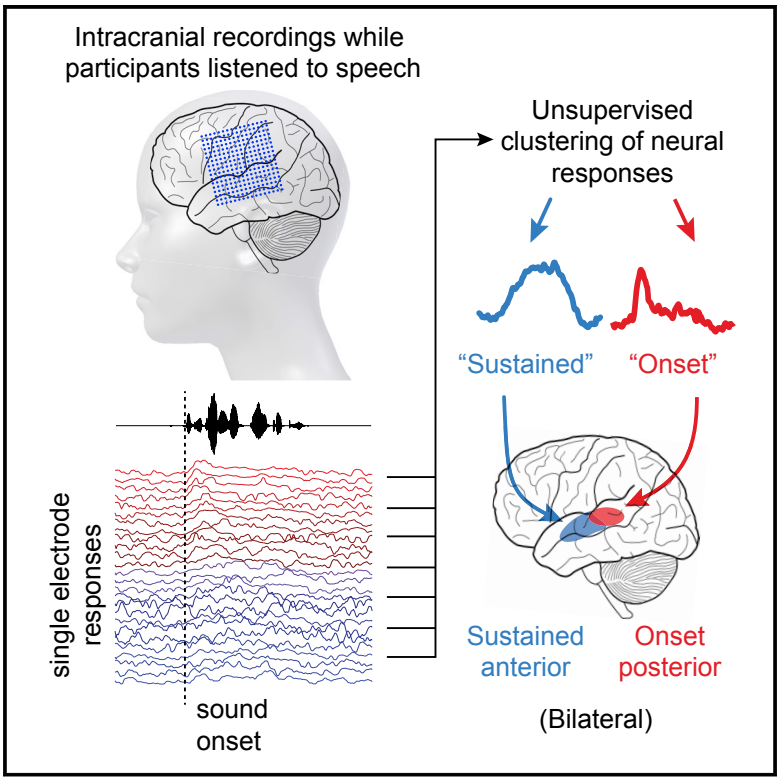
Article

Current Biology

A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus

Graphical Abstract



Authors

Liberty S. Hamilton, Erik Edwards, Edward F. Chang

Correspondence

edward.chang@ucsf.edu

In Brief

Hamilton, Edwards, and Chang use a combination of unsupervised and supervised methods on high-density intracranial recordings to reveal a spatially localized region of the posterior superior temporal gyrus that specifically parses acoustic onsets and an anterior region that exhibits sustained responses to speech.

Highlights

- Human STG is divided into two regions with onset and sustained responses to speech
- Onset selective regions are located posteriorly, and sustained are more anterior
- These response properties are the main organizing feature of the STG, not phonemes
- Onset and sustained electrodes determine sentence start and identity in a decoder

Hamilton et al., 2018, Current Biology *28*, 1860–1871
June 18, 2018
https://doi.org/10.1016/j.cub.2018.04.033

CellPress

CellPress

Current Biology
Article

A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus

Liberty S. Hamilton,[1,2,3,4] Erik Edwards,[1,4] and Edward F. Chang[1,5,*]
[1]Department of Neurological Surgery and Center for Integrative Neuroscience, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA
[2]Department of Communication Sciences and Disorders, Moody College of Communication, The University of Texas at Austin, 2504A Whitis Avenue (Stop A1100), Austin, TX 78712, USA
[3]Department of Neurology, Dell Medical School, The University of Texas at Austin, 1701 Trinity Street, Austin, TX 78705, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: edward.chang@ucsf.edu
https://doi.org/10.1016/j.cub.2018.04.033

SUMMARY

To derive meaning from speech, we must extract multiple dimensions of concurrent information from incoming speech signals. That is, equally important to processing phonetic features is the detection of acoustic cues that give structure and context to the information we hear. How the brain organizes this information is unknown. Using data-driven computational methods on high-density intracranial recordings from 27 human participants, we reveal the functional distinction of neural responses to speech in the posterior superior temporal gyrus according to either onset or sustained response profiles. Though similar response types have been observed throughout the auditory system, we found novel evidence for a major spatial parcellation in which a distinct caudal zone detects acoustic onsets and a rostral-surround zone shows sustained, relatively delayed responses to ongoing speech stimuli. While posterior onset and anterior sustained responses are used substantially during natural speech perception, they are not limited to speech stimuli and are seen even for reversed or spectrally rotated speech. Single-electrode encoding of phonetic features in each zone depended upon whether the sound occurred at sentence onset, suggesting joint encoding of phonetic features and their temporal context. Onset responses in the caudal zone could accurately decode sentence and phrase onset boundaries, providing a potentially important internal mechanism for detecting temporal landmarks in speech and other natural sounds. These findings suggest that onset and sustained responses not only define the basic spatial organization of high-order auditory cortex but also have direct implications for how speech information is parsed in the cortex.

INTRODUCTION

A fundamental goal in the neurobiology of language is to understand how acoustic information in speech is transformed into meaningful linguistic content. Speech is thought to be serially processed through the hierarchical structure of the auditory system, from acoustic to phonemic to word and higher order representations [1–5]. Accordingly, most traditional approaches have been model based, usually examining the relationship between a well-defined stimulus feature and neural activity. For example, the basic cochlear decomposition of different sound frequencies is reflected in the tonotopically organized maps found throughout the ascending auditory system, including the primary auditory cortex and adjacent areas [6–9]. In contrast, in higher order auditory cortex, including the superior temporal gyrus (STG), there is evidence for the encoding of acoustic-phonetic features [10–13]. While productive, these approaches often require *a priori* knowledge of potential acoustic (e.g., spectrotemporal) or linguistic (e.g., phoneme and syllable) features.

A major limitation of such model-based approaches is that we do not yet fully know all the potential stimulus features. Debates persist in the linguistics literature regarding the role of phonemes, syllables, and other theorized cognitive representations in the neural processing of speech [14–16]. Predicting neural responses from reduced sets of features represents a major challenge for characterizing high-order sensory cortices, where neural responses are driven more strongly by complex natural stimuli than their component features. Indeed, recent evidence suggests spectrotemporal modulation tuning to speech in the human STG, yet non-speech control stimuli designed specifically to probe modulation features did not drive strong responses [12, 13].

For these reasons, we used an unbiased, data-driven approach to discover the major patterns of variability in auditory cortex to natural continuous speech. This model-independent strategy allowed us to identify functional response types across participants without imposing assumptions about which features or dimensions of speech are most relevant or about their localization. We examined a large dataset of direct cortical recordings from human participants with high-density intracranial recordings for surgical treatment of epilepsy. Participants listened to

natural sentences while recordings were made from the STG, middle temporal gyrus, and related prefrontal and motor areas also involved in speech perception.

We first applied unsupervised non-negative matrix factorization (NMF) to neural responses from 1,906 speech-responsive electrodes across 27 participants listening to natural sentences. NMF is a dimensionality reduction method that can be used to uncover underlying statistical structure in data [17]. This method has been used in neuroscience to study object representations [18], identify brain tumors from spectroscopic imaging [19], and to solve problems in automatic speech recognition [20]. Here, we use NMF to uncover profiles of neural responses that are observed across patients listening to the same stimuli without specifying the features (acoustic, phonetic, or otherwise) assumed to be driving the response. We discovered two canonical response profiles that divided speech-responsive cortex into spatially distinct processing zones: a localized caudal zone dominated by strong responsivity to stimulus onsets and a more spatially widespread rostral zone that was not onset driven and showed generally sustained activity throughout the stimulus. The response profiles of STG electrodes were also seen in other areas of speech-responsive cortex, including prefrontal and motor areas. Segmental phonetic features were represented locally at single electrodes and were embedded equally in each zone. Together, these regions define a striking pattern of temporal dynamics that govern the auditory processing of speech.

RESULTS

Human STG Is Partitioned into Two Zones with Distinct Sentence-Level Response Profiles

Participants listened passively to 499 naturally spoken sentences from the Texas Instruments and Massachusetts Institute of Technology (TIMIT) acoustic-phonetic corpus, spoken by 402 male and female talkers. We applied an unsupervised soft clustering algorithm, convex non-negative matrix factorization (cNMF), on recordings from 1,906 electrodes across these patients, using the high gamma time series from all speech-responsive electrodes throughout the recording session (see STAR Methods). This analysis was designed to define the electrode response profiles that were similar across patients and did not rely on the identification of any acoustic or phonetic segmentation or knowledge of spatial location or anatomical area of the recordings. Our analysis showed that two dominant response profiles characterized the activity of the electrodes.

To understand the differences between response types, we began by visually inspecting the single-electrode responses to each sentence (Figure 1A; also see Figure S1). We observed a striking difference in responses: one group showed very strong responses to sentence onset, and the other appeared to have responses that were sustained or had broad peaks at various times throughout the sentence. We then examined the response across the entire population by plotting the cluster-weighted average responses to single sentences, aligned by sentence onset and sorted by length (Figure 1B). Because the cluster-weighted time series is collapsed across all electrodes within a cluster, only the overall shape of population activity is observed. This "onset" and "sustained"-like response profile is thus a general characterization of the two populations of electrodes. At the single electrode, a variety of response types were seen within onset and sustained electrodes (Figure S1). Some onset electrodes were very strongly responsive at sentence onset only, whereas others responded strongly at the onset and then would respond to other onsets within a sentence, though at a lower magnitude. Sustained electrodes usually showed even peaks throughout the sentence and did not exhibit the highly adaptive profile of the onset electrodes.

Although our unsupervised analysis was designed to uncover similarities in functional structure across brain areas and subjects, we also wanted to examine whether these functional properties were spatially localized. The onset-sensitive cluster was mainly localized to caudal or posterior STG, while the other was more spatially distributed across rostral or anterior and middle STG (Figure 1C). This spatial segregation was not a requirement of our clustering algorithm, which was performed on all subjects simultaneously without *a priori* information about electrode locations. Spatial organization of response types was clearly seen in both left (N = 14 subjects) and right (N = 13 subjects) auditory cortices across all participants (Figure 1D). We henceforth refer to these clusters as onset (for the first, more caudal cluster) and sustained (for the second, more rostral cluster). In addition to the large number of rostral sustained electrodes, we also observed a limited number of sustained-type electrodes posterior to the onset region; however, this was more variable across individuals.

