Neural mechanisms of the transition from planning to execution in speech production

Suseendrakumar Duraivel[1], Shervin Rahimpour[2,3], Katrina Barth[1], Chia-Han Chiang[1], Charles Wang[1], Stephen C. Harward[2], Shivanand P. Lad[2], Daniel P. Sexton[2], Allan H. Friedman[2], Derek G. Southwell[1,2,4,5], Saurabh R. Sinha[6], Jonathan Viventi[1,2,4,7, #], Gregory Cogan[1,2,4,7,8,9, #]

[#]Corresponding authors

Author affiliations:

1 Department of Biomedical Engineering, Duke University, Durham, NC, USA

2 Department of Neurosurgery, Duke School of Medicine, Durham, NC, USA

3 Department of Neurosurgery, Clinical Neuroscience Center, University of Utah, Salt Lake City, UT, USA

4 Department of Neurobiology, Duke School of Medicine, Durham, NC, USA

5 Department of Neurology, Duke School of Medicine, Durham, NC, USA

6 Penn Epilepsy Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

7 Duke Comprehensive Epilepsy Center, Duke School of Medicine, Durham, NC, USA

8 Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

9 Center for Cognitive Neuroscience, Duke University, Durham, NC, USA

Correspondence to:

Gregory B. Cogan

Department of Neurology

Duke South, Blue Zone 4519, Durham, NC 27710

gregory.cogan@duke.edu

Jonathan Viventi

Biomedical Engineering

1149 CIEMAS, 101 Science Dr, Durham, NC 27705

j.viventi@duke.edu

Running title: Neural speech sequencing

Keywords: Speech, IEEG, syllable sequences, speech planning, speech production

Abstract

The neural basis of speech production involves the rapid transition from abstract planning of speech units such as syllables and phonemes, to the motor execution of speech sounds. Although a distributed network of brain regions has been implicated in speech production overall, it is unclear how the brain transitions from planning to execution for speech production. Leveraging the high spatio-temporal resolution of intracranial recordings, we find evidence for neural mechanisms that operate in space and time across the prefrontal and premotor cortices to facilitate the transition from planning to execution. During this execution, we show evidence for motor sequencing from neural activity that tracks both phonological units as well as the transition between them, suggesting both discrete elements from planning as well as continuous motor transitions. We demonstrate temporally-resolved neural mechanisms for the transition between planning and execution to facilitate speech production.

Word Count: 142

Introduction

Speech production enables verbal communication from planning to motor execution of speech sound sequences[1–3]. Planning occurs by turning mental ideas to words[4–6], which are mapped to smaller speech units to generate planned sequences for motor execution[7,8]. From the psycholinguistic literature[9–11], these planned motor sequences are hypothesized to operate at two distinct levels: a framing level in which a framing plan is created that corresponds to the syllable, and a content level in which phonemes are combined within the syllable frame[12–16]. Behavioral evidence for these elements of planning comes from speech error data such as slips of the tongue and phoneme exchange errors during everyday speech and in aphasia, suggesting neural mechanisms for distinct stages in speech planning[9,17,18]. These combined sequenced plans are then fed into a motor execution system that: 1) generates commands to control vocal tract gestures for articulation, and 2) updates the commands from sensory feedback[12–15,19,20]. Together, representations for syllables and phonemes combine during planning and execution to allow the speaker to effortlessly utter any sequence of speech sounds within the phonological rules of their proficient language[9–11]. Despite the importance of planning and execution for speech production, no work has been done to characterize the mechanisms that enable the transition between them ***(Fig. 1a, b)***.

Neuroimaging studies of speech production have identified a distributed neuroanatomical network that includes inferior frontal gyrus (IFG), rostral and caudal middle frontal gyrus (rMFG & cMFG), pre-central and post-central gyrus (PCG & POG; sensorimotor cortex), supra-marginal gyrus (SMG),

inferior parietal cortex (IPC), insula, superior temporal gyrus (STG), and primary auditory cortex (PAC)[1,12,21–30]. Within this network, prefrontal regions have primarily been associated with planning, sensorimotor circuits (PCG, POG, SMG and IPC) were associated with motor execution,[3,31] and superior temporal regions and the insula associated with monitoring based on sensory feedback (auditory and somatosensory)[3,32–34], as well as internal error correction[35,36]. Significant progress has been made to understand how speech is executed by sensorimotor circuits through coordinated activity to produce speech sounds[4,24,29], as well as how speech is monitored through neural suppression of self-generated speech sounds in the superior temporal gyrus (STG)[37–40]. For studies that examine planning, neuroimaging studies have been limited by their lack of precise timing to delineate planning from execution. However, studies that have manipulated the syllable frame (order, complexity, and novelty of syllables), found activation in the posterior inferior frontal gyrus (pars-opercularis)[10,11,41–43], thereby supporting the idea that the left IFG could be involved in syllable planning for speech production.

Computational models of speech production have been developed based on the neuroimaging literature to integrate psycholinguistic (syllable-phoneme) and motor-control (planning-execution) theories that model the accurate planning, execution, and monitoring of sequences of speech sounds[14,16,27]. Simulations from these models using this two-tiered hierarchical organization resulted in accurate generation of phoneme sequences, with a higher order planning state coding speech at the syllabic level and a lower order planning state at the phoneme level.[13,14] These computational studies suggest that during planning, the two distinct speech representations are composed hierarchically and in serial order, with syllables planned prior to phoneme plans. Empirical evidence for this temporal ordering for planning, is, however, lacking due to limitations of recording technology to resolve neural signals accurately in both space and time.

Intracranial recordings using electrocorticography (ECoG) and stereo-electroencephalography (SEEG) record neural activity directly from cortex with high temporal resolution (milliseconds) with precise anatomical localization[44–47]. This technology has been further advanced through the development of high-density micro-electrocorticographic (µECoG; 1.33 – 1.72 mm inter-electrode distance) arrays that can measure micro-scale cortical activity with millimeter scale spatial resolution[48–52]. Neural activations from ECoG and SEEG in the high-gamma band (HG: 70-150 Hz) have been shown to correlate with multi-unit firing rates, making this signal a good measure of local neural activity[53–57]. During speech production, studies have shown HG activations in the ventral sensorimotor cortex that encode articulatory properties during speech motor execution[19,20,58–63], however, it is not known how the brain enables the transition from speech planning to this motor execution.

Surprisingly, there has been very little intracranial work examining speech planning, and no work examining the transition between planning and execution. One previous study used overt sentence completion to examine lexical access in the IFG prior to articulation[64], but this activity could be related to either speech perception or production, making its role in speech production unclear. A more recent study used an unstructured experimental setting to examined the timing of planning in relation to turn taking in conversation[65] but it did not examine the content of planning or provide evidence for how speech plans enable motor execution. These studies do however, highlight the promise of intracranial recordings for studying different stages of speech production via the measurement of spatially specific neural representations. In particular, the high temporal and spatial resolution of intracranial recordings presents a possibility to delineate the neural mechanisms associated with the transition from speech planning to execution. Understanding this transition could have important implications for treating neurological conditions that affect both higher order cognition and speech production, as well as in developing more advanced neural prosthetic technology.

To elucidate the neural mechanisms for planning and execution in speech production, we performed intracranial recordings with 52 patients while they verbally articulated constructed pseudowords. We show distinct neural activations for planning and execution that represent both syllables and phonemes. The neural activations demonstrate distinct spatial and temporal patterns that reflect the transition from speech planning to execution. During this transition, we elucidate neural mechanisms that transform syllables and phonemes from plans to motor execution. In execution, we show evidence for motor sequencing through sequential patterns of phonemes as well as their transitions. Together, these findings provide evidence for distinct neural mechanisms that enable the transition from speech planning to sequential motor execution for the support of speech production.

1 Results

We conducted a pseudo-word repetition task with 52 patients (n = 52, mean age = 33.1 years, 23 females, Supplementary table 1) in the epilepsy clinical monitoring unit, and with 3 patients intra-operatively (n = 3, mean age = 64.3 years, 0 females, Supplementary table 1). Patients in the monitoring unit were implanted with clinical intracranial electrodes (electrocorticography [ECoG: 2 patients] or stereo-electroencephalography [SEEG: 50 patients]) for the localization of epileptic tissue to guide surgical resection or neuromodulation targeting (**Fig. 1c**). Patients recorded intra-operatively were awake as part of their surgical procedure for tumor resection (1 patient) or deep brain stimulator (DBS) placement for the treatment of movement disorders (2 patients), and these neural recordings were obtained at micro-scale cortical level at high spatial density using µECoG electrode arrays[63].

Pseudo-words were presented auditorily through speakers connected to a presentation laptop running Psychtoolbox[66]. After the pseudo-word was presented, there was a short delay (1.2 – 1.6 s), followed by a prompt (using a visual "Speak" cue), during which the patients were instructed to repeat back the presented pseudo-word stimulus (**Fig. 1d**). Each pseudo-word was either composed with a monosyllabic (consonant-vowel-consonant: CVC) or a disyllabic (vowel-consonant-vowel: VCV) frame, that contained three phonemes (out of a set of 9 total available phonemes - **Fig. 1e,** Supplementary table 2). Each patient completed four task blocks that contained 52 unique tokens each. Due to the shorter availability of time, patients in the intra-operative setting underwent a similar but modified version of the pseudo-word repetition task in which they repeated the pseudo-words immediately after the auditory presentation. The intraoperative patients completed three to four repetitions of task blocks.

1.1 Behavior code for syllable sequencing

On average, patients repeated the pseudo-words after 0.58 seconds (± 0.33 s) from the visual "Speak" cue and with an average duration of 0.65 seconds (± 0.16 s). Mean response times (RT) and response durations (RD) were (mean ± SD): CVC frames: RT = 0.56 ± 0.35 s, RD = 0.58 ± 0.17 s; and VCV frames: RT = 0.6 ± 0.32 s, RD = 0.68 ± 0.14 s (**Fig. 1d**).

To quantify the effect of differences in syllabic frame on behavioral performance, we performed linear mixed effects (LME) modeling on each trial, with fixed effects modeling the syllabic frame identity linked to each pseudo-word and random effects modelling the patient identity. VCV frame utterances had longer durations compared to CVC frames [LME: $t_{(8696)} = 10.7$, $\beta = 0.098$, $p = 2.1e\text{-}26$, 95% CI = 0.08 to 0.1, **Fig. 1f,** Supplementary Fig. 1a], in keeping with their longer articulatory and acoustic duration (2 vs. 1 syllable). Interestingly, response times were significantly later for disyllabic VCV frames as compared to monosyllabic CVC frames [LME: $t_{(8698)} = 3.7$, $\beta = 0.045$, $p = 2.2\ e\text{-}04$, 95% CI = 0.02 to 0.07, **Fig. 1g,** Supplementary Fig. 1b], suggesting that speech motor planning reflects the composition of syllabic frames. Further, this result also held with utterance length as a predictor, suggesting that this behavioral difference could reflect the abstract planning of syllabic frames, in addition to the planning effects from overall length of the utterance [LME: $t_{(9084)} = 2.75$, $\beta = 0.103$, $p = 0.006$, 95% CI = 0.03 to 0.18, Supplementary Fig. 1c].

1.2 Neural representations during speech production

To investigate the neural mechanisms involved in speech production, we first extracted neural activations in the high-gamma band (HG: 70 – 150 Hz) that have previously been shown to be an index of local neural computation and a correlate of multi-unit firing[54,56,57]. Only correctly articulated trials with response times greater than 50 ms (from "Speak"/go cue) were used for neural analysis.

Out of 8106 preprocessed IEEG electrodes from 52 patients in the clinical monitoring unit, we observed significant increases in HG power in 3534 electrodes during speech production as compared to a pre-stimulus baseline (one-sided permutation test with post-hoc FDR correction per patient, $p < 0.05$, **Fig. 2a,** Supplementary table 2). Speech neural activations were present in both the left (1701 electrodes) and right (1665 electrodes) hemisphere (**Fig. 2b**). We further selected electrodes within anatomical regions of interest (ROIs) that have been shown to be active for speech production[4,16,19,65,67]. The selected ROIs in both hemispheres included: inferior frontal gyrus (IFG: 92 significant electrodes), rostral middle frontal gyrus (rMFG: 137), caudal middle frontal gyrus (cMFG: 186), precentral and postcentral gyrus (PCG & POG: 351), inferior parietal cortex (IPC: 268), insula (172), anterior superior temporal gyrus (aSTG: 91), posterior superior temporal gyrus (pSTG: 90), superior temporal sulcus (STS: 88), and primary auditory cortex (PAC: 43).

