Transformation of a temporal speech cue to a spatial neural code in human auditory cortex

Neal P. Fox[a], Matthew K. Leonard[a], Matthias J. Sjerps[b,c] & Edward F. Chang[a,d]*

[a] Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, California, 94158, USA
[b] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University, Kapittelweg 29, Nijmegen, 6525 EN, The Netherlands
[c] Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, 6525 XD, Netherlands
[d] Weill Institute for Neurosciences, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, California, 94158, USA

* Please address correspondence to: edward.chang@ucsf.edu

ABSTRACT

In speech, listeners extract continuously-varying spectrotemporal cues from the acoustic signal to perceive discrete phonetic categories. Spectral cues are spatially encoded in the amplitude of responses in phonetically-tuned neural populations in auditory cortex. It remains unknown whether similar neurophysiological mechanisms encode temporal cues like voice-onset time (VOT), which distinguishes sounds like /*b*/-/*p*/. We used direct brain recordings in humans to investigate the neural encoding of temporal speech cues with a VOT continuum from /*ba*/ to /*pa*/. We found that distinct neural populations respond preferentially to VOTs from one phonetic category, and are also sensitive to sub-phonetic VOT differences within a population's preferred category. In a simple neural network model, simulated populations tuned to detect either temporal gaps or coincidences between spectral cues captured encoding patterns observed in real neural data. These results demonstrate that a spatial/amplitude neural code underlies the cortical representation of both spectral and temporal speech cues.

KEYWORDS

speech perception; electrocorticography (ECoG); human auditory cortex; temporal processing; voice-onset time (VOT); categorical perception; sub-phonetic detail

INTRODUCTION

During speech perception, listeners must extract acoustic cues from a continuous sensory signal and map them onto discrete phonetic categories, which are relevant for meaning(1, 2). Many such cues to phonological identity are encoded within the fine temporal structure of speech(3–5). For example, voice-onset time (VOT), defined as the interval between a stop consonant’s release and the onset of vocal fold vibration (acoustically, the *burst* and the *voicing*), is a critical cue that listeners use to distinguish *voiced* (e.g., /*b*/, /*d*/, /*g*/) from *voiceless* (e.g., /*p*/, /*t*/, /*k*/) stop consonants in English(6, 7). When the burst and voicing are roughly coincident (short VOT; ~0ms), listeners perceive a bilabial stop as a /*b*/, but when voicing follows the burst after a temporal gap (long VOT; ~50ms), listeners hear a /*p*/.

Recent evidence from human electrocorticography (ECoG) has shown that information about a speech sound’s identity is encoded in the amplitude of neural activity at phonetically-tuned cortical sites in the superior temporal gyrus (STG)(8). Distinct neural populations in this region respond selectively to different classes of phonemes that share certain spectral cues, such as the burst associated with stop consonants or the characteristic formant structure of vowels produced with specific vocal tract configurations. However, it is unclear whether phonetic categories distinguished by temporal cues (e.g., voiced vs. voiceless stops) are represented within an analogous spatial encoding scheme. If so, this would entail that local neural populations are tuned to detect not merely the presence of certain spectral cues (the burst and voicing), but also their timing relative to one another.

In addition to distinguishing phonetic categories, the exact VOT of a given utterance of a /*b*/ or a /*p*/ will vary considerably depending on numerous factors such as speech rate, phonetic context, and speaker accent(9–15). Although only categorical phonetic identity (e.g., whether a particular VOT is more consistent with a /*b*/ or a /*p*/) is strictly necessary for understanding meaning, sensitivity to fine-grained sub-phonetic detail (e.g., whether a particular /*p*/ was pronounced with a 40ms vs. a 50ms VOT) is also crucial for robust speech perception, allowing listeners to flexibly adapt and to integrate multiple cues to phonetic identity online in noisy, unstable environments(16–21). However, the neurophysiological mechanisms that support listeners’ sensitivity(22–28) to such detailed speech representations are not known. We tested whether sub-phonetic information might be encoded in the neural response amplitude of the same acoustically-tuned neural populations that encode phonetic information in human auditory cortex.

To address these questions, we recorded neural activity directly from the cortex of seven human participants using high-density ECoG arrays while they listened to and categorized syllables along a VOT continuum from /*ba*/ (0ms VOT) to /*pa*/ (50ms VOT). We found that the amplitude of cortical responses in STG simultaneously encodes both phonetic and sub-phonetic information about a syllable’s initial VOT. In particular, spatially discrete neural populations respond preferentially to VOTs from one category (either /*b*/ or /*p*/). Furthermore, peak response amplitude is modulated by stimulus VOT within each population’s preferred – but not its non-preferred – voicing category (e.g., stronger response to 0ms than to 10ms VOT in voiced-selective [/*b*/-selective] neural populations). This same encoding scheme emerged in a computational neural network model simulating neuronal populations as leaky integrators tuned to detect either temporal coincidences or gaps between distinct spectral cues. Our results provide direct evidence that phonetic and sub-phonetic information carried by VOT are represented within spatially discrete, phonetically-tuned neural populations that integrate temporally-distributed spectral cues in speech. This represents a crucial step towards a unified model of cortical speech encoding, demonstrating that both spectral and temporal cues and both phonetic and sub-phonetic information are represented by a common (spatial) neural code.

RESULTS

Participants listened to and categorized speech sounds from a digitally synthesized continuum of consonant-vowel syllables that differed linearly only in their voice-onset time (VOT) from /*ba*/ (0ms VOT) to /*pa*/ (50ms VOT). This six-step continuum was constructed by manipulating only the relative timing of the spectral burst and the onset of voicing while holding all other acoustic properties of the stimuli constant (**Figures 1A/B**; see **Methods**)(29). Analysis of participants' identification behavior confirmed that stimuli with longer VOTs were more often labeled as /*pa*/ (mixed effects logistic regression: $\beta_{\mathrm{VOT}} = 0.19$, $t = 17.78$, $p = 5.6*10^{-63}$; data for example participant in **Figure 1C**; data for all participants in **Figure 1-figure supplement 1**). Moreover, and consistent with past work, listeners' perception of the linear VOT continuum was sharply non-linear, a behavioral hallmark of categorical perception(30–32). A psychophysical category boundary between 20ms and 30ms divided the continuum into stimuli most often perceived as voiced (/*b*/: 0ms, 10ms, 20ms VOTs) or as voiceless (/*p*/: 30ms, 40ms, 50ms VOTs).

Temporal cues to voicing category are encoded in spatially distinct neural populations

To investigate neural activity that differentiates the representation of speech sounds based on a temporal cue like VOT, we recorded high-density electrocorticography in seven participants while they listened to the VOT continuum. We examined high-gamma power (70-150 Hz)(33–36), aligned to the acoustic onset of each trial (burst onset), at every speech-responsive electrode on the lateral surface of the temporal lobe of each patient (n = 346 electrodes; see **Methods** for details of data acquisition, preprocessing, and electrode selection).

We used nonparametric correlation analysis (Spearman's ρ) to identify electrodes where the peak high-gamma amplitude was sensitive to stimulus VOT. Across all participants, we found 49 VOT-sensitive sites, primarily located over the lateral mid-to-posterior STG, bilaterally. Peak response amplitude at these VOT-sensitive electrodes reliably discriminated between voicing categories, exhibiting stronger responses to either voiced (/*b*/; VOT = 0-20ms; n = 33) or voiceless (/*p*/; VOT = 30-50ms; n = 16) stimuli (**Figure 1D**; locations of all sites shown in **Figures 2A** and **1-figure supplement 2**). We observed that, within individual participants, electrodes spaced only 4mm apart showed strong preferences for different voicing categories, and we did not observe any clear overall regional or hemispheric patterns in the prevalence or selectivity patterns of VOT-sensitive electrodes (see **Methods** for additional information).

Robust category selectivity in voiceless-selective (V-) and voiced-selective (V+) neural populations emerged as early as 50-150ms post-stimulus onset and often lasted for several hundred milliseconds (example electrodes in **Figure 1E**). Across all VOT-sensitive electrodes, voicing category selectivity was reliable whether a trial's voicing category was defined based on the psychophysically-determined category boundary (0-20ms vs. 30-50ms VOTs; V- electrodes: $z = 3.52$, $p = 4.4\text{x}10^{-4}$; V+ electrodes: $z = -5.01$, $p = 5.4\text{x}10^{-7}$; Wilcoxon signed-rank tests) or based on the actual behavioral response recorded for each trial (V- electrodes: $p = 4.9\text{x}10^{-4}$; V+ electrodes: $p = 6.1\text{x}10^{-5}$; Wilcoxon signed-rank tests).

These results show that spatially distinct neural populations in auditory cortex are tuned to speech sound categories defined by a temporal cue. Critically, if individual neural populations only responded to spectral features (e.g., to the burst or to the onset of voicing), we would not have observed overall amplitude differences in their responses to /*b*/ versus /*p*/ categories.

Given this pattern of spatial tuning, we tested whether the voicing category of single trials could be reliably decoded from population neural activity across electrodes. For each participant, we trained a multivariate pattern classifier (linear discriminant analysis with leave-one-out cross validation) to predict trial-by-trial voicing category using high-gamma activity across all speech-responsive electrodes on the temporal lobe during the peak neural response (150-250ms after stimulus onset; see **Methods**). We found that, across participants, classification accuracy was significantly better than chance (Wilcoxon signed-rank test: $p = 0.016$; **Figure 1F**, leftmost box plot), demonstrating that spatially and temporally distributed population neural activity during the peak response contains information that allows for decoding of a temporally-cued phonetic distinction in speech.

Peak neural response amplitude robustly encodes voicing category

Next, we asked which features of the population neural response encode voicing category. Specifically, we evaluated three alternatives for how temporally-cued voicing category is encoded by high-gamma responses in cortex during the peak neural response: (1) the spatial pattern of peak response amplitude across electrodes, (2) the temporal patterns of evoked responses across electrodes during the peak response, or (3) both amplitude and timing of neural activity patterns. We tested these hypotheses by selectively corrupting amplitude and/or temporal neural features that were inputs for the classifier. As with the previous analyses, and following prior work on speech sound encoding(8), these analyses (**Figure 1F**) focused on cortical high-gamma activity during the peak response window (150-250ms after stimulus onset; but see **Figure 3** for analyses of an earlier time window).

To corrupt temporal information, we randomly jittered the exact timing of the neural response for each trial by shifting the 100ms analysis window by up to ±50ms. Because the uniform random jitter was applied independently to each trial, this procedure disrupts any temporal patterns during the peak neural response that might reliably distinguish trials of different voicing categories, such as precise (millisecond-resolution) timing of the peak response at an electrode or the dynamics of the evoked response during the peak window, including *local* temporal dynamics (during a single electrode's peak response) or *ensemble* temporal dynamics (the relative timing of responses of spatially-distributed electrodes in the same participant). To corrupt amplitude information, we eliminated any condition-related differences in the peak response amplitude at every electrode. For each electrode, the evoked high-gamma response to all trials within a given voicing category were renormalized so that the average responses to both voicing categories had identical amplitudes at the peak, but could still vary reliably in the timing and dynamics during the peak window. These techniques allowed us to examine the relative contributions of temporal and amplitude information contained within the peak neural response window to the classification of voicing category (see **Methods** for detailed description of this approach).

Across participants, we found that, when the classifiers had access to amplitude information but not timing information (+Amplitude/-Timing) during the peak response, performance was significantly better than chance (Wilcoxon signed-rank test: $p = 0.016$; **Figure**

1F). Furthermore, despite the profound corruption of temporal information in the neural responses, classification accuracy was statistically comparable to the model that had access to both amplitude and timing information (+Amplitude/+Timing; Wilcoxon signed-rank test: p = 0.69; **Figure 1F**), suggesting that amplitude information alone is sufficient for classifying a trial's voicing category.

In contrast, when amplitude information was corrupted and only temporal patterns in the peak response window were reliable (-Amplitude/+Timing), classifier performance was not different from chance (Wilcoxon signed-rank test: p = 0.69; **Figure 1F**) and was worse for every participant compared to the model with both types of information (Wilcoxon signed-rank test: p = 0.016). Finally, we compared the model with only timing information to a model where both amplitude and timing information during the peak window were corrupted (-Amplitude/-Timing). We found that preserving timing information alone had no effect on classification performance compared to the most impoverished model (-Amplitude/-Timing; Wilcoxon signed-rank test: p = 0.58; **Figure 1F**), which also failed to perform better than chance (Wilcoxon signed-rank test: p = 0.94; **Figure 1F**). Together, these results constitute evidence for a spatial/amplitude code for speech categories that differ in a temporal cue. Thus, localized peak high-gamma response amplitude spatially encodes voicing of single trials in STG, analogous to other spectrally-cued phonetic features(8). Note that, while spatial (and not temporal) patterns of high-gamma responses robustly encode voicing during this critical peak window, we later describe additional analyses that address possible temporal encoding patterns in the local field potential (**Figure 1-figure supplements 3** and **4**) and in an earlier time window (**Figure 3**).

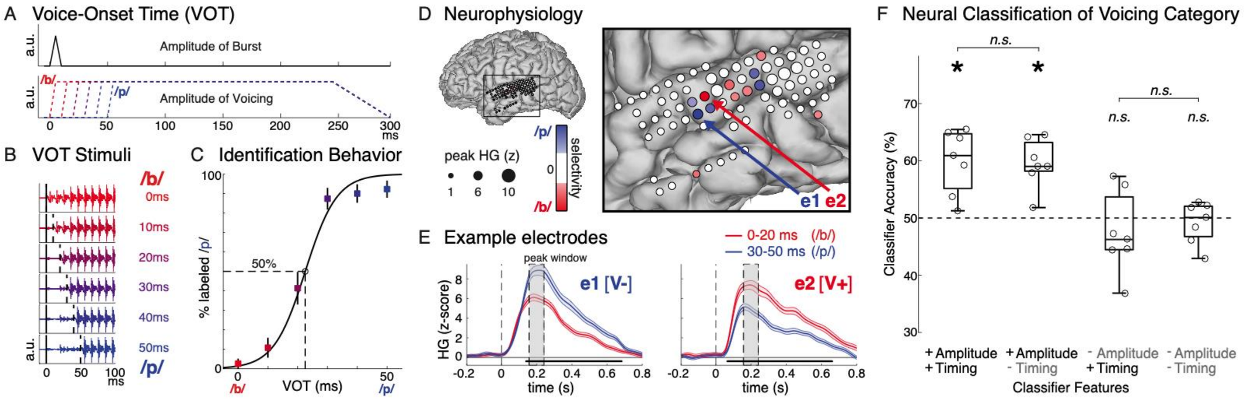

Fig. 1. Speech sound categories that are distinguished by a temporal cue are spatially encoded in the peak amplitude of neural activity in distinct neural populations. A. *Stimuli varied only in voice-onset time (VOT), the duration between the onset of the burst (**top**) and the onset of voicing (**bottom**) (a.u. = arbitrary units). **B.** Acoustic waveforms of the first 100ms of the six synthesized stimuli. **C.** Behavior for one example participant (mean ± bootstrap SE). Best-fit psychometric curve (mixed effects logistic regression) yields voicing category boundary between 20-30ms (50% crossover point). **D.** Neural responses in the same representative participant show selectivity for either voiceless or voiced VOTs at different electrodes. Electrode size indicates peak high-gamma (HG; z-scored) amplitude at all speech-responsive temporal lobe sites. Electrode color reflects strength and direction of selectivity (Spearman's ρ between peak HG amplitude and VOT) at VOT-sensitive sites (p < 0.05). **E.** Average HG responses (± SE) to voiced (0-20ms VOTs; red) and voiceless (30-50ms VOTs; blue) stimuli in two example electrodes from **D**, aligned to stimulus onset (e1: voiceless-selective, V-; e2: voiced-selective, V+). Horizontal black bars indicate timepoints with category discriminability (p < 0.005). Grey boxes mark average peak window (± SD) across all VOT-sensitive electrodes (n = 49). **F.** Population-based classification of*

voicing category (/p/ vs. /b/) during peak window (150-250ms after stimulus onset). Chance is 50%. Boxes show interquartile range across all participants; whiskers extend to best- and worst-performing participants; horizontal bars show median performance. Asterisks indicate significantly better-than-chance classification across participants ($p < 0.05$; n.s. = not significant). Circles represent individual participants.

