participant), we found that both models performed equally well at original and $^1/_2$ speech rate. However, with additional slowing, the peakRate model became increasingly better than the peakEnv model [linear effect of speech rate: $b_{rate} = 0.03$, SE = 0.007, $t(218) = 3.9$, $P = 10^{-4}$; Fig. 2, G and H]. We also examined the average evoked responses aligned to peakEnv and peakRate events at different speech rates. We predicted that responses should be reliably time-locked to the preferred landmark event at all speech rates. The average responses across electrodes aligned to peakEnv events (Fig. 2J, left) revealed a neural peak that shifted backward with speech slowing. Crucially, when speech was slowed by a factor of 3 or 4, neural peaks in high gamma amplitude occurred concurrent with or even before peakEnv events [test against 0: rate 3: $b = 0$, $P = 1$; rate 4: $b = -0.02$, SE = 0.01, $t(39) = 2.02$, $P = 0.05$], providing clear evidence against encoding of the peakEnv landmark. In contrast, when aligned to peakRate, neural responses peaked at the same latency at all speech rates (Fig. 2J, right), as summarized in Fig. 2K [interaction effect between speech rate and alignment: $F(1, 424) = 16.6$, $P < 10^{-4}$; effect of speech rate on alignment to peakEnv: $F(1, 209) = 20.13$, $P < 10^{-10}$; main effect of speech rate on alignment to peakRate: $F(1, 215) = 1.6$, $P = 0.2$]. Three further analyses supported this result. First, a comparison of a model including binary peakRate predictors and a model that includes peakRate magnitude showed that including peakRate magnitude increased model R^2 by up to 10% (fig. S1B). Second, a comparison of neural response alignments to peakRate versus peakEnv in natural speech supported the peakRate over the peakEnv model (fig. S2). Third, a comparison between the peakRate and an envelope trough (minEnv) model showed that peakRate events predicted neural data better than minEnv events (fig. S3). Moreover, the change in latency between acoustic and neural peaks also refutes the continuous envelope model because this model assumes a constant latency between the acoustic stimulus and corresponding points in the neural response. Notably, the fact that neural response occurred at the same time relative to stimulus onset at all speech rates (at time of peakRate event) refutes the possibility of qualitatively different processing of the natural and slowed speech stimuli, particularly that of increased top-down processing of the slowed speech. Together, the slow speech data show that STG neural responses to the speech amplitude envelope encode discrete events in the rising slope of the envelope, namely, the maximal rate of amplitude change, rejecting the alternative models of instantaneous envelope representation and peakRate encoding.

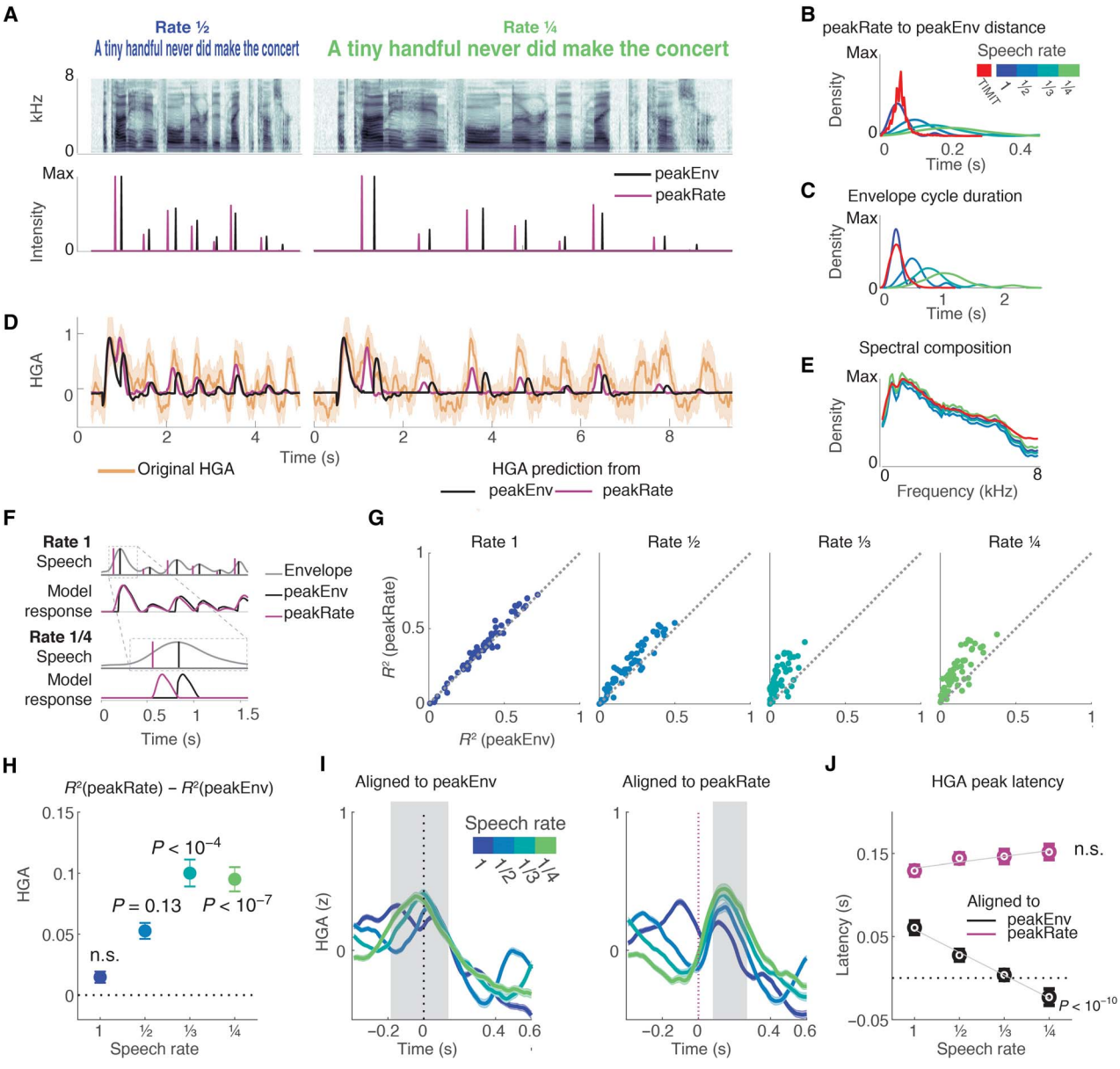

Fig. 2. Neural responses to slowed speech demonstrate selective encoding of peakRate events. (**A**) Top: Example sentence spectrogram at slowed speech rates of $^1/_2$ and $^1/_4$. Bottom: Example sentence peakEnv and peakRate events for both speech rates. (**B**) Distribution of latency between peakRate and subsequent peakEnv events, across all slowed speech task sentences and in full TIMIT stimulus set. Slowing increases time differences, and events become more temporally dissociated. (**C**) Distribution of envelope cycle durations by speech rate, across all slowed speech task sentences and in full TIMIT stimulus set. Sentence slowing makes envelope cycles more variable, increasing discriminability. (**D**) HGA response (orange) to an example sentence and neural responses predicted by sparse peakEnv (black) and peakRate (purple) models. Neural responses precede predicted responses of peakEnv model but are aligned with predicted responses of peakRate model accurately. (**E**) Average spectral composition is similar for stimuli at different speech rates and the full set of TIMIT stimuli. (**F**) Predicted neural responses for tracking of peakEnv events (black) and peakRate events (purple) for normally paced speech (top) and for slow speech (bottom). At rate 1, the models are indistinguishable. At rate $^1/_4$, the models predict different timing of evoked responses. (**G**) Comparison of test R^2 values for peakRate and peakEnv models by speech rate in all speech-responsive STG electrodes. As speech rate is slowed, peakRate model explains neural responses better than peakEnv model. Each dot represents a single speech-responsive electrode. (**H**) Mean (SEM) difference in R^2 between peakEnv and peakRate models. The peakRate model significantly outperforms the peakEnv model at $^1/_3$ and $^1/_4$ rates. (**I**) Average HGA after alignment to peakEnv (left) and peakRate (right) events. Gray area marks window of response peaks across all speech rates, relative to event occurrence. When aligned to peakEnv events, response peak timing becomes earlier for slower speech. When aligned to peakRate events, response peak timing remains constant across speech rates. (**J**) Mean (error bar, SEM across electrodes) HGA peak latency by speech rate and alignment. Speech slowing leads to shortening of the response latency relative to peakEnv events only, such that it occurs before peakEnv events at the slowest speech rate.