To characterize the timing of activation between ROIs, we estimated the peak onsets of HG activation for each electrode relative to the utterance onset (**Fig. 2c**). There were significant differences in peak activations between ROIs in both left and right hemispheres (one-way ANOVA; left hemisphere: $F_{9,717} = 10.87$, $p = 6.1e\text{-}16$, right hemisphere: $F_{9,781} = 6.69$, $p = 2.8e\text{-}9$. **Fig. 2d**). Based on neuro-anatomical studies that investigated speech articulation[4,16,19,65,67], we divided these ROIs into three distinct networks: planning (IFG, rMFG, and cMFG), articulation (PCG & POG, and IPC), and monitoring (aSTG, pSTG, STS, and PAC). Using a second level one-way ANOVA, we found that in the left-hemisphere, electrodes in the planning network were activated early, followed by the electrodes the articulation network, and then the monitoring network (one-way ANOVA: $F_{2,724} = 30.7$, $p = 1.6e\text{-}13$, **Fig. 2e**). Pair-wise post-hoc t-tests showed that the mean values were significantly different between all three groups ($p < 0.05$, $T_{Planning} < T_{Articulation}$, $t_{(365)} = 2.938$, $p = 0.0018$; $T_{Articulation} < T_{Monitoring}$, $t_{(507)} = 9.25$, $p = 3e\text{-}13$, $T_{Planning} < T_{Monitoring}$, $t_{(394)} = 9.84$, $p = 7.2e\text{-}21$, **Fig. 2e**). Differences were also significant in the right hemisphere ($F_{2,788} = 8.86$, $p = 1.6e\text{-}4$, $T_{Articulation} < T_{Monitoring}$, $t_{(528)} = 7.53$, $p = 2.2e\text{-}13$, $T_{Planning} < T_{Monitoring}$, $t_{(385)} = 2.36$, $p = 0.0188$ **Fig. 2e**). However, the prefrontal-planning activations occurred later in time than the articulation regions ($T_{Planning} > T_{Articulation}$, $t_{(511)} = 2.3$, $p = 0.0214$), which is consistent with the role of the right hemisphere in error monitoring[68,69]. A similar organization existed when the neural data were aligned to the go-cue (Supplementary Fig. 2). These results clearly demonstrate the presence of three distinct speech networks that are predominately in the left-hemisphere: 1) a planning network in the pre-frontal regions to initiate speech, 2) an articulation network that generates motor commands and executes planned speech, and 3) a monitoring network for auditory feedback tracking.

1.3 Neural syllabic code is anatomically organized

We next sought to investigate the neural mechanisms for speech representation associated with the three levels of processing (planning, articulation, and monitoring) at each stage (**Fig. 3a**). Based on prior computational work, we theorized that speech production involves the recruitment of syllabic frames during the planning stage[4,11,14,41], followed by their execution during the produced utterance (articulation and monitoring)[14,27]. To assess the specific timing of neural activations that reflect differences in the syllable frame, we examined differences in ROI-grouped HG activations between syllabic frames (ΔHG; VCV and CVC pseudo-word trials, see Methods), at every time point around the response epoch (**Fig. 3b**). With phonemes controlled at each position within the syllabic frame, these HG contrasts (ΔHG) can reveal activation patterns selective to the syllabic frame alone. Results show that neural activations associated with syllabic frame showed consecutive time points with significant syllable frame coding (time clusters) for each ROI in the left hemisphere (**Fig. 3c,** except STS; $p < 0.05$, one-sided permutation test with cluster correction). The earliest time clusters for syllable frame contrasts were observed in left-hemispheric planning regions (cluster start time & mean peak onsets): 1) IFG: -0.74s & -0.3s, 2) rMFG: -0.68s & -0.05s, and 3) cMFG: -0.07s & -0.17s. Speech articulatory ROIs demonstrated relatively late timing: 1) PCG & POG: -0.6 & 0.83s, and 2) IPC: -0.1s & 0.3s. Finally, the monitoring ROIs exhibited timing closer to voice-onset times: 1) Insula: -0.27s & 0.32s, 2) aSTG: -0.22s & 0.37s, 3) pSTG: -0.16s & 0.26s, and 4) PAC: -0.16s & 0.31s. We further performed a one-way ANOVA to assess the temporal properties of syllabic coding, which demonstrated significant differences in peak coding between individual ROIs (**Fig. 3d,** except STS, $F_{8,489} = 18.1$, $p = 8.9e\text{-}24$). To directly investigate this anatomical division of syllabic frame coding, we repeated the analysis on electrodes grouped within each speech network (**Fig. 3e,** planning, articulation, and monitoring). We observed the earliest coding (ΔHG) in the prefrontal-planning network (cluster onset: -0.73s, peak onset: -0.16s), followed by the articulatory network (cluster onset: -0.22s, peak onset: 0.23s), and monitoring network (cluster onset: -0.16s, peak onset: 0.32s (one-way ANOVA, $F_{2,495} = 64.3$, $p = 1.5e\text{-}25$, **Fig. 3f**). Post-hoc t-tests showed that the differences showed pairwise significance between the three areas ($p < 0.05$), except for between the articulatory and the monitoring areas ($T_{Planning} < T_{Articulation}$, $t_{(333)} = 3.3$, $p = 0.0011$; $T_{Articulation} < T_{Monitoring}$, $t_{(403)} = 1.3$, $p = 0.2$ (not significant), $T_{Planning} < T_{Monitoring}$, $t_{(352)} = 3.78$, $p = 1.7e\text{-}4$, **Fig. 3f**). In the right hemisphere, significant differences in syllable neural codes were observed between the articulation and monitoring network (negative t-statistic on comparisons against planning network, Supplementary Fig. 3), supporting the right hemisphere's involvement in execution. Similar temporal progression is also observed when neural data was aligned with the go-cue onset, showing that the transition from planning to execution of syllable frames is also preserved at the commencement of planning with respect to the instructional cue (Supplementary Fig. 4). These, results are further present in a

controlled analysis that preserved initial syllabic structure (CV'CVC vs. CVC, Supplementary Fig. 5), further validating the existence of a neural syllable code that is independent of the first position phonemes (C vs. V). Our results show a clear temporal progression from planning to execution of neural activations associated with the syllable frame. Further, this neural activation in the planning stage prior to execution supports the existence of the syllable as a unit of representation for speech planning in left-hemispheric prefrontal-planning regions.

We next sought to more closely investigate the transition from planning to execution for speech production. Based on prior intracranial studies[65], neuroimaging[11], and theoretical work[14,16], we posited that neural responses associated with the transition of syllabic frame coding (VCV or CVC) would be present in distinct neural populations in the IFG (pars-opercularis, Supplementary Fig. 6 & 7) and PCG.

To more closely investigate the anatomical and temporal characteristics of the neural populations associated with syllable frame coding for planning and execution, we used high-channel count, high-density µECoG electrode arrays to record speech neural activations from two subjects. We used two versions of the µECoG arrays to record neural activity from IFG and PCG (**Fig. 4a**): subject S1, during an awake craniotomy, was acutely implanted with a 256-channel subdural array (12 x 22 matrix; inter-electrode distance: 1.72 mm), and subject S2, during an awake DBS procedure, was acutely implanted with a 128-channel subdural array (8 x 16 matrix; inter-electrode distance: 1.33 mm). The implantation procedure for S1 enabled simultaneous recordings from the pars-opercularis portion of the IFG and the ventral portion of PCG, with the array's longest side oriented parallel to the sylvian fissure. Similarly, the electrode array for S2 was primarily positioned over the pars-opercularis, with the distal end stretching to PCG. Both subjects completed 4 blocks of the modified speech repetition task. We observed neural activations on individual electrodes with significant modulation in the HG band in both subjects (**Fig. 4b,** S1: 201/256 significant electrodes, S2: 63/128 electrodes). Neural activations showed spatially varying temporal characteristics associated with both planning and execution. 18.9% of significant electrodes (**Fig. 4c,** S1: 12/201, and S2: 38/63) reached their peak signal power prior to the onset of the utterance. These results indicate the presence of spatially distinct neural population sites that code for speech planning at a micro-scale level recorded with our high-resolution µECoG electrode arrays.

To examine the coding for syllabic frames, we repeated our syllabic discrimination framework analysis on an electrode specific level, by contrasting HG power between disyllabic (VCV) and monosyllabic (CVC) pseudo-word trials. A one-sided permutation test followed by a second-level cluster correction ($p < 0.05$, 1000 shuffles) revealed electrodes that exhibited significant contrasts in

activations for the two different syllable frames, providing further evidence for syllabic coding at the micro-scale (**Fig. 4d, e,** S1: 100/201, and S2: 37/63). Interestingly, these syllable contrasting electrodes were observed to be spatially organized and they exhibited distinct temporal profiles identified by onset and peak syllabic coding (**Fig. 4d, e**). For both subjects, we observed spatially varying temporal coding with earlier coding in the anterior portion of the array (pars opercularis) and later coding towards the posterior side (PCG: spatial linear model for S1 (F-statistic test against constant model, $F_{2,92}$ = 6.02, *p = 0.016*, **Fig. 4f**) and peak coding ($F_{2,98}$ = 10.9, *p = 1.3e-3,* **Fig. 4f**). Comparable prediction results were obtained for S2 with peak coding demonstrating a significant spatial anatomical gradient ($F_{2,28}$ = 5.68, *p = 0.024,* **Fig. 4f**), and for onset coding ($F_{2,28}$ = 6.02, *p = 0.07,* **Fig. 4f**). Taken together, the presence of a spatial anatomical gradient suggest the existence of populations of temporally tuned neurons that activate in succession from pars-opercularis to the pre-motor cortex. These populations could form the basis for a neural mechanism for the transition from planning to execution for speech production.

1.4 Two-tiered hierarchical organization during utterance.

The ROI-grouped analysis demonstrated how syllable frames are composed for planning and execution during speech production. However, as syllable frames are incrementally composed, the speech production system must incorporate appropriate phonemic codes into these frames before generating complete motor articulatory commands during execution[4,13,70]. In other words, the syllable frame (e.g. CVC) must be filled in with its constituent phonemes (e.g. CVC → /kaet/). We propose that this hierarchical organization is reflected in distinct activation for the syllable frame and phoneme units in both planning and execution for speech production. The syllable frames are composed prior to phonemes, which reflect their position at higher level in the speech compositional hierarchy. Conversely, phonemes are encoded later as they reflect a lower level of phonological unit for speech production[14](**Fig. 5a**).

We demonstrated this hierarchical representation by performing temporally resolved neural decoding of speech features (syllables and phonemes) in our speech production tasks from our in-unit patients. We performed a 10-fold, cross-validated decoding (SVD-LDA; see Methods) from HG neural activations (-0.5 s to 0.5 s around utterance onset) using speech significant electrodes pooled across subjects (2895 total electrodes across both hemispheres, 69 mean electrodes per subject). Significant decoding performance were observed for both syllables (2-way: CVC vs. VCV, average accuracy= 93%, chance = 50%, **Fig. 5b**) and phonemes (9-way: /a/, /ae/, /i/, /u/, /b/, /p/, /v/, /g/, and /k/, average accuracy = 38.3%, chance = 11.11%, **Fig. 5b**). Having established significant decoding for both syllable frames and phonemes, we sought to characterize the temporal evolution of syllable frames and phonemes separately for left and right hemispheric electrodes. Using time-resolved

decoding (200 ms window; 10 ms step size), we obtained stable decoding for both syllable frames and phonemes (temporal permutation cluster test, $p < 0.05$, one-sided permutation test against 250 shuffled decoding models). We observed sustained decoding for syllable frames and phonemes in the left-hemisphere, with syllable frames (cluster onset: -1.3s) detectable at earlier times compared to phonemes (cluster onset: -0.17s, $Onset_{Syllable} < Onset_{Phoneme}$, one-sided Mann-Whitney U test, $p = 3.14e\text{-}8$, **Fig. 5c**), providing evidence that syllable frames are coded prior to phonemes. We also demonstrated the same temporal order of syllable frame followed by phoneme within individual ROIs and observed similar stable decoding for both syllable frames and phonemes (**Fig. 5d** except rMFG that exhibited coding only for syllables). For both planning and articulation, we observed syllable frames represented earlier in time as compared to phonemes (**Fig. 5e, f**), suggesting that this representational hierarchy is preserved both within and across brain areas and speech stages.