We quantified functional and anatomical clustering strength using the silhouette index, which measures the degree of within-cluster and between-cluster similarity. A silhouette index near 1 indicates good clustering. Functional and anatomical clustering within onset and sustained regions was significantly higher than chance, suggesting that electrodes belonging to each group were distinct processing zones (Figure 1E; $p < 0.001$; Wilcoxon signed rank tests). Still, anatomical localization was relatively stronger in the caudal onset electrodes, whereas sustained responses were seen both anteriorly and surrounding the onset area.

These clusters represent the major source of variance within our dataset. Similar clustering was observed regardless of clustering method (for example, K-means and other factor analytic methods showed similar results; data not shown). Across all subjects, the two clusters explained 16.9% of the variance in the data. Adding more clusters explained only marginally more variance (Figure S2A). More importantly, within the additional clusters, we observed the same onset and sustained response types, mostly further subdivided according to response magnitude (Figures S2B, S2C, and S2D).

At a global level, these results suggest that distinct regions of the human STG are sensitive to important temporal cues in sentences, such as onsets and ongoing speech. However, we also know from previous work that local encoding in STG is sensitive to spectrotemporal and phonetic feature cues in speech [11–13, 21]. To connect these findings, we next asked how processing in each zone relates to other acoustic and phonetic feature representations.

Acoustic Representations in Onset and Sustained Areas

We fit spectrotemporal receptive field (STRF) models to each electrode separately to determine which combinations of

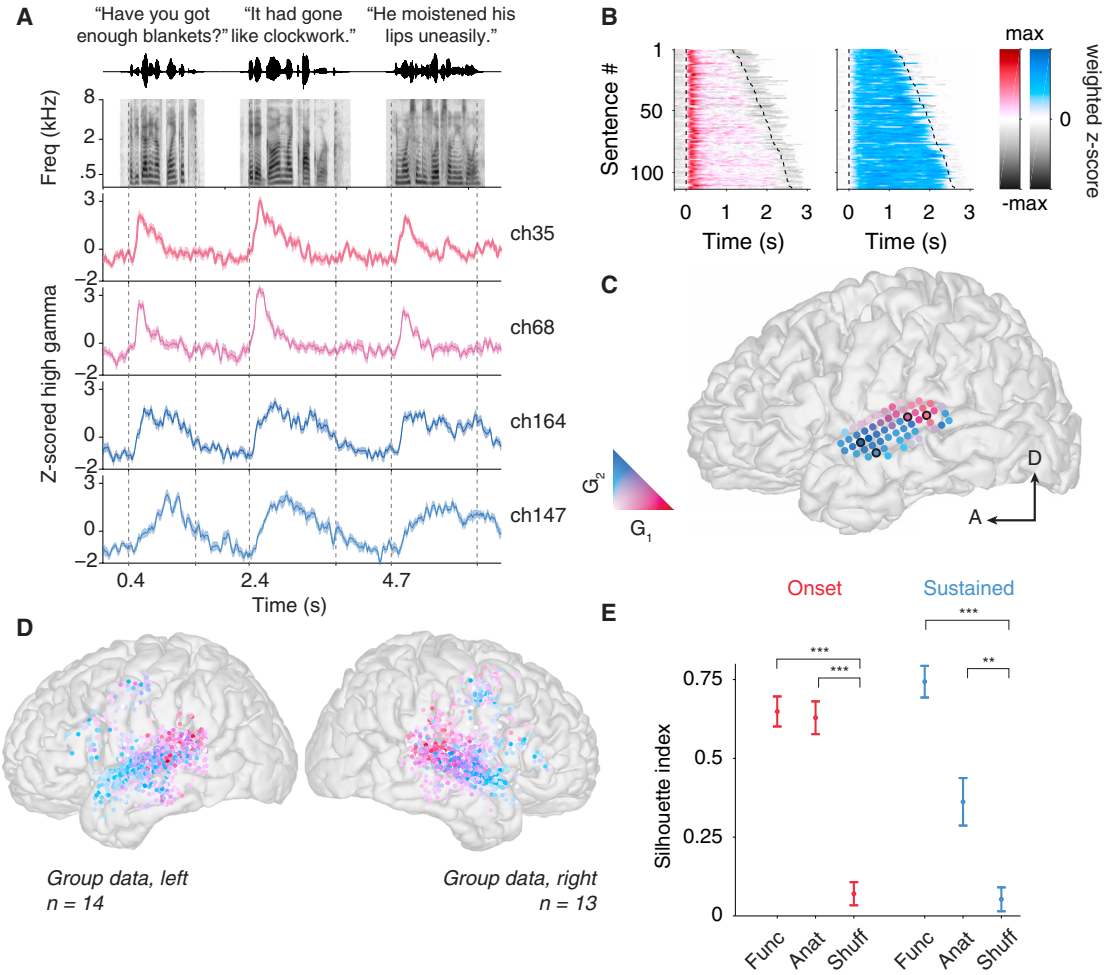


Figure 1. The Superior Temporal Gyrus and Middle Temporal Gyrus Can Be Split into Two Spatially Distinct Regions Showing Differential Temporal Responsivity to Sentences

(A) Example mean responses to sentences from single electrodes in the first (red) and second (blue) clusters, colored according to their NMF activation weights. The waveform and spectrogram for each sentence are shown at the top. Electrodes in the first cluster ("onset") showed a strong response at sentence onset, sometimes followed by lower amplitude responses to other features within the sentence. Electrodes in the second cluster ("sustained") showed differing responses throughout the sentence and were not strongly selective for onsets. Sentence onsets and offsets are marked as dashed lines. Shaded error bars indicate SEM.

(B) Average cluster time series across the onset and sustained populations reveals that the two main distinguishing features of the neural response are (1) fast responses with strong activity at sentence onset and (2) slow responses, with weak responses at sentence onset and more sustained activity throughout the sentence. Each subplot shows the responses to all overlapping sentences across all subjects projected onto the NMF bases and sorted by sentence length. Sentence onset and offset are marked by dashed lines.

(C) NMF activation weights on electrodes from one example left hemisphere subject, colored as in (A). Onset responses were observed in caudal STG close to the auditory core, whereas Sustained responses were found more rostrally. Outlined electrodes identify the electrodes plotted in (A).

(D) Activation weights for all left and right hemisphere subjects plotted on an average Montreal Neurological Institute (MNI) brain. Results from clustering on each individual subject separately are shown in Figures S4 and S5.

(E) Evaluation of functional and anatomical clustering goodness of fit using the silhouette index (mean ± SE across subjects). Functionally, both onset and sustained electrodes show tight clustering that was significantly higher than chance (onset: $p = 9.3 \times 10^{-6}$; sustained: $p = 1.5 \times 10^{-5}$; Wilcoxon signed rank test). Anatomically, onset electrodes are close to one another in space and tend to be far away from sustained electrodes, as evidenced by a high silhouette index that was significantly greater than a null shuffled distribution ($p = 1.3 \times 10^{-5}$; Wilcoxon signed rank test). Sustained electrodes are still significantly anatomically clustered ($p = 1.5 \times 10^{-3}$; Wilcoxon signed rank test), though less so than the onset electrodes.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. See also Figures S1, S2, S4, and S5.

spectrotemporal acoustic features would strongly elicit neural responses from these areas. Both onset and sustained electrodes were well described by these models (NMF weighted average: $r_{Onset} = 0.26$ and $r_{Sustained} = 0.34$). The weighted average of STRFs from onset and sustained clusters is shown in Figure 2A. Both regions exhibited variable spectral selectivity (narrow and broad tuning) and integrated sound information over relatively long timescales (up to 600 ms for excitatory and inhibitory response). However, their temporal response profiles were substantially different. Onset electrodes strongly responded to

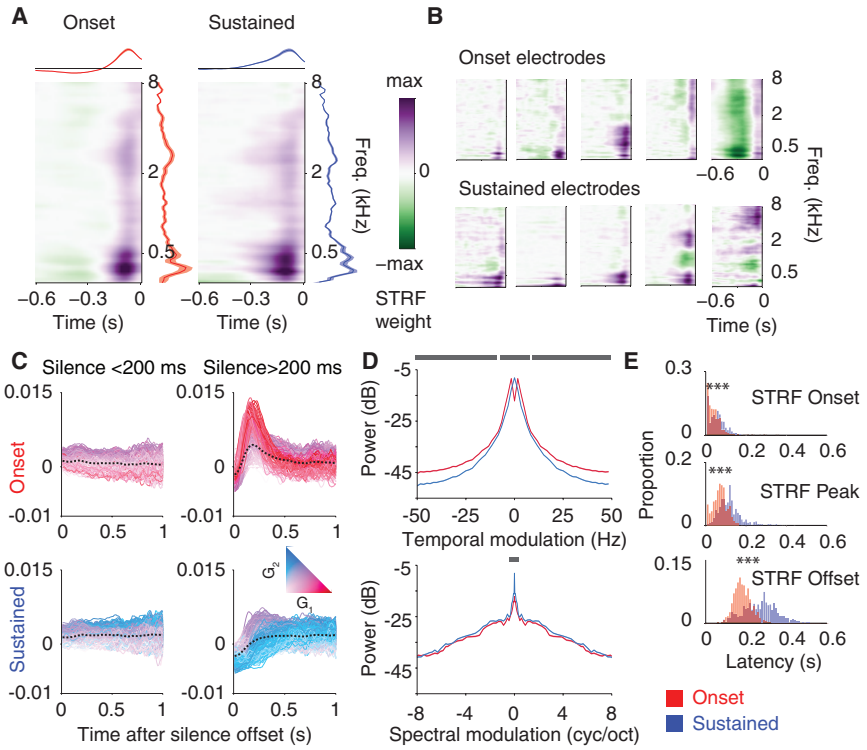


Figure 2. Onset Electrodes Detect Contrasts and Show Selectivity to Fast Temporal Modulations Present in Speech

Sustained electrodes are longer temporal integrators and are insensitive to speech onsets.

(A) Onset (left) and sustained (right) weighted average spectrotemporal receptive fields (STRFs). Onset electrodes show short latency responses and integrate over short temporal windows, whereas sustained electrode responses are slower and longer. Temporal responses collapsed across frequency features are shown on the top; spectral responses collapsed across time are shown at right (mean ± SE).

(B) Example single electrode STRFs in onset (top) and sustained (bottom) zones. Overall, onset and sustained electrodes showed spectral selectivity over similar ranges but differed in their temporal response profile.