Speech analyses show that peakRate cues the phonological structure of syllables

Having identified peakRate as the envelope feature that is encoded in the STG, we aimed to understand how peakRate as acoustically defined temporal landmark relates to the linguistically defined syllabic structure of speech. Syllables are considered the temporal building blocks of words that carry the speed, prosody, and stress patterns of speech (*24*).

Figure 3A shows the first line of Shakespeare's sonnet XVIII (Shakespeare, 1609) annotated for lexical stress, syllabic boundaries, and the linguistically defined internal structure of syllables: the onset and the rhyme (composed of nucleus and coda) (*25*). The syllabic onset is the consonant or consonant cluster that precedes the syllabic vowel nucleus, and the rhyme includes the vowel nucleus and any consonant sounds (coda) that follow. A universal feature of syllables is that the speech amplitude (sonority) peaks locally on syllabic nuclei (Fig. 3B), even if the nucleus is a consonant sound, as in some language (*26*). We thus hypothesized that peakRate events would mark the transition between the syllable onset and its rhyme. Note that the term "syllable onset" here is distinct from our use of acoustic onsets described previously, which refer to the beginnings of sentences or phrases following long silences.

To test this, we analyzed the speech signal around peakRate events in our stimulus set. In the example in Fig. 3A, the syllable /sum/ in the word "summer's" has a delay between the syllable boundary and the peakRate event to accommodate the fricative onset consonant /s/, whereas the peakRate event is concurrent with the vowel onset. Across sentences, sound intensity increased rapidly at peakRate events (Fig. 3C, top), which was due to peakRate events occurring nearly concurrently with the linguistically defined transition from syllable onset to syllable nucleus (latency between peakRate and vowel nucleus onset: median = 0 ms, mean = 11 ms, SD = 50 ms; Fig. 3B, bottom). This relation was highly reliable, as more than 90% of vowel onsets were within 40 ms of a peakRate event. On the contrary, the latency between peakRate events and syllable boundaries was significantly larger and more variable (mean = 90 ms, SD = 60 ms, $t = -64$, $P < 0.001$; Fig. 3C).

In comparison, peakEnv events mark the midpoint of syllabic nuclei that are cued by peakRate events, as they occur after the peakRate events and within vowels (Fig. 3D). PeakEnv is, however, significantly less precise in cueing the consonant-vowel (C-V) transition than peakRate (latency between peakEnv and C-V transition: mean = 72, SD = 55 ms; comparison to peakRate bootstrap $P < 0.05$ for difference in means and variances; Fig. 3E). PeakEnv events thus inform the syllabic structure of a sentence by marking syllabic nuclei but are not informative with regard to the internal onset-rhyme structure of syllables.

In addition to serving as a reliable temporal landmark for syllables, we also found that the magnitude of peakRate events was important for distinguishing between unstressed and stressed syllables. Lexical stress carries lexical information in many languages, including English (i.e., distinguishing between different word meanings such as in ínsight versus incíte), supports the segmentation of continuous speech in words (*27*, *28*), and is the basis of poetic meter. Despite the frequent reduction of syllables in continuous natural speech, peakRate events marked more than 70% of nucleus onsets overall and 89% of stressed syllable nuclei (Fig. 3F and see fig. S4 for analysis of unmarked stressed syllable nuclei). The magnitude of peakRate was larger for stressed syllables than for unstressed syllables (sensitivity: $d' = 1.06$; Fig. 3G). PeakRate events thus provide necessary information to extract the timing of syllabic units from continuous speech, the critical transition from onset to rhyme within a syllable, and the presence of syllabic stress. While

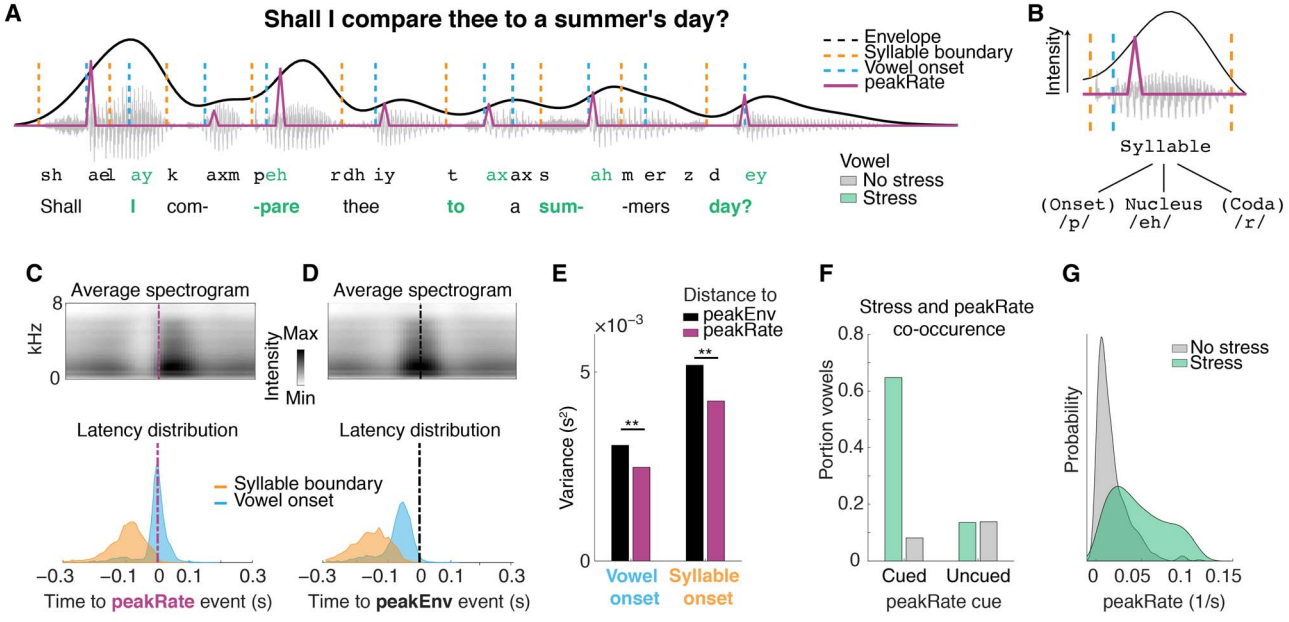

Fig. 3. peakRate events cue the transition from syllabic onset consonants to nucleus vowels. (**A**) Waveform of an example sentence with lexical stress, syllabic boundaries, vowel onsets, and peakRate events. peakRate events are concurrent with vowel onsets but not with syllabic boundaries. Middle: Schematic of syllabic structure in the example sentence, marking stressed and unstressed syllables. (**B**) Schematic of the envelope profile for a single syllable and the linguistic structure of a syllable. Intensity peaks on the syllabic nucleus relative to onset and coda. (**C** and **D**) Average speech spectrogram aligned to peakRate (C) and peakEnv (D) events. Top: Average speech spectrogram aligned to discrete event. peakRate events occur at time of maximal change in energy across frequency bands, whereas peakEnv events occur at times of maximal intensity across frequency bands. Bottom: Distribution of latencies of syllable boundaries and vowel (syllable nucleus) onsets relative to discrete event occurrence. Nucleus onsets are aligned to peakRate events more than syllable boundaries. For peakEnv, both distributions are wider than for peakRate alignment. (**E**) Variance in relative timing of syllable and vowel onsets and temporal landmarks. Smaller variance indicates that peakRate is a more reliable cue to vowel onsets that peakEnv, $^{**}P < 0.05$. (**F**) Co-occurrence of peakRate and vowels for stressed and unstressed syllables separately in the TIMIT stimulus set. PeakRate is a sensitive cue for C-V transitions, particularly to stressed syllables. (**G**) Distribution of peakRate magnitudes in stressed and unstressed syllables. Above a peakRate value of 0.05, a syllable has a 90% chance of being stressed.

many theories have posited the role of envelope for syllabic segmentation, i.e., detecting syllable boundaries, our results provide neurophysiological evidence for the alternative that peakRate is a landmark for the onset of the syllabic vowel nucleus, the importance of which for the cognitive representation of the syllabic structure of speech (*29*, *30*) and for detection of the most informative portions of the speech signal has been shown behaviorally (*31*). Notably, the relation between peakRate and nucleus onset also held in two other languages, Spanish and Mandarin Chinese (fig. S5).