The encoding of stop consonant voicing in the amplitude of evoked high-gamma responses in STG suggests that the representation of temporally-cued phonetic features may be explained within the same neural coding framework as the representation of spectrally-cued phonetic features. However, previous work on the cortical representation of voicing has identified a role for temporal information in the local field potential (LFP) (37, 38), which is dominated by lower- frequencies (39, 40).

To link our results with this existing literature, we conducted a series of exploratory analyses of the neural responses to our stimuli using the raw voltage (LFP) signal. For each VOT-sensitive electrode (defined in the high-gamma analysis), we estimated the correlations between VOT and peak latency and between VOT and peak amplitude for 3 peaks in the auditory evoked potential (AEP) occurring approximately 75-100 ms (P_α), 100-150 ms (N_α), and 150-250 ms (P_β) after stimulus onset (**Figure 1-figure supplement 3**)(41, 42). We found that some VOT-sensitive electrodes encoded VOT in the latency of these peaks (e.g., **Figure 1-figure supplement 4**, **panels E/I/M**), replicating previous results (43). However, among electrodes that encode VOT in peak high-gamma amplitude, there exist many more electrodes that *do not* encode VOT in these temporal features of the AEP, and many that also encode VOT in the amplitude of these AEP peaks (**Figure 1-figure supplements 3** and **4**). This further supports the prominent role that amplitude information plays in the neural representation of voicing and VOT, both in high-gamma and in the LFP. Therefore, subsequent analyses focus on the high-gamma amplitude. (For detailed descriptions of these LFP analyses and their results, see **Methods** and **Figure 1-figure supplements 3** and **4**.)

Peak response amplitude encodes sub-phonetic VOT information within preferred category

Next, we assessed whether VOT-sensitive neural populations (**Figure 2A**), which reliably discriminate between phonetic categories (voiced vs. voiceless), also encoded within-category sub-phonetic detail in the peak response amplitude. Specifically, the cortical representation of stimuli from the same voicing category but with different VOTs (e.g., 30, 40, and 50ms VOTs that all correspond to /*p*/) could be either categorical (i.e., all elicit the same peak response amplitude) or graded (i.e., peak response amplitude depends on within-category VOT).

We examined the average responses to each of the six VOTs separately in the voiceless-selective electrodes (V-; **Figure 2B**) and the voiced-selective electrodes (V+; **Figure 2C**). We observed clear differences in activity evoked by different VOTs at the peak response (~200ms after stimulus onset), even within the same voicing category, consistent with sensitivity to sub-phonetic detail(44–47). However, the discriminability of responses to within-category VOTs depended on the preferred voicing category of a given electrode.

To quantify this observation, at each electrode, we computed the rank-based correlation (Spearman's ρ) between stimulus VOT and peak response amplitude separately for each voicing category (0-20ms and 30-50ms VOTs). This procedure resulted in two correlation coefficients for each VOT-sensitive site ($\rho_{0\text{-}20}$, $\rho_{30\text{-}50}$) and corresponding test statistics reflecting the strength of within-category amplitude encoding of stimulus VOT in each voicing category. These test statistics (one per voicing category per VOT-sensitive electrode) then served as the input data for a series of signed-rank statistical tests to assess overall within-category encoding properties of groups of electrodes (e.g., of all V- electrodes) (see **Methods** for details). For example, consider V- electrodes, which exhibit stronger responses, overall, for voiceless stimuli (30-50ms VOTs) compared to voiced stimuli (0-20ms VOTs). Across V- electrodes, we found that voiceless stimuli with longer VOTs (i.e., closer to the preferred category's 50ms endpoint VOT) also elicit increasingly stronger responses (Wilcoxon signed-rank test: z = 3.52, p = 4.4x10^{-4}). At the same V- sites, however, within-category VOT does not reliably predict response amplitude among (non-preferred) voiced stimuli (Wilcoxon signed-rank test: z = -1.60, p = 0.11; **Figure 2B**: differences among solid blue lines but not dashed red lines). Across all V- and V+ electrodes, peak high-gamma response amplitude encoded stimulus VOT within the preferred category (Wilcoxon signed-rank test: z = 6.02, p = 1.7x10^{-9}), but not the nonpreferred category (Wilcoxon signed-rank test: z = 1.31, p = 0.19). While V- electrodes encoded sub-phonetic VOT more robustly within the voiceless category than within the voiced category (**Figure 2D**, **left**; Wilcoxon signed-rank test: z = 3.00, p = 2.7x10^{-3}), the opposite pattern emerged for V+ electrodes, which encoded sub-phonetic VOT more robustly within the voiced category than within the voiceless category (**Figure 2D**, **right**; Wilcoxon signed-rank test: z = 3.78, p = 1.6x10^{-4}).

Together, these analyses revealed two key results: (1) VOT encoding in human STG is not purely categorical, but also (2) the relationship between response amplitude and VOT is not linear across the entire continuum (**Figure 2D**). These results suggest that, even at the level of STG, the brain maintains information about the specific, sub-phonetic details of individual speech sounds. The asymmetrical pattern of within-category encoding suggests that individual neural populations in human auditory cortex encode information about both the category identity of a speech sound and its more fine-grained acoustic properties, or its "category goodness."(22, 44, 48)

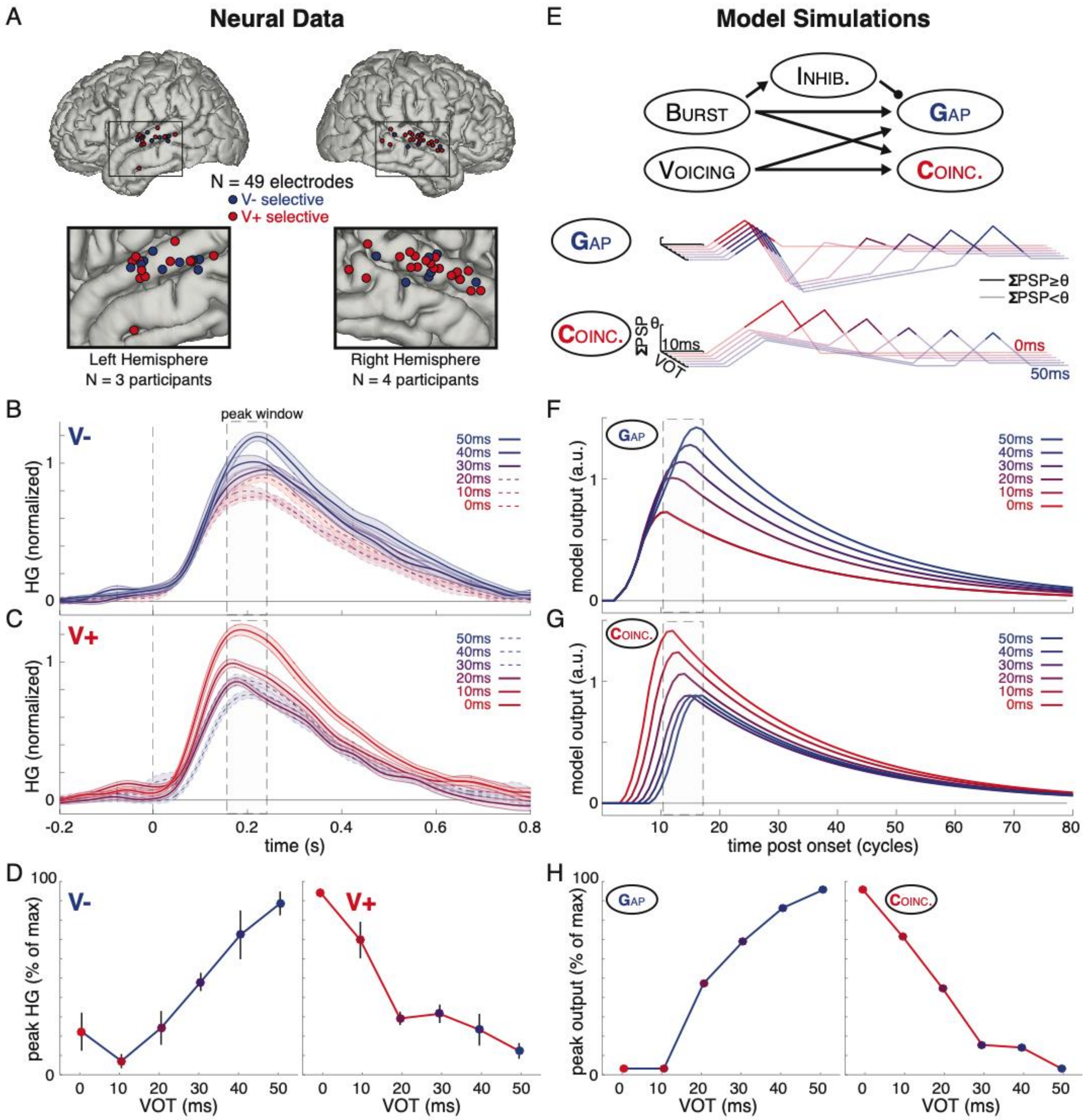

Fig. 2. Human auditory cortex encodes both phonetic (between-category) and sub-phonetic (within-category) information in peak response amplitude, which can be modeled by a simple neural network that implements temporal gap and coincidence detection. A. *Spatial distribution of VOT-sensitive electrodes across all (on standardized brain).* ***B.*** *Average (± SE) normalized HG response to each VOT across all voiceless-selective (V-) electrodes, aligned to stimulus onset. Line style denotes category membership of a given VOT (solid: preferred category; dashed: non-preferred category). Grey box marks average peak window (± SD) across all VOT-sensitive electrodes.* ***C.*** *Average (± SE) normalized response to each VOT across all voiced-selective (V+) electrodes.* ***D.*** *Average (± SE) peak response to each VOT stimulus for V- electrodes (**left**) and V+ electrodes (**right**) (see **Methods**).* ***E.*** *A simple neural network model (**top**) comprised of five leaky integrator nodes was implemented to examine computational mechanisms that could account for the spatial encoding of a temporal cue (VOT). Arrows and circle represent excitatory and inhibitory connections between nodes. See **Methods** for details on model*

*parameters. Postsynaptic potentials (PSPs) illustrate the internal dynamics of the gap detector (**Gap, middle**) and coincidence detector (**Coinc., bottom**) in response to simulated VOT stimuli (line color). Outputs (**panels F/G**) are triggered by suprathreshold instantaneous PSPs ($\Sigma PSP \geq \theta$, dark lines) but not by subthreshold PSPs ($\Sigma PSP < \theta$; semitransparent lines). **F.** Model outputs (a.u. = arbitrary units) evoked by simulated VOT stimuli for **Gap** (1 cycle = 10ms). Note that outputs for 0ms and 10ms VOTs are overlapping. No error bars shown because model simulations are deterministic. Grey box marks average peak window (across **panels F/G**); width matches peak window of real neural data (**panels B/C**). **G.** Model outputs for **Coinc.** **H.** Peak response to each simulated VOT stimulus for **Gap** (**left**) and **Coinc.** (**right**).*

A simple neural network model of VOT encoding in STG

Thus far, we have demonstrated that a temporal cue that distinguishes speech sounds is represented by a spatial/amplitude code(49, 50) in human STG. To understand how this could be implemented computationally in the brain, we built an architecturally minimalistic neural network (**Figure 2E**, **top**). The network was designed to implement a small set of basic computations, motivated by well-established models of temporal processing(51–57). Specifically, our model employs discrete integrator units that detect temporal gaps or coincidences between distinct spectral events by incorporating canonical neurophysiological mechanisms that allow current input to modulate a unit's sensitivity to subsequent input in highly specific ways.

The entire model is comprised of just five localist units: a burst detector, a voicing detector, a gap detector (***Gap***), a coincidence detector (***Coinc.***), and an inhibitory unit. Conventional leaky integrator dynamics governed continuously varying activation values of each rectified linear unit within the model(58, 59), with the activity $a_i(t)$ of a given unit i at time t depending on its prior activity $a_i(t-1)$, the weighted sum of its excitatory and inhibitory inputs $\sum_j w_{ji} * a_j(t-1)$, and unit-specific activation parameters (e.g., propagation threshold [θ], decay rate). To illustrate intuitively how time-dependent neuronal properties can give rise to spatially-localized temporal cue processing, model parameters and connection weights were set manually (see **Methods** for details; **Figure 2-figure supplement 1; Supplementary File 2**). We presented the network with simplified inputs mimicking the spectral and temporal properties of the six VOT stimuli used in the ECoG experiment (**Figure 1A**; see **Methods**; **Supplementary File 3**). Presentation of burst and voicing inputs triggered propagation of activation that spread through the network, and our analyses assessed how the resulting activation dynamics differed depending on VOT.

The simulated responses of ***Gap*** and ***Coinc.*** to VOTs of 0-50ms are shown in **Figures 2F/G**. We observed striking qualitative similarities between ***Gap***'s simulated outputs (**Figure 2F**) and the real neural responses of V- electrodes (**Figure 2B**), and between ***Coinc.***'s outputs (**Figure 2G**) and the V+ electrodes (**Figure 2C**). By design, voicing category is clearly distinguished in both ***Gap*** and ***Coinc.***, with ***Gap*** responding more strongly to longer (voiceless) VOTs (30-50ms), and ***Coinc.*** responding more strongly to shorter (voiced) VOTs (0-20ms). This demonstrates that spatial encoding of temporal cues (gaps vs. coincidences) can arise naturally within a simple, biologically-inspired neural network(51–57).

Perhaps more surprisingly, we also found that both ***Gap*** and ***Coinc.*** detector units exhibit sensitivity to within-category VOT distinctions (**Figure 2H**). These partially graded activations mirror the pattern observed in the neural data (**Figure 2D**), where V- electrodes and ***Gap*** units are only sensitive to differences among long (voiceless) VOTs, and V+ electrodes and ***Coinc.*** units are only sensitive to differences among short (voiced) VOTs.

These relatively sophisticated dynamics are the natural result of well-established computational and physiological mechanisms. Within the model, the burst and voicing detector units are tuned to respond independently to distinct spectral cues in the simulated acoustic input. Hence, the relative timing of their responses, but not their amplitudes, differ as a function of VOT. Both the gap (***Gap***) and the coincidence (***Coinc.***) detector units receive excitatory input from both the burst and voicing detector units, but ***Gap*** and ***Coinc.*** differ in how they integrate these inputs over time. Specifically, as described below, while initial excitatory input (from the burst detector) temporarily *decreases* the sensitivity of ***Gap*** to immediate subsequent excitatory input (from the voicing detector), the opposite is true of ***Coinc***.