Next, we sought to determine whether the temporal relationship between syllable frames and phonemes is dependent on speech stages from planning to execution. We observed a significant difference in decoding onsets for syllables and phonemes from planning to execution, with syllables occurring earlier than phonemes (syllables: one-way ANOVA; onset $p = 1.7e\text{-}9$, peak $p = 3.5e\text{-}54$, **Fig. 5e, f**, phonemes: one-way ANOVA; onset $p = 2.5e\text{-}16$, peak $p = 3.6e\text{-}13$, **Fig. 5e, f**). Post-hoc Tukey's HSD test showed that the differences were also pairwise significant ($p < 0.05$) between the three networks, except for peak phoneme decoding between the articulatory and the monitoring networks which approached significance ($p = 0.0785$).

To assess the change in relative timing of the neural response between syllables and phonemes, we then compared their activation timing within each ROI group. We reasoned that higher order planning areas that code for abstract planning representations would show a greater temporal delay between syllable and phoneme representations as compared to execution stages which would show a more concurrent activation pattern. Our results support this hypothesis: there was a maximum delay in the prefrontal-planning network (**Onset**: syllable frame mean = -0.43s, phoneme mean = -0.22s, difference = 0.25s, $p = 0.0385$, one-sided Mann Whitney U test; **Peak**: syllable frame mean = -0.0125s, phoneme mean = 0.3s, difference = 0.32s, $p = 3.11e\text{-}8$, **Fig. 5e, f**), followed by the articulation network (**Onset**: syllable frame mean = -0.22s, phoneme mean = -0.13s, difference = 0.13s, $p = 3.5e\text{-}5$; **Peak**: syllable frame mean = 0.24s, phoneme mean = 0.35s, difference = 0.11s, $p = 2.7e\text{-}8$, **Fig. 5e, f**), and then minimally within the monitoring network (**Onset**: syllable frame mean = 0.039s, phoneme mean = 0.055, difference = 0.017s, $p = 0.082$; **Peak**: syllable frame mean = 0.39s, phoneme mean = 0.37s, difference = -0.025s, $p = 0.9$, **Fig. 5e, f**). The mean differences were pairwise significant between all three networks (pairwise t-test, planning vs. articulation, **onset**: $t_{(798)} = -10.7$, $p = 5.5e\text{-}25$, **peak**: $t_{(798)} = 71$, $p = 0$; articulation vs. monitoring: **onset**: $t_{(798)} = 18.2$, $p = 1.9e\text{-}62$,

peak: $t_{(798)} = 101$, $p = 0$; planning vs. monitoring: **onset**: $t_{(798)} = -8.04$, $p = 3.3e\text{-}15$, **peak**: $t_{(798)} = 117.8$, $p = 0$, **Fig. 5g, h**). Repeating the analysis with electrodes pooled from right-hemispheric ROIs, we observed significant decoding for syllables and phonemes, but we did not observe temporal differences in support of the syllable-phoneme hierarchy (Supplementary Fig. 8). Overall, these results highlight the hierarchical organization between the syllable frame and phoneme representation from planning to execution. The differences in hierarchical representation at each speech stage indicates a transition from abstract (separate) representations of syllables and phonemes for speech planning in the prefrontal cortex to execution (integrated) coding in the articulation (motor regions) and monitoring (temporal cortex and insula). This temporal transition of activation could serve as a neural mechanism for the integration of representational units for speech production (**Fig. 5i**).

1.5 Phonological units are sequentially articulated only during execution.

After planning, the speech production system must successfully articulate the correct ordering of phonological sequences (phonemes) in real-time for the correct execution of the utterance (**Fig. 6a**). To show neural evidence for this sequencing, we performed temporal decoding (9-way; chance 11.11%) to predict the positional phonological units within the pseudo-word utterance in time (e.g. /abae/: /a/→/b/→/ae/, P1 = /a/, P2 = /b/, P3 = /ae/). We observed stable decoding and the correct ordering of sequences (P1→P2→P3) in the ROIs specific to the articulation and monitoring networks. Surprisingly, this ordering was absent in the planning network ROIs ($p < 0.05$, one-sided permutation test against 250 shuffled decoding models, **Fig. 6b**), indicating that the abstract phoneme codes retrieved during planning are not explicitly sequenced at this stage, rather, they are sequenced in real-time in the motor cortex. In the articulation ROIs, we observed pairwise significant differences between phoneme positions in both decoding onset and peak times (one-sided Mann-Whitney U test, $p < 0.05$, **Fig. 6c, d**). The monitoring ROIs exhibited similar onsets for all positions (no significant differences between phonological sequencing), however, this network displayed positional phoneme coding in peak times (one-sided Mann-Whitney U test, $p < 0.05$, **Fig. 6c, d**), which demonstrate continuous monitoring from auditory feedback. Similar ordering of phonological sequencing was also observed in the articulation and the monitoring network in the right hemisphere (Supplementary Fig. 9). Taken together, the results show that the articulation network generates positional phonological unit codes for the execution of speech in their correct order while the monitoring network examines the phonological sequence of the utterance through feedback.

Lastly, we further investigated the neural mechanisms involved in speech sequencing during articulation. If articulation involves not only the discrete stable generation of phonological units in time but also continuously codes this phonological information, then there should also be evidence for both discrete and continuous processing in the motor cortex during speech production. To demonstrate

both discrete and continuous processing, we investigated the presented neural activation associated with the transitions between phonemes in the sequence. These transitions reflect the distribution of phoneme co-occurrences within a language. We propose that the sequences that co-occur more frequently could demonstrate different neural motor transition properties as compared to transitions that co-occur less frequently and that this mechanism would be reflected in continuous activation in the motor cortex[71,72](**Fig. 7a**).

We recorded micro-scale cortical activations in one subject (S3) using the 128-channel high-density µECoG electrode array over speech motor cortex (SMC) (**Fig. 7b**). Our results show significant HG activations (111/128 electrodes, see also patient S1 in our previous work[63]). To assess whether the transitions between phonemes were tracked in the SMC, we focused on trials with monosyllabic frames (CVC; to remove the confound of two syllables), by measuring the ability to decode the forward phonotactic transitions (P_{fwd1}: P(V2|C1) & P_{fwd2}: P(C3|V2)). The forward transitions were estimated by calculating the conditional probability of one phoneme given the previous, based on positional and co-occurrence counts. The resultant transition probabilities were quantized into three evenly distributed classes producing trials with low, medium, and high transitions scores (P_{fwd1}: low – 0.0024 to 0.034, medium – 0.034 to 0.1031, high – 0.1031 to 0.1791; P_{fwd2}: low – 0.0083 to 0.0314, medium – 0.0314 to 0.0543, high – 0.0543 to 0.0814). Time resolved decoding of phonological units (C: 5-way, V: 4-way) within the CVC trials, resulted in correct ordering of phonological unit positions in time (C1→V2→C3, **Fig. 7c**).

Next, we repeated a similar decoding analysis to predict the forward transition scores (3-way) and computed these values in time. We found that the decoding for P_{fwd1} occurred earlier in time compared to $P_{fwd2.}$ Within features, we observed significant temporal ordering of phonological units ($T_{C1} < T_{V2} < T_{C3}$; one-sided Mann Whitney U test, $p < 0.05$; $T_{C1} < T_{V2,}$ p = 1.6e-08; $T_{V2} < T_{C3}$, p = 4.2e-07, **Fig. 7d**) and phonotactic transitions ($T_{Pfwd1} < T_{Pfwd2,}$ p = 6.6e-08). Across features, the forward transitions were significantly embedded between the phonological unit positions ($T_{C1} < T_{Pfwd1} < T_{V2} < T_{Pfwd2} < T_{C3}$; pairwise one-sided Mann-Whitney U test, $p < 0.05$; $T_{C1} < T_{Pfwd1,}$ p = 3.7e-08; $T_{Pfwd1} < T_{V2}$, p = 6.6e-08, $T_{V2} < T_{Pfwd2}$, p = 6.2e-08, $T_{Pfwd2} < T_{C3}$, p = 7e-07, **Fig. 7d**). Taken together, these results indicate that the neural basis of speech sequential processing during articulation reflects both discrete activation of phonological units, as well as the continuous processing of the phonotactic transition between these units (**Fig. 7e**).

2 Discussion

In this work, we sought to quantify the neural representations for speech planning and execution (**Fig. 1a,b**)[11,13]. We performed a speech repetition task with constructed pseudo-words to isolate and

characterize planning and execution circuits involved in speech production (**Fig. 1d**). By using pseudo-words, we could remove lexical and semantic confounds, and focus on the sub-lexical stages of speech planning and execution. Behaviorally, we found that utterances with two syllables demonstrated longer spoken durations, which is consistent with their syllabic structure. Further, we also found that these utterances were associated with longer response times, suggesting that syllable complexity reflects part of the planning process prior to speech execution (**Fig. 1f, g**).

Through the examination of speech neural activations in the HG band, we characterized a broad network of neuro-anatomical ROIs in the frontal, parietal, and temporal regions that were active during speech production (**Fig. 2**). Our ability to directly record cortical activations at higher temporal resolution, allowed us to precisely track the timing of these activations and separate planning from execution processes (**Fig. 2b**). Specifically, we identified sequential neural activations (time to reach maximum HG power) in three networks of regions corresponding to their planning, motor articulation, and monitoring during speech production (**Fig. 2c**). Planning activations were observed in pre-frontal regions that included the inferior frontal gyrus (IFG), and the rostral and caudal middle frontal gyrus (MFG). In previous literature, the IFG and MFG were primarily implicated in different aspects of speech planning including conversation engagement[65], and retrieval of lexical information[64]. In this current study, using controlled stimuli in a single pseudo-word repetition task, we show that planning neural activations reflect the encoding of the speech components necessary for execution. This controlled setting allowed us to study the neural operations involved in planning sub-lexical speech features (syllables and phonemes) for execution and isolate them from higher order language processes such as lexical, semantic, and syntactic information. Neural activations associated with articulation were observed later in time, in the sensorimotor (precentral and postcentral gyrus) and inferior parietal cortices. These regions have previously been shown to code for phoneme articulatory sequences to control vocal tract gestures[14,19,58,73], as well as during phonological processing and internal speech representation[74,75]. These activations therefore reflect the motor commands necessary to execute phonological sequences. Neural activity in these motor articulatory regions have also been used to drive brain-computer interfaces; therefore, effective characterization of articulation processes is essential to achieve for accurate speech neuro-prostheses[60,63,76–80]. Finally, late activations associated with monitoring were localized to temporal regions that included primary auditory cortex, posterior and anterior superior temporal gyrus, as well as the insula. The relatively late activations are consistent with studies that examined the role of monitoring regions to generate feedback signals in order to reduce the sensitivity to self-generated sounds and to update motor commands by anticipating the sound and motor/somatosensory information produced during speech[67,81–83]. In addition to monitoring, neural activations in STG and STS have been hypothesized

to code for auditory targets during planning which is in keeping with the early activations of STS prior to response onset[84,85]. Overall, HG activations demonstrate a hierarchy of prefrontal, sensorimotor, and temporal regions that reflect the planning, articulation, and monitoring stages of speech (**Fig. 2e**).

We next analyzed different types of speech units for planning and execution. Focusing on the syllable, we estimated the time-epochs with significant syllable coding (**Fig. 3a**) as measured by contrasting HG responses between monosyllabic and disyllabic utterances. This allowed us to identify time-points with significant contrasts indicative of syllabic coding. We show that syllable information is present in the planning network regions (IFG – pars opercularis, rMFG, cMFG) at approximately 730 ms prior to articulation, demonstrating that syllables are coded as part of the representation for speech planning. This pre-articulatory syllable code is observed prior to the instructional go-cue, suggesting that the syllable frame can be a sustained process during verbal working memory (Supplementary Fig. 4). We demonstrated that the syllabic code is independent of phonemic confounds using a control analysis performed to rule out the differences in initial phoneme class (Supplementary Fig. 5). Our study differed from previous neuro-imaging research on syllable sequencing as we explicitly showed that syllable coding is present in the planning stage prior to speech execution (**Fig. 3e-f)**[11,41,43]. This finding is in accordance with previous modeling studies, which predicted the origin of syllabic codes in pre-frontal regions[4,11,86]. Lastly, the syllabic coding in articulation and monitoring network regions occurred later in time across both hemispheres, indicating that the flow of syllabic information occurs across different speech stages.