Continuous speech: Topographic organization of temporal feature encoding on STG

Previous research described the encoding of sentence and phrase onset from silence in the posterior STG and encoding of spectrotemporal patterns corresponding to phonetic features, in particular vowel formants, in ongoing speech in the middle STG (*18*, *19*). We thus aimed to understand how peakRate encoding fits within this global organization and to verify that peakRate is encoded in addition to phonetic features and onset in STG. To this end, we fit the neural data with an extended time-delayed regression model that included binary sentence onset predictors, consonant phonetic feature predictors (plosive, fricative, nasal, dorsal, coronal, and labial), and vowel formants (F1, F2, F3, and F4), in addition to peakRate (Fig. 4A). We found that 80% of electrodes significantly responded to at least two features and that peakRate was most frequently encoded by itself (21 electrodes) or coencoded with either sentence onsets (73 electrodes) or vowel formants (70 electrodes; Fig. 4B).

Anatomically, encoding of peakRate was most prominent in middle STG in both hemispheres (left: $r = 0.18$, $P < 0.05$ and right: $r = 0.26$, $P < 10^{-5}$; Fig. 4, C and E). This pattern was distinct from the anatomical distribution of onset responses, which were strongest in posterior STG (left: $r = -0.29$, $P = 0.001$; right: $r = -0.37$, $P < 10^{-10}$; Fig. 4, C and E), consistent with our previous work (*18*).

We found no difference between the left and right hemispheric encoding of peakRate, suggesting bilateral encoding of this feature (Fig. 4D). However, because none of our patients had bilateral coverage, more subtle differences in temporal processing between hemispheres might exist. Together, these results indicate that encoding of temporal features spans a map from onsets in posterior STG to peakRate in middle STG, whereas the spectral structures that corresponds to phonetic content are encoded throughout STG (*19*).

Amplitude-modulated tones: Amplitude-rise dynamics alone drive neural responses to peakRate

Temporal and spectral changes in natural speech are inherently correlated. As a result, one potential confound is that peakRate encoding actually reflects the spectral changes that occur at the C-V transition in syllables (*3*, *32*). We therefore asked whether STG responses to peakRate reflect amplitude rise dynamics in the absence of concurrent spectral variation. To this end, we designed a set of nonspeech amplitude-modulated harmonic tone stimuli for a subset of eight participants.

Amplitude-modulated tone stimuli contained amplitude ramps rising from silence (ramp-from-silence condition) or from a "pedestal" at baseline amplitude of 12 dB below the ramp peak amplitude (ramp-from-pedestal condition), as shown in Fig. 5A. These two conditions were designed to broadly resemble amplitude rises at speech onset and within an ongoing utterance, respectively, but without any spectral modulations (such as vowel formant transitions) and variation in peak amplitude (such as amplitude differences between unstressed and stressed vowels) that are correlated with the amplitude rises in speech. Ramp durations and peak amplitude were kept constant across all stimuli, whereas rise times were parametrically varied (10 to 15 values between 10 and 740 ms; see table S1 for all rise time values) under both silence and pedestal conditions. The stimuli had complementary rising and falling slopes, together ensuring equal stimulus durations. To simplify the analyses across the silence and pedestal conditions, we describe these stimuli in terms of amplitude rate of change [(peak amplitude − baseline amplitude)/rise time, i.e., amplitude rise slope; Fig. 5B]. Because

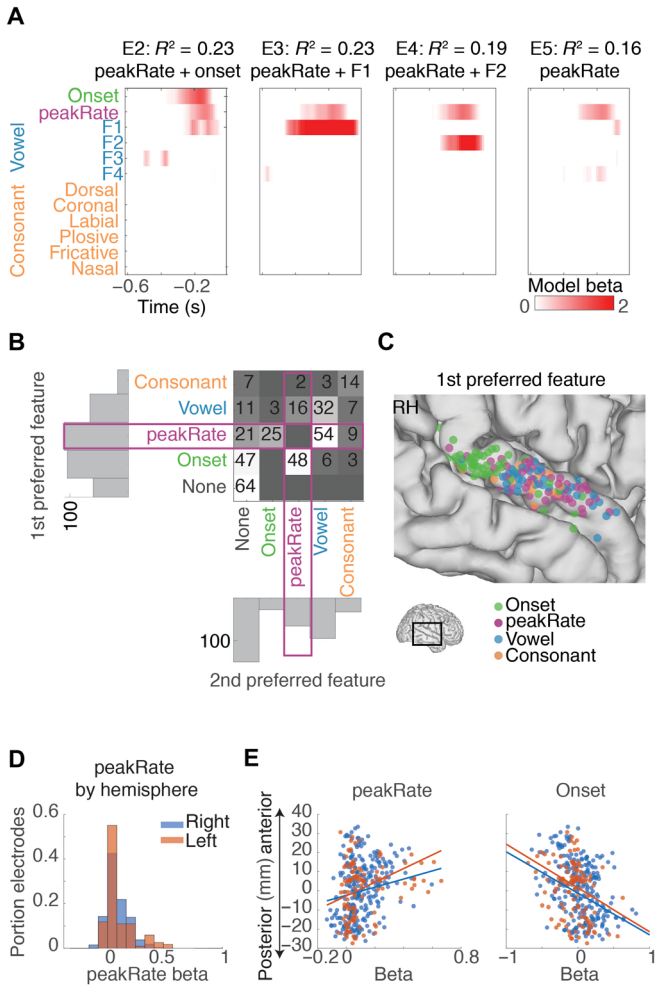

Fig. 4. Independent and joint encoding of peakRate and other speech features. (**A**) Linear weights from an encoding model with phonetic features and peakRate events for four example electrodes. Different electrodes show encoding of different features alongside peakRate. (**B**) Number of electrodes with different combinations of the two significant features with the largest linear weights across STG electrodes. Vowel formant predictors (blue) and consonant predictors (orange) are each combined for visualization purposes. Onset and peakRate are blank along the diagonal because they contain one predictor only. peakRate encoding co-occurs with different phonetic features [e.g., E2 to E4 in (A)] but can also occur in isolation [E5 in (A)]. (**C**) Anatomical distribution of electrodes with primary encoded onset, peakRate, vowel, or consonant features across all right hemisphere electrodes. Onset encoding is clustered in posterior STG, and peakRate encoding is predominant in middle STG. RH, right hemisphere. (**D**) Distribution of model beta values for peakRate in left and right hemisphere. (**E**) Left: Correlation between electrode position along STG and peakRate beta. Right: Correlation between electrode positions along STG and onset beta. Onset beta values are largest in posterior STG, and peakRate beta values are largest in middle STG.