(C) Aligned responses to speech sounds after short (<200 ms, left) and long (>200 ms, right) silences. Onset electrodes respond robustly after long silences, which can occur within a sentence or before sentence onset. Sustained electrodes respond to speech after short and long silences in a sustained manner.

(D) Average onset and sustained temporal modulation (top) and spectral modulation (bottom) power. Shaded error bars indicate SE across all electrodes in each group. Onset electrodes show higher temporal modulation selectivity, whereas sustained electrodes show marginally higher spectral modulation selectivity. Gray lines indicate significant differences at Bonferroni-corrected $p < 0.05$ (Wilcoxon rank sum test).

(E) Onset, peak, and offset response latencies calculated from onset and sustained STRFs. Latencies were significantly greater in sustained compared to onset electrodes ($p < 0.001$; Wilcoxon sign rank test), indicating longer temporal integration times for sustained electrodes.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

silence followed by sound onset, consistent with onset sensitivity observed at the sentence level. The excitatory part of the STRF was short in duration, while preceded by a relatively long inhibitory period. In contrast, sustained electrodes had a long-duration excitatory component only. Single electrodes in onset and sustained regions both show selectivity for low-, mid-, and high-frequency content (Figure 2B). Some onset electrodes exhibited broadband, non-selective onset responses, whereas sustained electrodes showed more complex spectral tuning, such as adjacent excitatory and inhibitory sidebands.

We were interested to know whether similar responses were found within sentences after naturally occurring silent pauses between phrases or for general onsets within sentences. We plotted time-aligned onset and sustained responses to speech following short (<200 ms) or long (>200 ms) silences (Figure 2C). Strong onset responses occurred only after longer silences, consistent with the STRFs described above. Longer silences usually appeared at phrase boundaries in natural speech, and these findings suggest a similar response profile that strongly encodes these temporal landmarks in speech.

While low-level auditory areas perform a time-frequency decomposition of incoming sounds, higher auditory areas are often sensitive to more complex combinations of features, including joint temporal and spectral amplitude changes or modulations [22, 23]. Speech comprehension relies on encoding a relatively narrow set of spectral and temporal modulations within the spectra of all natural sounds [24, 25]. We thus measured selectivity to joint spectrotemporal modulations using the modulation transfer function (MTF), which describes whether these electrodes follow changes in spectral content, temporal content, or both [25]. In agreement with previous work [12, 13, 21, 26], we found higher temporal and lower spectral modulation selectivity in caudal onset electrodes, whereas rostral sustained electrodes showed low temporal/high spectral modulation selectivity (Figure 2D). Onset electrodes had higher temporal modulation selectivity across the whole range of possible temporal modulations (Bonferroni corrected $p < 0.05$; Wilcoxon rank sum test), whereas sustained electrodes had higher spectral modulation selectivity around core spectral modulations (between −0.22 cycles (cyc)/octave (oct) and 0.22 cyc/oct; gray bars indicate modulations for which Bonferroni corrected $p < 0.05$; Wilcoxon rank sum test).

The temporal integration profiles of brain areas can provide insights as to the type of information that is being encoded. To quantify the difference in latencies and temporal integration profile in onset and sustained electrodes, we calculated the onset, peak, and offset latencies of the excitatory component in each STRF. Onset electrodes exhibited earlier onset, peak, and offset latencies compared to sustained electrodes ($p < 0.001$; Wilcoxon rank-sum test; Figure 2E), but the difference was most pronounced at offset, where the difference in average offset latencies was 98 ms. The duration of the excitatory response

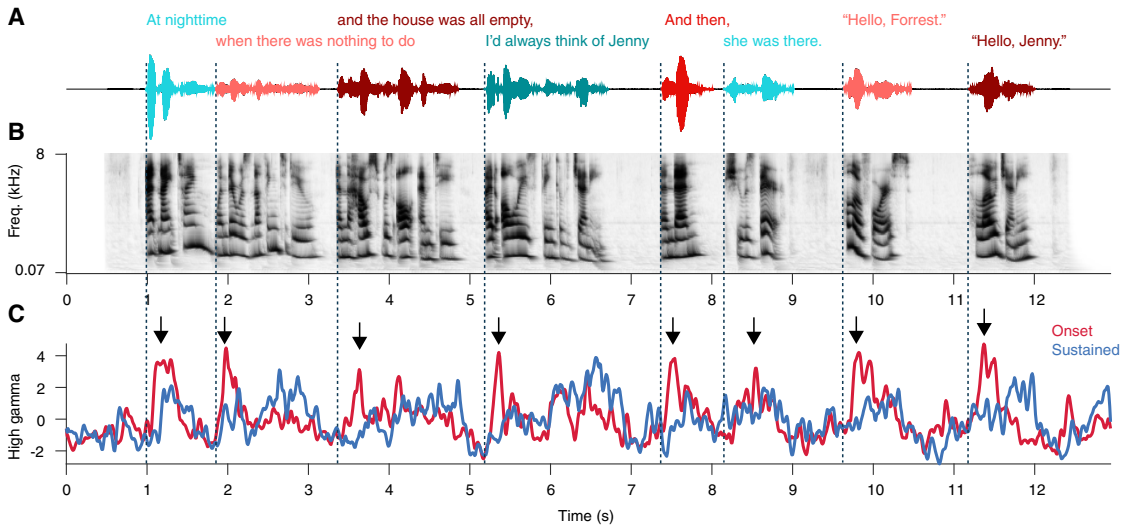


Figure 3. The Caudal Onset Zone Identifies Onsets in Natural, Continuous Speech

(A) Sound waveform and transcription for an example stimulus from a reading of the movie script for *Forrest Gump*. The stimulus contains multiple sentences and phrases, with natural pauses in between utterances rather than defined silences as used for the Texas Instruments and Massachusetts Institute of Technology (TIMIT) sentence-listening task. Parts of the sentence are shown in color for ease of alignment with the sentence waveforms.

(B) Spectrogram of the example stimulus shown in (A).

(C) Example activity for an onset (red) and a sustained electrode (blue) from SL04. The onset electrode activity demarcates acoustic cues for sentence and phrase-level pauses, such as after "at nighttime" or after "all empty". Longer pauses, such as those in between the dialog, also elicit strong onset responses from this electrode.

See also Figure S3.

was 138 ± 5 ms in onset and 211 ± 9 ms in sustained (mean ± SD). These temporal properties are consistent with previous findings [27–29] and have been interpreted as evidence of serial processing within the "ventral stream," with the idea that caudal/posterior areas are lower order. However, the results here suggest that the representations are fundamentally different between the zones. First, many onset electrodes are non-selective in the spectral domain, whereas sustained electrodes respond to spectrally complex sounds. Second, if the onset zone were low level, it would be activated throughout the sentence, not primarily driven by the onset. Third, and most importantly, how the two zones integrate sound information over time is completely different.

The evidence for onset-like and sustained-like activity was found using responses to isolated TIMIT sentences, which are arguably not as natural as continuous speech in, for example, a narrative context. To determine whether this result was specific to TIMIT sentences, which were separated in time by 400-ms pauses, we also looked at responses to continuous speech taken from a reading of the movie *Forrest Gump* (see STAR Methods). As with TIMIT, we found obvious onset and sustained activity in the same electrodes (Figure 3). Onset electrodes responded at the onsets of sentences and phrases and after short pauses. Sustained electrodes were found to respond throughout these utterances, as before.

Onset and sustained responses were observed even when stimuli were played backward or were spectrally rotated to remove phonetic and lexical content (Figures S3A and S3B). Pure tone stimuli could also drive responses in onset, but not sustained, electrodes (Figure S3C). Thus, although our STRF analysis replicates our findings of high temporal modulation representation in pSTG, the most parsimonious explanation of these data appears to be that onset regions are onset selective rather than simply selective for high temporal modulations. This also explains our previous results whereby modulated ripple noises did not elicit strong activity in STG beyond an initial onset response [12].

Our acoustic analysis indicated that caudal onset electrodes are onset detectors with varied spectral selectivity and short temporal integration profiles and relatively high temporal modulation selectivity, whereas rostral sustained electrodes are long temporal integrators that are not sensitive to onsets and encode spectral modulations important for speech comprehension [24]. Next, we wanted to examine whether onset responses were specific to sentence onset and how these acoustic properties related to phoneme feature representations in each zone.

Local Phoneme Feature Embedding in Onset and Sustained Zones

Phonetic features, such as plosive, nasal, and fricative, describe how the sounds that define different categories of phonemes are produced by the vocal tract [30]. Previously, we showed that the human STG exhibited selectivity for phonetic features [11]; however, no consistent spatial map for these features was found across subjects.

We speculated that plosives characterized by silence followed by broadband burst (e.g., /ba/, /pa/, and /ta/) might be selectively processed in the onset zone because of its short temporal integration time, broadband spectral selectivity, and selectivity for high temporal modulations. Conversely, we predicted that

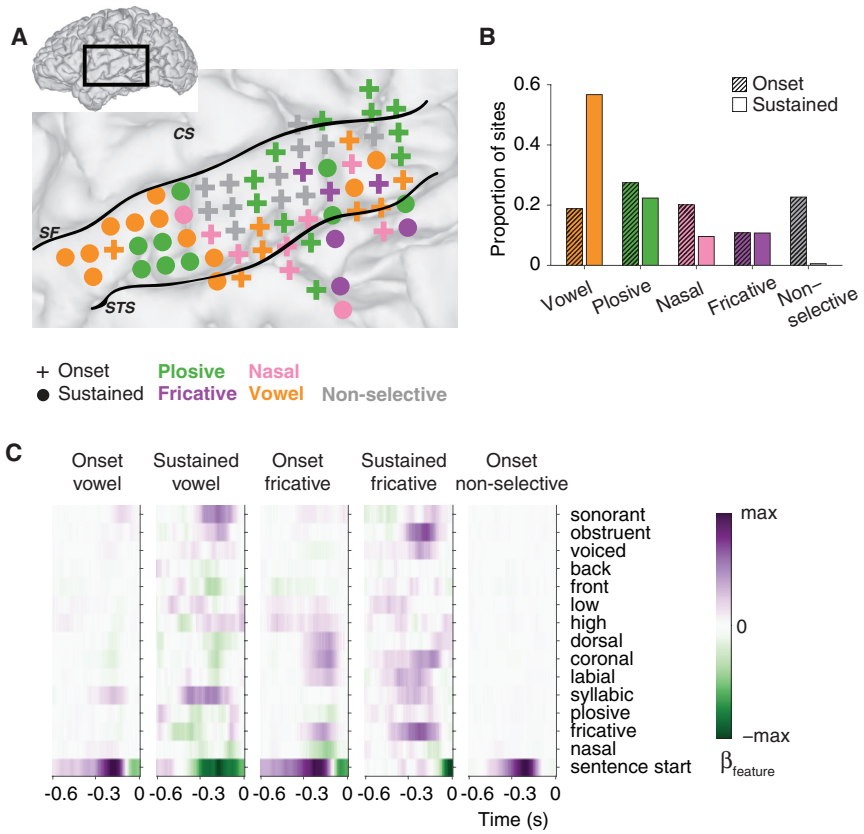


Figure 4. Onset and Sustained Electrodes Exhibit Overlapping Acoustic-Phonetic Feature Selectivity

(A) Example feature selectivity maps for one subject (SL04). Electrodes are colored according to their acoustic-phonetic feature selectivity (non-selective, nasal, vowel, plosive, or fricative). Onset electrodes are denoted by the + symbol and sustained electrodes by a circle. CS, central sulcus; SF, Sylvian fissure; STS, superior temporal sulcus.