In particular, prior work has shown that one computational implementation of gap detection involves configuration of a *slow inhibitory postsynaptic potential* (*IPSP*) microcircuit (**Figure 2E**, **middle**)(51, 52, 60, 61). In our model, activity in the burst detector following burst onset elicits fast suprathreshold *excitatory postsynaptic potentials* (*EPSPs*) in both ***Gap*** and the inhibitory unit, immediately followed by a longer-latency ("slow") IPSP in ***Gap***. This slow IPSP renders ***Gap*** temporarily insensitive to subsequent excitatory input from the voicing detector, meaning that voicing-induced excitation that arrives too soon (e.g., 10ms) after the burst input, when inhibition is strongest, is not able to elicit a second suprathreshold EPSP in ***Gap***. Consequently, all short VOTs (below some threshold) elicit uniformly weak responses in ***Gap*** that reflect only the initial excitatory response to the burst (see, e.g., indistinguishable responses to 0ms and 10ms VOTs in **Figure 2F**). However, as ***Gap*** gradually recovers from the burst-induced slow IPSP, later-arriving voicing input (i.e., longer VOTs) tends to elicit suprathreshold responses that grow increasingly stronger with longer gaps, until ***Gap*** has reached its pre-IPSP (resting) baseline. In this way, our implementation of gap detection naturally captures three key patterns observed across V- electrodes (**Figure 2H**, **left**; **Figure 2D**, **left**): (1) amplitude encoding of a temporally cued category (selectivity for gaps over coincidences); (2) amplitude encoding of within-category differences in the preferred category (amplitude differences among gaps of different durations); and (3) no amplitude encoding of differences within the non-preferred category (uniformly lower amplitude responses to short VOTs of any duration).

In contrast, coincidence detection(54–56, 62–64) (**Figure 2E**, **bottom**) emerges in the model because activity in the burst detector evokes only a subthreshold EPSP in ***Coinc.***, temporarily increasing ***Coinc.***'s sensitivity to immediate subsequent excitatory input (from the voicing detector). During this period of heightened sensitivity, voicing-induced excitatory input that arrives simultaneously or after short lags can elicit larger amplitude (additive) EPSPs than could voicing-induced excitatory input alone. Because the magnitude of the initial burst-induced EPSP gradually wanes, the summation of EPSPs (from the burst and voicing) is greatest (and hence elicits the strongest response) for coincident burst and voicing (0ms VOT), and the magnitude of ***Coinc.***'s response to other voiced stimuli (e.g., 10-20ms VOTs) becomes weaker as the lag between burst and voicing increases. Finally, in voiceless stimuli, since voicing arrives late enough after the burst (30+ ms) that there is no residual boost to ***Coinc.***'s baseline post-synaptic potential, elicited responses are entirely driven by a suprathreshold voicing-induced EPSP that reaches the same peak amplitude for all voiceless stimuli. Thus, our implementation of coincidence detection captures three key patterns observed in V+ electrodes (**Figure 2H**, **right**; **Figure 2D**, **right**): (1) amplitude encoding of a temporally cued category (selectivity for coincidences over gaps); (2) amplitude encoding of within-category differences in the preferred category (amplitude differences among stimuli with short VOTs); and (3) no amplitude encoding of differences within the non-preferred category (uniformly lower amplitude responses to long VOTs of any duration).

In summary, the neurophysiological dynamics underlying local STG encoding of VOT can be modeled using a simple, biologically-inspired neural network. The computational model captures both the between-category (phonetic) and within-category (sub-phonetic) properties of observed neural representations via well-established physiological mechanisms for gap and coincidence detection(51–57).

Mechanisms that explain local category-selectivity also predict early temporal dynamics

Thus far, we have focused on the encoding of speech sounds that differ in VOT based on activity patterns around the peak of the evoked response. However, in comparing the real and simulated neural data (**Figure 2**), we also observed a qualitative resemblance with respect to the onset latencies of evoked responses. Specifically, the timing of the evoked neural responses (relative to burst onset) appeared to depend on stimulus VOT in V+ electrodes and in the coincidence detector (***Coinc.***) unit (**Figures 2C/G**), but not in V- electrodes or in the gap detector (***Gap***) unit (**Figure 2B/F**). This pattern could suggest that early temporal dynamics of the evoked response contribute to the pattern of category selectivity observed at the peak.

We examined the neural activity evoked by each VOT stimulus in V- and V+ electrodes at the onset of the response, typically beginning approximately 75-125ms after stimulus (burst) onset. In the same two example electrodes from **Figure 1E**, we observed clear differences in the relationship between response onset latency and VOT (**Figure 3A**). To quantify the onset latency for each electrode to each VOT stimulus, we found the first timepoint after stimulus onset where the evoked high gamma response exceeded 50% of the electrode's overall peak amplitude (grand mean across conditions). The rank correlation between VOT and response onset latency for e1 (a V- electrode) was substantially lower (Spearman's ρ = 0.42) than for e2 (a V+ electrode; ρ = 0.89).

A bootstrapped rank-based correlation coefficient was computed for each V- and V+ electrode (1000 resamples; see **Methods**). We found that response onset latency was strongly associated with VOT for V+, but not V-, electrodes (Wilcoxon signed-rank tests: V+, $p = 1.6\text{x}10^{-6}$; V-, p = 0.57), and this difference between the two electrode types was highly reliable (Mann-Whitney rank-sum test: $p = 1.7\text{x}10^{-5}$) (**Figure 3B**).

The association between VOT and response latency also differed in ***Gap*** versus ***Coinc.*** units in the model simulations (**Figures 2F/G**), with VOT-dependent response latencies emerging for ***Coinc.***, but not ***Gap***. Closer examination of the model's internal dynamics reveals how the same time-dependent mechanisms that give rise to peak amplitude encoding of VOT are also responsible for these early temporal dynamics. As described above, the category selectivity of ***Gap*** (voiceless) and ***Coinc.*** (voiced) results from how each unit's subsequent activity is modulated after detection of the burst. While the burst always elicits a fast suprathreshold response in ***Gap*** (irrespective of VOT), ***Coinc.***'s response to the burst alone is subthreshold (**Figure 2E**, **middle** vs. **bottom**). Consequently, ***Gap***'s initial response is evoked by the burst of any VOT stimulus, so the response onset latency (when aligned to burst onset) does not depend on VOT (**Figure 2F**). Conversely, ***Coinc.***'s earliest suprathreshold response is triggered by the onset of voicing, so the response onset latency (relative to burst onset) is later for longer VOTs (**Figure 2G**). Thus, the same well-established physiological mechanisms that give rise to peak amplitude encoding of temporally-cued voicing categories also predict the early temporal dynamics we observe in real neural data.

Finally, **Figure 3** shows that, unlike during the peak response window (150-250ms after stimulus onset; **Figure 1F**), temporal information does encode VOT during an earlier window around the neural response onset in some neural populations. Indeed, both sub-phonetic and phonetic category-level information are carried by the onset latency of V+ electrodes, with evoked responses arising later at these sites for stimuli with progressively longer VOTs. Critically, the modeling results indicate that both the amplitude encoding patterns during the peak window and the temporal encoding patterns during the earlier onset window are captured by the same canonical neurophysiological mechanisms.

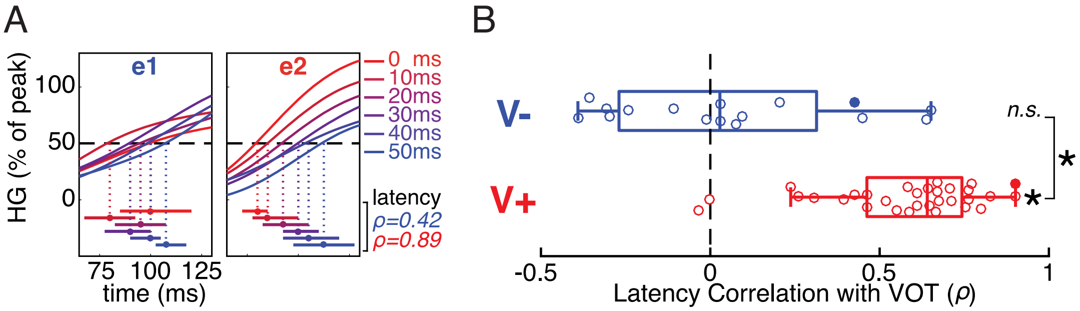

Fig. 3. Early temporal dynamics of stimulus-evoked neural responses differ between voiceless-selective (V-) and voiced-selective (V+) electrodes. A. *Normalized trial-averaged HG responses to each VOT stimulus (line color) in two example electrodes (e1 and e2; same electrodes shown in* ***Figures 1D/E****). The time window (x-axis) is relative to onset of the burst and precedes the peak response. Horizontal bars show estimates (bootstrapped mean ± SE) of response onset latency for each VOT (first timepoint exceeding 50% of electrode's average peak HG). Mean bootstrapped rank-based correlation (Spearman's ρ) between VOT and response onset latency shown for e1 (blue) and e2 (red).* ***B.*** *Across all V- electrodes, the bootstrapped correlation coefficients did not differ significantly from 0, suggesting that onset latency was time-locked to the burst. In contrast, across all V+ electrodes, the bootstrapped correlation coefficients were reliably positive (longer latencies for longer VOTs), and greater than for V- electrodes. Circles represent individual electrodes (filled: example electrodes in* ***A****). Boxes show interquartile range; whiskers extend to maximum/minimum of each group (excluding 2 outlier V+ electrodes); vertical bars are medians. Asterisks indicate significance ($p < 10^{-4}$; n.s. = not significant).*

DISCUSSION

This study investigated how voice-onset time (VOT), a temporal cue in speech, is represented in human auditory cortex. Using direct intracranial recordings, we found discrete neural populations located primarily on the bilateral posterior and middle STG that respond preferentially to either voiced sounds, where the onset of voicing is coincident with the burst or follows it after a short lag (20ms or less), or voiceless sounds, where the onset of voicing follows the burst after a temporal gap of at least 30-50ms.

Past work has also found that phonetic information about speech sounds is encoded in the amplitude of evoked neural responses at spatially localized cortical sites(8). In that work, however, STG activity was shown to encode the spectral properties of speech sounds most robustly, such as whether a phoneme is a vowel or a consonant and whether a consonant's spectrum is broadband (as in plosives, like /*b*/ and /*p*/) or is dominated by acoustic energy at high frequencies (as in fricatives, like /*f*/ and /*s*/).

The present results extend these earlier findings in a critical way, suggesting that the cortical representation of both spectral and temporal cues in speech follow a common spatial coding scheme. This result is also consistent with prior reports that neural response amplitude depends on VOT(8), but such results have often involved natural speech stimuli where voicing categories varied along many other spectral acoustic dimensions besides the temporal cue(65–68). Here, the digitally synthesized VOT stimuli were tightly controlled to vary only in the relative timing of two invariant spectral cues (burst and voicing), thereby demonstrating that this temporal speech cue is encoded in the peak high-gamma response amplitude of spatially distinct neural populations in human STG.

While the present results clearly implicate a spatial/amplitude code in the cortical representation of VOT, other work has described VOT-dependent temporal response patterns that can also be used to encode voicing categories(69–71). For instance, Steinschneider and colleagues have observed neurons and neuronal populations in primate and human auditory cortices in which short VOTs elicit a single-peaked neural response, while longer VOTs elicit a double-peaked response(37, 38, 43, 72–75). Under this "local" temporal coding model, the precise temporal dynamics of the response evoked at a single cortical site could distinguish voiced from voiceless VOTs. Our examination of the timing and amplitude of three peaks in the auditory evoked local field potentials of VOT-sensitive electrodes confirmed that such patterns do appear in some electrodes (**Figure 1-figure supplements 3** and **4**), clearly demonstrating that temporal and amplitude codes for VOT are not mutually exclusive (see also temporal encoding patterns in onset latencies of V+ electrodes; **Figure 3**). However, as with spectrally-defined phonetic contrasts (e.g., plosive vs. fricative(8)), it clear that the amplitude of the peak high-gamma (and, in many cases, of the LFP) response emerged as a robust representation of voicing category and of VOT.

VOT could also be encoded in the relative timing of responses in spatially-distributed, spectrally-tuned burst- and voicing-selective neural populations. Under this "ensemble" temporal coding model(76, 77), the pattern of neural activity evoked by voiced VOTs (characterized by roughly coincident burst and voicing cues) would differ from the pattern evoked by voiceless VOTs in the precise temporal latency of the response in a vowel-selective neural population (a voicing detector) compared to the response in a plosive-selective neural population (a burst detector). However, the fact that we found cortical sites in every participant that exhibited robust category-dependent differences in their peak response amplitude rules out the possibility that at least these neural populations are merely responding to spectral cues in the burst or voicing alone.

Notably, if either (or both) of these models – a local or ensemble temporal code – were primarily responsible for the neural representation of VOT in the high-gamma range, then the selective corruption of temporal information in a classifier (**Figure 1F**) should have reduced neural decoding of voicing category to chance levels, while corrupting peak amplitude information should have had little or no effect. We found the opposite pattern of results: corrupting peak amplitude information had a devastating effect on the decoding of voicing category, while corrupting the fine temporal patterns that could have discriminated between voicing categories had no measurable impact on classifier performance. To be clear, our work does not rule out the possibility that local or ensemble temporal codes may also play a role in the cortical representation of VOT. However, it does highlight spatially-localized peak neural response amplitude as a robust code for VOT. Thus, in contrast to prior work theorizing parallel, but fundamentally different, coding schemes for spectrally- and temporally-cued phonetic features(37, 38), we demonstrate evidence for a shared representation of both by high-gamma in the human superior temporal lobe.

In order to explicitly test potential computational and physiological mechanisms that could give rise to the observed spatial coding scheme, we implemented an architecturally simple neural network model. Although it is well known that spectral information is represented by a spatial neural code from the earliest stages of auditory transduction in the cochlea(78, 79), the emergence of a spatial code for the representation of temporally-distributed cues in a transient acoustic signal poses a nontrivial computational problem. Our model highlights one parsimonious approach by which selectivity for either temporal gaps or coincidences could be implemented by biologically-inspired neurophysiological microcircuits(51–57).

We found that, just like in the neural data, gap and coincidence detector units responded to simulated voiced (/*b*/) and voiceless (/*p*/) stimuli with different response amplitudes. As such, we need not invoke any specialized temporal code to account for the representation of temporally cued phonetic features. Rather, our results provide evidence implicating a common neural coding scheme in the neural representation of behaviorally relevant speech features, whether they are embedded within the instantaneous spectrum or the fine temporal structure of the speech signal. Recent ECoG evidence suggests an even more expansive view of the fundamental role of spatial coding in cortical speech representation(80) in which different neural populations also encode pitch(81) and key properties of the speech envelope such as onsets and auditory edges(82, 83).