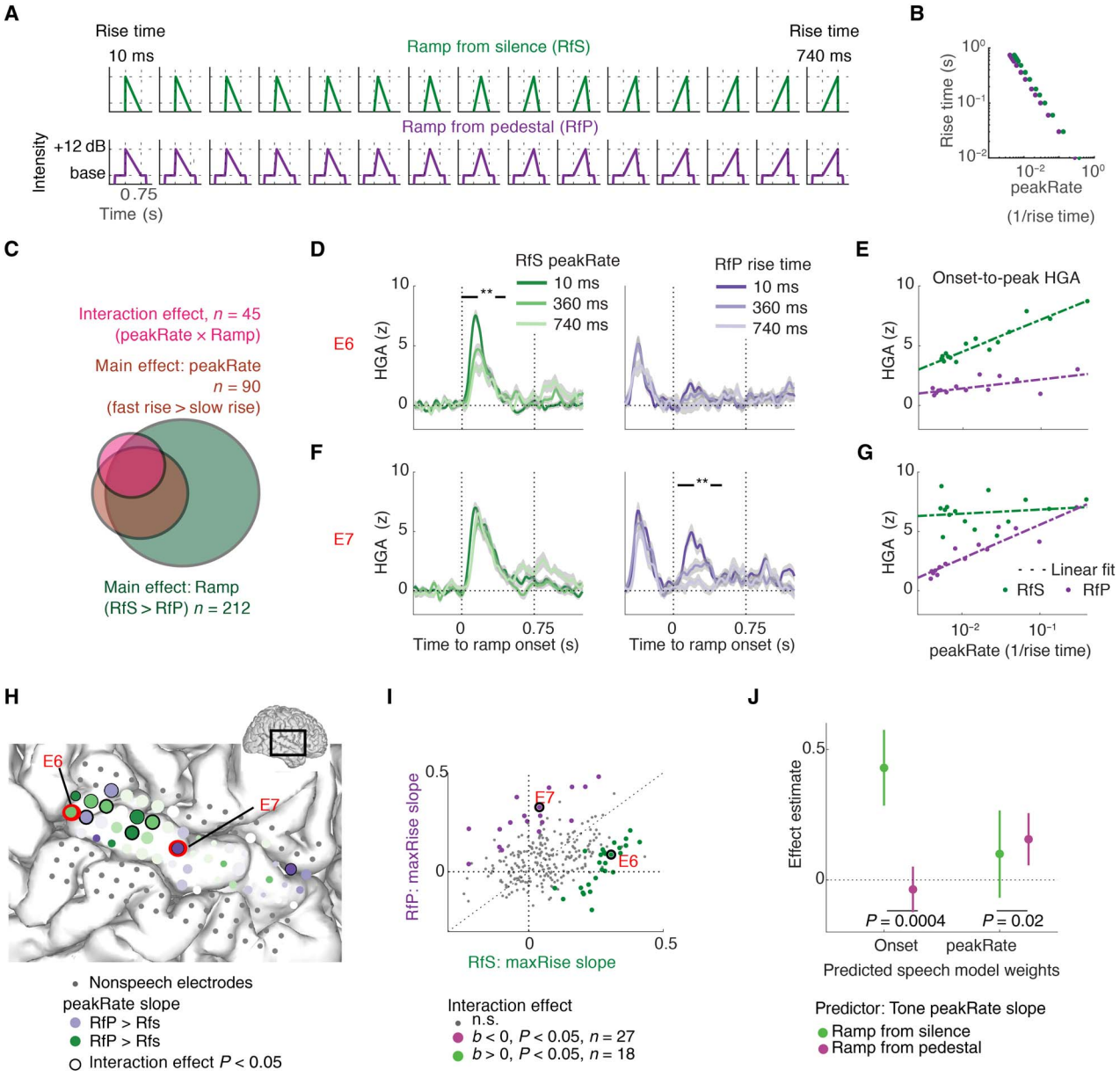

Fig. 5. STG encoding of amplitude modulations in nonspeech tones in onsets and in ongoing sounds. (**A**) Tone stimuli used in the nonspeech experiment. Rate of amplitude rise is manipulated parametrically, but peak amplitude and total tone duration are matched. (**B**) Relationship between ramp rise time and peakRate defined as for the speech stimuli. The peakRate value was reached immediately at ramp onset, as ramp amplitude rose linearly. (**C**) Effect distribution across all electrodes. Eighteen percent of all electrodes showed a significant interaction effect between ramp type and peakRate, in addition to 72% showing a main effect of ramp type and 36% showing a main effect of peakRate. (**D**) HGA responses to tones with three selected ramp rise times under ramp-from-silence (RfS; left) and ramp-from-pedestal (RfP; right) conditions in example electrode E6, $^{**}P < 0.05$. (**E**) Onset-to-peak HGA in electrode E6 as function of ramp peakRate, separately for ramp-from-silence and ramp-from-pedestal conditions. E6 codes for amplitude rate of change under ramp-from-silence condition but not under ramp-from-pedestal condition. (**F**) Same as (C), for example electrode E7, $^{**}P < 0.05$. (**G**) Same as (D), for example electrode E7. E7 codes for amplitude rate of change under ramp-from-pedestal condition but not under ramp-from-silence condition. (**H**) Temporal lobe grid from an example patient, with example electrodes E6 and E7 marked in red. Electrode color codes for relative magnitude of the peakRate effect on peak HGA under tone conditions. The purple electrodes' HGA was more affected by peakRate under ramp-from-pedestal condition, and the green electrodes' HGA was correlated with peakRate values under ramp-from-silence condition more than under ramp-from-pedestal condition. Electrode size reflects maximal onset-to-peak HGA across all conditions. (**I**) Slopes of peakRate effects on peak HGA, separately for each ramp condition. In colored electrodes, the ramp condition × peakRate interaction was significant. Two distinct subsets of electrodes code for rate of amplitude change under one of the two conditions only. (**J**) Linear weights from a multiple regression model that predicted onset and peakRate linear weights in the speech model from peakRate slopes in tone model across electrodes. Representation of amplitude modulations at onsets and in ongoing sounds is shared in speech and in nonspeech tones. Encoding of peakRate for envelope rises from silence is dissociated from peakRate encoding in ongoing sounds, in speech and in nonspeech tone stimuli.

the amplitude rose linearly, the maximal rate of amplitude change (peakRate) occurred at ramp onset and was consistent throughout the rise.

Analyses were focused on the same electrodes that were included in analyses of the speech task ($n = 226$ electrodes across eight patients, with 11 to 41 electrodes per patient). Of these electrodes, 95% showed evoked responses to tone stimuli [FDR-corrected for multiple comparisons $P < 0.05$ for at least one of the effects under the ramp condition × rise time analysis of variance (ANOVA) analysis of peak amplitudes]. Different rates of change were associated with differences in HG responses, which stereotypically started immediately after ramp onset and peaked at ~120 ms. In particular, for stimuli with intermediate and slow rates of change, the neural HGA response peak preceded the peak amplitude in the stimulus (fig. S6A). This result further corroborates peakRate, and not peakEnv, as the acoustic event that drives neural responses (*33*). Moreover, neural responses to ramp-from-pedestal tones returned to baseline between stimulus and ramp onsets, despite an unchanged level of tone amplitude (signed-rank test between HGA of 0 to 200 ms after stimulus onset and HGA of 300 to 500 ms after stimulus onset, $P < 10^{-10}$). This provides additional direct evidence for the encoding of amplitude rises and not the continuous envelope or amplitude peaks on STG.

In addition, in a control experiment, we tested whether neural responses would be different if the rate of amplitude change varied throughout the rising slope of the envelope. Neural responses to linearly rising ramps did not differ from responses to stimuli with a nonlinear amplitude rise dynamics, for which amplitude rate of change was maximal at ramp onset and then slowed down throughout the rising slope of the envelope ($n = 2$ patients). The neural responses to both stimuli were qualitatively identical and determined by the peakRate values (see fig. S7).

Amplitude-modulated tones: Distinct encoding of onsets and amplitude modulations in tone stimuli

Next, we wanted to test how the rate of amplitude rise would alter the magnitude of neural responses and whether neural responses would differentiate between preceding contexts, that is whether the ramp started from silence (analog-to-speech onsets) or from a pedestal (as in ongoing speech). We focused the following analyses on the effect of amplitude rise dynamics on the onset-to-peak magnitude of HGA responses, defined as the difference between the HGA at the time of ramp onset and at HG peak. We tested how peak HGA depended on the ramp condition (ramp from pedestal versus ramp from silence) and peakRate values by fitting a general linear model with predictors tone condition, peakRate, and their linear interaction, separately for each electrode.

Tone stimuli evoked robust responses in electrodes located in posterior and middle STG (see Fig. 5G for example electrode grid), with stronger responses to ramps starting from silence (mean b = 0.3212 of 243 electrodes with $P < 0.05$; exact binomial test against chance level of observing the effect on 5% of electrodes, $P < 10^{-4}$; Fig. 5C). Moreover, on a subset of STG electrodes, peak HGA was modulated by peakRate, with larger neural responses to fast rising ramps (mean b = 0.2 on 90 of 243 electrodes with $P < 0.05$; exact binomial test against chance level of observing the effect on 5% of electrodes, $P < 10^{-4}$; Fig. 5C). Similar to our findings in speech, some electrodes encoded peakRate under one of the two ramp conditions only, resulting in a significant interaction effect on 45 electrodes (18% of channels; exact binomial test against chance level of observing the effect on 5% of electrodes, $P < 10^{-4}$).

Electrodes E6 (Fig. 5D) and E7 (Fig. 5F) exemplify the two response patterns that drove this interaction effect, with a negative interaction effect in E6 and a positive interaction effect in E7. The amplitude of evoked responses in electrode E6 decreased with peakRate under the ramp-from-silence condition ($b = 0.3$, $P < 0.05$) but was not affected by peakRate under the ramp-from-pedestal condition [$b = 0.08$, $P > 0.05$; Fig. 5, C (right) and E for peak HGA under all rise time conditions; linear interaction of ramp condition × peakRate: $b = -0.29$, $P < 0.05$]. Electrode E7 showed the opposite pattern, with a decrease in HGA for lower peakRate values under the ramp-from-pedestal condition ($b = 0.32$, $P < 0.05$; Fig. 5F, right) but no effect of peakRate under the ramp-from-silence condition [$b = 0.04$, $P > 0.05$; Fig. 5, F (left) and G for peak HGA under all rise time conditions; linear interaction of ramp condition × peakRate: $b = 0.21$, $P < 0.05$]. Overall, neural activity on electrodes with a negative interaction effect ($n = 27$; green in Fig. 5J) encoded peakRate under the ramp-from-silence condition but not under the ramp-from-pedestal condition, whereas electrodes with a positive interaction effect ($n = 18$; purple in Fig. 5J) encoded peakRate under the ramp-from-pedestal condition only (onset and peakRate encoding were similarly independent in the speech data; see fig. S8).