(B) The distribution of phonetic class selectivity is similar across onset and sustained zones ($p = 0.22$; chi-square test). The proportion of sites exhibiting selectivity for each of the five phonetic features is shown for onset (dark bars, left) and sustained (light bars, right) sites.

(C) Example feature model weights for electrodes with similar phonetic feature selectivity in onset or sustained zones. Examples of vowel selectivity and fricative selectivity are shown for onset and sustained electrodes, as well as an onset non-phonetically selective electrode that responded primarily at sentence start.

(D) Spatial map for non-selective onset (+) or non-selective sustained (●) sites. Only non-selective (mostly onset) responses were localized to a particular area of speech cortex—most of these responses were in posterior STG.

sustained electrodes would be sensitive to spectral modulation content in vowels. We fit a linear model to predict electrode activity, this time employing a reduced binary feature matrix to represent the presence or absence of phonetic features (nasals, fricatives, plosives, high or front or low or back vowels, etc.) in the sentence stimuli (see STAR Methods for details). We also included a feature for sentence start in order to model the non-linearities present in onset electrodes.

Contrary to our expectations, we observed overlapping phonetic feature representation in onset and sustained electrodes—that is, we did not find segregation of consonants and vowels in these areas. Instead, we found evidence of single-electrode selectivity for all phonetic feature classes, vowels, nasals, fricatives, and plosives, in both the onset and sustained zones. Figure 4A illustrates this diversity of phonetic feature representation, with onset electrodes demarcated by + and sustained electrodes by circles. Both onset and sustained show selectivity for vowels, plosives, fricatives, and nasals. The proportion of sites tuned to a particular phonetic feature did not significantly differ across onset and sustained zones ($p = 0.22$; chi-square test; Figure 4B). Examples of the feature-temporal receptive fields are shown in Figure 4C. We saw evidence of onset electrodes that jointly encoded sentence start and particular phonetic features, such as vowels or fricatives. Sustained electrodes encoded the same features without an enhanced response at sentence start.

The pronounced differentiation of sentence onset feature representation for onset electrodes was strongly spatially localized (Figure 4D). Electrodes with a high response at sentence start were located in the posterior STG, in a distinct region largely overlapping with our previously defined onset zone. Mean onset weights were strongly positively correlated with onset cluster 1 NMF weight (Spearman rho = 0.87; $p < 0.001$) and strongly negatively correlated with sustained cluster 2 NMF weights (Spearman rho = −0.33; $p < 0.001$). We found no evidence for a reliable or consistent spatial map of phonetic feature selectivity (data not shown).

Altogether, these results demonstrate that phonetic feature encoding is differentiated at the local scale of individual electrodes, whereas temporal parameters (onset versus sustained) are a global-scale organizational property that partitions the STG. One important caveat is that we are unable to model some neural changes that may result from natural co-articulations, reductions, or elisions in speech, although by incorporating temporal delays into our feature-encoding models, we can control for some temporal correlations within the stimuli.

Because these feature models are based on average responses to many instantiations of the same phonemes, this may reduce our ability to infer the responses to more variable pronunciations.

Responses to passive speech were also observed in the sensorimotor cortex and inferior frontal gyrus, as shown by our group [31, 32] and others [33–37]. In these electrodes, we again found a separation of responses into onset and sustained response types (see Figure S1C; SL04-ii and SL06-iv). Although the overall magnitude of the cluster weights in this area was lower than in the classical auditory areas, the response profile for single electrodes within these regions was similar, as was the overall structure of the STRFs (data not shown). Because of their ability to be predicted by the spectrotemporal model and their strong responses during listening, these responses are likely auditory in nature, despite appearing outside of classical auditory sensory cortex. Onset and sustained distinctions thus apply not only to the STG but also appear to be a fundamental organizing response feature across the entire auditory-responsive speech cortex. The distribution of onset and sustained response types is shown for all speech-responsive electrodes for each subject separately in Figures S4 and S5. In most single subjects, a distinct onset zone was localized to the posterior STG, and while the sustained zone was usually observed anteriorly, with some types of coverage, we also observed multiple sustained zones anteriorly and posteriorly.

Decoding Sentence Onsets and Identity from Neural Activity

Our encoding analyses showed that onset and sustained electrodes show overlapping acoustic-phonetic feature selectivity but that only onset electrodes show enhanced responses at sentence and phrase onsets. We next investigated whether activity in onset electrodes could provide an internal temporal reference point for speech analysis. In speech, the onset is a critical feature to initiating computation of the following acoustic and linguistic information. To test this, we first used the neural activity projected into onset or sustained electrodes to determine the accuracy of detecting sentence onsets during single trials (see STAR Methods). Critically, these neurally detected onsets did not rely on any outside knowledge of the stimulus transcription or acoustics. Examples of detected onsets for single trials for four sentences are shown in Figure 5A, where onsets detected from onset electrode activity are shown in the middle panel and onsets detected from sustained electrodes are shown as arrows in the bottom panel. The spectrograms for these four sentences are shown in the top panel. The dashed black lines represent the true onset times from the stimulus transcription. Given the natural lag between stimulus presentation and a subsequent neural response, neurally detected onsets usually occurred at a delay compared to the actual stimulus onset. Critically, although onset and sustained populations were defined using responses to TIMIT sentences, using these populations to detect onsets in a separately recorded naturally spoken narrative (from *Forrest Gump*, as above) was extremely similar, with onset populations able to precisely detect onsets of sentences and phrases (Figure S6A). We calculated the accuracy of onset detection from onset and sustained populations across all participants, where correct detection was specified as an onset being detected within 0–150 ms of the actual stimulus onset. This window was chosen according to the average STRF peak excitatory response across all sites (onset and sustained). The accuracy of onset detection was significantly higher for the onset population (for 150 ms window: mean ± SE: 61.8% ± 4.8%; max accuracy 98%) compared to the sustained population (mean ± SE: 5.1% ± 1.3%; max accuracy 17%; $p = 4.4 \times 10^{-4}$; $Z_{15} = 3.52$; Wilcoxon signed rank test; Figure 5B). The error between neurally detected onsets and the actual stimulus onset is shown in Figure 5C, with the +150 ms boundary for "accurate" detections marked as a dashed line. Because one could argue that the sustained electrodes might still be able to detect sentence onsets but at a longer delay relative to the onset electrodes, we also repeated this analysis with windows up to 600 ms. Even with this window—arguably too long for a reliable onset detector—results were similar, with onset electrodes always predicting onsets significantly better than sustained (Figure S6B). Sustained electrode onset estimates were highly variable (Figure 5A), owing to the relative prominence of acoustic-phonetic feature and other selectivity rather than strong responses at sentence onset.

Having demonstrated that onset electrode responses are a highly reliable marker for the start of each sentence, we next wanted to show the implication of such a temporal "reference frame" [38] for decoding the subsequent sentence information. We performed a template-matching classification analysis [39] after aligning single-trial neural responses to either (1) onsets detected by onset electrodes, (2) onsets detected by sustained electrodes, or (3) the actual stimulus onset (Figure 5D). This analysis yielded two interesting findings: first, that onset electrodes can be used to align single-trial responses for decoding almost as well as using the actual stimulus onset and, second, that sustained populations, when appropriately aligned, provide higher sentence classification accuracy when used in the template decoder than onset electrodes and perform similarly to using all electrodes simultaneously (Figure 5E). When onset electrodes are used in the classifier with the true stimulus alignment, only the beginning of the sentence can be classified and then performance worsens to chance (middle panel in Figure 5E). These results were similar, though with poorer accuracy, when using suprasylvian (inferior frontal or vSMC) electrodes for alignment and classification (data not shown). This suggests that the strong onset-selectivity observed in the onset region can provide a marker of when speech starts, whereas the strong feature selectivity and even response of sustained electrodes throughout each stimulus allows for identification of which sentence was heard. Such interactions between onset detectors and ongoing spectral analysis may also be critical for parsing auditory scenes in single- and multi-speaker environments [40, 41].