Crucially, although the neural network was only designed to discriminate between categories (i.e., gaps vs. coincidences), we also observed graded amplitude differences in response to different VOTs (**Figure 2H**), but only in an electrode's preferred category. These within-category patterns emerged naturally from the same computational properties that allowed the network to capture basic between-category encoding: (1) the relative responsiveness of each temporal integrator unit (***GAP***, ***COINC.***) to its various inputs (burst, voicing, and inhibition); (2) the time-dependent properties inherent to neuronal activation dynamics (e.g., decay of postsynaptic potentials towards a unit's resting activation level); and (3) the nonlinear transformation of postsynaptic inputs into response outputs (rectified linear activation function controlled by a unit's propagation threshold).

This asymmetric within-category encoding scheme closely resembled the pattern observed in real neurophysiological data, where peak response amplitude to VOTs within the same voicing category only differed within a neural population's preferred category (**Figure**

2D). This result clearly demonstrates that human nonprimary auditory cortex maintains a robust, graded representation of VOT that includes the sub-phonetic details about how a particular speech token was pronounced(44–47). Even though sub-phonetic information is not strictly necessary for mapping sound to meaning in stable, noise-free listening environments, this fine-grained acoustic detail has demonstrable effects on listeners' behavior(22–28), and modern theories of speech perception agree that perceptual learning (e.g., adaptation to accented speakers) and robust cue integration would be impossible if the perception of speech sounds were strictly categorical(16–20, 84–87). Crucially, these data suggest that the same spatial/amplitude code that is implicated in the representation of *phonetic* information (from spectral or temporal cues) can also accommodate the representation of *sub-phonetic* information in the speech signal.

The onset latency results (**Figure 3**) established an entirely novel correspondence between the real and simulated results that extended beyond the peak response window. Response onset latencies of V- electrodes were time-locked to the burst (**Figures 2B** and **3**), while responses of V+ electrodes were time-locked to voicing onset (**Figures 2C** and **3**). These highly reliable neurophysiological results neatly match specific predictions of our parsimonious model without the need to postulate additional mechanisms (**Figures 2F/G**).

The correspondence between simulated and real neural data in the onset latency results may also have implications for the question of whether the observed temporal integration is occurring locally in STG or is inherited from earlier levels of auditory processing (e.g., from midbrain or primary auditory cortex). The model's gap and coincidence detectors (***GAP***, ***COINC.***) are designed to directly simulate neural populations in the STG. Their inputs from the burst and voicing detectors are only spectrally processed, so, in the model, the temporal onset latency dynamics (**Figures 2F/G**) first arise in ***GAP*** and ***COINC.*** As such, the fact that the model's prediction is borne out in the neural data in STG (**Figures 2B/C** and **3**) is consistent with local temporal integration in STG. While these modeling results do not definitively rule out temporal integration at lower levels of the ascending auditory pathway, its potentially local emergence in high-order auditory cortex illustrates how even relatively simple computational models can be used to generate novel hypotheses, which can ultimately be tested in real neurophysiological data.

Overall, the results of these model simulations illustrate how the same network properties that transform temporal cues into a spatial code are also able to naturally explain at least three additional patterns observed within category-selective neural populations: (1) the graded encoding of VOT within a population's preferred category; (2) the lack of graded encoding of VOT within a population's non-preferred category; and (3) the early temporal dynamics of neural responses, which depend on a population's category-selectivity. Thus, the model provides an explicit, mathematical account of multiple seemingly disparate observations about the neurophysiological data, all of which arise directly from a parsimonious implementation of gap- and coincidence-detection with well-established, theoretically-motivated neuronal circuits.

The model we present is just one of many possible architectures that could capture these interesting properties of the neural response. For example, mechanisms like temporal delay lines (54, 56) could also be used to implement gap detection. Broadly, we chose to implement a simple hand-tuned neural network model to maximize our ability to explore the detailed dynamics we observed in the neural data. Our approach follows a rich history of using these types of hand-tuned models to explain a wide array of cognitive and perceptual phenomena (including the perception of VOT in speech), as exemplified by the influential TRACE model of speech perception(84). An alternative approach to modeling VOT perception is to train a neural network to distinguish voiced from voiceless sounds based on distributed activation dynamics within biologically-grounded spectral processing maps(88). Our model borrows aspects of these two approaches (hand-tuning; biological plausibility) and it extends this past work by directly modeling the time-dependent mechanisms that could give rise to continuously-varying neural responses in STG.

While the model captured several notable features of the neural data (including some for which it was not explicitly designed), we observed at least one inconsistency between the simulated and real neural responses. The model predicted VOT-dependence in the latency of the *peak* response in both ***Gap*** and ***Coinc.*** units (**Figures 2F/G**), but we did not find evidence for these fine-grained patterns in the high-gamma data (**Figures 2B/C**; see also lack of category-dependent temporal patterns during peak window: **Figure 1F**). However, it is unclear whether this is a false prediction of the model, or whether we did not observe the effect in the neural data because of, for example, poor signal-to-noise ratio for this effect. Regardless of whether the discrepancy arises from the model or the real data, it represents a gap in our mechanistic understanding of the processing of this phenomenon, and should therefore be a target for further research.

Although topographic functional organization is pervasive among many spatial neural coding schemes described in sensory neuroscience, including for the representation of spectral and temporal acoustic cues in audition (e.g., tonotopy in mammalian auditory cortex(78, 79) or chronotopy in bats(89, 90)), this functional organization seems not to extend to the spatial code for speech on the lateral temporal cortex in humans. As with tuning for spectrally-cued phonetic features(8, 82) (e.g., plosives vs. fricatives), VOT-sensitive neural populations in the present study were scattered throughout posterior and middle superior temporal gyrus with no discernible topographical map of selectivity or evidence for lateralized asymmetries(71, 91), although data limitations prevent us from ruling out this possibility entirely (for detailed results, see **Methods**).

Most of the present analyses focused on the high-gamma component of the neural response, but this work does not discount a potential role for lower-frequency oscillations in speech perception(92, 93) or in the perception of phonemes(94, 95). Indeed, it is clear from the exploratory analyses of auditory evoked local field potentials (**Figure 1-figure supplements 3** and **4**) that there do exist complex associations between VOT and the amplitude/temporal information carried in lower-frequency ranges. Future work should systematically investigate the relationship between high-gamma and other neural signals (such as the local field potential), their relative contributions to the perceptual experience of and neural representation of speech, and the importance of detailed temporal information in each (see, e.g., 42).

Finally, it is critical to distinguish our results from studies describing neural correlates of categorical speech perception, *per se* (e.g., 96). Neural responses to different VOT tokens that are members of the same voicing category can only be considered truly categorical if the responses are indiscriminable (e.g., 30, 97). In our results, acoustically distinct members of the same phonetic category *are* distinguishable in neural populations that are selective for that voicing category (**Figure 2**). In light of this graded VOT representation, the present results are best interpreted as elucidating neural mechanisms of category perception, but not necessarily categorical perception, of voiced vs. voiceless stop consonants. While limited coverage beyond the superior temporal lobe precludes us from ruling out the influence of top-down categorical perception (98–100) (possibly originating in frontal cortex (101–104)) on our results, it is notable that the model we present (which does not posit top-down effects) suggests that top-down effects may not be a necessary condition for explaining the observed non-linear encoding patterns (see also 84, 85, 105–107).

In conclusion, the present results show that spatially-discrete neural populations in human auditory cortex are tuned to detect either gaps or coincidences between spectral cues, and these sites simultaneously represent both phonetic and sub-phonetic information carried by VOT, a temporal speech cue found in almost all languages(7, 108). This demonstrates a common (spatial) neural code in STG that accounts for the representation of behaviorally relevant phonetic features embedded within the spectral and temporal structure of speech. From a simple model that transforms a temporal cue into a spatial code, we observed complex dynamics that show how a highly variable, continuous sensory signal can give rise to partially abstract, discrete representations. In this way, our findings also add to a growing body of work highlighting the critical role of human STG as a sensory-perceptual computational hub in the human speech perception system(80, 81, 96, 102, 109–112).

METHODS

Data and code availability. All data and code associated with this study and necessary for replication of its results are available under a Creative Commons license at the associated Open Science Framework project page (https://osf.io/9y7uh/).(113)

Participants. A total of seven human participants with self-reported normal hearing were implanted with high-density (128 or 256 electrodes; 4 mm pitch) multi-electrode cortical ECoG surface arrays as part of their clinical treatment for epilepsy. Placement of electrode arrays was determined based strictly on clinical criteria. For all patients who participated in this study, coverage included peri-Sylvian regions of the lateral left (n = 3) or right (n = 4) hemisphere, including the superior temporal gyrus (STG). All participants gave their written informed consent before the surgery and affirmed it at the start of each recording session. The study protocol was approved by the University of California San Francisco Committee on Human Research. Data from two additional participants were excluded from analyses because of excessive epileptiform activity (artifacts) during recording sessions.

Imaging. Electrode positions (**Figure 1D** and **Figure 1-figure supplement 2**) were determined from post-surgical computed tomography (CT) scans and manually co-registered with the patient's MRI. Details of electrode localization and warping to a standardized brain (MNI; **Figure 2A**) are described elsewhere(114).

Stimuli. Stimuli (**Figure 1B**) were generated with a parallel/cascade Klatt-synthesizer KLSYN88a using a 20-kHz sampling frequency (5ms frame width in parameter tracks). All stimulus parameters were identical across stimuli, with the exception of the time at which the amplitude of voicing began to increase (in 10ms steps from 0ms to 50ms after burst onset; **Figure 1A**). The total duration of each stimulus was 300ms regardless of VOT. The onset noise-burst was 2ms in duration and had constant spectral properties across all stimuli. The dominant frequency ranges for the vowel were: F0 = 100 Hz; F1 = 736 Hz; F2 = 1221 Hz; F3 = 3241 Hz (consistent with a vocal tract length of 13.5 cm). Formant transitions always began at 30ms. The vowel's amplitude began ramping down 250ms after stimulus onset. The stimuli are made available among this study's supplementary materials and at the associated Open Science Framework page.(113)

Behavioral Procedure. During ECoG recording, the VOT stimuli were presented monaurally over free-field loudspeakers at a comfortably listening level via a custom MATLAB script(113) in a blocked pseudorandom order. Four of seven participants simultaneously performed a behavioral task wherein they indicated on each trial whether they heard "ba" or "pa" using a touchscreen tablet (programmed using a custom MATLAB GUI). In these recording sessions, the onset of the next trial began 500ms after a response was registered or 5 seconds after the end of the stimulus (if no response was registered). In sessions where participants chose to listen to the stimuli passively (instead of participating in the behavioral task), the onset of the next trial began approximately 1000ms after the end of the previous trial. **Supplementary File 1** reports number of trials per participant.

Behavioral Analysis. For the four participants who participated in the behavioral identification task, individual trials were excluded from behavioral analysis if a participant did not make a response or if the participant's reaction time was more than 3 standard deviations from the participant's mean reaction time.

Behavioral response data were submitted to mixed effects logistic regression with a fixed effect of VOT (coded as a continuous variable) and random intercepts for participants, allowing individual participants to vary in their voicing category boundary. Using the best-fit model estimates, we calculated the overall voicing category boundary across all participants (χ = 21.0ms; **Figure 1-figure supplement 1**, **panel A**) and in the each individual participant (after adjusting for random intercept fit for each participant; **Figure 1-figure supplement 1**, **panel B**, and **Figure 1C**) as follows(115), where β_0 is the best-fit intercept and β_{VOT} is the best-fit effect of slope:

$$\chi = -\frac{\beta_0}{\beta_{VOT}}$$

ECoG signal processing.

Recording and preprocessing. Voltage fluctuations were recorded and amplified with a multichannel amplifier optically connected to a digital signal acquisition system (Tucker-Davis Technologies) sampling at approximately 3051.78 Hz. Line noise was removed via notch filtering (60 Hz and harmonics at 120 and 180 Hz) and the resulting time series for each session was visually inspected to exclude channels with excessive noise. Additionally, time segments with epileptiform activity were excluded. The time series data were then common-average referenced (CAR) to included electrodes either across an electrode's row in a 16x16 channel grid or across the entire grid depending on the technical specifications of the amplifier used for a given participant.

High-gamma extraction. The analytic amplitude of the high-gamma (HG; 70-150Hz) frequency band was extracted by averaging across eight logarithmically-spaced bands with the Hilbert transform as described elsewhere(8, 112). The HG signal was down-sampled to 400 Hz, providing temporal resolution to observe latency effects on the order of <10ms (the spacing of the VOTs of among the six experimental stimuli).

Trial alignment and extraction. Trial epochs were defined as 500ms before to 1000ms after each stimulus onset. Trials were excluded for all channels if the epoch window contained any time segments that had been marked for exclusion during artifact rejection. The HG signal for each trial was z-scored based on the mean and standard deviation of a baseline window from 500ms to 200ms before stimulus onset. A 50ms moving average boxcar filter was applied to the HG time series for each trial.

Local field potential extraction. Data for analyses of auditory evoked local field potentials consisted of the same raw voltage fluctuations (local field potential), preprocessed with identical notch filtering, CAR, artifact/channel rejection, and down-sampling (to 400 Hz). Trial epochs (500ms before to 1000ms after each stimulus onset) were not z-scored.

Electrode selection.

Speech-responsive electrodes. An electrode was included in our analyses if (1) it was anatomically located on the lateral temporal lobe (either superior or middle temporal gyrus), and (2) the electrode's grand mean HG (across all trials and timepoints during a window 100-300ms after stimulus onset) exceeded one standard deviation of the baseline window's HG activity. Across all seven participants, 346 electrodes met these criteria (*speech-responsive electrodes*; **Supplementary File 1; Figure 1-figure supplement 2**).

Peak neural response. The timepoint at which each speech-responsive electrode reached its maximum HG amplitude (averaged across all trials, irrespective of condition) was identified as that electrode's peak, which was used in the subsequent peak encoding analyses. Because we were focused on auditory-evoked activity in the temporal lobe, the search for an electrode's peak was constrained between 0 and 500ms after stimulus onset. Electrode size in **Figure 1D** and **Figure 1-figure supplement 2** corresponds to this peak HG amplitude for each speech-responsive electrode.

VOT-sensitive electrodes. To identify electrodes where the peak response depended on stimulus VOT (*VOT-sensitive electrodes*), we computed the nonparametric correlation coefficient (Spearman's ρ) across trials between VOT and peak HG amplitude. Because nonparametric (rank-based) correlation analysis measures the monotonicity of the relationship between two variables, it represents an unbiased ("model-free") indicator of amplitude-based VOT encoding, whether the underlying monotonic relationship is categorical, linear, or follows some other monotonic function (Bishara & Hittner, 2012). This procedure identified 49 VOT-sensitive electrodes across all seven participants ($p < 0.05$; **Figure 2A** and **Figure 1-figure supplement 2**; **Supplementary File 1**). Electrode color in **Figure 1D** and **Figure 1-figure supplement 2** corresponds to the correlation coefficient at each electrode's peak (min/max $\rho = \pm 0.35$), thresholded such that all speech-responsive electrodes with non-significant ($p > 0.05$) correlation coefficients appear as white.

This set of VOT-sensitive sites was then divided into two sub-populations based on the sign of each electrode's correlation coefficient (ρ): voiced-selective (V+) electrodes (n = 33) had significant $\rho < 0$, indicating that shorter (more /*b*/-like; voiced) VOTs elicited stronger peak HG responses; voiceless-selective (V-) electrodes (n = 16) had significant $\rho > 0$, indicating that longer (more /*p*/-like; voiceless) VOTs elicited stronger peak HG responses.