These results demonstrate that neural populations on STG encode amplitude rises independent from other co-occurring cues in speech. By parametrically varying peakRate in isolation from other amplitude parameters, these data strongly support the notion that the STG representation of amplitude envelopes reflects encoding of discrete auditory edges, marked by time points of fast amplitude changes. These data also revealed a notable double dissociation between the contextual encoding of peakRate in sounds that originate in silence and the encoding of peakRate in amplitude modulations of ongoing sounds, which indicates that dedicated neural populations track onsets after silences, e.g., sentence and phrase onsets, and intrasyllabic transitions.

Comparison between onset and peakRate encoding in continuous speech and amplitude-modulated tones

Amplitude rate-of-change encoding is similar in speech and nonspeech tones

In a final analysis, we tested whether encoding of peakRate events in nonspeech tones reflects the same underlying computations as detection of peakRate events in speech. We reasoned that if neural populations encoded amplitude rises in tones and speech stimuli similarly, then neural responses on electrodes that preferentially encode the dynamics of amplitude rises for amplitude ramps that start in silence (e.g., Fig. 5, D and E) would also respond to sentence onsets in speech (as indicated by high beta values for the onset predictor in the speech encoding model). Conversely, we expected that electrodes that encode the dynamics of amplitude rises for ramps in ongoing tones (ramp-from-pedestal condition; e.g., Fig. 5, F and G) would also encode peakRate events within sentences [as indicated by high beta values for peakRate in the speech TRF model]. To test this, we assessed whether speech model beta values for onset and peakRate could be predicted from the same electrodes' peakRate beta values under the ramp-from-silence and ramp-from-pedestal conditions. Two separate linear multiple regressions were fit to predict speech model beta values from the peakRate beta values in the tone task (Fig. 5K).

We found that responses to sentence onsets in speech were significantly predicted by encoding of peakRate in tone ramps starting from silence ($b = 0.64$, SD = 0.11, $P < 10^{-7}$), but not by tracking of peakRate in tone ramps within ongoing tones ($b = 0.06$, SD = 0.14, $P = 0.7$), and this difference was significant (permutation test of regression estimate

equality, $P = 0.003$). Likewise, encoding of peakRate events after sentence onset in speech was not related to encoding of tone amplitude rise from silence ($b = 0.02$, SD = 0.06, $P = 0.7$), but it was significantly predicted by encoding of amplitude rise dynamics in ongoing tones ($b = 0.16$, SD = 0.07, $P = 0.02$). Crucially, this difference was also significant (permutation test of regression estimate equality, $P = 0.02$). This analysis shows a robust overlap between the neural computations underlying the tracking of sound onset and amplitude modulation dynamics in speech and tones. Moreover, it corroborates the functional and anatomical dissociation between tracking of amplitude modulations in two distinct dynamic ranges—at onset and in ongoing sounds.

DISCUSSION

Our findings demonstrate that a defined region of the human auditory speech cortex, the middle STG, detects a specific envelope feature: the acoustic onset edge (peakRate). It does not linearly process the moment-by-moment modulation of the ongoing speech envelope or other events such as amplitude peaks, valleys, or offsets. In this way, the middle STG represents the speech envelope as a series of temporally discrete events, and the cortical response magnitude reflects the velocity of envelope rises. Edge detection emerges as a flexible computational mechanism for encoding the structure of continuous speech across speech rates (*34*, *35*), providing a framework to the temporal organization of the speech stream and discretizing it into a series of amplitude-based events.

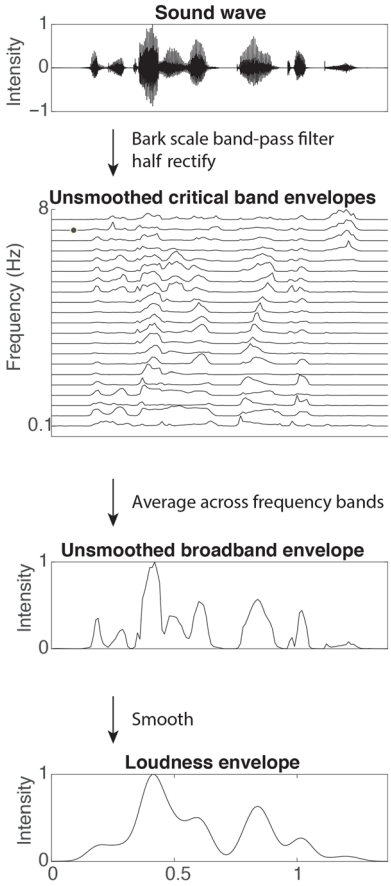

Fig. 6. Schematic of envelope extraction method.

According to a prominent view in speech neuroscience, the envelope allows speech to be parsed into chunks defined by syllabic boundaries (*36*). However, definitive evidence for neural encoding of syllable boundaries, which more closely correspond to troughs in the amplitude envelope (*37*), has been elusive (*2*, *38*). Instead, peakRate provides a temporal structure for the organization of the speech signal around the intrasyllabic transition between onset and nucleus within each syllable and, unlike syllabic boundaries, conveys essential phonologically relevant information, such as the timing of the onset-rhyme transition, speech rate, and syllabic stress patterns (*27*, *28*).

Our findings are consistent with landmark-based theories of speech recognition (*39*, *40*) that have posited that amplitude-based landmarks are a necessary level of speech analysis. This is supported by behavioral psychophysics: For example, the introduction of fast amplitude rises alone to vowel-like harmonic tones can induce the perception of a consonant-vowel sequence (*41*, *42*), and amplitude rises are critical for the correct perception of the temporal order of phonetic sequences (*43*).

The peakRate landmark model links a relatively simple auditory encoding property (edge detection) to linguistic properties of the syllable. In phonology, the relative intensity of neighboring phonemes (called sonority) within a syllable follows a universal structure. Sonority always peaks on the syllabic nucleus, such that peakRate marks the onset of the nucleus. In English, this is equivalent to marking vowel onsets. However, in languages with consonants in the syllabic nucleus [e.g., Czech, Tashlhiyt Berber (*26*)], peakRate would still mark the nucleus because those consonantal nucleus sounds are the most sonorous. That is, this syllable landmark is based on amplitude, not the spectral features of a vowel or consonant.

The onset-rhyme transition emerges from this STG representation of the envelope as a major aspect of syllabic structure. This is well in line with behavioral findings. For instance, while listeners often disagree in their placement of syllabic boundaries, they easily agree on the number and stress of syllables in an utterance (*44*). Moreover, there is strong behavioral evidence for the perceptual distinctiveness of an onset and rhyme in a syllable across many languages (the overwhelming majority of human languages adheres to the onset-rhyme distinction in syllables. However, even for languages with a different structure, e.g., mora languages such as Japanese, where the onset and the vowel form a unit, peakRate events might contribute by marking the time of the mora, and the magnitude of peakRate might inform about the number of morae in the syllable), and detection of a landmark at this transition may support this. For example, speech confusions often occur at the same syllable position (e.g., onsets are exchanged with other onsets), the similarity between words is more readily recognized if it occurs along the onset-rhyme distinction (*45*), and the ability to distinguish between onset and rhyme is a predictor of successful reading acquisition (*46*).

The amplitude envelope is an important feature of sounds, and amplitude envelope dynamics are encoded throughout the auditory system of different animal models (*47*). Single-unit recordings along the auditory pathway up to secondary auditory cortices showed that the timing of single neural spikes and their firing rate reflect the dynamics of envelope rises (*48*–*50*). Envelope encoding in human STG possibly emerges from amplitude envelope representations at lower stages of the auditory system. It is thus likely not unique to speech processing but rather a universal acoustic feature with a direct link to the linguistic structure of speech.

Edge detection is also a central principle of processing in vision. Previous work demonstrated that similar computational principles, namely, tracking of first and second derivatives of the signal intensity,

can account for perceptual and physiological aspects of edge detection in audition and vision (*51*). In addition, our results raise the possibility that distinct neural populations might be dedicated to detection of acoustic edges within different dynamic ranges, such as onsets and changes within a stimulus.