We then used high-resolution µECoG recordings (1.3 – 1.7 mm inter-electrode distance) to independently quantify this flow of syllable information at a patient specific level. Simultaneous recordings from IFG and PCG in two patients revealed distinct population sites within caudal IFG (pars-opercularis) that showed earlier syllabic coding during planning, followed by progressively later activations towards and into PCG during execution (**Fig. 4**). This ‘spatial gradient’ provides a potential mechanism for the transition from planning to execution through the sequential activation of neural populations with syllable information. Previous planning studies using ECoG identified similar spatial structures within IFG and MFG and revealed differing levels of speech selectivity[65,81]. Our results identified a similar network selective to syllable coding and their transition from planning to execution, providing novel evidence for the role of the syllable as a unit of representation in the neurobiology of speech production. Future work will focus on extracting functional connectivity between these sites and quantifying the relevance of syllable coding to these connections.

Speech production is not only dependent on syllabic structure, but it also must involve the encoding of phonemic content within the syllabic frame for the successful execution of speech during motor

articulation. Speech computational models have theorized the presence of a hierarchical organization between syllabic and phoneme activations which reflects the sequential execution of speech sounds (**Fig. 5a**)[13,14,16,27]. We show the existence of phoneme information that follows syllabic coding by using neural decoding models to resolve the emergence of phonemes in time for speech planning and execution. We find that at the network level and within individual ROIs, syllabic information is present prior to phoneme information for speech production. This temporal relationship indicates that speech neuro-anatomical areas operate at both the syllabic and phonemic levels (**Fig. 5b-d**). Interestingly, we also found that this temporal relationship exhibited compression along the three anatomical networks, with the planning network regions showing a large temporal lag between syllable and phoneme activations within the planning network (250-350 ms lag), followed by a shorter lag for the articulation network (110-130 ms), and nearly concurrent temporal activation in the monitoring regions (17-25 ms). Taken together, this difference in temporal lags suggests that 'temporal compression' may act as a mechanism for the integration of different speech representations (syllables and phonemes) in the transition from speech planning to execution (**Fig. 5e-f**). Future modeling studies could further this decoding work by using temporal generalization maps to better identify neural architectures involved in this hierarchical encoding (e.g. maintenance vs. evolution of speech neural features)[87]. These neural architectures could also potentially reveal latent neural features corresponding to other segments of the hierarchical model, such as the efference copy[14].

Speech articulation involves the precise and correct ordering of articulatory movements, and consequently the phonological units must be dynamically encoded sequentially during speech execution (**Fig. 6a**)[13]. We demonstrated the sequential generation of phonological units by accurately decoding phoneme sequences in time. These phonological sequence representations were found in the articulation and monitoring network in both hemispheres but were absent in the planning network, supporting the idea of more abstract phoneme coding during planning (**Fig. 6b,** Supplementary Fig. 9). The presence of sequencing in the articulation regions is in keeping with previous studies that examined phoneme articulatory networks in laryngeal motor cortex[58,73] (ventral) and vocalization properties in the mid pre-central gyrus[88] (dorsal), suggesting that phonological sequencing could be spatially distributed. Although our results focused on how the entire sensorimotor system is involved in generating phonological unit sequences, future studies could involve high-resolution electrodes to delineate dorsal-ventral circuity to better explain the discrete-linguistic and continuous motor-control properties of speech. The spatio-temporal distinctive characteristics of sensorimotor systems can further inform the design of speech-BCI algorithms to simultaneously model the spatial gradient[62] and temporal sequencing[76–78,89] of speech elements for accurate speech neuro-prostheses. The presence

of neural activations sensitive to sequencing in the monitoring network regions could be attributed to both perception of self-generated sounds[67] and comparison of sounds with the generation of auditory efferent copies for intended speech[14]. The separation of these two elements could be examined in future work by identifying causal connections between monitoring and articulation regions, and using these connections to help explain how the current state of phonological unit generation is modulated by inputs from auditory feedback.

In this work, we also sought to characterize one of the mechanisms that can facilitate articulatory sequencing during speech production. Over years of language practice through speaking, adult speakers have entrained sensorimotor systems to enable them to fluently articulate linguistically familiar speech sounds[4,8,16,43,72]. This linguistic tuning can be further quantified by measuring the statistical properties between phonemes: phonotactics[72]. Although the sensory (auditory) encoding of phonotactic distributions have been examined during speech perception[71], the motor component of internalized phonotactics has not yet been established. In this study, the use of phonologically constructed pseudo-words allowed us to calculate the statistical transitionary properties between phonemes. Using a decoding approach from high-resolution recordings in sensorimotor cortex, we observed robust and stable neural activation patterns associated with phonotactic transitions. Further, these neural activations occurred between their phoneme positions (**Fig. 7)**. These findings suggest evidence for neural tracking of speech motor activations through phonotactics, that enable continuous and fluent execution of speech sounds with years of native language use. Future studies could combine high-resolution recordings and stimuli with more complex syllabic structures to further study how sensorimotor systems plan and execute speech.

Overall, this work highlights potential mechanisms for the neural basis of the transition from speech planning to execution. The usage of intracranial and high-resolution recordings enabled us to identify spatio-temporal activations across and within speech networks that were specific to different levels of processing, as well as their transitional states. This work could have important implications for the development of therapeutic interventions for patients with speech and language disorders, where the ability to plan and implement speech coordinated activity is disrupted.

3 Methods

This study was performed at the Duke University Health System with approval from the Institutional Review Board under the protocol IDs:

Pro00065476: Network dynamics of human cortical processing.

Pro00072892: Studying human cognition and neurological disorders using µECoG electrodes.

3.1 Participants

Intracranial recordings with standard clinical intracranial recording systems were obtained from 52 patients with pharmaco-resistant epilepsy at Duke University Medical Center (DUMC) (mean age = 33.1 years, 23 Female, Supplementary table 1). Patients were implanted with clinical intracranial electro-encephalography (IEEG) systems as part of their work up for the neurosurgical treatment of pharmacologically resistant epilepsy. Recordings using high-density micro-scale cortical recordings were obtained in an intraoperative setting from 3 adult patients (mean age = 64.3 years, 0 Female), during awake neuro-surgical procedures. Patients were awoken as part of their surgical procedure and the decision to have them awake was entirely clinically motivated. Two of the three patients (S1, and S3) underwent the placement of a deep brain stimulator (DBS) for the treatment of movement disorders. The other patient (S2) underwent surgical resection for tumor removal. All patients were native English speakers and written informed consent was obtained in accordance with the Institutional Review Boards at the DUMC.

3.2 Electrode implantation, localization and neural recordings

Cortical recordings were obtained using electrodes from either one of the three recording systems: 1) macro-electrocorticographic grids (macro-ECoG: 2 participants), 2) stereo-electroencephalographic probes (SEEG: 50 participants), and 3) liquid crystal polymer thin-film (LCP-TF) micro-electrocorticographic arrays (µECoG: 3 participants). Macro-ECoG grid recordings were done using Ad-Tech electrode grids (48 or 64 electrodes; Ad-Tech Medical Instrument Corporation) with 10 mm electrode spacing and 2.3 mm exposed diameter, surgically implanted onto the cortex of the patients. SEEG placement was done via robotically guided surgery (ROSA Robotics) using either PMT (PMT Corporation), Ad-Tech, or Dixi (Dixi Medical) SEEG electrodes with 3.5 mm to 5 mm inter-electrode spacing (0.8 to 0.86 mm exposed diameter). µECoG recording included LCP-TF µECoG arrays (1.33 – 1.72 mm electrode spacing; 200 µm diameter), acutely implanted either through a burr hole (DBS surgery) or directly over the cortical surface (awake craniotomy).

Pre-operative anatomical 3T MRI were obtained for all patients, and post-operative CT images were obtained for both macro IEEG and DBS patients to identify electrode locations. Patients undergoing DBS surgery underwent a routine intraoperative CT for assessment of DBS electrode placement with the subdural ECoG array in place. For the tumor resection patient, BrainLab coordinates[90,91] were obtained at the corner of the electrode array and this information was then mapped to three-dimensional reconstruction of the patient's brain obtained from their preoperative MRI. For each patient, electrode locations were manually marked using BioImage Suite, and were mapped on to the reconstructed T1 volume generated using Freesurfer[92,93]. Electrodes were then assigned to specific regions of interest (ROIs), based on the cortical parcellations obtained from connectivity based

Brainnetome Atlas[94]. For the electrodes that were not centered at specific ROIs (e.g. depth electrodes from SEEG probes near/in white matter), a sphere with a 10 mm radius was constructed around the electrode's center, and the relative density of ROIs was determined through the calculation of the percentage proportion of ROI voxels within this sphere. Electrodes were then assigned to a specific ROI with the maximum density if the proportion exceeded 5% (otherwise they were labelled as white matter contacts). For visualization, electrode coordinates were also transformed to a population averaged MNI brain and color-coded for ROIs.

IEEG recordings in the clinical monitoring unit (ECoG and SEEG) were analog filtered between 0.01 Hz and 1000Hz and digitized at 2,048 samples per second, using a Natus Quantum LTM amplifier (with Natus Neuroworks EEG software). Neural recordings from µECoG electrodes (intra-operative) were analog filtered between 0.1 Hz and 7500 Hz and digitized at 20,000 samples per second (with Intan RHX data acquisition software, v3.0; Intan Technologies, Inc).

3.3 Delayed speech repetition task

Patients performed a delayed pseudo-word repetition task in the clinical monitoring unit. In each trial, they listened and repeated one of 52 pseudo-words that were modelled on the phonology of American English. Pseudo-words were equally distributed between monosyllabic and di-syllabic frames (consonant-vowel-consonant - CVC: 26) and vowel-consonant-vowel (VCV: 26 tokens) and included a total of 9 phonemes (4 vowels: /a/, /ae/, /i/, /u/ and 5 consonants: /b/, /p/, /v/, /g/, /k/). The phonemes were arranged within the pseudo-words (CVC or VCV) by selecting biphones (CV or VC) with a positive likelihood of occurring in the English language corpus[95]. The third phoneme was selected with the constraint that the occurrence probability of the triphone sequence was zero, resulting in a constructed pseudo-word. For each pseudoword, forward biphoneme transition probabilities were calculated based on the occurrence probabilities of individual phonemes [P(C) or P(V)] and co-occurring biphones [P(CV) and P(VC)] as follows:

$P(V|C) = \frac{P(CV)}{P(C)}$, and $P(C|V) = \frac{P(VC)}{P(V)}$.

The resultant transitional probability scores across the pseudoword tokens were quantized into three classes: low, medium, and high transitions scores (P(V|C): low – 0.0024 to 0.034, medium – 0.034 to 0.1031, high – 0.1031 to 0.1791; P(C|V): low – 0.0083 to 0.0314, medium – 0.0314 to 0.0543, high – 0.0543 to 0.0814).

The constructed pseudo-words were auditorily presented and subsequently cued for the patients to repeat via a visual Go Cue ('speak'). Each trial lasted between 5.1 – 6.1 seconds, starting with a visual "Listen" cue followed by the auditory stimulus presentation (mean stimulus duration: 0.5

second). After a 1.5 s delay (150 ms jitter), the “speak” (Go) cue was visually presented and the patients were given 3 s (200 ms jitter) to respond, before the next trial began. Trials were presented in four consecutive blocks with 52 unique pseudo-word tokens shuffled in each block and the overall task time with 208 speech trials took less than 40 minutes in the monitoring unit. A subset of patients underwent a similar delayed repetition task where they repeated pseudo-words with di-syllabic frames (CVCVC). The patients in the intra-operative setting underwent a similar version of pseudo-word repetition task, but with the delay and “speak” cue removed for simplicity. Intraoperative patients completed 3 to 4 blocks of 52 trials (156 – 208 trials). The speech task was designed using Psychtoolbox scripts in MATLAB R2014a.