Across VOT-sensitive electrodes, the mean peak occurred 198.8ms after stimulus onset (SD = 42.3ms). The semi-transparent grey boxes in **Figures 1E** and **2B/C** illustrate this peak window (mean peak ± 1 SD).

Analysis of VOT-sensitive electrodes.

Encoding of voicing category. Electrodes that exhibit a monotonic relationship between VOT and peak HG amplitude should also be likely to exhibit a categorial distinction between shorter (voiced) and longer (voiceless) VOTs. We conducted two analyses that confirmed this expectation. In each analysis, we computed a nonparametric test statistic describing the discriminability of responses to voiced vs. voiceless stimuli at each electrode's peak (*z*-statistic of Mann-Whitney rank-sum test) and then tested whether the population of test statistics for each group of electrodes (V- and V+) differed reliably from zero (Wilcoxon signed-rank tests). In the first analysis, voicing category was defined based on the psychophysically determined category boundary (voiced: 0-20ms VOTs; voiceless: 30-50ms VOTs), which allowed us to include all VOT-sensitive electrodes (n = 49) in the analysis, including electrodes from participants who did not complete the behavioral task (3/7 participants).

In the second analysis, a trial's voicing category was determined based on the actual behavioral response recorded for each trial (irrespective of VOT), so this analysis was not dependent on the assumption that the VOT continuum can be divided into two categories based on the average boundary calculated across participants. This analysis examined the subset of trials with behavioral responses and the subset of VOT-sensitive electrodes found in the four participants with behavioral data (n = 27; 12 V- electrodes, 15 V+ electrodes) (**Supplementary File 1**).

Given the strong correspondence between the categorically defined VOT stimulus ranges (0-20ms vs. 30-50ms VOTs) and identification behavior (e.g., **Figure 1C**), the agreement between these results was expected.

Significance bars for the two example STG electrodes in one participant (e1 and e2; **Figure 1E**) we computed to illustrate the temporal dynamics of category selectivity. In these electrodes, we conducted the test of between-category encoding (Mann-Whitney rank-sum test; first analysis) at every timepoint during the trial epoch (in addition to the electrodes' peaks). Bars plotted for each electrode in **Figure 1E** begin at the first timepoint after stimulus onset where the significance level reached $p < 0.005$ and ends at the first point thereafter where significance fails to reach that threshold (e1: 140 to 685ms post onset; e2: 65 to 660ms post onset).

Encoding of VOT within voicing categories. Because VOT-sensitive electrodes were identified via nonparametric correlation analysis (Spearman's ρ) across all VOTs, the monotonic relationship between VOT and peak HG amplitude at these sites could be driven by the observed phonetic (between-category) encoding of voicing without any robust sub-phonetic (within-category) encoding of VOT. To assess sub-phonetic encoding of VOT in the peak response amplitude of VOT-sensitive electrodes, we computed the rank-based correlation (Spearman's ρ) between VOT and HG amplitude at each electrode's peak separately for trials in each voicing category (0-20ms vs. 30-50ms VOTs). The statistical reliability of within-category encoding was summarized by computing a test-statistic (t) for every correlation coefficient ($\rho_{0\text{-}20}$ and $\rho_{30\text{-}50}$ for each VOT-sensitive electrode) as follows:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

where n is the number of trials with VOTs in a given voicing category. The resulting set of test statistics (one per voicing category per VOT-sensitive electrode) served as the basis for the following analyses of peak within-category encoding.

For each group of electrodes (V- and V+), we tested whether the encoding of VOT within each voicing category differed reliably from 0 (Wilcoxon signed-rank tests). We also conducted a Wilcoxon signed-rank test for each electrode group that compared the within-category correlation t-statistics for voiceless and voiced categories.

The above tests addressed the encoding properties of one electrode group at a time (either V- or V+ electrodes). Finally, a pair of Wilcoxon signed-rank tests combined across the full set of VOT-sensitive electrodes (n = 49) to summarize the within-category VOT encoding results within electrodes' (1) preferred and (2) non-preferred categories. In order to conduct this "omnibus" test, we multiplied the correlation *t*-statistics for all V+ electrodes (for tests within each category) by -1. This simple transformation had the consequence of ensuring that positive correlation statistics always indicate stronger peak HG responses to VOTs that were closer to the endpoint of an electrode's preferred category.

Visualizations of within-category VOT encoding. To visualize the pattern of within-category encoding of VOT in the peak HG amplitude of V- and V+ electrodes, we computed a normalized measure of the peak response amplitude to each VOT stimulus for each VOT-sensitive electrode. **Figures 2B** and **2C** show the full time series of the average (± SE) evoked responses of V- and V+ electrodes to all six VOT stimuli. To show encoding patterns across electrodes with different peak amplitudes, each electrode's activity was normalized by its peak HG (grand mean across all VOTs). **Figure 2D** shows the amplitude of the average response evoked by a given VOT at a given electrode's peak relative to the average response evoked by the other VOT stimuli, or *peak HG (% of max),* averaged across electrodes in each group (V-, **left**; V+, **right**) and participants (± SE). For each electrode, the mean HG amplitude evoked by each VOT at the peak was scaled and normalized by subtracting the minimum across all VOTs and dividing by the maximum across all VOTs after scaling.

Neural response latency. The normalized HG responses used for **Figures 2B/C** were also used for the analysis of onset latency effects (**Figure 3**): *HG (normalized)* (**Figures 2B/C**) and *HG (% of peak)* (**Figure 3A**) are computationally equivalent. Neural response onset latency for an electrode was defined as the first timepoint at which its average response to a given VOT stimulus exceeded 50% of its peak HG (based on the peak of the grand average response across all VOTs). A bootstrapping with resampling procedure was employed to estimate the onset latencies of responses to different VOTs at each electrode and to assess any possible relationship between onset latency and VOT. During each sampling step in this procedure (1000 bootstrap samples), we computed the average time series of the normalized HG response to each VOT, the onset latency for the response to each VOT, and the nonparametric correlation (Spearman's ρ) between onset latency and VOT. Wilcoxon signed-rank tests asked whether the population of bootstrapped correlation coefficient estimates for each electrode group reliably differed from zero. A Mann-Whitney rank-sum test compared the VOT-dependency of response onset latency between electrode groups. Color-coded horizontal bars below the neural data in **Figure 3A** show onset latency estimates (mean ± bootstrap standard error) for responses to each VOT at two example electrodes. All electrodes were included in the analyses, but the bootstrapped correlation coefficient estimates for two V+ electrodes that were outliers (>3 SDs from median) were excluded from the visualized range of the box-plot's whiskers in **Figure 3B**.

Population-based neural classification. For each participant, we trained a set of multivariate pattern classifiers (linear discriminant analysis with leave-one-out cross validation) to predict trial-by-trial voicing category (/*b*/: 0-20ms VOTs vs. /*p*/: 30-50ms VOTs) using HG activity across all speech-responsive electrodes on the temporal lobe during a time window around the peak neural response. The peak window was defined as beginning 150ms and ending 250ms after stimulus onset, selected based on the average and standard deviation of the peaks across all VOT-sensitive electrodes. We created four separate classifiers for each participant that allowed us to evaluate the contribution of amplitude and temporal structure to voicing category encoding (**Figure 1F**).

To corrupt the reliability of any spatially-localized amplitude information about whether the VOT stimulus presented to a participant on a given trial was a /*b*/ or a /*p*/, the neural responses at every electrode on every trial were normalized so that the average response to a /*b*/ and the average response to a /*p*/ reached the same amplitude at each electrode's peak. Specifically, for each electrode, we found its peak (timepoint where the grand average HG time series across all trials reached its maximum), calculated the mean HG amplitude across all trials for VOTs within each category at that peak, and divided the HG values for every timepoint in a trial's time series by the peak HG amplitude for that trial's category. This amplitude normalization procedure forces the average amplitude of the neural response across all trials of /*b*/ and of /*p*/ to be equal at each electrode's peak, while still allowing for variation in the amplitude of any individual trial at the peak.

To corrupt the reliability of any timing information during the peak response window about whether the VOT stimulus presented to a participant on a given trial was a /*b*/ or a /*p*/, the timing of the neural response on every trial (across all electrodes) was randomly shifted in time so that the trial could begin up to 50ms before or after the true start of the trial. Specifically, for each trial, a jitter value was drawn from a discrete (integer) uniform random distribution ranging between -20 to 20 (inclusive range) ECoG time samples (at 400 Hz, this corresponds to ±50ms, with a mean jitter of 0ms), and the HG time series for all electrodes on that trial was moved backward or forward in time by the number of samples dictated by the trial's jitter value. This temporal jittering procedure has the effect of changing whether the peak response window for a given trial is actually drawn from 100-200ms after stimulus onset, 200-300ms after stimulus onset, or some other window in between.

Crucially, this procedure will misalign any reliable, category-dependent differences in peak timing or temporal dynamics within individual electrodes or temporal patterns or relationships that exist across distributed electrodes. For instance, the peak window overlaps with a window during which past work examining intracranial auditory evoked local field potentials found evidence of waveform shape differences between responses of single electrodes to voiced and voiceless stimuli (single- vs. double-peaked responses (see, e.g., Fig. 10 of 43). If similar temporal differences in waveform shape existed in the present high-gamma data, the temporal jittering procedure would detect a contribution of temporal information to decoding. Moreover, to the extent that the peak of a trial's evoked high-gamma response occurs during or close to the peak window (either within one electrode ["local" temporal code] or across multiple electrodes in the same participant ["ensemble" temporal code]), the temporal jittering procedure would disrupt the reliability of this information to reveal the contribution of peak latency information to decoding accuracy. On the other hand, if the peak responses to stimuli from distinct voicing categories differ in the amplitude of the HG response at VOT-sensitive cortical sites, and if these differences persist throughout much of the peak window, then this temporal jittering procedure is unlikely to prevent the classifier from learning such differences.

For each participant, we trained one classifier where neither amplitude nor timing information were corrupted (+Amplitude/+Timing), one where only timing information was corrupted (+Amplitude/-Timing), one where only amplitude information was corrupted (-Amplitude/+Timing), and one where both were corrupted (-Amplitude/-Timing; here, amplitude normalization preceded temporal jittering). With each of these datasets, we then performed dimensionality reduction to minimize overfitting using spatiotemporal principal component analysis on the ECoG data for every electrode and all timepoints within the peak window (retaining PCs accounting for 90% of the variance across trials of all VOTs). Finally, training and testing of the linear discriminant analysis classifiers were conducted iteratively, holding out a single trial, training a classifier to predict voicing category using all other trials, and then predicting the voicing category of the held-out trial. For each participant and for each classifier, accuracy was the proportion of held-out trials that were correctly labeled. Wilcoxon signed-rank tests assessed and compared accuracy levels (across participants) achieved by the different models.

Computational neural network model.

Overview of architecture and dynamics. A simple five-node, localist neural network (**Figure 2E**) was hand-connected to illustrate how time-dependent properties of neuronal units and their interactions can transform a temporal cue into a spatial code (responses of different amplitudes to different VOTs at distinct model nodes). A gap detector received excitatory input from both a burst detector and voicing detector, as well as input from an inhibitory node that only received excitatory input from the burst detector. This represented an implementation of a slow inhibitory postsynaptic potential (slow IPSP) circuit(51, 52, 60, 61). A coincidence detector received excitatory input from the burst and voicing detectors.

Network Connectivity. Weights between units in this sparsely connected, feedforward network were set according to a minimalist approach. All excitatory connections from the burst detector (to the inhibitory node, the gap detector, and the coincidence detector) had identical weights. All excitatory connections from the voicing detector (to the gap detector and the coincidence detector) had identical weights (stronger than from burst detector). **Figure 2-figure supplement 1** indicates all nonzero connection weights between the network's nodes, as illustrated in **Figure 2E**.

Leaky-integrator dynamics. At the start of the model simulations, prior to the onset of any stimulus ($t = 1$), the activation level $a_i(t)$ of each node i was set to its resting level (ρ_i). Simulations ran for 100 cycles, with 1 cycle corresponding to 10ms. On each subsequent cycle ($t \in [2,100]$), activation levels of every node in the model were updated iteratively in two steps, as described in the following algorithm:

(1) **Decay:** For every node i with prior activation level $a_i(t-1)$ that differs from ρ_i, $a_i(t)$ decays towards ρ_i by its decay rate (λ_i) without overshooting ρ_i.

(2) **Sum Inputs:** For every node i, the total excitatory and inhibitory inputs are summed. This includes both model-external (clamped) inputs (i.e., from stimuli presented to the model) on the current cycle t and model-internal inputs from other nodes based on their activation level on the prior cycle $a_j(t-1)$. Inputs from a presynaptic node j can only affect the postsynaptic node i if its prior activation $a_j(t-1)$ exceeds the presynaptic node's propagation threshold (θ_j). Summation of model-internal inputs within i is

weighted by the connection weights from the various presynaptic nodes (**Figure 2-figure supplement 1**): $\sum_j w_{ji} * a_j(t-1)$. The new activation level $a_i(t)$ is bounded by the node's minimum (m_i) and maximum (M_i) activation levels, irrespective of the magnitude of the net effect of the inputs to a node.

All activation parameters for all nodes are listed in **Supplementary File 2**. Minimum, maximum, and resting activation levels were identical across all units. Decay rates and propagation thresholds were identical across the burst and voicing detectors and the inhibitory node. The integrator units (gap and coincidence detectors) decayed more slowly than the other units, which could only affect other model nodes during one cycle. Activation levels in the coincidence detector had to reach a higher level (propagation threshold) to produce model outputs than in the gap detector, a difference which allowed the gap detector to register the fast suprathreshold response characteristic of slow IPSP circuits and allowed the coincidence detector to register a coincidence only when both burst and voicing were detected simultaneously or at a short lag.

Model inputs. Two inputs were clamped onto the model in each simulation, representing the onset of the burst and of voicing (**Figure 1A**). The voicing input was only clamped onto the voicing detector at the onset of voicing. **Supplementary File 3** illustrates vectors describing each of the simulated VOT inputs.

Sensitivity of model dynamics to variations in hand-tuned model parameters. Although most of the parameters of the model are theoretically uninteresting and were set to default levels (see **Supplementary File 2**), analysis of parameter robustness for the model revealed four primary sensitivities based on the relative values set for certain specific parameters. (1) and (2) below involve the propagation thresholds [θ] of the temporal integrator units (***Gap***, ***Coinc.***), which allow the model to achieve gap and coincidence detection. (3) and (4) below involve the rate of decay of activation [λ] of the temporal integrator units, which dictate where along the VOT continuum the boundary between voicing categories lies.

(1) **Propagation threshold [$\boldsymbol{\theta}$] of coincidence detector unit (*Coinc.*):** In our model, coincidence detection is achieved by preventing the coincidence detector (***Coinc.***) from propagating an output in response to the burst until the voicing has arrived (hence responding with a higher-than-minimum peak amplitude only when the voicing is coincident with or arrives shortly after the burst). Thus, the propagation threshold for ***Coinc.*** ($\theta_{\boldsymbol{Coinc.}}$) must be <u>greater than</u> the connection weight from the burst-detector to ***Coinc.*** ($w_{Burst \rightarrow Coinc.}$).