We recently reported that the entire posterior STG encodes the speech onsets from silence. Here, we reproduce this finding and also show that encoding for another amplitude cue within ongoing speech, peakRate, is localized to the middle STG area where it induces evoked responses on more than half of speech-responsive electrode sites. Local neural populations within each zone can be cotuned to specific phonetic features (*19*). In summary, our results establish a cortical map for a landmark-based temporal analysis of speech in the human STG, which lays the foundation for perception of the temporal dynamics of speech, including stress patterns in everyday speech and poetic meter and rhyme.

METHODS

Participants

Twelve (two female) patients were implanted with 256-channel, 4-mm electrode distance, subdural ECoG grids as part of their treatment for intractable epilepsy. Electrode grids were placed over the peri-Sylvian region of one of patients' hemispheres (five left and six right hemisphere grids). Grid placement was determined by clinical considerations. Electrode positions were extracted from postimplantation computer tomography scans, coregistered to the patients' structural magnetic resonance imaging and superimposed on three-dimensional reconstructions of the patients' cortical surfaces using a custom-written imaging pipeline (*52*). All participants had normal hearing and left-dominant language functions. Ten participants were native speakers of English. Two participants were native speakers of Spanish with no knowledge of English. As we saw no difference between their results and the data of English native speakers, their data were included in all analyses. The study was approved by the University of California, San Francisco Committee on Human Research. All participants gave informed written consent before experimental testing. All patients participated in the speech experiment, a subset of four patients participated in the slow speech experiment, and a subset of eight patients participated in the amplitude modulated tone experiment (table S1).

Stimuli and procedure

All stimuli were presented at a comfortable ambient loudness (~70 dB) through free-field speakers (Logitech) placed approximately 80 cm in front of the patients' head using custom-written MATLAB R2016b (MathWorks, www.mathworks.com) scripts. Speech stimuli were sampled at 16,000 Hz, and tone stimuli were sampled at 48,000 Hz for presentation in the experiment. Participants were asked to listen to the stimuli attentively and were free to keep their eyes open or closed during stimulus presentation.

Continuous speech (TIMIT)

Participants passively listened to a selection of 499 English sentences from the TIMIT corpus (*20*), spoken by a variety of male and female speakers with different North American accents. Data in this task were recorded in five blocks of approximately 4-min duration each. Four blocks contained distinct sentences, presented only once across all four blocks, and one block contained 10 repetitions of 10 sentences. This latter block was used for validation of TRF models (see below). Sentences were 0.9 to 2.4 s long and were presented with an intertrial interval of 400 ms. Acoustic analyses of Spanish (*53*) and Mandarin Chinese (*54*) corpora shown in fig. S5 followed the same amplitude extraction methods as for the TIMIT corpus.

Slowed speech

The slowed speech stimulus set consisted of four sentences selected from the repetition block of the TIMIT stimulus set presented at four different speech rates: original, $^1/_2$, $^1/_3$, and $^1/_4$. Participants listened to the stimuli in blocks of 5-min duration, which contained three repetitions of each stimulus with an intertrial interval of 800 ms. Each participant listened to three to five blocks of slowed speech, resulting in 9 to 15 stimulus repetitions per participant. Slowed speech stimuli were created using the PSOLA (Pitch Synchronous Overlap and Add) algorithm, as implemented in the software Praat (*55*), which slows down the temporal structure of the speech signal, while keeping its spectral structure constant (*56*).

Amplitude-modulated tones

In this nonspeech tone experiment, participants passively listened to harmonic tones that contained an amplitude ramp starting either from silence (ramp-from-silence condition) or from clearly audible baseline amplitude (ramp-from-pedestal condition; Fig. 4A). The total duration of the amplitude ramp was 750 ms. Under the ramp-from-pedestal condition, the ramp was preceded by 500 ms and followed by 250 ms of the tone at baseline amplitude (12 dB below the peak amplitude). The peak amplitude of the ramp was the same across conditions. Ramp amplitude increased linearly from baseline/silence and then immediately fell back to baseline/silence for the remainder of the ramp duration. Ramp rise times took 10 to 15 different values between 10 and 740 ms dependent on the patient (full set: 10, 30, 60, 100, 140, 180, 270, 360, 480, 570, 610, 650, 690, 720, and 740). Under the ramp-from-silence condition, the stimuli were harmonic tones with fundamental frequency of 300 Hz and five of its harmonics (900, 1500, 2100, 2700, and 3300 Hz). Under the ramp-from-pedestal condition, half of the stimuli had the same spectral structure as in the ramp-from-silence background, and half the stimuli were pure tones of either 1500 or 2700 Hz. C-weighted amplitude was equalized between harmonics. Because neural responses to the ramp did not differ between harmonic and pure ramp stimuli, we report all analyses pooled across these stimuli. Patients passively listened to 10 repetitions of each stimulus. Stimulus order was pseudorandomized, and the whole experiment was split into five equal blocks of approximately 5 min each.

For a comparison between conditions, we converted ramp rise times to rate of amplitude rise, calculated as

$$\text{Rate of change}\left(\frac{1}{s}\right) = \frac{P_{\text{peak}} - P_{\text{base}}}{\text{rise time}} \quad (1)$$

where P_{peak} and P_{base} are sound pressure at ramp peak and at baseline, respectively. Because of the linear rise dynamics, the rate of amplitude rise reached its maximum at ramp onset and remained constant throughout the upslope of the ramp, so that peakRate was equal to the rate of amplitude rise.

Data analysis

All analyses were conducted in MATLAB R2016b (MathWorks, www.mathworks.com) using standard toolboxes and custom-written scripts.

Neural data acquisition and preprocessing

We recorded ECoG signals with a multichannel PZ2 amplifier, which was connected to an RZ2 digital signal acquisition system [Tucker-Davis Technologies (TDT), Alachua, FL, USA], with a sampling rate of 3052 Hz. The audio stimulus was split from the output of the presentation computer and recorded in the TDT circuit time aligned with the ECoG signal. In addition, the audio stimulus was recorded with a microphone and also input to the RZ2. Data were online referenced in the amplifier. No further re-referencing was applied to the data.

Offline preprocessing of the data included (in this order) downsampling to 400 Hz, notch-filtering of line noise at 60, 120, and 180 Hz, exclusion of bad channels, and exclusion of bad time intervals. Bad channels were defined by visual inspection as channels with excessive noise. Bad time points were defined as time points with noise activity, which typically stemmed from movement artifacts, interictal spiking, or nonphysiological noise. From the remaining electrodes and time points, we extracted the analytic amplitude in the high-gamma frequency range (70 to 150 Hz, HGA) using eight band-pass filters [Gaussian filters, logarithmically increasing center frequencies (70 to 150 Hz) with semilogarithmically increasing bandwidths] with the Hilbert transform. The high-gamma amplitude was calculated as the first principal component of the signal in each electrode across all eight high-gamma bands, using principal components analysis. Last, the HGA was down-sampled to 100 Hz and z-scored relative to the mean and SD of the data within each experimental block. All further analyses were based on the resulting time series.

Initial electrode selection

Analyses included electrodes located in the higher auditory and speech cortices on the STG, which showed robust evoked responses to speech stimuli, defined as electrodes for which a linear spectrotemporal encoding model (*22*) explained more than 5% of the variance in the test dataset (see below for model fitting procedure, which was identical to the TRF fitting procedure). Analyses contained 384 electrodes, 11 to 56 within single patients.

Continuous speech experiment (TIMIT)

Acoustic feature extraction

We extracted the broad amplitude envelope of speech stimuli using the specific loudness method introduced by Schotola (*57*), which is qualitatively identical to other widely used amplitude extraction methods (*58*, *59*). This method extracts the analytic envelope of the speech signal filtered within critical bands based on the Bark scale (*60*) by square-rectifying the signal within each filter bank, averaging across all bands and band-pass filtering between 1 and 10 Hz (Fig. 6). We then calculated the derivative of the resulting loudness contours as a measure of the rate of change in the amplitude envelope. Last, we extracted the sparse time series of local peaks in the amplitude envelope (peakEnv) and in its derivative (peakRate). This procedure resulted in a set of features for each cycle of the amplitude envelope (defined as the envelope between two neighboring local troughs; Fig. 1A, inset): peakEnv and peakRate amplitudes, their latencies relative to preceding envelope trough, and the total duration of the cycle. Note that we did not apply any thresholding to definition of troughs or peaks; however, we retained the magnitude of the envelope and its derivative at local peaks for all model fitting, such that models naturally weighted larger peaks more than small peaks. We also compared this envelope extraction method to a 10-Hz low-pass filtered broadband envelope of the speech signal, which produced the same qualitative results throughout the paper.