3.4 Audio recording and behavioral representation

The speech repetition task included both auditory and speaking components. The auditory stimulus wave files were presented using a laptop connected to a powered stereo-speaker (amazonbasics A100 series) through a USB DAC and audio amplifier (Fiio). The beginning of each trial and the subsequent stimulus onsets were marked by flashing white circles on the stimulus laptop screen which was detected by a photodiode (Thorlabs, Inc.) attached to an auxiliary BNC-to-mono TS cable connected to the Natus Recording system. The audio waveforms were recorded using a condenser microphone (Marantz Professional, MPM-1000) which was connected to pre-amplifier (Behringer) and digitized at 44100 Hz on the recording laptop. The auditory stimulus, the spoken pseudo-word onsets, and the offsets were either manually identified using the Audacity software package, with the built-in Mel-spectrogram visualization feature or automated through the Montreal Forced Aligner (Montreal Corpus Tools).

For behavior analysis, each trial was characterized by its response time (with respect to Go cue onset) and its response duration. For the neural analysis, the trials were categorically coded for their syllabic frame (CVC or VCV), the phonemes (9-way) at each position, and the corresponding forward transition category (3-way) for each biphoneme pair within its frame. Trials with no intelligible responses were removed from the analysis.

3.5 Behavioral representations of syllabic coding

A linear mixed effects (LME) model was used to predict the effects of syllabic frame on response time and response duration. The models incorporated fixed effects for categorical syllabic IDs (one vs. two syllables) and random effects for patient and an intercept. Two sets of models were utilized separately to predict response time and response duration respectively, and in each case the estimate of syllabic frame (beta coefficient for fixed effect) was computed. The significance of the

beta estimates was computed using a permutation test by randomly shuffling syllabic frames (1000 shuffles) and repeating the LME prediction for each shuffle.

3.6 Neural preprocessing and signal analysis

The neural preprocessing steps were performed offline. The IEEG recordings were filtered for line-noise with multi-taper band-stop notches centered at 60, 120, 180, and 240 Hz. The following categories of electrodes were removed from subsequent analyses: 1) electrodes that were outside the cortex (determined from localization), 2) electrodes that exhibited muscle artifacts during the utterance (as visualized by a wavelet scalogram), and 3) electrodes with recording power greater than 3 * log-RMS. From a total of 9679 electrodes, 1573 were excluded using the above steps. The remaining set of 8106 electrodes were subjected to common-average-referencing (CAR) and utilized for further processing.

3.6.1 Neural feature extraction

Analyses were performed to extract neural high-gamma signals (HG: 70 – 150 Hz) from each electrode. HG signals have previously been shown to characterize speech activations from cortical regions as well as demonstrating a high correlation with multi-unit firing[54,56,57]. The pre-processed neural signals from each IEEG electrode were band-pass filtered into 8 center frequencies (logarithmically arranged) between 70 and 150 Hz[96]. The envelope of each frequency band was extracted using the Hilbert transform and was averaged across the bands. The resultant envelope was then downsampled to 200 Hz using an anti-aliasing filter and z-scored with respect to a baseline period (-500 ms to 0 s before trial start). For each trial, we defined the response epoch as the time intervals from -1500 ms to 1500 ms with respect to the utterance onset.

3.6.2 Neural representations of syllabic frame coding

A non-parametric statistical test was performed to test whether HG power from each ROI could differentiate between syllabic frames (CVC vs. VCV) at each time-point. For each electrode, the absolute difference in HG power (ΔHG) was computed between VCV trials and CVC trials. The resultant difference (across-frame) was averaged across electrodes within a given ROI (within patient followed by across patient mean) and was considered as the test statistic. For a control analysis, ΔHG was computed within CVC trials and within VCV trials, and the average within-frame ΔHG was determined for each ROI. The null distribution was generated from 1000 iterations of the within-frame statistic for each ROI, and the significance was determined by calculating the proportion of null-distribution (within-frame) less than the test-statistic (across-frame) ($p < 0.05$). The resultant p-value time-series were subjected to a second-level non-parametric cluster correction to determine consecutive sequence of time-points that exhibited significant syllabic frame coding.

To rule out the confounding of initial phonemes (C vs. V) in the neural syllable code, a control analysis was performed in a subset of patients who underwent a modified version of speech repetition task, where they repeated pseudowords with di-syllabic frames (CVCVC). Similar to the above method, the absolute difference in HG power was computed between CVCVC trials and CVC trials as the test statistic and was subjected to second-level non-parametric cluster correction.

3.6.3 Linear classification model for speech decoding

A previously established linear-classification model with 10-fold cross-validation, was used to decode speech units from normalized HG[63]. First, we performed a singular value decomposition (SVD) to compress HG to low-dimensional features. We chose the number of kept components as the total number of components that explained 80% of the variance. We utilized a supervised linear-discriminant (LDA) model to predict speech units. The decoding performance was estimated by calculating the average prediction accuracy in the test set across all folds.

Multivariate decoding was performed by pooling HG epochs from electrodes across patients. To perform this pooling, trials with similar tokens across patients were concatenated together. In instances where there was a mismatch in the number of trials across patients for a particular token, a synthetic trial was generated using a Mix-Up technique through the linear combination of two randomly selected trials for that token[97]. This Mix-Up approach was performed separately on the training and the testing set to avoid data leakage.

To infer temporal characteristics of speech features, decoding was performed in the temporal domain by training and testing at each time-segment (200 ms window; 10 ms step size) within the epoch[87]. The decoding accuracy was calculated at each time-segment to characterize temporally evolving representations of speech features in the response epoch. Separate temporal models were evaluated per ROI to decode syllabic frames (2-way, chance: 50%), phoneme units (9-way, chance: 11.11%) in each position (P1, P2, and P3), and phoneme transition probabilities (3-way, chance: 33.33%) for each biphone (P_{fwd1}, and P_{fwd2}) within the pseudo-word. The resultant accuracy (Acc_{Test}), as a percentage, was normalized (Acc_{Norm}) with respect to theoretical chance (Acc_{Chance}) for visualization purposes as follows:

$$Acc_{norm} = \frac{Acc_{Test} - Acc_{Chance}}{100 - Acc_{Chance}}$$

3.6.4 Statistical analysis

Multiple statistical tests were conducted to validate IEEG recordings for speech representations. To identify electrodes with significant speech-neural activations, signal power in the HG band was

compared between the response (-500 ms to 500 ms) and the baseline period using a 1000-iteration one-sided permutation test (corrected for false discovery rate).

To quantify the neural morphology of HG responses for each ROI, the above non-parametric test was extended to determine the significance at each time point (against baseline), with HG power averaged across electrodes within the ROI. In addition, the HG onset times, and peak times were calculated for each electrode. The onset times were estimated by performing a change-point analysis which determines the time-point at which the mean of the HG time-series changes most significantly (findchangepts algorithm in MATLAB). Peak times were determined by either finding the time-point where HG activations achieved the maximum value or by computing the weighted mean of HG activations over time. The distribution of HG onset and peak times were used to quantify the evoked temporal characteristics of a speech responsive electrode. Significant differences in onset or peak times were evaluated using pairwise Mann-Whitney U test, or student t-tests.

For speech decoding, the prediction accuracies were validated against the null distribution obtained from random shuffling of the predictor labels (250 iteration). ROI characteristic decoding onsets were obtained using a similar change-point analysis. The resultant time-series was subjected to second-level non-parametric temporal cluster correction.

References

1. Guenther, F. H., Hampson, M. & Johnson, D. A theoretical investigation of reference frames for the planning of speech movements. *Psychol. Rev.* **105**, 611 (1998).
2. Houde, J. F. & Jordan, M. I. Sensorimotor adaptation in speech production. *Science (80-.).* **279**, 1213–1216 (1998).
3. Brandi, M.-L., Wohlschläger, A., Sorg, C. & Hermsdörfer, J. The Neural Correlates of Planning and Executing Actual Tool Use. *J. Neurosci.* **34**, 13183 LP – 13194 (2014).
4. Indefrey, P. & Levelt, W. J. M. The spatial and temporal signatures of word production components. *Cognition* **92**, 101–144 (2004).
5. Levelt, W. J. M., Roelofs, A. & Meyer, A. S. A theory of lexical access in speech production. *Behav. Brain Sci.* **22**, 1–38 (1999).
6. Roelofs, A. The WEAVER model of word-form encoding in speech production. *Cognition* **64**, 249–284 (1997).
7. Fiez, J. A., Balota, D. A., Raichle, M. E. & Petersen, S. E. Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading. *Neuron* **24**, 205–218 (1999).
8. Mechelli, A., Friston, K. J. & Price, C. J. The effects of presentation rate during word and pseudoword reading: a comparison of PET and fMRI. *J. Cogn. Neurosci.* **12**, 145–156 (2000).
9. Fromkin, V. A. The non-anomalous nature of anomalous utterances. *Language (Baltim).* **47**, 27–52 (1971).

10. Shattuck-Hufnagel, S. The role of word structure in segmental serial ordering. *Cognition* **42**, 213–259 (1992).

11. Rong, F., Isenberg, A. L., Sun, E. & Hickok, G. The neuroanatomy of speech sequencing at the syllable level. *PLoS One* **13**, e0196381 (2018).

12. Geschwind, N. The organization of language and the brain. *Science* **170**, 940–944 (1970).

13. Hickok, G. The architecture of speech production and the role of the phoneme in speech processing. *Lang. Cogn. Neurosci.* **29**, 2–20 (2014).

14. Hickok, G. Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* **13**, 135–145 (2012).

15. Houde, J. F. & Nagarajan, S. S. Speech production as state feedback control. *Front. Hum. Neurosci.* **5**, 82 (2011).

16. Bohland, J. W., Bullock, D. & Guenther, F. H. Neural representations and mechanisms for the performance of simple speech sequences. *J. Cogn. Neurosci.* **22**, 1504–1529 (2010).

17. Baars, B. J., Motley, M. T. & MacKay, D. G. Output editing for lexical status in artificially elicited slips of the tongue. *J. Verbal Learning Verbal Behav.* **14**, 382–391 (1975).

18. Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M. & Gagnon, D. A. Lexical access in aphasic and nonaphasic speakers. *Psychol. Rev.* **104**, 801 (1997).

19. Cogan, G. B. *et al.* Sensory-motor transformations for speech occur bilaterally. *Nature* **507**, 94–8 (2014).

20. Cogan, G. B. *et al.* Manipulating stored phonological input during verbal working memory. *Nat. Neurosci.* **20**, 279–286 (2017).

21. Brown, S., Ingham, R. J., Ingham, J. C., Laird, A. R. & Fox, P. T. Stuttered and fluent speech production: an ALE meta-analysis of functional neuroimaging studies. *Hum. Brain Mapp.* **25**, 105–117 (2005).

22. Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J. & Fedorenko, E. Functional characterization of the human speech articulation network. *Cereb. Cortex* **28**, 1816–1830 (2018).

23. Broca, P. Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bull. Mem. la Soc. Anat. Paris* **6**, 330–357 (1861).

24. Tourville, J. A., Nieto-Castañón, A., Heyne, M. & Guenther, F. H. Functional parcellation of the speech production cortex. *J. Speech, Lang. Hear. Res.* **62**, 3055–3070 (2019).

25. Wise, R. J. S., Greene, J., Büchel, C. & Scott, S. K. Brain regions involved in articulation. *Lancet* **353**, 1057–1061 (1999).

26. Turkeltaub, P. E., Eden, G. F., Jones, K. M. & Zeffiro, T. A. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* **16**, 765–780 (2002).

27. Tourville, J. A. & Guenther, F. H. The DIVA model: A neural theory of speech acquisition and production. *Lang. Cogn. Process.* **26**, 952–981 (2011).

28. Indefrey, P. The spatial and temporal signatures of word production components: a critical update. *Front. Psychol.* **2**, 255 (2011).

29. Guenther, F. H. *Neural control of speech*. (Mit Press, 2016).

30. Eickhoff, S. B., Heim, S., Zilles, K. & Amunts, K. A systems perspective on the effective connectivity of overt speech production. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **367**, 2399–2421 (2009).

31. Berger, M., Agha, N. S. & Gail, A. Wireless recording from unrestrained monkeys reveals motor goal encoding beyond immediate reach in frontoparietal cortex. *Elife* **9**, 305334 (2020).

32. Gale, D. J., Flanagan, J. R. & Gallivan, J. P. Human somatosensory cortex is modulated during motor planning. *J. Neurosci.* **41**, 5909–5922 (2021).

33. Sober, S. J. & Sabes, P. N. Flexible strategies for sensory integration during motor planning. *Nat. Neurosci.* **8**, 490–497 (2005).