(2) **Propagation threshold [$\boldsymbol{\theta}$] of gap detector unit (*Gap*):** On the other hand, the propagation threshold for the gap detector [***Gap***] (θ_{Gap}) must be <u>less than</u> the connection weight from the burst-detector to ***Gap*** ($w_{Burst \rightarrow Gap}$) to register the fast suprathreshold response characteristic of slow IPSP circuits.

The primary factor affecting the location of the boundary between voiced (short VOTs) and voiceless (long VOTs) categories is the time-dependent rate of decay of postsynaptic potentials in ***Gap*** and ***Coinc.*** towards the unit's resting activation level.

(3) **Rate of decay of activation [$\boldsymbol{\lambda}$] in *Coinc.* in comparison to connection weights from inputs to *Coinc.*:** For ***Coinc.***, the boundary is the VOT value after which there is no longer any additional boost to its peak amplitude from the initial burst, and this requires

the decay rate of ***Coinc.*** ($\lambda_{Coinc.}$) and the connection weight from the burst-detector to ***Coinc.*** ($w_{Burst \rightarrow Coinc.}$) to be in balance. Increasing $\lambda_{Coinc.}$ or decreasing $w_{Burst \rightarrow Coinc.}$ (independently) will move the boundary earlier in time.

(4) **Rate of decay of activation [λ] in *Gap* in comparison to connection weights from inputs to *Gap*:** Similarly, for ***Gap***, the category boundary is the VOT value before which the remaining influence of the initial inhibition is still so strong that the arrival of voicing input cannot exceed θ_{Gap}. Increasing λ_{Gap}, decreasing $w_{Inhib. \rightarrow Gap}$, or increasing $w_{Voicing \rightarrow Gap}$ (independently) would each move the boundary earlier in time. All three of these parameters are in balance in these hand-tuned parameter settings.

It is critical to note that, for all of these cases where the hand-tuned parameter settings are in balance, the balance is required for the model to achieve gap and coincidence detection and/or to determine the position of the VOT boundary between categories. This was all the model was designed to do. No parameters were hand-tuned to achieve the other response properties (e.g., asymmetric within-category encoding, onset latency dynamics).

Analysis of auditory evoked local field potentials.

Identification of key LFP peaks. We identified 3 peaks of the grand mean auditory evoked local field potential (AEP), which were consistent with AEP peaks previously described in the literature(41, 42): P_α (positive deflection approximately 75-100 ms after stimulus onset), N_α (negative deflection approximately 100-150 ms after stimulus onset), and P_β (positive deflection approximately 150-250 ms after stimulus onset) (see **Figure 1-figure supplements 3** and **4**).

Bootstrapping approach. For each VOT-sensitive electrode (speech-responsive electrodes whose peak high-gamma amplitude was correlated with VOT), a bootstrapping with resampling procedure was used to estimate the latencies and amplitudes of each peak of the AEP elicited by trials from each VOT condition. During each sampling step in this procedure (1000 bootstrap samples), we computed the average time series of the AEP for each VOT (**Figure 1-figure supplement 4**, **panels I-L**), the ECoG samples of the time series during each of three time-ranges with the maximum (for positive peaks) or minimum (for the negative peak) mean voltage values for each VOT, and six correlation coefficients (Pearson's *r* between VOT and amplitude/latency for each peak; see **Figure 1-figure supplement 4**, **panels M-T**).

Details of peak-finding. P_α was defined as the maximum mean voltage from 0-150 ms after stimulus onset, N_α was defined as the minimum mean voltage from 75-200 ms after stimulus onset, and P_β was defined as the maximum mean voltage from 150-250 ms after stimulus onset. To aid peak detection and enforce sequential ordering of the peaks, time ranges for the latter two peaks (N_α, P_β) were further constrained on a per-sample basis by setting the minimum bound of the search time range to be the time of the previous peak (i.e., the earliest possible times for N_α and P_β were P_α and N_α, respectively). For a given sample, if a peak occurred at either the earliest possible or latest possible time, it was assumed that the peak was either not prominent or did not occur during the defined time range for this electrode/VOT, so that sample was ignored in the analysis for that peak and any subsequent peaks. Because correlation coefficients for each peak were computed over just 6 VOTs in each sample, exclusion of a peak latency/amplitude value for one VOT condition resulted in exclusion of the all conditions for that peak for that sample. Finally, if more than 50% of the bootstrap samples were excluded for a given peak in a given electrode, no samples for that electrode/peak pair were not included in the analysis (see, e.g., P_β for e4 in **Figure 1-figure supplement 4**, **panels H/P/T**).

Analysis of bootstrapped correlation estimates. For each remaining VOT-sensitive electrode/peak pair, we determined whether or not the latency and/or amplitude of the peak was significantly associated with VOT by evaluating whether the 95% confidence interval (95% CI) across all included bootstrapped estimates of the correlation coefficient excluded 0 (taking the highest density interval of the bootstrapped statistics) (**Figure 1-figure supplement 3**, **panel B**). These exploratory analyses did not undergo multiple comparison correction.

Detailed results of analysis of AEPs. The exploratory analyses of correlations between VOT and the latency and/or amplitude of three peaks of the AEP in all VOT-sensitive electrodes revealed four overall conclusions:

1. Comparison of the AEPs evoked by different VOTs shows that there exist associations between stimulus VOT and the amplitude/temporal information in local field potential (LFP). Among electrodes that robustly encode voicing in their peak high-gamma amplitude (i.e., VOT-sensitive electrodes), these associations between VOT and LFP features are complex and highly variable (**Figure 1-figure supplements 3** and **4**).
2. Replicating prior results regarding VOT encoding by AEPs (e.g., 43), we find that some electrodes (e.g., e1 in **Figure 1-figure supplement 4**, **panels E/I**) exhibit temporal encoding of VOT in the latency of various peaks of the AEP. In some electrodes, the nature of this temporal code is straightforward (e.g., in e1, the latency of N_α is delayed by ~10ms for every additional 10ms of VOT duration; **Figure 1-figure supplement 4**, **panel M**), but – more often – the relationship between VOT and peak latency is less direct (**Figure 1-figure supplement 4**, **panels N-P**).
3. Among electrodes that encode VOT in their peak high-gamma amplitude, there exist many more electrodes that *do not* encode VOT in these temporal features of the AEP (**Figure 1-figure supplement 3**), supporting a prominent role for the peak high-gamma amplitude in the neural representation of voicing and of VOT.
4. Besides the timing of the various AEP peaks, there also exist many electrodes that encode VOT in the amplitude of those peaks (**Figure 1-figure supplement 3**). The encoding patterns are often visually similar to the encoding patterns observed in high-gamma (i.e., graded within the electrode's preferred voicing category; see **Figure 1-figure supplement 4**, **panels Q-S**). However, there are also many electrodes that do encode VOT in their peak high-gamma amplitude but *not* in these amplitude features of the LFP (**Figure 1-figure supplement 3**, **panel B**; compare, e.g., **Figure 1-figure supplement 4**, **panels D** vs. **H**).

Supplementary analyses of spatial patterns of VOT effects. Of the 49 VOT-sensitive electrodes, 76% were located posterior to the lateral extent of the transverse temporal sulcus (defined as $y \geq 6$ in MNI coordinate space based on projection of the sulcus onto the lateral STG in the left hemisphere). This is the same region that is densely populated with neural populations that are tuned for other phonetic features (e.g., manner of articulation(8, 82)). Mann-Whitney rank-sum tests showed that there was no significant difference in the localization of voiceless-selective (V-) versus voiced-selective (V+) electrodes along either the anterior-posterior axis (*y*-

dimension in MNI coordinate space; U = 342, z = -1.23, p = 0.22) or the dorsal-ventral axis (z-dimension in MNI coordinate space; U = 414, z = 0.29, p = 0.77).

Although no regional patterns were visually apparent, we tested for hemispheric differences in relative prevalence of VOT-sensitive sites or in voicing category selectivity. Of the seven participants (all of whom had unilateral coverage), four had right hemisphere coverage (57%), and these four patients contributed 28 of the 49 VOT-sensitive electrodes identified in this study (57%) (see **Figure 2A** and **Figure 1-figure supplement 2**; **Supplementary File 1**). Pearson's χ^2 tests confirmed there was no difference in the rate of VOT-sensitive sites ($\chi^2(1) = 0.15$, $p = 0.70$) or in the proportion of VOT-sensitive sites that were selective for each category ($\chi^2(1) = 1.74$, $p = 0.19$) as a function of hemisphere. Thus, consistent with past ECoG work examining spatial patterns of STG encoding for other phonetic features(e.g., 82), we found no evidence that the observed spatial/amplitude code reflected any topographical organization nor any lateralized asymmetries in the encoding of VOT, although data limitations prevent us from ruling out this possibility entirely.

ACKNOWLEDGEMENTS

We are grateful to John Houde, who provided the stimuli used in this work, and to all members of the Chang Lab for helpful comments throughout this work. This work was supported by European Commission grant FP7-623072 (M.J.S.); and NIH grants R01-DC012379 (E.F.C.) and F32-DC015966 (N.P.F.). E.F.C. is a New York Stem Cell Foundation-Robertson Investigator. This research was also supported by The William K. Bowes Foundation, the Howard Hughes Medical Institute, The New York Stem Cell Foundation and The Shurl and Kay Curci Foundation.

REFERENCES

1. Stevens KN (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *J Acoust Soc Am* 111(4):1872–1891.
2. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. *Psychol Rev* 74(6):431–461.
3. Shannon R V, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. *Science* 270(5234):303–4.
4. Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci* 336(1278):367–73.
5. Klatt DH (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J Acoust Soc Am* 59(5):1208–1221.
6. Liberman AM, Delattre PC, Cooper FS (1958) Some Cues for the Distinction Between Voiced and Voiceless Stops in Initial Position. *Lang Speech* 1(3):153–167.
7. Lisker L, Abramson AS (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word J Int Linguist Assoc* 20(3):384–422.
8. Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science (80-)* 343(6174):1006–1010.
9. Miller JL, Green KP, Reeves A (1986) Speaking Rate and Segments: A Look at the Relation between Speech Production and Speech Perception for the Voicing Contrast. *Phonetica* 43(1–3):106–115.
10. Kessinger RH, Blumstein SE (1997) Effects of speaking rate on voice-onset time in Thai, French, and English. *J Phon* 25(2):143–168.
11. Klatt DH (1975) Voice onset time, frication, and aspiration in word-initial consonant clusters. *J Speech Hear Res* 18:686–706.
12. Lisker L, Abramson AS (1967) Some effects of context on voice onset time in English stops. *Lang Speech* 10(1):1–28.
13. Allen JS, Miller JL, DeSteno D (2003) Individual talker differences in voice-onset-time. *J Acoust Soc Am* 113(1):544.
14. Flege JE, Eefting W (1986) Linguistic and Developmental Effects on the Production and Perception of Stop Consonants. *Phonetica* 43(4):155–171.
15. Fox NP, Reilly M, Blumstein SE (2015) Phonological neighborhood competition affects spoken word production irrespective of sentential context. *J Mem Lang* 83:97–117.
16. Miller JL, Volaitis LE (1989) Effect of speaking rate on the perceptual structure of a phonetic category. *Percept Psychophys* 46(6):505–512.
17. Clayards MA, Tanenhaus MK, Aslin RN, Jacobs RA (2008) Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108(3):804–809.
18. Kleinschmidt DF, Jaeger TF (2015) Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol Rev* 122(2). doi:10.1037/a0038695.
19. McMurray B, Jongman A (2011) What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychol Rev* 118(2):219–246.
20. Toscano JC, McMurray B (2010) Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn Sci* 34(3):434–464.

21. Fox NP, Blumstein SE (2016) Top-down effects of syntactic sentential context on phonetic processing. *J Exp Psychol Hum Percept Perform* 42(5):730–741.

22. Kuhl PK (1991) Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Percept Psychophys* 50(2):93–107.

23. Carney AE, Widin GP, Viemeister NF (1977) Noncategorical perception of stop consonants differing in VOT. *J Acoust Soc Am* 62(4):961–970.

24. Pisoni DB, Tash J (1974) Reaction times to comparisons within and across phonetic categories. *Percept Psychophys* 15(2):285–290.

25. Massaro DW, Cohen MM (1983) Categorical or continuous speech perception: A new test. *Speech Commun* 2:15–35.

26. Andruski JE, Blumstein SE, Burton MW (1994) The effect of subphonetic differences on lexical access. *Cognition* 52(3):163–187.

27. McMurray B, Tanenhaus MK, Aslin RN (2002) Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86(2):B33–B42.

28. Schouten B, Gerrits E, van Hessen A (2003) The end of categorical perception as we know it. *Speech Commun* 41(1):71–80.

29. Klatt DH (1980) Software for a cascade/parallel formant synthesizer. *J Acoust Soc Am* 67(3):971–995.

30. Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol* 54(5):358–368.

31. Liberman AM, Harris KS, Kinney JA, Lane H (1961) The discrimination of the relative onset time of the components of certain speech and non-speech patterns. *J Exp Psychol* 61:379–388.

32. Kronrod Y, Coppess E, Feldman NH (2016) A unified account of categorical effects in phonetic perception. *Psychon Bull Rev* 23(6):1681–1712.

33. Chang EF (2015) Towards Large-Scale, Human-Based, Mesoscopic Neurotechnologies. *Neuron* 86(1):68–78.

34. Crone N, et al. (2001) Induced electrocorticographic gamma activity during auditory perception. *Clin Neurophysiol* 112:565–582.

35. Steinschneider M, Fishman YI, Arezzo JC (2008) Spectrotemporal Analysis of Evoked and Induced Electroencephalographic Responses in Primary Auditory Cortex (A1) of the Awake Monkey. *Cereb Cortex* 18(3):610–625.

36. Ray S, Maunsell JHR (2011) Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biol* 9(4):e1000610.

37. Steinschneider M, Volkov IO, Noh MD, Garell PC, Howard MA (1999) Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *J Neurophysiol* 82(5):2346–2357.

38. Steinschneider M, Nourski K V., Fishman YI (2013) Representation of speech in human auditory cortex: Is it special? *Hear Res* 305(1):57–73.

39. Buzsáki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat Rev Neurosci* 13(6):407–420.

40. Einevoll GT, Kayser C, Logothetis NK, Panzeri S (2013) Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nat Rev Neurosci* 14(11):770–785.

41. Howard MA, et al. (2000) Auditory cortex on the human posterior superior temporal gyrus. *J Comp Neurol* 416(1):79–92.

42. Nourski K V, et al. (2015) Sound identification in human auditory cortex: Differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain Lang* 148:37–50.

43. Steinschneider M, et al. (2011) Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb Cortex* 21(Cv):2332–47.

44. Blumstein SE, Myers EB, Rissman J (2005) The perception of voice onset time: an fMRI investigation of phonetic category structure. *J Cogn Neurosci* 17(9):1353–66.

45. Toscano JC, Mcmurray B, Dennhardt J, Luck SJ (2010) Continuous perception and graded categorization: electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychol Sci* 21(10):1532–1540.

46. Toscano JC, Anderson ND, Fabiani M, Gratton G, Garnsey SM (2018) The time-course of cortical responses to speech revealed by fast optical imaging. *Brain Lang* 184:32–42.

47. Frye RE, et al. (2007) Linear coding of voice onset time. *J Cogn Neurosci* 19(9):1476–1487.

48. Myers EB (2007) Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. *Neuropsychologia* 45(7):1463–1473.

49. Ferster D, Spruston N (1995) Cracking the neuronal code. *Science (80-)* 270(5237):756–757.

50. Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. *Curr Opin Neurobiol* 4(4):569–579.