Encoding of Temporal Landmarks in Speech Dynamics

We identified distinct pathways of the speech-responsive cortex that appear to respond differentially to onset and non-onset components of speech and that integrate over short and long timescales. These observations suggest that the auditory system is highly sensitive to the temporal dynamics intrinsic to the overall structure of phrases and sentences. To visualize how onset and sustained populations contribute to the temporal dynamics of natural speech processing, we performed a cortical state-space analysis [42–45], in which responses to sentences

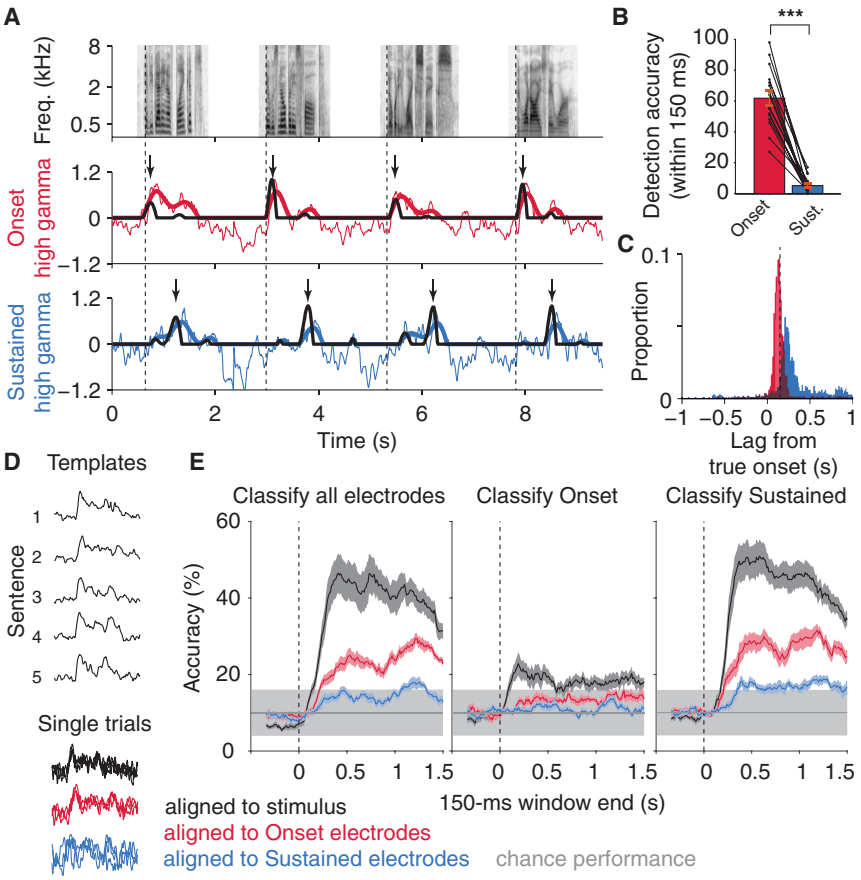


Figure 5. Detection and Alignment of Ongoing Speech Information by the Caudal Onset Zone

(A) Four example sentence spectrograms are shown at the top along with onsets detected from neural activity in onset (middle panel) and sustained (bottom panel) zones in subject SL04. In both the middle and bottom panels, the NMF-projected high gamma data are shown as the thin red (onset) or blue (sustained) lines. This represents the population activity across all onset or sustained electrodes in this patient. The bold red and blue lines show the 2-Hz low-pass-filtered and half-wave-rectified signal. The derivative of this low-pass-filtered signal is then taken and used for onset detection (black line). Detected onsets are indicated with black arrows. The dashed line indicates the true stimulus onset. While onsets detected from the caudal zone were very close to the actual stimulus onset, onsets detected by the rostral sustained zone occurred at random time points during the sentence.

(B) Accuracy of detection from activity in onset electrodes and sustained electrodes. The onset zone showed significantly higher accuracy compared to the sustained zone (Wilcoxon signed rank test). Bars show mean ± SE (in light gray) across N = 16 subjects with at least 10 repeats per sentence (see STAR Methods). Single points represent accuracy for a single subject.

(C) Difference between stimulus onset and neurally detected onsets in onset (red) and sustained (blue) electrodes. Most onset zone-detected onsets fell within <150 ms of the actual stimulus onset, as shown by dashed line.

(D) Schematic of classification of single trials matched to "template" responses calculated from the average across all repetitions of a stimulus. Templates are shown as the average response of one electrode to 5 example sentences. For simplicity, only one electrode's activity is shown, although classification was performed on the single-trial population responses (see STAR Methods). Single trials were taken from data aligned to the stimulus (black, representing a true alignment), onset zone electrodes (red), or sustained zone electrodes (blue). Alignments using sustained zone-detected onsets were poor, as observed by the lack of a consistent response across trials. These single trials were matched to the templates using the lowest Euclidean distance metric.

(E) Classification of sentence responses aligned to stimulus, onset activity, or sustained activity. Input to the classifier included activity from all electrodes (left), onset electrodes (center), or sustained electrodes (right) aligned to onsets detected from the stimulus (black), onset (red), or sustained (blue). Classifiers were calculated over a fixed 150-ms window starting at −0.5 s pre-onset and ending at up to 1.5 s post-onset. In all cases, alignment to the stimulus resulted in the highest classifier performance. Using sustained electrode activity (aligned to onsets detected from the caudal onset zone) in the classifier resulted in higher performance than using onset activity. Using sustained electrode activity to classify the stimulus showed similar performance to including all possible electrodes. Chance classification is shown in gray. Shaded error bars indicate mean ± SE.

*p < 0.05, **p < 0.01, ***p < 0.001. See also Figure S6.

were projected onto NMF components. We first examined the relationship between onset and sustained responses across all anatomical sites (Figure 6A; Video S1). Immediately following sentence onset, the state space trajectory was dominated by onset electrode responses and then moved toward sustained response types. For single sentences, the trajectories through this state space show a stereotyped response profile. The sentence "How on earth do you manage it?" shows a sweep into the onset caudal zone, followed by coactivation of sustained electrodes (Figure 6B). A natural sentence containing a substantial pause in the middle ("Then he—then what?") results in two rotations through this state space ("then he" in yellow hues; "then what?" in yellow and orange hues). Notably, when marking the position of these features (for example, vowels, fricatives, or plosives) within this state space, these features did not occupy a defined region of the global dynamics represented here but rather were distributed throughout the sentence trajectory (Figures 6C–6E).

DISCUSSION

Unsupervised clustering of human cortical responses to speech identified a distinct onset region that was remarkably consistent across 27 participants. This onset region represents both low-level features found throughout the auditory system, such as onsets, but also higher level features, including acoustic-phonetic features that are shared with the sustained region. We also found that onset responses in onset electrodes could be used for decoding and alignment of ongoing speech signals and approached the accuracy of the true stimulus alignment (Figure 5).

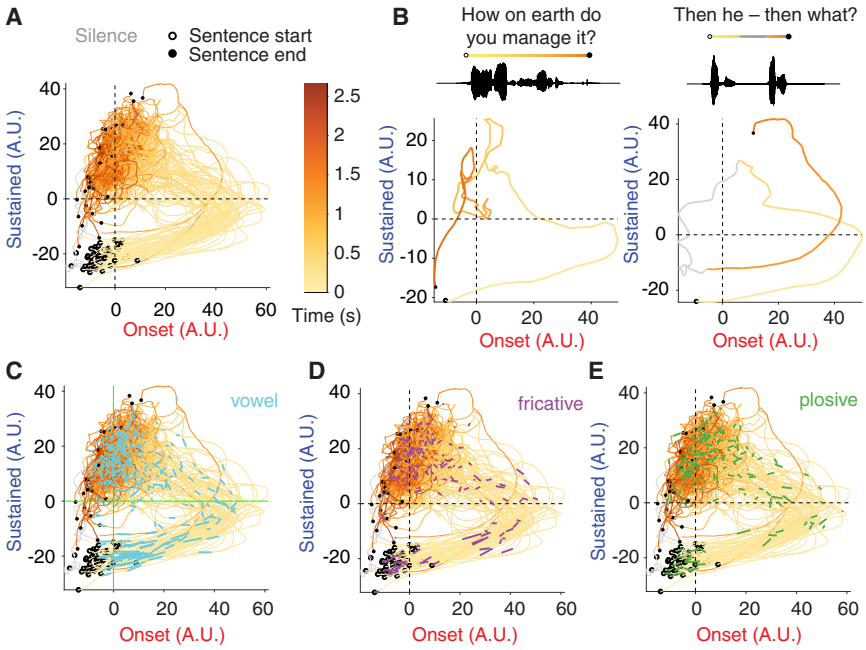


Figure 6. State Space Trajectories for Onset and Sustained Electrodes Encode Temporal Landmarks in Speech

(A) State-space representation for all electrodes (n = 1,906) projected onto onset and sustained population activity during single sentences. Each trajectory within the space represents one sentence. Each sentence starts and ends with silence (marked in gray). The open circle indicates speech onset and is followed by a colored line that progresses from light yellow to orange as time progresses in the sentence. Sentence offset is marked as the filled circle. Across all sentences, neural activity follows a stereotyped sweep through the caudal onset zone and then coactivating the sustained zone.

(B) Single-sentence trajectories for a sentence without pauses (left) and a sentence broken by a brief pause (right). In the case of a pause, neural activity traverses the space twice—once for each segment of the sentence.

(C–E) Responses to phonetic features are widely distributed within global state space trajectories. State space trajectories are marked for the presence of vowels (C), fricatives (D), and plosives (E), which, unlike sentence and phrase onsets, occupy multiple regions of the cortical state space.

See also Video S1.

Additionally, neural activity in onset electrodes may be used to temporally align information processed in sustained electrodes. Together, our findings demonstrate a potential neural code for demarcating the timing or position of important events in natural speech, thereby providing contextual information to the segmental acoustic-phonetic feature processing locally within each large region.

While onset and sustained properties are not speech specific, and in fact occur in response to reversed and spectrally rotated speech (Figure S3), these response types clearly have relevance for segmenting acoustic boundaries important for parsing sentences and phrases in natural speech. With our current stimulus set, it is not possible to completely decouple acoustic and phonetic selectivity, because they are tightly correlated. Phrase boundaries in our stimulus set may likely covary with the particular amplitude profile to which the onset electrodes respond, which may not generalize to languages with other markers of phrase boundaries. However, the frequency specificity of some onset electrodes (see Figure 4) also points to a mechanism for detecting onsets within sentences and phrases, which could be important for speech segmentation in multi-talker and noisy environments [46, 47]. The sustained responses seen for incomprehensible spectrally rotated and reversed speech stimuli indicate that these sites may respond to fluctuations in temporal envelope, which has also been seen for high rates of envelope fluctuations in primary auditory cortex [48] and is consistent with reports of envelope following, even in the absence of linguistic content [49], but may be enhanced during comprehension [50].

Our work uncovered two major response types with onset and sustained responsivity to sentences. Similar "phasic" and "tonic" response types have been observed in single-unit electrophysiological recordings throughout the auditory system, including the temporal lobe auditory cortex [23, 51, 52], auditory brain stem [53, 54], and even prefrontal cortex, where they may be involved in decision making and object identification within the ventral stream [2, 37, 51, 55]. Caudal onset responses are likely related to the strongly adapting onset responses also seen in animal models, while sustained responses are non-adapting and more linear in the sense that they are not as sensitive to temporal context [56]. Previous studies, however, have not clearly documented the spatial segregation of these response types in auditory cortex. Thus, it was a surprise to discover that such response distinctions can be consistently regionalized at a macro-anatomical level. To our knowledge, only limited fMRI evidence has suggested transient and sustained cortical responses in humans [57–59], although those responses were over much longer timescales (seconds and minutes), largely confined to the temporal plane, and in the context of synthetic stimuli or scanner noise rather than natural speech.