34. Crevecoeur, F., Cluff, T. & Scott, S. H. Computational approaches for goal-directed movement planning and execution. (2014).

35. Perkell, J. *et al.* Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Commun.* **22**, 227–250 (1997).

36. Perkell, J. S. Movement goals and feedback and feedforward control mechanisms in speech production. *J. Neurolinguistics* **25**, 382–407 (2012).

37. McGuire, P. K., Silbersweig, D. A. & Frith, C. D. Functional neuroanatomy of verbal self-monitoring. *Brain* **119**, 907–917 (1996).

38. Hirano, S. *et al.* Cortical processing mechanism for vocalization with auditory verbal feedback. *Neuroreport* **8**, 2379–2382 (1997).

39. Wheeldon, L. R. & Levelt, W. J. M. Monitoring the time course of phonological encoding. *J. Mem. Lang.* **34**, 311–334 (1995).

40. Levelt, W. J. M. Monitoring and self-repair in speech. *Cognition* **14**, 41–104 (1983).

41. Bohland, J. W. & Guenther, F. H. An fMRI investigation of syllable sequence production. *Neuroimage* **32**, 821–841 (2006).

42. Segawa, J. A., Tourville, J. A., Beal, D. S. & Guenther, F. H. The neural correlates of speech motor sequence learning. *J. Cogn. Neurosci.* **27**, 819–831 (2015).

43. Peeva, M. G. *et al.* Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. *Neuroimage* **50**, 626–638 (2010).

44. Parvizi, J. & Kastner, S. Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.* **21**, 474–483 (2018).

45. Lachaux, J. P., Rudrauf, D. & Kahane, P. Intracranial EEG and human brain mapping. *J. Physiol.* **97**, 613–628 (2003).

46. Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E. & Crone, N. E. High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research. *Prog. Neurobiol.* **98**, 279–301 (2012).

47. Mullin, J. P. *et al.* Is SEEG safe? A systematic review and meta-analysis of stereo-electroencephalography–related complications. *Epilepsia* **57**, 386–401 (2016).

48. Chiang, C.-H. *et al.* Development of a neural interface for high-definition, long-term recording in rodents and nonhuman primates. *Sci. Transl. Med.* **12**, eaay4682 (2020).

49. Chiang, C.-H. *et al.* Flexible, high-resolution thin-film electrodes for human and animal neural research. *J. Neural Eng.* **18**, (2021).

50. Trumpis, M. *et al.* Sufficient sampling for kriging prediction of cortical potential in rat, monkey, and human $μ$ECoG. *Submiss.* (2020).

51. Sun, J. *et al.* Intraoperative microseizure detection using a high-density micro-electrocorticography electrode array. *Brain Commun.* **4**, fcac122 (2022).

52. Barth, K. J. *et al.* Flexible, high-resolution cortical arrays with large coverage capture microscale high-frequency oscillations in patients with epilepsy. *Epilepsia*.

53. Kellis, S. S., House, P. A., Thomson, K. E., Brown, R. & Greger, B. Human neocortical electrical activity recorded on nonpenetrating microwire arrays: applicability for neuroprostheses. *Neurosurg. Focus* **27**, E9 (2009).

54. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).

55. Rowland, N. C. *et al.* Task-related activity in sensorimotor cortex in Parkinson's disease and essential tremor: changes in beta and gamma bands. *Front. Hum. Neurosci.* **9**, (2015).

56. Dubey, A. & Ray, S. Cortical Electrocorticogram (ECoG) is a local signal. *J. Neurosci.* (2019).

57. Leonard, M. K. *et al.* Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* **626**, 593–602 (2024).

58. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).

59. Cheung, C., Hamilton, L. S., Johnson, K. & Chang, E. F. The auditory representation of speech sounds in human motor cortex. *Elife* **5**, e12577 (2016).

60. Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A. & Chang, E. F. The speech neuroprosthesis. *Nat. Rev. Neurosci.* (2024) doi:10.1038/s41583-024-00819-9.

61. Muller, L., Hamilton, L. S., Edwards, E., Bouchard, K. E. & Chang, E. F. Spatial resolution dependence on spectral frequency in human speech cortex electrocorticography. *J. Neural Eng.* **13**, (2016).

62. Angrick, M. *et al.* Speech Synthesis from ECoG using Densely Connected 3D Convolutional Neural Networks. *Biorxiv* (2018).

63. Duraivel, S. *et al.* High-resolution neural recordings improve the accuracy of speech decoding. *Nat. Commun.* **14**, 6938 (2023).

64. Wang, Y., Korzeniewska, A., Usami, K., Valenzuela, A. & Crone, N. E. The dynamics of language network interactions in lexical selection: an intracranial EEG study. *Cereb. Cortex* **31**, 2058–2070 (2021).

65. Castellucci, G. A., Kovach, C. K., Howard, M. A., Greenlee, J. D. W. & Long, M. A. A speech planning network for interactive language use. *Nature* **602**, 117–122 (2022).

66. Kleiner, M., Brainard, D. & Pelli, D. What's new in Psychtoolbox-3? (2007).

67. Flinker, A. *et al.* Single-Trial Speech Suppression of Auditory Cortex Activity in Humans. *J. Neurosci.* **30**, 16643 LP – 16650 (2010).

68. Mitchell, R. L. C. & Crow, T. J. Right hemisphere language functions and schizophrenia: the forgotten hemisphere? *Brain* **128**, 963–978 (2005).

69. Luks, T. L. *et al.* Atrophy in two attention networks is associated with performance on a Flanker

task in neurodegenerative disease. *Neuropsychologia* **48**, 165–170 (2010).

70. van Turennout, M., Hagoort, P. & Brown, C. M. Electrophysiological evidence on the time course of semantic and phonological processes in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* vol. 23 787–806 at https://doi.org/10.1037/0278-7393.23.4.787 (1997).

71. Leonard, M. K., Bouchard, K. E., Tang, C. & Chang, E. F. Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. *J. Neurosci.* **35**, 7203 LP – 7214 (2015).

72. Vitevitch, M. S., Luce, P. A., Charles-Luce, J. & Kemmerer, D. Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. *Lang. Speech* **40**, 47–62 (1997).

73. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–32 (2013).

74. Oberhuber, M. *et al.* Four Functionally Distinct Regions in the Left Supramarginal Gyrus Support Word Processing. *Cereb. Cortex* **26**, 4212–4226 (2016).

75. Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M. & Lœvenbruck, H. What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behav. Brain Res.* **261**, 220–239 (2014).

76. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).

77. Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nat. Neurosci.* **23**, 575–582 (2020).

78. Sun, P., Anumanchipalli, G. K. & Chang, E. F. Brain2Char: a deep architecture for decoding text from brain recordings. *J. Neural Eng.* **17**, 66015 (2020).

79. Wilson, G. H. *et al.* Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. *J. Neural Eng.* **17**, 66007 (2020).

80. Wandelt, S. K. *et al.* Representation of internal speech by single neurons in human supramarginal gyrus. *Nat. Hum. Behav.* (2024) doi:10.1038/s41562-024-01867-y.

81. Flinker, A. *et al.* Redefining the role of Broca's area in speech. *Proc. Natl. Acad. Sci.* **112**, 2871–2875 (2015).

82. Forseth, K. J., Hickok, G., Rollo, P. S. & Tandon, N. Language prediction mechanisms in human auditory cortex. *Nat. Commun.* **11**, 5240 (2020).

83. Hamilton, L. S., Edwards, E. & Chang, E. F. A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Curr. Biol.* **28**, 1860-1871.e4 (2018).

84. Nourski, K. V *et al.* Electrophysiology of the human superior temporal sulcus during speech processing. *Cereb. Cortex* **31**, 1131–1148 (2021).

85. Kriegstein, K. V & Giraud, A.-L. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* **22**, 948–955 (2004).

86. Khanna, A. R. *et al.* Single-neuronal elements of speech production in humans. *Nature* **626**, 603–610 (2024).

87. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).

88. Silva, A. B. *et al.* A Neurosurgical Functional Dissection of the Middle Precentral Gyrus during Speech Production. *J. Neurosci.* **42**, 8416 LP – 8426 (2022).

89. Livezey, J. A., Bouchard, K. E. & Chang, E. F. Deep learning as a tool for neural data analysis: Speech classification and cross-frequency coupling in human sensorimotor cortex. *PLOS Comput. Biol.* **15**, e1007091 (2019).

90. No Title. https://brainmappinglab.org/.

91. Gupta, D., Hill, N. J., Adamo, M. A., Ritaccio, A. & Schalk, G. Localizing ECoG electrodes on the cortical anatomy without post-implantation imaging. *NeuroImage. Clin.* **6**, 64–76 (2014).

92. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).

93. Iglesias, J. E. *et al.* SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Sci. Adv.* **9**, eadd3607 (2023).

94. Fan, L. *et al.* The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb. Cortex* **26**, 3508–3526 (2016).

95. Vitevitch, M. S. & Luce, P. A. A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behav. Res. Methods, Instruments, Comput.* **36**, 481–487 (2004).

96. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–10 (2014).

97. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. MixUp: Beyond empirical risk minimization. *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.* 1–13 (2018).