51. Buonomano D V., Merzenich MM (1995) Temporal information transformed into a spatial code by a neural network with realistic properties. *Science (80-)* 267(February):1028–1030.

52. Gao X, Wehr M (2015) A Coding Transformation for Temporally Structured Sounds within Auditory Cortical Neurons. *Neuron* 86(1):292–303.

53. Eggermont JJ (2000) Neural Responses in Primary Auditory Cortex Mimic Psychophysical, Across-Frequency-Channel, Gap-Detection Thresholds. *J Neurophysiol* 84(3):1453–1463.

54. Carr CE (1993) Processing of Temporal Information in the Brain. *Annu Rev Neurosci* 16(1):223–243.

55. Konishi M (2003) Coding of Auditory Space. *Annu Rev Neurosci* 26(1):31–55.

56. Rauschecker JP (2014) Is there a tape recorder in your head? How the brain stores and retrieves musical melodies. *Front Syst Neurosci* 8:149.

57. Rauschecker JP (1998) Cortical processing of complex sounds. *Curr Opin Neurobiol* 8(4):516–521.

58. McClelland JL, Rumelhart DE (1981) An interactive activation model of context effects in letter perception. *Psychol Rev* 88:375–407.

59. McClelland JL, Mirman D, Bolger DJ, Khaitan P (2014) Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition. *Cogn Sci* 38(6):1139–89.

60. Douglas RJ, Martin KA (1991) A functional microcircuit for cat visual cortex. *J Physiol* 440:735–69.

61. McCormick DA (1989) GABA as an inhibitory neurotransmitter in human cerebral cortex. *J Neurophysiol* 62(5):1018–27.

62. Margoliash D, Fortune ES (1992) Temporal and harmonic combination-sensitive neurons in the zebra finch’s HVc. *J Neurosci* 12(11):4309–26.

63. Peña JL, Konishi M (2001) Auditory Spatial Receptive Fields Created by Multiplication.

Science (80-) 292(5515):249–252.

64. Peña JL, Konishi M (2002) From Postsynaptic Potentials to Spikes in the Genesis of Auditory Spatial Receptive Fields. *J Neurosci* 22(13):5652–5658.
65. Lisker L (1986) “Voicing” in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang Speech* 29(1):3–11.
66. Soli SD (1983) The role of spectral cues in discrimination of voice onset time differences. *J Acoust Soc Am* 73(6):2150–2165.
67. Stevens KN, Klatt DH (1974) Role of formant transitions in the voiced-voiceless distinction for stops. *J Acoust Soc Am* 55(3):653–659.
68. Summerfield Q, Haggard M (1977) On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *J Acoust Soc Am* 62(2):435–448.
69. Eggermont JJ (1995) Representation of a voice onset time continuum in primary auditory cortex of the cat. *J Acoust Soc Am* 98(2):911–920.
70. Eggermont JJ, Ponton CW (2002) The neurophysiology of auditory perception: from single units to evoked potentials. *Audiol Neurootol* 7(2):71–99.
71. Liégeois-Chauvel C, De Graaf JB, Laguitton V, Chauvel P (1999) Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cereb Cortex* 9(5):484–496.
72. Steinschneider M, et al. (2005) Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic. *Cereb Cortex* 15:170–186.
73. Steinschneider M, Schroeder CE, Arezzo JC, Vaughan HG (1994) Speech-evoked activity in primary auditory cortex: effects of voice onset time. *Electroencephalogr Clin Neurophysiol* 92:30–43.
74. Steinschneider M, Schroeder CE, Arezzo JC, Vaughan HG (1995) Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: temporal response patterns. *Brain Lang* 48(3):326–340.
75. Steinschneider M, Fishman YI, Arezzo JC (2003) Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex. *J Acoust Soc Am* 114(1):307–321.
76. Theunissen FE, Miller JP (1995) Temporal encoding in nervous systems: A rigorous definition. *J Comput Neurosci* 2(2):149–162.
77. Engineer CT, et al. (2008) Cortical activity patterns predict speech discrimination ability. *Nat Neurosci* 11(5):603–8.
78. Eggermont JJ (2001) Between sound and perception: reviewing the search for a neural code. *Hear Res* 157(1–2):1–42.
79. Oxenham AJ (2018) How We Hear: The Perception and Neural Coding of Sound. *Annu Rev Psychol* 69(1):27–50.
80. Yi HG, Leonard MK, Chang EF (2019) The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron* 102(6):1096–1110.
81. Tang C, Hamilton LS, Chang EF (2017) Intonational speech prosody encoding in the human auditory cortex. *Science (80-)* 357(6353):797–801.
82. Hamilton LS, Edwards E, Chang EF (2018) A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Curr Biol* 28(12):1860-1871.e4.
83. Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human

superior temporal gyrus. *Sci Adv* 5(11):eaay6279.

84. McClelland JL, Elman JL (1986) The TRACE model of speech perception. *Cogn Psychol* 18(1):1–86.
85. Norris D, McQueen JM (2008) Shortlist B: A Bayesian model of continuous speech recognition. *Psychol Rev* 115(2):357–395.
86. Norris D, McQueen JM, Cutler A (2015) Prediction, Bayesian inference and feedback in speech recognition. *Lang Cogn Neurosci*:1–15.
87. Magnuson J, et al. EARSHOT: A minimal neural network model of incremental human speech recognition. doi:10.31234/OSF.IO/H7A4N.
88. Damper RI (1994) Connectionist models of categorical perception of speech. *Proceedings of ICSIPNN 1994 International Symposium on Speech, Image Processing and Neural Networks* (Institute of Electrical and Electronics Engineers Inc.), pp 101–104.
89. Kössl M, et al. (2014) Neural maps for target range in the auditory cortex of echolocating bats. *Curr Opin Neurobiol* 24:68–75.
90. Portfors C V., Wenstrup JJ (2001) Topographical distribution of delay-tuned responses in the mustached bat inferior colliculus. *Hear Res* 151(1–2):95–105.
91. Zatorre RJ, Belin P (2001) Spectral and temporal processing in human auditory cortex. *Cereb Cortex* 11(10):946–953.
92. Fries P (2009) Neuronal Gamma-Band Synchronization as a Fundamental Process in Cortical Computation. *Annu Rev Neurosci* 32(1):209–224.
93. Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15(4):511–517.
94. Kösem A, et al. (2018) Neural Entrainment Determines the Words We Hear. *Curr Biol* 28(18):2867-2875.e3.
95. Peelle JE, Davis MH (2012) Neural Oscillations Carry Speech Rhythm through to Comprehension. *Front Psychol* 3:320.
96. Chang EF, et al. (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13(11):1428–32.
97. Macmillan NA, Kaplan HL, Creelman CD (1977) The psychophysics of categorical perception. *Psychol Rev* 84(5):452–471.
98. Lee YS, Turkeltaub P, Granger R, Raizada RDS (2012) Categorical speech processing in Broca’s area: An fMRI study using multivariate pattern-based analysis. *J Neurosci* 32(11):3942–3948.
99. Myers EB, Blumstein SE, Walsh E, Eliassen J (2009) Inferior Frontal Regions Underlie the Perception of Phonetic Category Invariance. *Psychol Sci* 20(7):895–903.
100. Evans S, Davis MH (2015) Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cereb Cortex* 25:4772–4788.
101. Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. *J Neurosci* 32(25):8443–53.
102. Leonard MK, Baud MO, Sjerps MJ, Chang EF (2016) Perceptual restoration of masked speech in human cortex. *Nat Commun* 7:13619.
103. Cope TE, et al. (2017) Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nat Commun* 8(1):2154.
104. Park H, Ince RAA, Schyns PG, Thut G, Gross J (2015) Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in

Human Listeners. *Curr Biol* 25(12):1649–1653.

105. McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? *Trends Cogn Sci* 10(8):363–369.

106. McQueen JM, Norris D, Cutler A (2006) Are there really interactive processes in speech perception? *Trends Cogn Sci* 10(12). doi:10.1016/j.tics.2006.10.004.

107. Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. *Behav Brain Sci* 23(3):299–325.

108. Cho T, Ladefoged P (1999) Variation and universals in VOT: evidence from 18 languages. *J Phon* 27(2):207–229.

109. DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci U S A* 109(8):E505-14.

110. Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. *Trends Cogn Sci* 13(1):14–19.

111. Leonard MK, Chang EF (2014) Dynamic speech representations in the human temporal lobe. *Trends Cogn Sci* 18(9):472–479.

112. Sjerps MJ, Fox NP, Johnson K, Chang EF (2019) Speaker-normalized sound representations in the human auditory cortex. *Nat Commun* 10(1):2465.

113. Fox NP, Leonard MK, Sjerps MJ, Chang EF (2020) Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. *Open Sci Framew*. Available at: https://osf.io/9y7uh/.

114. Hamilton LS, Chang DL, Lee MB, Chang EF (2017) Semi-automated Anatomical Labeling and Inter-subject Warping of High-Density Intracranial Recording Electrodes in Electrocorticography. *Front Neuroinform* 11:62.

115. Feldman NH, Griffiths TL, Morgan JL (2009) The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychol Rev* 116(4):752–82.

FIGURE SUPPLEMENTS

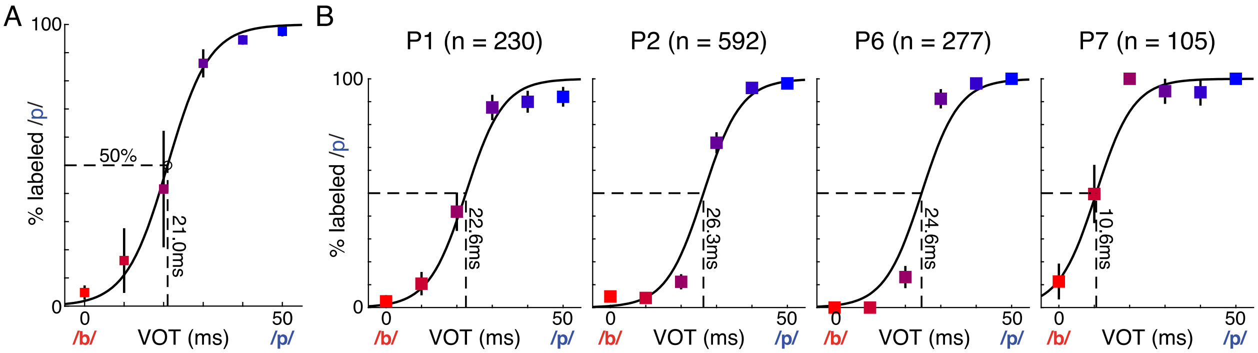

Figure 1-figure supplement 1. Identification behavior across all participants with behavioral data. A. *Mean (± SE across participants; n = 4 of 7 participants) percent /pa/ responses for each voice-onset time (VOT) stimulus. Best-fit psychometric curve (mixed effects logistic regression) yields voicing category boundary at 21.0ms (50% crossover point; see* ***Methods*** *for details).* ***B.*** *Behavior (mean ± bootstrap SE) for each individual participant (P1, P2, P6, P7). Total trials (n) listed for each participant (see* ***Supplementary File 1****). Best-fit psychometric curves and category boundaries were computed using the mixed effects logistic regression across all participants, adjusted by the random intercept fit by the model for each participant. Voicing category boundaries were subject-dependent, with 3 of 4 participants' occurring between 20-30ms. P1 is representative participant in* ***Figure 1C****.*

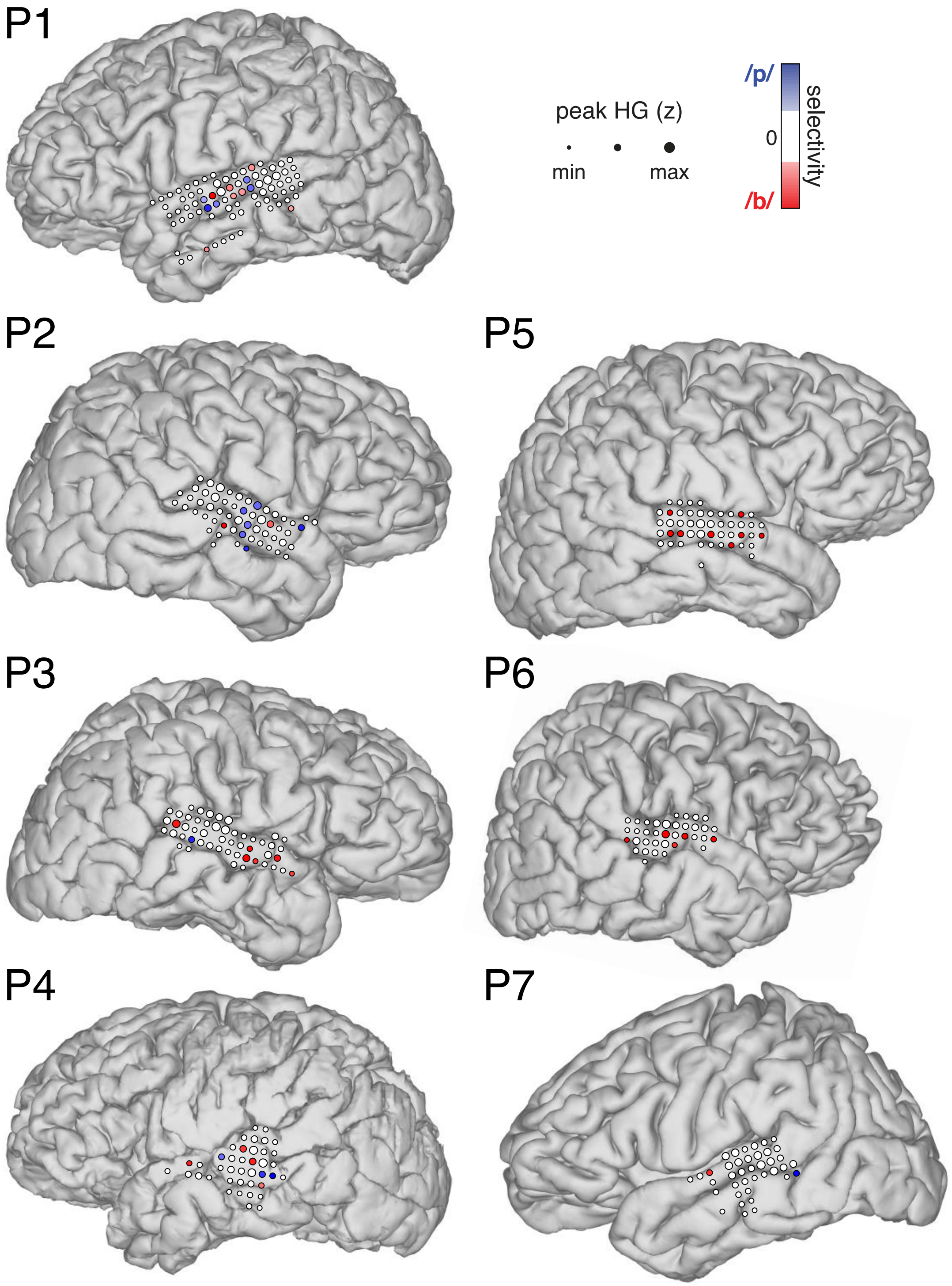

Figure 1-figure supplement 2. Locations of all speech-responsive and VOT-sensitive electrodes in each participant (P1-P7). *P1 is representative participant in* ***Figure 1D****. Electrode color reflects strength and direction of selectivity (Spearman's ρ between peak HG amplitude and VOT) at subset of VOT-sensitive sites ($p < 0.05$) for either voiceless VOTs (/p/; blue) or voiced VOTs (/b/; red). Electrode size indicates peak high-gamma (HG; z-scored) amplitude at all speech-responsive temporal lobe sites. Maximum and minimum electrode size and selectivity was calculated per participant for visualization.*