General model fitting and comparison approach

All models were fivefold cross-validated: Models were fit on 80% of the data and evaluated on the held-out 20% of the dataset, as Pearson's correlations of predicted and actual brain responses. There correlations were then squared to obtain R^2, a measure of the portion of variance in the signal explained by the model. Model comparisons were conducted on cross-validated R^2 values, averaged across all fivefolds, which were calculated separately for average neural responses (across 10 repetitions) for each test set sentence. The use of cross-validation and model testing on a held-out set allows model comparisons across models of different complexity, as between the continuous envelope and sparse peakRate models (*22*). Formal comparisons between R^2 values across electrodes were conducted using Wilcoxon rank sum test and a significance threshold of 0.05. To test the significance of each model for single electrodes, models were refit 1000 times on shuffled data, producing permutation-based null distributions of model R^2.

Representation of instantaneous amplitude envelope

To test whether neural data contain a representation of instantaneous amplitude envelope values, we calculated the maximum of the cross-correlation between the speech amplitude envelopes and HGA, restricted to positive lags (i.e., neural data lagging behind the speech envelope). The optimal lag was determined on the training set of the data, and model fit was then assessed on the independent test set (see above).

Time-delayed multiple regression model (TRF)

To identify which features of the acoustic stimulus electrodes responded to, we fit the neural data with linear temporal receptive field (TRF) models with different sets of speech features as predictors. For these models, the neural response at each time point [HGA(t)] was modeled as a weighted linear combination of features (f) of the acoustic stimulus (X) in a window of 600 ms before that time point, resulting in a set of model coefficients, $b_{1\ldots,d}$ (Fig. 1C) for each feature f, with $d = 60$ for a sampling frequency of 100 Hz and inclusion of features from a 600-ms window.

$$\sum_{k=1}^{d}\sum_{f=1}^{F} b(k,f)X(f,t-k) = \mathrm{HGA}(t) \quad (2)$$

The models were estimated separately for each electrode, using linear ridge regression on a training set of 80% of the speech data. The regularization parameter was estimated using a 10-way bootstrap procedure on the training dataset for each electrode separately. Then, a final value was chosen as the average of optimal values across all electrodes for each patient.

For all models, predictors and dependent variables were scaled to between −1 and 1 before entering the model. This approach ensured that all estimated beta values were scale free and could be directly compared across predictors, with beta magnitude being an index for the contribution of a predictor to model performance.

Feature receptive field models

To assess the extent of overlap between amplitude envelope tracking and phonetic feature encoding in STG electrodes, we also fit a time-delayed multiple regression model that included median values of the first four formants for all vowels and place and manner of consonant articulation, in addition to onset and peakRate predictors. Phonetic feature and formant predictors were timed to onsets of the respective phonemes in the speech signal. Comparisons of beta values for the different predictors were based on maximal beta values across time points. Phonetic features for this model were extracted from time-aligned

phonetic transcriptions of the TIMIT corpus and standard phonetic descriptions of American English phonemes. Vowel formants were extrapolated using the freely available software package Praat (*55*).

To assess the significance of predictors in TRF encoding models, we used a bootstrapping procedure. The model was refit 1000 times on a randomly chosen subset of the data. This was used to estimate distributions of model parameters. The significance of a single feature in the model was determined as at least 10 consecutive significant beta values for this feature ($P < 0.05$, with Bonferroni correction for multiple comparisons across electrodes).

Spatial distribution test

The spatial organization of electrodes encoding onsets, peakRate, and phonetic features on STG was tested by correlating the beta values (b) for each feature with the location of electrodes along the anterior-to-posterior axis (p) of the projection of single patients' electrode location on the MNI template (Montreal Neurological Institute). Positive correlations indicate stronger encoding in more anterior STG, whereas negative correlations indicate stronger encoding in more posterior STG.

Slowed speech experiment

TRF models

We tested the time-delayed multiple regression models that were fitted on the TIMIT training data on data from the four speech rate conditions. Note that all four sentences that were presented in this task were part of the TIMIT test set and, thus, features created for the TIMIT sentences were reused in this task with the appropriate adjustment of latencies. We used quality of model fits and the comparison between models at each speech rate as an indicator of whether STG retained a representation of the instantaneous amplitude envelope or of landmark events. We used a linear regression across electrodes to test whether the difference between peakEnv and peakRate model changed with speech rate.

Realignment to acoustic landmarks

Neural HGA data were segmented around peakEnv and peakRate landmark occurrence (400 ms before and 600 ms after each landmark) and averaged within each rate condition. The analysis included all landmark occurrences ($n = 21$), excluding sentence onsets. We extracted the latency of HG peaks relative to both landmarks for each electrode. Because latency estimations are highly sensitive to noise at low signal-to-low ratios, we only included electrodes from the upper quantile of response magnitudes to peakRate or peakRate, as estimated in the TRF models (5 to 20 per participant, 41 overall). The effect of speech rate and landmark onto HGA peak latencies was assessed using a two-way repeated-measures ANOVA with factor speech rate and landmark.

Amplitude-modulated tone experiment

Data acquisition and preprocessing

Data acquisition and preprocessing followed the same procedure as for the speech data. However, z scoring for the tone task was performed separately for each trial based on HG mean and variance during the 500 ms before stimulus onset. Responses were averaged across the repetitions of the same ramp condition and rise time combination before further analyses.

Responsiveness to tone stimuli and electrode selection

Because we were interested in characterizing how speech electrodes respond to nonspeech amplitude-modulated tones, analyses were performed on all electrodes that were included in the speech task. We quantified the response to ramp onsets as the trough-to-peak amplitude difference between the HGA in a 50-ms window around ramp onset and the maximal HGA in a 750-ms window after ramp onset.

General linear model of response amplitudes

We analyzed the effects of ramp type (ramp from silence versus ramp from background) onto response trough-to-peak amplitude for every electrode separately, using a general linear model with predictors ramp type, log-ramp rise time, and their linear interaction, with a significance threshold set to $P < 0.05$ (uncorrected).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/11/eaay6279/DC1

Fig. S1. Comparison between continuous envelope model and sparse landmark models.
Fig. S2. Segmentation of neural responses to naturally produced sentences (TIMIT) around peakEnv and peakRate events.
Fig. S3. Comparison between neural response predictions based on peakRate and minEnv models for slowed speech.
Fig. S4. Stressed vowels missed by peakRate.
Fig. S5. Cross-linguistic analysis of peakRate and vowel onset co-occurrence.
Fig. S6. Latency of neural response peaks as function of ramp rise time in amplitude-modulated tones.
Fig. S7. Single-electrode responses to linear and sigmoidal ramp tones.
Fig. S8. Independent and joint encoding of peakRate and other speech features.
Table S1. Participants' details.

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. R. Drullman, J. M. Festen, R. Plomp, Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* **95**, 2670–2680 (1994).
2. S. Rosen, temporal information in speech: Acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. Lond. B* **336**, 367–373 (1992).
3. A. J. Oxenham, J. E. Boucher, H. A. Kreft, Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy. *J. Acoust. Soc. Am.* **142**, EL264 (2017).
4. A. W. F. Huggins, On the perception of temporal phenomena in speech. *J. Acoust. Soc. Am.* **51**, 1279–1290 (1972).
5. J. E. Peelle, M. H. Davis, Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* **3**, 320 (2012).
6. E. Ahissar, S. S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, M. M. Merzenich, Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13367–13372 (2001).
7. K. V. Nourski, R. A. Reale, H. Oya, H. Kawasaki, C. K. Kovach, H. Chen, M. A. Howard, J. F. Brugge, Temporal envelope of time-compressed speech represented in the human auditory cortex. *J. Neurosci.* **29**, 15564–15574 (2009).
8. C. Liégeois-Chauvel, C. Lorenzi, A. Trébuchon, J. Régis, P. Chauvel, Temporal envelope processing in the human left and right auditory cortices. *Cereb. Cortex* **14**, 731–740 (2004).
9. S. J. Kayser, R. A. A. Ince, J. Gross, C. Kayser, Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J. Neurosci.* **35**, 14691–14701 (2015).
10. O. Ghitza, On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. *Front. Psychol.* **3**, 238 (2012).
11. P. Heil, H. Neubauer, Temporal integration of sound pressure determines thresholds of auditory-nerve fibers. *J. Neurosci.* **21**, 7404–7415 (2001).
12. S. Biermann, P. Heil, Parallels between timing of onset responses of single neurons in cat and of evoked magnetic fields in human auditory cortex. *J. Neurophysiol.* **84**, 2426–2439 (2000).
13. K. B. Doelling, L. H. Arnal, O. Ghitza, D. Poeppel, Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* **85**, 761–768 (2014).
14. R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech recognition with primarily temporal cues. *Science* **270**, 303–304 (1995).
15. P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, E. F. Chang, Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci.* **36**, 2014–2026 (2016).
16. M. Moerel, F. De Martino, K. Ugurbil, E. Yacoub, E. Formisano, Processing of frequency and location in human subcortical auditory structures. *Sci. Rep.* **5**, 17048 (2015).
17. A. Thwaites, J. Schlittenlacher, I. Nimmo-Smith, W. D. Marslen-Wilson, B. C. J. Moore, Tonotopic representation of loudness in the human cortex. *Hear. Res.* **344**, 244–254 (2017).