Many previous studies, including our own, may have overlooked these properties in search of more canonical acoustic and speech features, such as phonemes and syllables, in the human STG. Here, the data-driven approach combined with large-scale coverage, dense sampling, and real-time electrocorticography (ECoG) recordings contributed directly to the novel functional clustering observed here. Despite the limited previous demonstration of functional organization, substantial anatomical evidence exists for differentiating the caudal and rostral auditory cortex [1, 9, 51, 60–62]. Anatomical studies show parallel topographic projections from the inferior colliculus to the medial geniculate body [63]. These streams may be part of the "where" (caudal) and "what" (rostral) pathways described in non-human primates [2, 51, 61, 64]. As for the "dorsal" versus ventral stream

model [65], we believe this to be a separate distinction. In our data, the distinct onset versus sustained responders were found within the ventral stream regions of STG and MTG. Dorsal stream sites (e.g., motor cortex and supramarginal gyrus) can also be classified as onset or sustained types but typically less strongly than in ventral regions. The anatomical distinctions between onset and sustained electrodes are primarily rostral-caudal rather than dorsal versus ventral. While onset selectivity defines the posterior zone, we also observe spectrotemporal and acoustic-phonetic selectivity in this and the sustained zone, indicating that both areas may be a part of the ventral "what" stream. In addition, cortical stimulation experiments have shown that disrupting these areas likely disrupts perceptual processing [66–68].

Our results build upon previous work showing that STG is organized by its modulation sensitivity, with high temporal modulation selectivity and high temporal precision posteriorly and high spectral modulation selectivity with low temporal precision anteriorly [12, 13, 21]. However, these results go beyond temporal modulation selectivity in that they demonstrate a specific sensitivity to onsets. Onsets by nature have high temporal modulation content, but not all high temporal modulation sounds are onsets. For example, reversed speech induces an onset response at what was previously the offset of the sentence (Figure S3). Also, the dynamics of responses shown in the state space trajectories (Figure 6) would not be predictable from temporal modulations alone. This inherent asymmetry in responses reflects a critical difference and clarification from previous findings. While this property does not appear to be specific to speech sounds, it clearly has a major influence on the neural processing of speech.

Our results suggest that this parcellation is a fundamental aspect of auditory cortex organization and is likely not specific to speech processing [69]. Nevertheless, these basic response properties have potential implications for detecting the timing of linguistically important events. Sensitivity to onsets can play a critical role in parsing sentence and phrase boundaries using acoustic cues, in combination with other high-level syntactic cues [70]. The caudal onset detectors may be critical for auditory scene analysis [71], because spectral energy from single sources is generally temporally coherent [47, 72]. We observed very few responses to offsets compared to onsets, in accordance with neurophysiological and behavioral evidence that sound onsets are given greater perceptual weight [73].

We describe a major division of the auditory cortex that supports multiple levels of spectrotemporal, phonological, and linguistic representations. This defining property of auditory cortical organization suggests how neural populations combine dynamically to support speech perception and likely reflects a general mechanism for processing natural sounds. While these findings demonstrate the extraction of multiple dimensions of acoustic-phonetic and temporal cues in speech, a major challenge is to understand how such information is fully integrated at higher cortical levels to support language comprehension.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, one table, and one video and can be found with this article online at https://doi.org/10.1016/j.cub.2018.04.033.

A video abstract is available at https://doi.org/10.1016/j.cub.2018.04.033#mmc4.

ACKNOWLEDGMENTS

The authors would like to thank Christoph Schreiner, Keith Johnson, Michael Stryker, Brian Malone, Neal Fox, Yulia Oganian, and Matthew Leonard for helpful comments on the manuscript. This work was supported by grants from the NIH (F32 DC014192-01 Ruth L. Kirschstein postdoctoral fellowship from the National Institute on Deafness and Other Communication Disorders to L.S.H. and DP2-OD00862 and R01-DC012379 to E.F.C.). 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 McKnight Foundation, The Shurl and Kay Curci Foundation, and The William K. Bowes Foundation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

AUTHOR CONTRIBUTIONS

L.S.H., E.E., and E.F.C. conceived of and designed the experiment. L.S.H., E.F.C., and others collected the data. L.S.H. and E.E. analyzed the data. E.C. performed surgery and grid implantation. L.S.H., E.E., and E.F.C. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 14, 2017
Revised: March 4, 2018
Accepted: April 10, 2018
Published: May 31, 2018

REFERENCES

1. Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. *12*, 718–724.
2. Bizley, J.K., and Cohen, Y.E. (2013). The what, where and how of auditory-object perception. Nat. Rev. Neurosci. *14*, 693–707.
3. Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., and Rauschecker, J.P. (2001). Hierarchical organization of the human auditory

cortex revealed by functional magnetic resonance imaging. J. Cogn. Neurosci. *13*, 1–7.

4. Leaver, A.M., and Rauschecker, J.P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. *30*, 7604–7612.
5. DeWitt, I., and Rauschecker, J.P. (2012). Phoneme and word recognition in the auditory ventral stream. Proc. Natl. Acad. Sci. USA *109*, E505–E514.
6. Saenz, M., and Langers, D.R.M. (2014). Tonotopic mapping of human auditory cortex. Hear. Res. *307*, 42–52.
7. Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. J. Neurosci. *32*, 14205–14216.
8. Da Costa, S., van der Zwaag, W., Miller, L.M., Clarke, S., and Saenz, M. (2013). Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. J. Neurosci. *33*, 1858–1863.
9. Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. USA *97*, 11793–11799.
10. Howard, M.A., Volkov, I.O., Mirsky, R., Garell, P.C., Noh, M.D., Granner, M., Damasio, H., Steinschneider, M., Reale, R.A., Hind, J.E., and Brugge, J.F. (2000). Auditory cortex on the human posterior superior temporal gyrus. J. Comp. Neurol. *416*, 79–92.
11. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.
12. Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., and Chang, E.F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J. Neurosci. *36*, 2014–2026.
13. Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. USA *106*, 14611–14616.
14. Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. Lang. Cogn. Process. *29*, 2–20.
15. Sussman, H.M. (1984). A neuronal model for syllable representation. Brain Lang. *22*, 167–177.
16. Nearey, T.M. (2001). Phoneme-like units and speech perception. Lang. Cogn. Process. *16*, 673–681.
17. Ding, C., Li, T., and Jordan, M.I. (2010). Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. *32*, 45–55.
18. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.
19. Ortega-Martorell, S., Lisboa, P.J.G., Vellido, A., Simões, R.V., Pumarola, M., Julià-Sapé, M., and Arús, C. (2012). Convex non-negative matrix factorization for brain tumor delimitation from MRSI data. PLoS ONE *7*, e47824.
20. Bertrand, A., Demuynck, K., Stouten, V., and Van hamme, H. (2008). Unsupervised learning of auditory filter banks using non-negative matrix factorisation. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 4713–4716.
21. Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. *10*, e1003412.
22. Shamma, S. (2001). On the role of space and time in auditory processing. Trends Cogn. Sci. *5*, 340–348.
23. Eggermont, J.J. (2001). Between sound and perception: reviewing the search for a neural code. Hear. Res. *157*, 1–42.
24. Elliott, T.M., and Theunissen, F.E. (2009). The modulation transfer function for speech intelligibility. PLoS Comput. Biol. *5*, e1000302.
25. Singh, N.C., and Theunissen, F.E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am. *114*, 3394–3411.
26. Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., and Chang, E.F. (2012). Reconstructing speech from human auditory cortex. PLoS Biol *10*, e1001251.
27. Lerner, Y., Honey, C.J., Silbert, L.J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. *31*, 2906–2915.
28. Honey, C.J., Thesen, T., Donner, T.H., Silbert, L.J., Carlson, C.E., Devinsky, O., Doyle, W.K., Rubin, N., Heeger, D.J., and Hasson, U. (2012). Slow cortical dynamics and the accumulation of information over long timescales. Neuron *76*, 423–434.
29. Nourski, K.V., Steinschneider, M., McMurray, B., Kovach, C.K., Oya, H., Kawasaki, H., and Howard, M.A., 3rd. (2014). Functional organization of human auditory cortex: investigation of response latencies through direct recordings. Neuroimage *101*, 598–609.
30. Ladefoged, P., and Johnson, K. (2011). A Course in Phonetics, Sixth Edition (Boston, MA: Wadsworth, Cengage Learning).
31. Cheung, C., Hamilton, L.S., Johnson, K., and Chang, E.F. (2016). The auditory representation of speech sounds in human motor cortex. eLife *5*, e12577.
32. Edwards, E., Nagarajan, S.S., Dalal, S.S., Canolty, R.T., Kirsch, H.E., Barbaro, N.M., and Knight, R.T. (2010). Spatiotemporal imaging of cortical activation during verb generation and picture naming. Neuroimage *50*, 291–301.
33. Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., and Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. Nature *507*, 94–98.
34. Wilson, S.M., Saygin, A.P., Sereno, M.I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nat. Neurosci. *7*, 701–702.
35. Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. USA *103*, 7865–7870.
36. Romanski, L.M., and Averbeck, B.B. (2009). The primate cortical auditory system and neural representation of conspecific vocalizations. Annu. Rev. Neurosci. *32*, 315–346.
37. Romanski, L.M., and Goldman-Rakic, P.S. (2002). An auditory domain in primate prefrontal cortex. Nat. Neurosci. *5*, 15–16.
38. Panzeri, S., Ince, R.A.A., Diamond, M.E., and Kayser, C. (2014). Reading spike timing without a clock: intrinsic decoding of spike trains. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20120467.
39. Foffani, G., and Moxon, K.A. (2004). PSTH-based classification of sensory stimuli using ensembles of single neurons. J. Neurosci. Methods *135*, 107–120.
40. Schwartz, J.-L., Beautemps, D., Arrouas, Y., and Escudier, P. (1992). Auditory analysis of speech gestures. In The Auditory Processing of Speech: From Sounds to Words, M.E.H. Schouten (Mouton de Gruyter), pp. 239–252.
41. Malone, B.J., Scott, B.H., and Semple, M.N. (2015). Diverse cortical codes for scene segmentation in primate auditory cortex. J. Neurophysiol. *113*, 2934–2952.
42. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.
43. Mazor, O., and Laurent, G. (2005). Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. Neuron *48*, 661–673.
44. Kao, J.C., Nuyujukian, P., Ryu, S.I., Churchland, M.M., Cunningham, J.P., and Shenoy, K.V. (2015). Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. Nat. Commun. *6*, 7759.
45. Briggman, K.L., Abarbanel, H.D.I., and Kristan, W.B., Jr. (2005). Optical imaging of neuronal populations during decision-making. Science *307*, 896–901.
46. Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature *485*, 233–236.