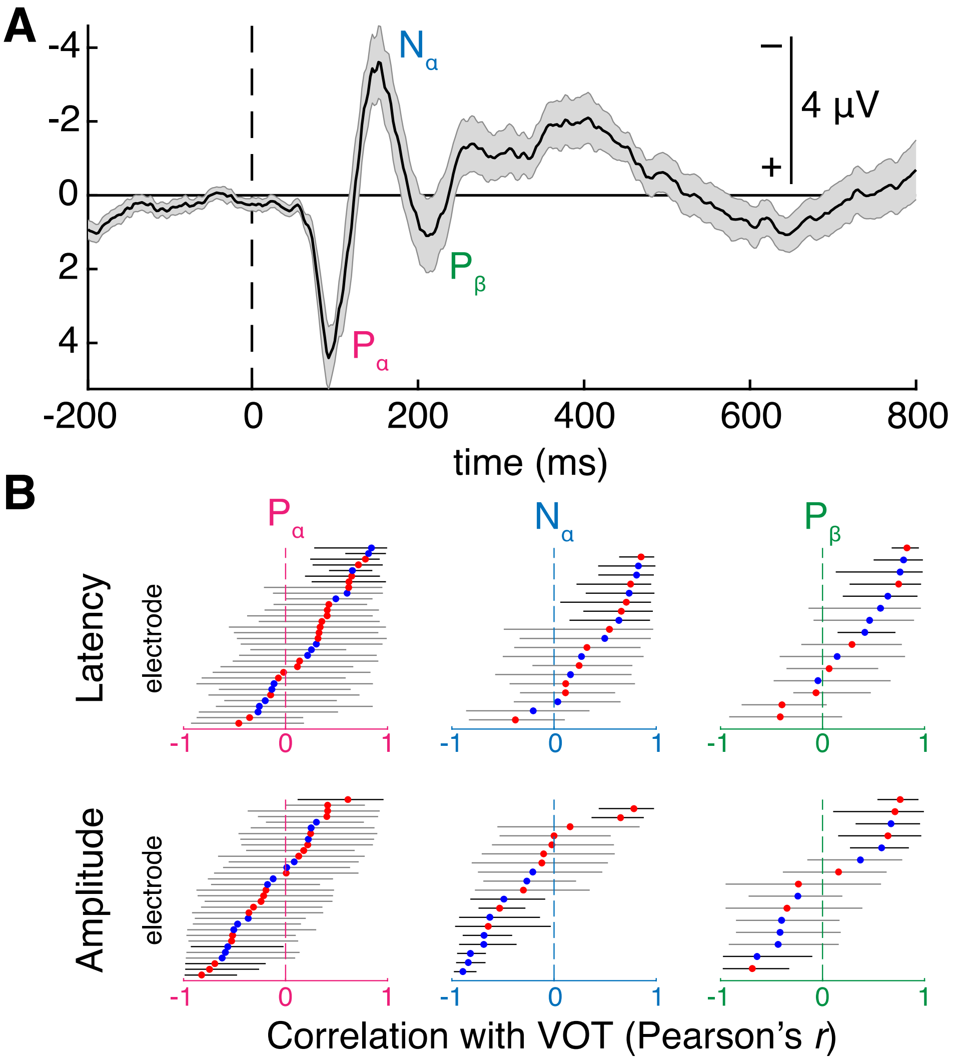

Figure 1-figure supplement 3. Analysis of evoked local field potentials reveals that some electrodes that encode VOT in their peak high-gamma amplitude also exhibit amplitude and/or temporal response features that are VOT-dependent. A. *Grand average auditory evoked potential (AEP) to all VOT stimuli. Evoked local field potentials (negative up-going) were averaged over all VOT-sensitive STG electrodes for one representative participant (P1) (mean ± SE, computed across electrodes). Three peaks of the AEP were identified for analysis: 75-100 ms (P_α), 100-150 ms (N_α), and 150-250 ms (P_β) after stimulus onset.* ***B.*** *Correlation coefficients (Pearson's r) quantifying association between VOT and latency (**top**) or amplitude (**bottom**) of each peak (P_α: **left**; N_α: **middle**; P_β: **right**) for each VOT-sensitive electrode for which that peak could be reliably identified (see **Figure 1-figure supplement 4** and **Methods** for details of this analysis). Horizontal bars represent bootstrapped estimate of correlation coefficient (mean and 95% CI) for each electrode (blue: voiceless-selective; red: voiced-selective; electrodes sorted by mean correlation value). Black bars around an electrode's mean indicate that encoding of VOT by the designated parameter (latency or amplitude of a given peak) was significant (95% CI excluded r = 0; grey bars: not significant). Later peaks were reliably identified for fewer electrodes (P_α: n = 32 of 49 electrodes; N_α: n = 19; P_β: n = 15).*

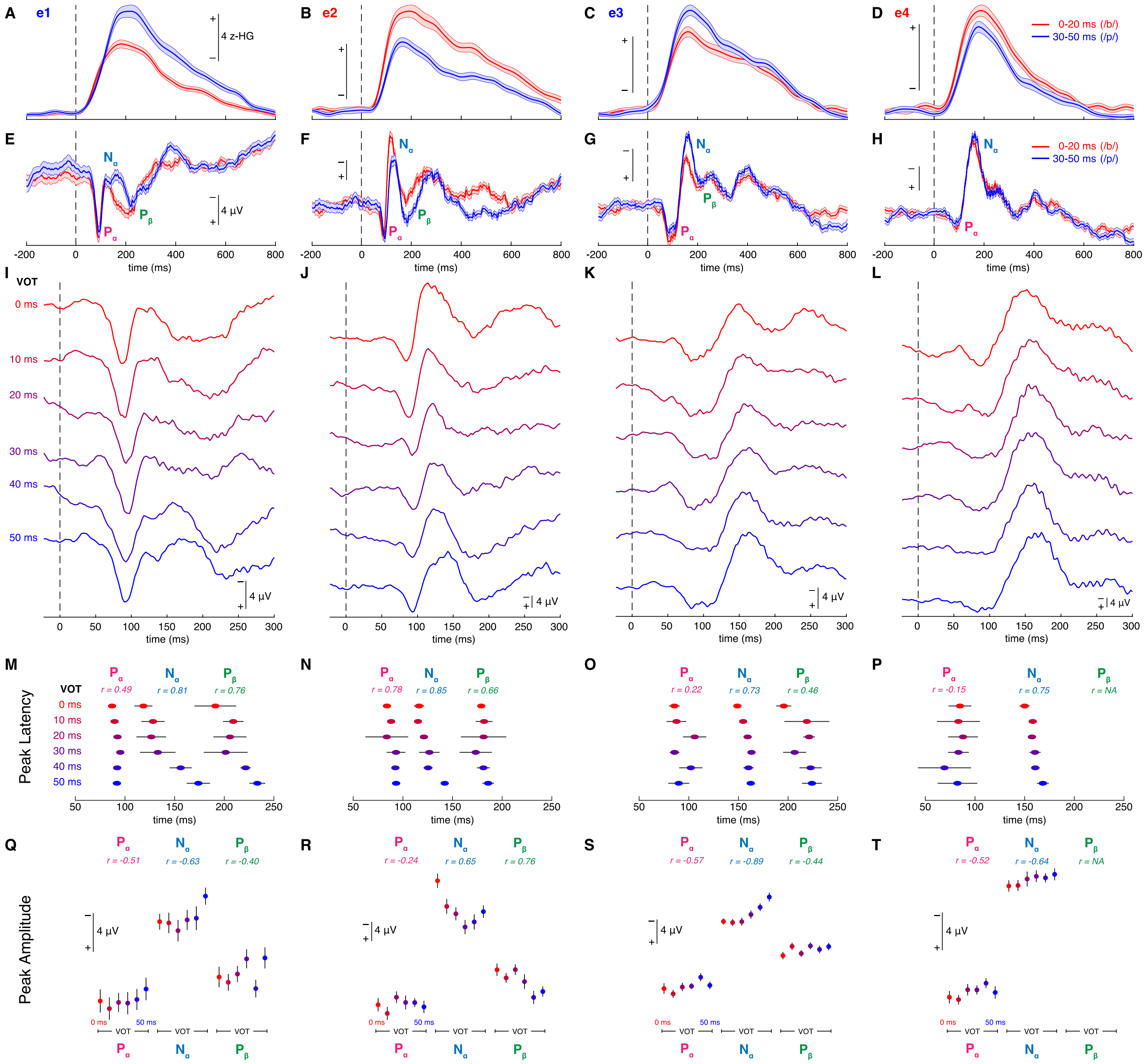
A e1
4 z-HG
+
–
B e2
C e3
D e4
0-20 ms (/b/)
30-50 ms (/p/)
E
F
G
H
N_α
P_α
P_β
4 µV
-200
0
200
400
600
800
time (ms)
I
J
K
L
VOT
0 ms
10 ms
20 ms
30 ms
40 ms
50 ms
0
50
100
150
200
250
300
time (ms)
M
N
O
P
Peak Latency
P_α r = 0.49
N_α r = 0.81
P_β r = 0.76
P_α r = 0.78
N_α r = 0.85
P_β r = 0.66
P_α r = 0.22
N_α r = 0.73
P_β r = 0.46
P_α r = -0.15
N_α r = 0.75
P_β r = NA
50
100
150
200
250
time (ms)
Q
R
S
T
Peak Amplitude
P_α r = -0.51
N_α r = -0.63
P_β r = -0.40
P_α r = -0.24
N_α r = 0.65
P_β r = 0.76
P_α r = -0.57
N_α r = -0.89
P_β r = -0.44
P_α r = -0.52
N_α r = -0.64
P_β r = NA
4 µV
0 ms
50 ms
VOT

Figure 1-figure supplement 4. Complex and variable associations between VOT and amplitude/temporal features of auditory evoked local field potentials (AEPs) exist in responses of electrodes that robustly encode voicing in their peak high-gamma amplitude. A** to **D. *Average high-gamma responses (± SE) to voiced (0-20ms VOTs; red) and voiceless (30-50ms VOTs; blue) stimuli in four representative VOT-sensitive STG electrodes, including two voiceless-selective (**A**: e1, **C**: e3) and two voiced-selective (**B**: e2, **D**: e4) electrodes, aligned to stimulus onset. Vertical bars indicate relative scaling of high-gamma (z-scored) in each panel. The two leftmost electrodes (e1, e2) correspond to e1 and e2 in main text (e.g., **Figure 1E**). **E** to **H.** Average local field potentials (± SE) evoked by voiced/voiceless stimuli in the same four electrodes, aligned to stimulus onset. Vertical bars (negative-upgoing) indicate relative scaling of voltage in each panel. The three peaks of the AEP that were identified for analysis are labeled for each electrode (P_α, N_α, P_β; see **Figure 1-figure supplement 3**). For a given electrode, peaks were omitted from this analysis if they could not be reliably identified across bootstrapped samples of trials from all six VOT conditions (e.g., P_β for e4). See **Methods** for details. **I** to **L.** Average local field potentials evoked by each VOT stimulus (line color) in the same four electrodes, aligned to stimulus onset. **M** to **P.** Mean latency (± bootstrap SE) of each AEP peak for each VOT stimulus for the same four electrodes. Mean bootstrapped correlation (Pearson's r) between VOT and peak latency shown for each peak/electrode. **Q** to **T.** Mean amplitude (± bootstrap SE) of each AEP peak for each VOT stimulus for the same four electrodes. Mean bootstrapped correlation (Pearson's r) between VOT and peak amplitude shown for each peak/electrode. Note that negative correlations are visually represented as rising from left to right. Correlation coefficients comprised the source data for summary representations in **Figure 1-figure supplement 3**.*

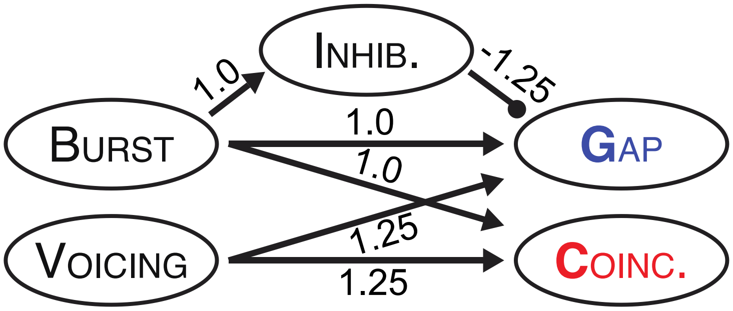

Figure 2-figure supplement 1. Connection weights between model nodes.

LEGENDS FOR SUPPLEMENTARY FILES

Participant	Hem	# trials (ECoG)	# trials (behavior)	# elecs (SR)	# elecs (VOT)	# elecs (V- / V+)
P1	LH	234	230	78	12	5 / 7
P2	RH	625	592	56	8	6 / 2
P3	RH	339	0	50	7	1 / 6
P4	LH	333	0	40	7	3 / 4
P5	RH	119	0	47	8	0 / 8
P6	RH	305	277	36	5	0 / 5
P7	LH	110	105	39	2	1 / 1

Supplementary File 1. Table of experimental summary statistics for each participant. *Each participant had ECoG grid coverage of one hemisphere (Hem), either left (LH) or right (RH). Participants completed as many trials as they felt comfortable with. Number of trials per participant for ECoG analyses indicate trials remaining after artifact rejection. Some participants chose to listen passively to some or all blocks, so three participants have no trials for behavioral analyses. See* ***Methods*** *for description of inclusion criteria for individual trials in ECoG and behavioral analyses. A subset of speech-responsive (SR) electrodes on the lateral surface of the temporal lobe had a peak amplitude that was sensitive to VOT, selectively responding to either voiceless (V-) or voiced (V+) stimuli. See* ***Methods*** *for details on electrode selection.*

		activation parameter				
		m	***M***	***ρ***	***λ***	***θ***
model node	**Burst**	-10	10	0	1	0
	Voicing	-10	10	0	1	0
	Inhibitor	-10	10	0	1	0
	Gap	-10	10	0	0.25	0.25
	Coincidence	-10	10	0	0.25	1

Supplementary File 2. Table of activation parameters for each model node. *m = minimum activation level. M = maximum activation level. ρ = resting activation level. λ = decay rate. θ = propagation threshold.*

VOT	0			BV												
	10			B	V											
	20			B		V										
	30			B			V									
	40			B				V								
	50			B					V							
		-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	…
		time post onset (cycles)														

Supplementary File 3. Table illustrating timing of 6 simulated model inputs. *The table is sparse, meaning that inputs to both Burst and Voicing detector units are 0 whenever a cell is blank. Inputs are clamped onto either Burst or Voicing detector units (always with strength = 1) for a given simulated VOT stimulus during the cycles that are labeled with a B or a V.*