18. L. S. Hamilton, E. Edwards, E. F. Chang, A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Curr. Biol.* **28**, 1860–1871.e4 (2018).
19. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
20. J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom." *Linguistic Data Consortium* (1993).
21. N. E. Crone, D. Boatman, B. Gordon, L. Hao, Induced electrocorticographic gamma activity during auditory perception. Brazier award-winning article, 2001. *Clin. Neurophysiol.* **112**, 565–582 (2001).
22. F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, J. L. Gallant, Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network* **12**, 289–316 (2001).
23. D. Malah, Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Trans. Acoust.* **27**, 121–133 (1979).
24. M. K. Gordon, *Phonological Typology* (Oxford Univ. Press, 2016).
25. B. Kessler, R. Treiman, Syllable structure and the distribution of phonemes in English. *J. Mem. Lang.* **37**, 295–311 (1997).
26. R. Ridouane, Syllables without vowels: Phonetic and phonological evidence from Tashlhiyt Berber. *Phonology* **25**, 321–359 (2008).
27. A. Cutler, S. Butterfield, Rhythmic cues to speech segmentation: Evidence from juncture misperception. *J. Mem. Lang.* **31**, 218–236 (1992).
28. A. Cutler, D. Norris, The role of strong syllables in segmentation for lexical access. *J. Exp. Psychol.* **14**, 113–121 (1988).
29. J. J. Ohala, H. Kawasaki, Prosodic phonology and phonetics. *Phonetics* **1**, 113–127 (1984).
30. R. Treiman, The division between onsets and rimes in English syllables. *J. Mem. Lang.* **25**, 476–491 (1986).
31. D. Fogerty, D. Kewley-Port, Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *J. Acoust. Soc. Am.* **126**, 847–857 (2009).
32. F. Chen, P. C. Loizou, Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise. *J. Acoust. Soc. Am.* **131**, 4104–4113 (2012).
33. P. Heil, Auditory cortical onset responses revisited . II. Response strength. *J. Neurophysiol.* , 2642–2660 (1997).
34. F. Cummins, Oscillators and syllables: A cautionary note. *Front. Psychol.* **3**, 364 (2012).
35. J. J. Ohala, in *Auditory Analysis and Perception of Speech*, G. Fant, M. A. A. Tatham, Eds. (Academic Press, 1975), pp. 431–453.
36. I. Hertrich, S. Dietrich, J. Trouvain, A. Moos, H. Ackermann, Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology* **49**, 322–334 (2012).
37. P. Mermelstein, Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am.* **58**, 880–883 (1975).
38. S. Greenberg, H. Carvey, L. Hitchcock, S. Chang, Temporal properties of spontaneous speech—A syllable-centric perspective. *J. Phon.* **31**, 465–485 (2003).
39. K. N. Stevens, Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* **111**, 1872–1891 (2002).
40. A. Salomon, C. Y. Espy-Wilson, O. Deshmukh, Detection of speech landmarks: Use of temporal information. *J. Acoust. Soc. Am.* **115**, 1296–1305 (2004).
41. L. A. Chistovich, E. A. Ogorodnikova, Temporal processing of spectral data in vowel perception. *Speech Commun.* **1**, 45–54 (1982).
42. L. V. Lesogor, L. A. Chistovich, Detecting consonants in tones. *Fiziol. Cheloveka* **4**, 213–219 (1978).
43. R. M. Warren, C. J. Obusek, R. M. Farmer, R. P. Warren, Auditory sequence: Confusion of patterns other than speech or music. *Science* **164**, 586–587 (1969).
44. R. Treiman, A. Zukowski, Toward an understanding of English syllabification. *J. Mem. Lang.* **29**, 66–85 (1990).
45. A. Cutler, J. Mehler, D. Norris, J. Segui, The syllable's differing role in the segmentation of French and English. *J. Mem. Lang.* **25**, 385–400 (1986).
46. C. Kirtley, P. Bryant, M. MacLean, L. Bradley, Rhyme, rime, and the onset of reading. *J. Exp. Child Psychol.* **48**, 224–24(1989).
47. P. X. Joris, C. E. Schreiner, A. Rees, Neural processing of amplitude-modulated sounds. *Physiol. Rev.* **84**, 541–577 (2004).
48. B. J. Malone, B. H. Scott, M. N. Semple, Temporal codes for amplitude contrast in auditory cortex. *J. Neurosci.* **30**, 767–784 (2010).
49. P. Heil, Auditory cortical onset responses revisited. I. First-spike timing. *J. Neurophysiol.* **77**, 2616–2641 (1997).
50. H. Neubauer, P. Heil, A physiological model for the stimulus dependence of first-spike latency of auditory-nerve fibers. *Brain Res.* **1220**, 208–223 (2008).
51. A. Fishbach, I. Nelken, Y. Yeshurun, Auditory edge detection: A neural model for physiological and psychoacoustical responses to amplitude transients. *J. Neurophysiol.* **85**, 2303–2323 (2001).
52. L. S. Hamilton, D. L. Chang, M. B. Lee, E. F. Chang, Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. *Front. Neuroinform.* **11**, 62 (2017).
53. L. A. Pineda, H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, L. Villaseñor, The Corpus DIMEx100: Transcription and evaluation. *Lang. Resour. Eval.* **44**, 347–370 (2010).
54. A. Li, M. Lin, X. Chen, Y. Zu, G. Sun, W. Hua, J. Yan, *Sixth International Conference on Spoken Language Processing* (ISCA Archive, 2000).
55. P. Boersma, D. Weenink, *Praat: Doing Phonetics by Computer* (2018).
56. E. Moulines, F. Charpentier, Pitch-synchronous waveform processing technizques for text-to-soeech synthesis using diphones. *Speech Commun.* **9**, 453–467 (1991).
57. T. Schotola, On the use of demisyllables in automatic word recognition. *Speech Commun.* **3**, 63–87 (1984).
58. C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, A. A. Ghazanfar, The natural statistics of audiovisual speech. *PLoS Comput. Biol.* **5**, e1000436 (2009).
59. N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, D. Poeppel, Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81**, 181–187 (2017).
60. E. Zwicker, E. Terhardt, Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.* **68**, 1523–1525 (1980).

Acknowledgments: We thank K. Johnson, C. Schreiner, S. Nagarajan, S. Greenberg, A. Breska, L. Hamilton, J. Downer, P. Hullett, M. Leonard, B. Malone, and C. Tang for helpful comments and feedback. We also thank L. Hamilton and E. Edwards for providing some of code for analysis of the TIMIT data and other members of the Chang Lab for help with data collection. **Funding:** This work was supported by grants from the NIH (R01-DC012379 to E.F.C.) and the German Research council (OG 105/1 to Y.O.). E.F.C. is a New York Stem Cell Foundation Robertson Investigator. This research was also supported by the New York Stem Cell Foundation, the Howard Hughes Medical Institute, the McKnight Foundation, The Shurl and Kay Curci Foundation, and The William K. Bowes Foundation. **Author contributions:** Y.O. and E.F.C. conceived and designed the experiments, collected the data, and wrote the paper. E.F.C performed the surgeries and grid implantations. Y.O. analyzed the data. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the corresponding author.

Submitted 4 July 2019
Accepted 16 September 2019
Published 20 November 2019
10.1126/sciadv.aay6279

Citation: Y. Oganian, E. F. Chang, A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci. Adv.* **5**, eaay6279 (2019).

Science Advances

A speech envelope landmark for syllable encoding in human superior temporal gyrus

Yulia Oganian and Edward F. Chang

Sci Adv **5** (11), eaay6279.
DOI: 10.1126/sciadv.aay6279

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.