47. Shamma, S.A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. Trends Neurosci. *34*, 114–123.
48. Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A., 3rd, and Brugge, J.F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. *29*, 15564–15574.
49. Millman, R.E., Johnson, S.R., and Prendergast, G. (2015). The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. J. Cogn. Neurosci. *27*, 533–545.
50. Peelle, J.E., Gross, J., and Davis, M.H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex *23*, 1378–1387.
51. Romanski, L.M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P.S., and Rauschecker, J.P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat. Neurosci. *2*, 1131–1136.
52. Malone, B.J., Scott, B.H., and Semple, M.N. (2015). Diverse cortical codes for scene segmentation in primate auditory cortex. J. Neurophysiol. *113*, 2934–2952.
53. Atencio, C.A., Sharpee, T.O., and Schreiner, C.E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. J. Neurophysiol. *107*, 2594–2603.
54. Zheng, Y., and Escabí, M.A. (2008). Distinct roles for onset and sustained activity in the neuronal code for temporal periodicity and acoustic envelope shape. J. Neurosci. *28*, 14230–14244.
55. Tsunada, J., Liu, A.S.K., Gold, J.I., and Cohen, Y.E. (2016). Causal contribution of primate auditory cortex to auditory perceptual decision-making. Nat. Neurosci. *19*, 135–142.
56. David, S.V., and Shamma, S.A. (2013). Integration over multiple timescales in primary auditory cortex. J. Neurosci. *33*, 19154–19166.
57. Seifritz, E., Esposito, F., Hennel, F., Mustovic, H., Neuhoff, J.G., Bilecen, D., Tedeschi, G., Scheffler, K., and Di Salle, F. (2002). Spatiotemporal pattern of neural processing in the human auditory cortex. Science *297*, 1706–1708.
58. Harms, M.P., and Melcher, J.R. (2003). Detection and quantification of a wide range of fMRI temporal responses using a physiologically-motivated basis set. Hum. Brain Mapp. *20*, 168–183.
59. Werner, S., and Noppeney, U. (2011). The contributions of transient and sustained response codes to audiovisual integration. Cereb. Cortex *21*, 920–931.
60. Rauschecker, J.P., Tian, B., Pons, T., and Mishkin, M. (1997). Serial and parallel processing in rhesus monkey auditory cortex. J. Comp. Neurol. *382*, 89–103.
61. Kaas, J.H., and Hackett, T.A. (1999). 'What' and 'where' processing in auditory cortex. Nat. Neurosci. *2*, 1045–1047.
62. Bendor, D., and Wang, X. (2007). Differential neural coding of acoustic flutter within primate auditory cortex. Nat. Neurosci. *10*, 763–771.
63. Cant, N.B., and Benson, C.G. (2007). Multiple topographically organized projections connect the central nucleus of the inferior colliculus to the ventral division of the medial geniculate nucleus in the gerbil, Meriones unguiculatus. J. Comp. Neurol. *503*, 432–453.
64. Rauschecker, J.P. (2012). Ventral and dorsal streams in the evolution of speech and language. Front. Evol. Neurosci. *4*, 7.
65. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.
66. Roux, F.E., Miskin, K., Durand, J.B., Sacko, O., Réhault, E., Tanova, R., and Démonet, J.F. (2015). Electrostimulation mapping of comprehension of auditory and visual words. Cortex *71*, 398–408.
67. Leonard, M.K., Cai, R., Babiak, M.C., Ren, A., and Chang, E.F. (2016). The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. Brain Lang. Published online July 19, 2016. https://doi.org/10.1016/j.bandl.2016.06.001.
68. Boatman, D., Lesser, R.P., and Gordon, B. (1995). Auditory speech processing in the left temporal lobe: an electrical interference study. Brain Lang. *51*, 269–290.
69. Steinschneider, M., Nourski, K.V., and Fishman, Y.I. (2013). Representation of speech in human auditory cortex: is it special? Hear. Res. *305*, 57–73.
70. Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. *19*, 158–164.
71. Bregman, A.S. (1994). Auditory Scene Analysis (MIT Press).
72. O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. J. Neurosci. *35*, 7256–7263.
73. Phillips, D.P., Hall, S.E., and Boehnke, S.E. (2002). Central auditory onset responses, and temporal asymmetries in auditory perception. Hear. Res. *167*, 192–205.
74. Hamilton, L.S., Chang, D.L., Lee, M.B., and Chang, E.F. (2017). Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. Front. Neuroinform. *11*, 62.
75. Edwards, E., Soltani, M., Kim, W., Dalal, S.S., Nagarajan, S.S., Berger, M.S., and Knight, R.T. (2009). Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. J. Neurophysiol. *102*, 377–386.
76. Ray, S., and Maunsell, J.H.R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol. *9*, e1000610.
77. Moses, D.A., Mesgarani, N., Leonard, M.K., and Chang, E.F. (2016). Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. J. Neural Eng. *13*, 056004.
78. Fischl, B., Sereno, M.I., Tootell, R.B.H., and Dale, A.M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. *8*, 272–284.
79. Garofolo, J.S., Lamel, L.F., Fisher, W.F., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., and Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguist (Philadelphia: Data Consortium).
80. Yuan, J., and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. J. Acoust. Soc. Am. *123*, 3878.
81. Blesser, B. (1972). Speech perception under conditions of spectral transformation. I. Phonetic characteristics. J. Speech Hear. Res. *15*, 5–41.
82. Ding, C., He, X., and Horst, D. (2008). On the Equivalence of Nonnegative Matrix Factorization and K -means — Spectral Clustering. Computational Statistics & Data Analysis *52*, 3913–3927.
83. Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network *12*, 289–316.
84. Slaney, M. (1998). Auditory toolbox, version 2 (Purdue Engineering).
85. Schnupp, J.W.H., Hall, T.M., Kokelaar, R.F., and Ahmed, B. (2006). Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. J. Neurosci. *26*, 4785–4795.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB R2017b	The Mathworks	https://www.mathworks.com
Anaconda python v2.7 with MKL acceleration libraries	Anaconda Cloud	https://anaconda.org/anaconda/python
Freesurfer	MGH Harvard	https://surfer.nmr.mgh.harvard.edu/
img_pipe electrode localization software	[74]	https://github.com/changlabucsf/img_pipe
Other		
Electrode grids	AdTech	https://adtechmedical.com

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Edward Chang (edward.chang@ucsf.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

Participants included 27 patients (13M/14F, age: 34 ± 12 years) implanted with high-density subdural intracranial electrode grids (AdTech 256 channels, 4mm center-to-center spacing and 1.17mm diameter) either chronically as part of their clinical evaluation for epilepsy surgery, or in an acute intraoperative setting for tumor resection. All procedures were approved by the University of California, San Francisco Institutional Review Board, and all patients provided informed written consent to participate. 14 subjects were implanted with left hemisphere grids, and 13 subjects were implanted with right hemisphere grids. Details of implantation (hemisphere, sex, handedness, language dominance, and seizure focus) are in Table S1.

Neural recordings

Electrophysiological recordings were acquired at a sampling rate of 3051.8 Hz using a 256-channel PZ2 amplifier or 512-channel PZ5 amplifier connected to an RZ2 digital acquisition system (Tucker-Davis Technologies, Alachua, FL, USA). We recorded the local field potential from each electrode, notch-filtered the signal at 60 Hz and harmonics (120 Hz and 180 Hz) to reduce line-noise related artifacts, and re-referenced to the common average across channels sharing the same connector to the preamplifier [31]. We then used the log-analytic amplitude of the Hilbert transform to bandpass signals in the high gamma range (70-150 Hz), using 8 logarithmically spaced center frequency bands and taking the first principal component across these bands to extract stimulus-related neural activity [75–77]. High gamma signals were then downsampled to 100 Hz for further analysis. Signals were z-scored relative to the mean and standard deviation of activity across each recording session.

Electrode localization

We localized electrodes on each individual's brain by co-registering the preoperative T1 MRI with a postoperative CT scan containing the electrode locations, using a normalized mutual information routine in SPM12. Pial surface reconstructions were created using Freesurfer. For visualization of electrode coordinates in MNI space, we performed nonlinear surface registration using a spherical sulcal-based alignment in Freesurfer, aligning to the cvs_avg35_inMNI152 template [78]. This nonlinear alignment ensures that electrodes on a gyrus in the subject's native space remain on the same gyrus in the atlas space but does not maintain the geometry of the grid. Full electrode localization procedures are described in [74].

Stimuli

TIMIT sentences

Participants listened passively to 499 sentences taken from the TIMIT acoustic-phonetic corpus [79], spoken by 286 males and 116 females from different regions of the United States of America. Most sentences were repeated 1-2 times; for the majority of patients (23/27), a subset of 10 sentences was repeated 10-22 times. Stimuli were played through free-field speakers (Logitech), and presentation was controlled using custom MATLAB software on a Windows Laptop. Sentences were presented in pseudorandom order with 0.4 s of silence in between each.

Natural speech from "Forrest Gump"

In addition to the TIMIT sentences, three participants heard naturally spoken sentences from a reading of the movie *Forrest Gump* (referred to here as "Gump"). In brief, it consisted of re-enacted natural speech samples from Robert Zemeckis's *Forrest Gump* by one male and one female talker. We included data from 116 dialog (4.5–19.9 s duration) speech samples (an example of one of these is shown in Figure 3, and another in Figure S6). Each stimulus was presented at least one time to each participant. Phonetic transcription was performed using forced alignment with the Penn Phonetics Lab Forced Aligner [80], followed by manual segmentation in Praat. This dataset was used previously by our group to investigate decoding of words and phonemes from continuous, naturally spoken speech [77].