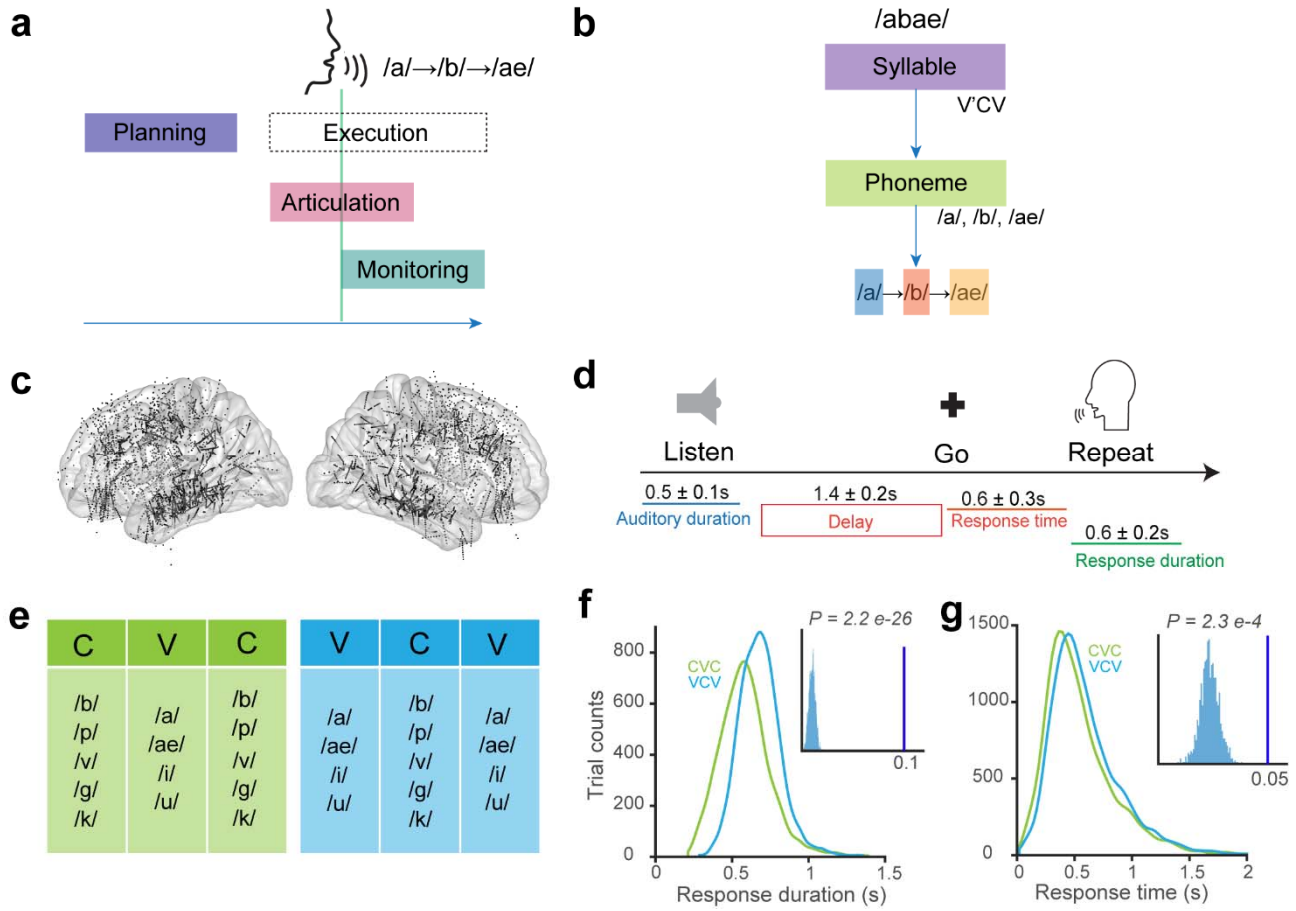


Figure 1. **Pseudo-word repetition task to study planning and execution for speech production. a.** Graphical representatio of speech neural mechanisms from planning of intended speech to execution of speech sounds. Articulatory representations fror planning to execution, which involves motor articulation followed by auditory monitoring of vocal feedback. **b.** Speech plannin and execution at the sublexical level (post word formation) involves the composition of syllable frames and the incorporation c phonemes within these frames for the execution of the ordered speech sequence. **c.** Location of all 8106 recorded intracrania electrodes from 52 patients in both left and right hemispheres. **d.** Schematic of pseudo-word repetition task. Color bars indicat the duration of the auditory stimulus (blue), response time from the Go cue (orange), and response duration (green). **e.** Pseudc word tokens used for the speech repetition task differed on their syllabic frame: consonant-vowel-consonant (CVC, green) an vowel-consonant-vowel (VCV, blue). The phonemes were chosen from a fixed for distribution for both consonants (5) and vowel (9) at each position within the pseudo-word. **f.** Histogram distributions of response duration for both CVC and VCV trials. Th durations of VCV trials were significantly longer than the CVC trials (linear mixed effects [LME] modeling, *p = 2.2e-26*). The dar blue line indicates the beta estimate of the actual LME model, and the histograms (light blue) indicate the estimated distribution from permuted models (1000 permutations). **g.** Histogram distributions of response times for both CVC and VCV trials. Th response time of CVC trials were significantly faster than the VCV trials (linear mixed effects [LME] modeling, *p = 2.3e-4*). Th dark blue line indicates the estimate of the actual LME model, and the histograms (light blue) indicate the estimate distribution from permuted models (1000 permutations).

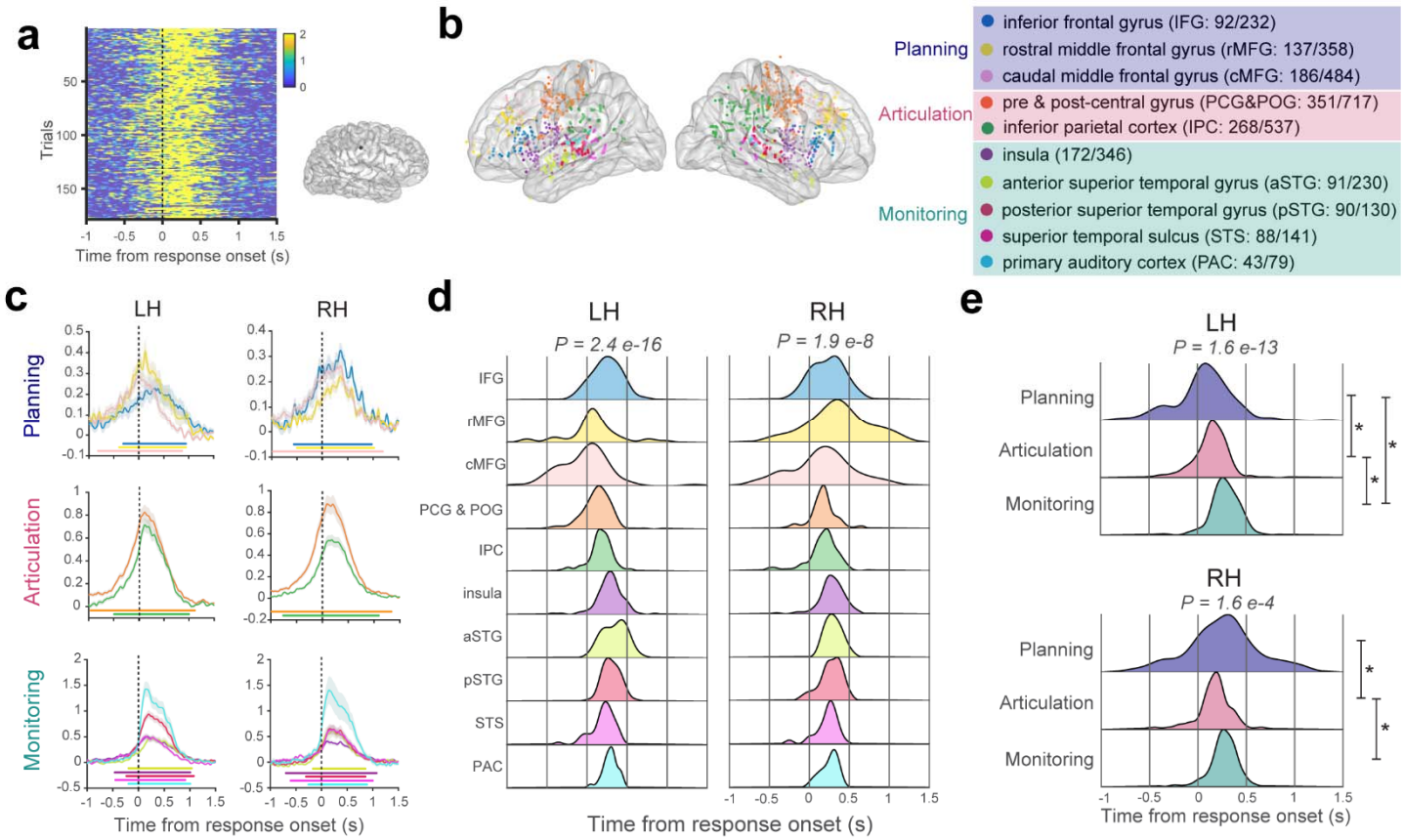


Figure 2. **Cortical activations from intracranial recordings during speech production. a.** Speech neural activations in the high-gamma band (HG: 70 – 150 Hz; z-score normalized with respect to baseline) of an example electrode in the pre-central gyrus that showed significant modulation during the speech utterance at the single trial level (FDR-corrected non-parametric permutation test). **b.** Location of all 3534 intracranial electrodes in both left (1701) and right (1605) hemisphere, with significant HG modulation during the speech utterance. The electrodes were color-coded with respect to anatomical regions-of-interest (ROIs), and were assigned to planning, articulation, and monitoring networks based on previous literature. The numbers in the brackets indicate the number of significant electrodes with respect to the total number of recorded electrodes per ROI. **c.** Mean HG activations of ROIs during response onset for both left and right hemispheres, averaged within patient. Shaded area denotes standard error. Colored horizontal bars indicate time-periods of significant activation above baseline (one-sided permutation test with cluster correction, $p<0.05$). **d.** Electrode distributions of peak HG time-points per ROI for each hemisphere. Significant differences in peak activations were observed in both left and right hemispheres, and there was a general trend of temporal progression of neural activations across ROIs around the speech utterance (one-way ANOVAs: left: $F_{9,717}$ = 10.87, $p = 2.4e\text{-}16$, right: $F_{9,781}$ = 6.69, $p = 1.9e\text{-}08$). **e.** Electrode distributions of peak HG time-points per ROI groups for each hemisphere. This temporal progression was also observed via significant differences in peak activations in both left and right hemispheres (one-way ANOVAs: left: $F_{2,724}$ = 30.7, p = $1.6e\text{-}13$, right: $F_{2,788}$ = 8.86, p = $1.6e\text{-}4$). Post-hoc t-tests indicate pairwise significance between the speech network: left hemisphere: $T_{Planning} < T_{Articulation}$, $t_{(365)} = 2.938$, $p = 0.0018$; $T_{Articulation} < T_{Monitoring}$, $t_{(507)} = 9.25$, $p = 3e\text{-}13$, $T_{Planning} < T_{Monitoring}$, $t_{(394)} = 9.84$, $p = 7.2e\text{-}21$, right hemisphere: $T_{Planning} > T_{Articulation}$, $t_{(511)} = 2.3$, $p = 0.0214$; $T_{Articulation} < T_{Monitoring}$, $t_{(528)} = 7.53$, $p = 2.2e\text{-}13$, $T_{Planning} < T_{Monitoring}$, $t_{(385)} = 2.36$, $p = 0.0188$. These analyses demonstrate the temporal progression of neural responses for planning, articulation, and monitoring for speech production in the left hemispheric speech network.