Sentence control stimuli

In addition to natural sentences, a subset of subjects (n = 4) were presented with a set of control stimuli that were synthesized from 10 of the original TIMIT sentences. These 10 sentences were the subset that were repeated 10-20 times and included 5 male and 5 female speakers. Control conditions included time-reversed sentences and spectrally rotated sentences. Time-reversed sentences were constructed by flipping the stimulus waveforms such that each sentence was played backward in time from its original version. Spectrally rotated sentences were constructed according to methods described by Blesser [81].

Pure tone stimuli

For n = 5 subjects, we also played pure tone stimuli, synthesized as 50-ms duration 5-ms cosine ramped sine wave tones with mel-spaced center frequencies that matched our sentence spectrograms. These center frequencies ranged from 74.5 Hz to 8kHz. Pure tones were played at 3 intensity levels at 10 dB spacing, with the lowest intensity calibrated to be minimally audible in the hospital room. Each pure tone frequency/intensity pair was repeated 3 times, and inter-stimulus intervals were jittered (range 0.28 s minimum ISI – 0.5 s maximum ISI).

Electrode selection

We identified electrodes with robust responses to speech sounds that were well-predicted by a linear spectrotemporal model ($r > 0.1$ on a held-out dataset, see Receptive Field estimation). This metric was used rather than simply testing for significant responses during speech compared to silence, since in practice the short time period of onset responses during speech sometimes led to false exclusion of onset electrodes. This selection procedure resulted in a total of 1,906 speech-responsive electrodes across the 27 patients.

Unsupervised clustering of time series data

We used convex non-negative matrix factorization (NMF) [17] to uncover functional areas based on correlated activity during a natural speech listening task. In brief, we estimated the time series X [n time points x p electrodes] with the following factorization:

$$X \approx \hat{X} = FG^{\top},$$

where

$$F = XW$$

The *G* matrix [p electrodes x k clusters] represents the spatial weighting of an electrode on a given cluster, and the *W* matrix [p electrodes x k clusters] represents the weights on each of the electrode time series. Restricting *F* to be a convex combination of the electrode time series allows us to compute a time series "centroid" – that is, the weighted time series *XW* for each cluster k will give us the prototypical time series for that cluster (see [17] for proof of this concept, and Figure 1B for cluster time series).

We concatenated the z-scored time series for all 27 subjects and performed the clustering analysis on all subjects simultaneously to find patterns of activity that were consistent across subjects. This resulted in a matrix X of 31,625 time points by 1,906 electrodes. We restricted this analysis to the sentences that were heard by all subjects, which included a total of 113 sentences. Sentence stimuli were aligned only across subjects; onset alignment across sentences (as shown in Figure 1B) was performed after clustering, not before. We initialized the W matrix by first performing an eigenvalue decomposition on the unit-normed covariance matrix $X^{\top}X$, followed by a varimax rotation and rectification. The G matrix was then initialized according to Ding et al., 2008 [82], followed by alternating updates of W and G as described in Ding et al., 2010 [17] until convergence was achieved.

To evaluate the number of clusters, we calculated the percent variance explained when projecting the data onto the computed NMF clusters. For this, we used the following equation:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum \left(X - \hat{X}\right)^2}{\sum \left(X_i - \overline{X}\right)^2}$$

We then plotted the additional percent variance explained for k = 2 to k = 32 clusters.

Silhouette index

To evaluate cluster separability, we employed the silhouette index *s(i)*, which describes how well each electrode *i* is matched to its own cluster compared to the non-match cluster. This takes the form:

$$s(i) = \frac{d_{across}(i) - d_{within}(i)}{\max\{d_{within}(i), d_{across}(i)\}}$$

where $d_{across}(i)$ is the lowest average dissimilarity of electrode *i* to the cluster for which it is not a member (as measured by the squared Euclidean distance), and $d_{within}(i)$ is the average dissimilarity of electrode *i* with all other electrodes in the same cluster. The silhouette index was calculated within each subject separately. For the functional clustering, we calculated the dissimilarity as the squared Euclidean distance between the NMF activation weights G for each cluster. For anatomical clustering, the dissimilarity was calculated using the physical pairwise distance between electrodes within or across clusters.

Trajectory analysis

State-space trajectory analyses were performed by projecting data from all electrodes onto NMF basis functions. This is equivalent to the calculation of the cluster time series "centroid" *F*, described above. For the Onset zone, this was the weighted time series XW_1, where only the first column of W was used, and for the Sustained zone, this was the weighted time series XW_2, with only the second column of W. For analyses that were restricted by subject, we used only the columns of X and rows of W corresponding to electrodes within the subject of interest, and calculated XW for those electrode subsets.

Receptive field estimation

To model acoustic and phonetic transformations in speech-sensitive cortex, we used linear encoding models to describe the high gamma activity recorded at each electrode as a weighted sum of stimulus features over time. This model is known in the literature as the spectrotemporal receptive field, and is widely used to describe selectivity for natural stimuli [83]. The models were of the form:

$$\hat{x}(t) = \sum_{f}\sum_{\tau}\beta(\tau, f)S(f, t-\tau)$$

Where x is the neural activity recorded at a single electrode, $\beta(\tau, f)$ contains the regression weights for each feature *f* at time lag τ, and *S* is the stimulus representation. In this analysis, we used all sentences that a subject heard to the fit the model. We estimated models using two representations of the data: (1) a spectrogram-based representation, and (2) a phoneme feature-based representation. For the spectrotemporal stimulus representation, we used the mel-band spectrogram as in our previous work [11]. The mel band frequencies ranged from approximately 75 Hz to 8 kHz, using an auditory filter bank with a cosine transform that gives a representation of spectral power over time that mimics the filtering performed by the human auditory system [84].

For the phoneme feature representation, we constructed a binary phoneme feature matrix describing each sentence as a set of features (1 for the presence of a feature, and 0 for its absence) describing phonetic or other linguistic content. Based on previous work showing that the STG responds to phonetic features rather than single phonemes [11], we included features for sonorant, obstruent, voiced, back, front, low, high, dorsal, coronal, labial, syllabic, plosive, fricative, and nasal. To model response nonlinearities at the beginning of sentences and after pauses, we also included a sentence onset feature to mark the first phoneme of each sentence. In the selectivity maps shown in Figure 4, we collapsed across the relevant features for plosives, nasals, fricatives, and vowels for ease of viewing.

We fit receptive fields using time delays of up to 600 ms in order to account for the longer responses observed in the rostral STG. β weights were fit using ridge regression, where the ridge parameter was estimated using a bootstrap procedure in which the training set was randomly divided into 80% prediction and 20% ridge testing sets. The ridge parameter was chosen as the parameter that gave the best average performance across electrodes as assessed by correlation between the predicted and ridge test set performance. The final performance of the model was computed on a final held out set not included in the ridge parameter selection. Performance was measured as the correlation between the predicted response on the model and the actual high gamma measured for sentences in the test set.

Modulation transfer function analysis

We calculated the modulation transfer function (MTF) of each STRF as the 2D Fourier transform of the STRF [25]. After taking the 2D Fourier transform, values were squared, log transformed, and multiplied by 10 to convert units to power (dB).

Response latency analysis

To calculate the response latencies from the STRF, we calculated a temporal kernel by taking the mean across all frequencies in the STRF matrix. We calculated the onset latency as the time at which the derivative of the temporal kernel reached its maximum. The peak latency was the peak of this temporal kernel, and the offset latency was the time after the peak at which the derivative of the temporal kernel was maximally negative.

Neural onset detection analysis

To calculate onsets from the neural data without incorporating knowledge about the stimulus, we projected single trial population high gamma responses onto the NMF components for the Onset and Sustained zone in each participant separately, using only participants for which at least 10 repeats of a subset of 10 sentences were available (N = 16). Next, these projected high gamma data were lowpass filtered at 2 Hz using a 3rd order zero-phase Butterworth filter and half-wave rectified to set all negative values to 0 (see bold traces in Figure 5A). We then squared this signal and took the derivative to detect the timing of strong changes in the high gamma signal (black traces, Figure 5A). The local maxima of this signal were detected and the top *n* peak times (where n = the number of stimuli) were marked as the detected onsets (Figure 5A, arrows).

To determine whether neutrally detected onsets were accurate, we calculated the absolute value of the difference between the actual stimulus onset (from the TIMIT transcription) and the detected onset. If the detected onset was within 0 – 150 ms of the actual stimulus onset (where *t* varied from 50 to 600ms, in 50 ms steps), it was counted as a correct detection. This window was chosen according to the average STRF peak excitatory response across all sites (onset and sustained). Accuracies were calculated as the percentage of correct detections for each stimulus. Onset electrodes always outperformed the sustained electrodes in accuracy even for windows up to 600 ms. In our classifier analysis, we used these neutrally detected onsets as well as the true stimulus onset as inputs to a template-matching based classifier (as detailed in [39]). In brief, we took single trials from subjects for whom we had a full set of the 10 sentence stimuli that were repeated 10 times, which included 5 sentences spoken by male speakers and 5 sentences spoken by female speakers (n = 16 subjects). In these subjects, we used the Onset electrodes, Sustained electrodes, or the stimulus to define single trial onsets as detailed above. Then, we calculated the Euclidean distance between each single trial population response to a stimulus and a ''template'' from the average of repeated trials of that stimulus that did not include the single trial to be matched. This classifier analysis was done using a sliding 150 ms window starting at 0.5 s before sentence onset. Electrodes included in the population response were either all electrodes, only Onset electrodes, or only Sustained electrodes, aligned to either the stimulus, Onset electrode onsets, or Sustained electrode onsets as specified in Figure 5. Chance performance was calculated by creating 10,000 random confusion matrices for the 10 sentence stimuli using methods defined in [85]. We then calculated the 95% confidence intervals for accuracies assessed from these randomly generated confusion matrices, which is shown in gray in Figure 5C.

QUANTIFICATION AND STATISTICAL ANALYSIS

For data that deviated from normality, we used nonparametric Wilcoxon rank sum (for unpaired data) or signed rank tests (for paired data). In some cases, a bootstrap t test was used. All tests were performed in MATLAB (vR2017b) or python (v2.7).

DATA AND SOFTWARE AVAILABILITY

Data and code available upon request to the Lead Contact, Edward F. Chang (edward.chang@ucsf.edu).