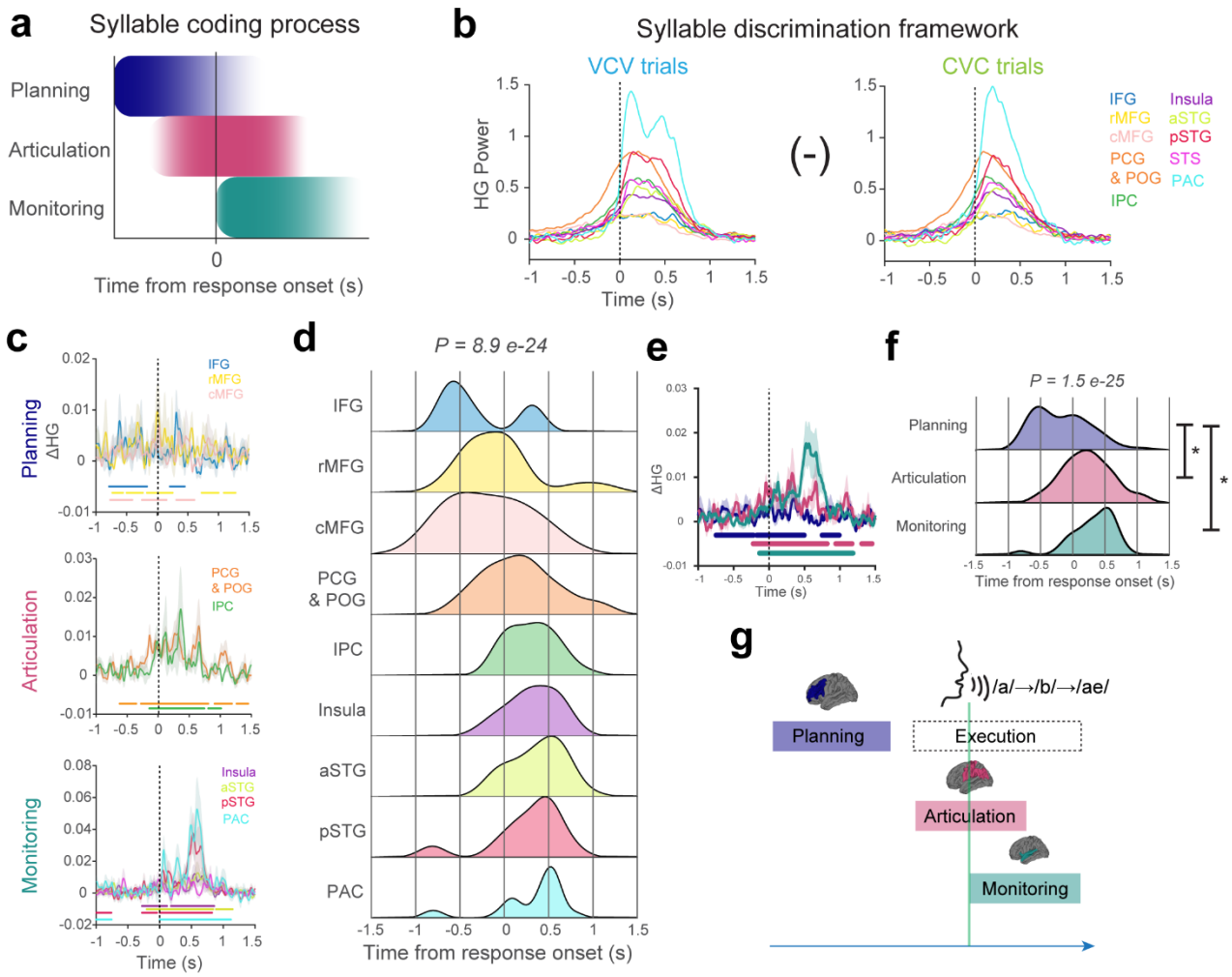


Figure 3. **Neural code for syllable frames in the left hemisphere. a.** Graphical schematic of the hypothesized temporal progression of the neural activation of syllable frames for speech planning, articulation, and monitoring. Syllable frames are theorized to be composed during planning and transferred to articulation and monitoring networks during speech execution. **b.** We analyze the neural code for syllable frames through the use of a syllable discrimination framework: ROI-grouped HG activations of VCV trials (left) were contrasted (subtracted) with CVC trials (right). Each colored trace indicates mean neural activations obtained per syllabic frame for each ROI in the left hemisphere, averaged within subject. Standard error for each ROI was omitted for visual clarity. **c.** Contrasting syllabic frames per ROI indicate time points with significant syllable discrimination around the speech utterance. Plots show mean ΔHG contrasts of VCV versus CVC frame for each ROI aligned to the response onset. Shaded area denotes standard error. Horizontal bars indicate durations of significant contrasts (across vs. within frame, one-sided permutation test with cluster correction, $p<0.05$). **d.** Histograms indicate electrode distributions of peak ΔHG contrasts per ROI. Significant differences in peak ΔHG were found (one-way ANOVA, $F_{8,489}$ = 18.1, p = *8.9e-24*). These differences followed a general temporal progression of syllable activations corresponding to syllable frames, around the utterance **e.** Contrasting syllabic frames per ROI group indicate time points with significant syllable discrimination for planning, articulation and monitoring networks. Plots show mean ΔHG contrasts of VCV versus CVC frame for each network aligned to the response onset. Shaded area denotes standard error. Horizontal bars indicate durations of significant contrasts (across vs. within frame, one-sided permutation test with cluster correction, $p<0.05$). **f.** Histograms show electrode distributions of peak ΔHG contrasts per network: planning, articulation, and monitoring. Significant differences in peak ΔHG (one-way ANOVA, $F_{2,495}$ = 64.3, p = *1.5e-25*), followed by post-hoc t-tests indicate pairwise significance in syllabic frame contrasts (ΔHG) between the networks indicate a temporal progression of syllable activations across ROI networks ($T_{Planning} < T_{Articulation}$, $t_{(333)} = 3.3$, $p = 0.0011$; $T_{Articulation} < T_{Monitoring}$, $t_{(403)} = 1.3$, $p = 0.2$ (not significant), $T_{Planning} < T_{Monitoring}$, $t_{(352)} = 3.78$, $p = 1.7e\text{-}4$). These analyses that the neural coding of the syllabic frame occurs in a temporal progression from planning to execution in left-hemispheric speech ROIs. **g.** This progression supports a model for planning and execution where syllable frames are first enacted in the planning stage, before being executed via articulatory speech networks.

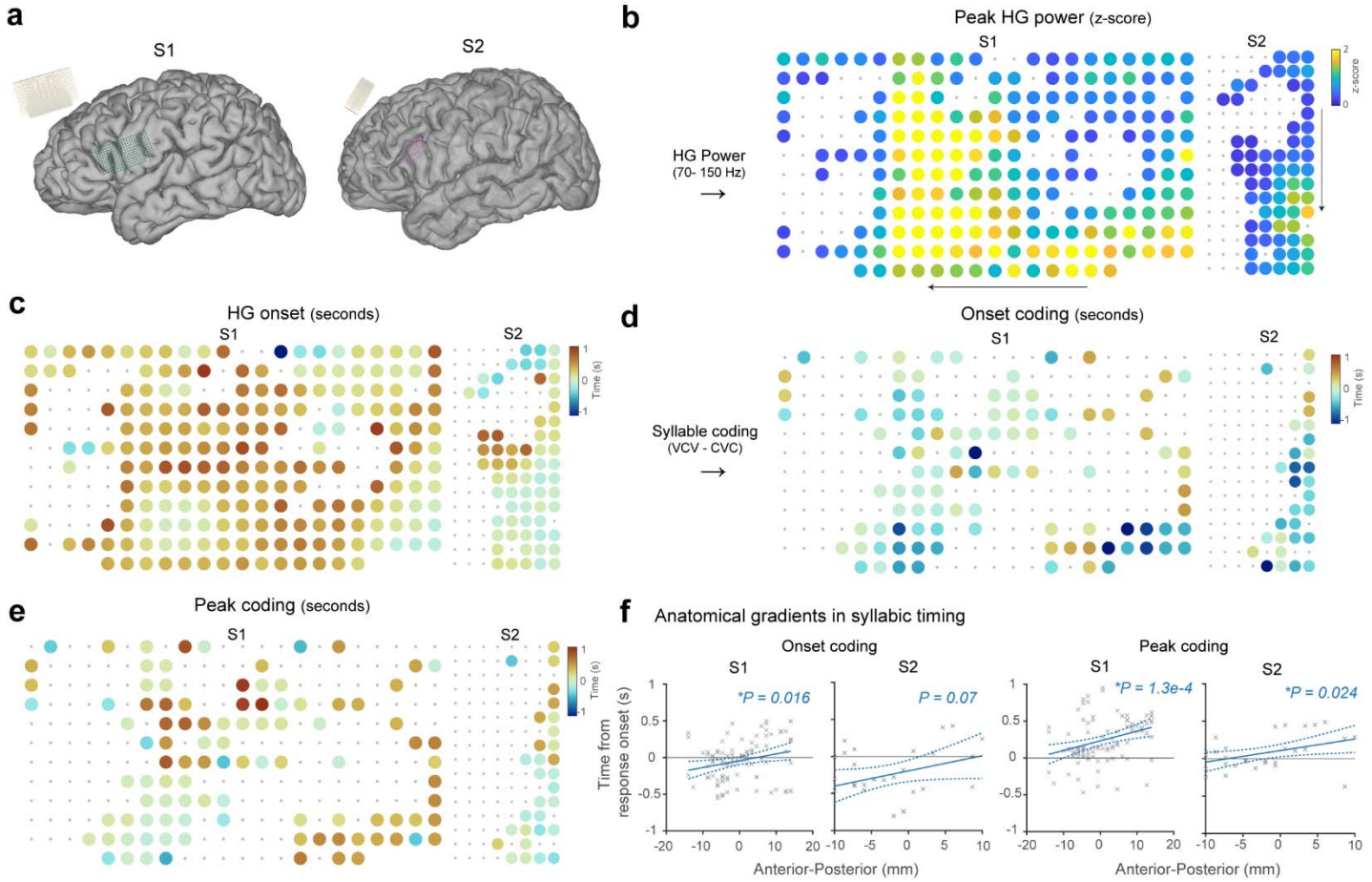
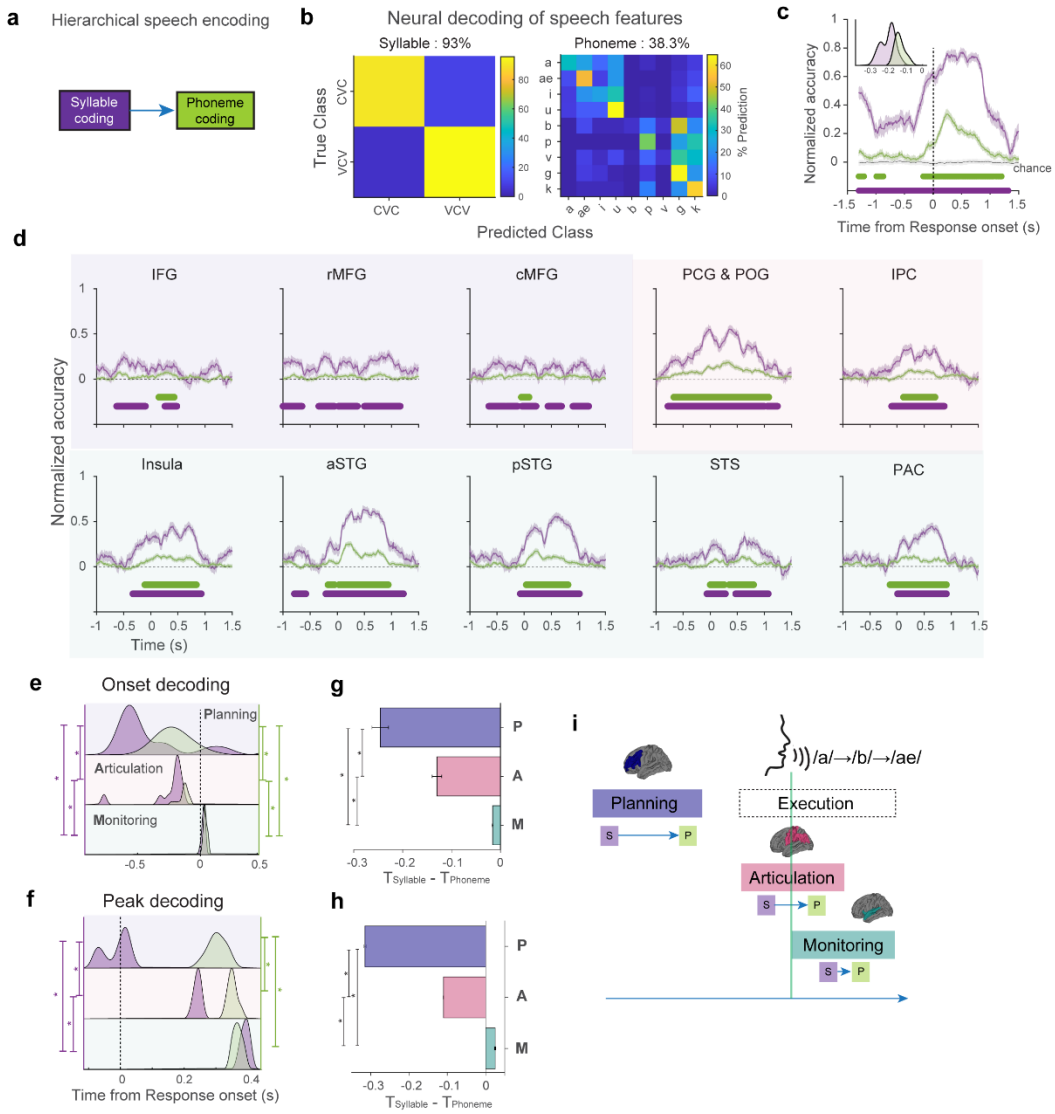


Figure 4. **µECoG captures spatio-temporally selective syllabic coding sites for speech planning to execution. a.** High density µECoG electrode arrays were implanted over inferior frontal gyrus (IFG) and pre-central gyrus (PCG) in two awake patients. Electrode arrays had either 256 electrodes with 1.72 mm inter-electrode distance for S1 (green) or 128 electrodes with 1.33 mm inter-electrode distance for S2 (pink). **b.** Peak HG activations show fine-scale spatio-temporal patterns across IFG and PCG during response onset. Large circles indicate electrodes with significant HG activations with heatmap colors corresponding to signal power (FDR-corrected permutation test, $p < 0.05$, S1: 201/256 significant/total electrodes, S2: 63/128). Non-significant electrodes are denoted by grey dots. Arrows indicate electrode array orientation from anterior to posterior. **c.** Spatial layouts of HG onsets show spatially varying temporal characteristics for both S1 and S2; 18.9% of significant electrodes reached peak signal power prior to response onset, indicating neural activations for speech planning at the micro-scale electrode level (S1: 12/201, S2: 38/63). **d, e.** Neural code for the syllable frame was extracted by contrasting HG activations per syllabic frame (ΔHG: VCV – CVC) for each electrode. Heat maps show spatial activation patterns of onset and peak times of ΔHG activations for electrodes with significant syllable contrasts. Large circles reflect electrodes with significant syllable contrasts (one-side permutation test with cluster correction, $p < 0.05$, S1: 100/201, S2: 37/63). Non-significant electrodes are indicated by grey dots. These significant contrasts indicate syllable frame coding on an electrode specific level and demonstrate spatially selectivity on the underlying neuroanatomy. **f.** A spatial correlation analysis of the syllable contrast onsets along the anterior-posterior axis shows a spatial progression of temporal coding for syllable codes starting in pars opercularis and moving towards the pre-central gyrus (S1: R^2 = 0.05, $F_{2,92}$ = 6.02, $p = 0.016$, S2: R^2 = 0.08, $F_{2,28}$ = 6.02, $p = 0.07$). This spatio-temporal progression was also seen in the peak times (F-statistic test against the constant model, S1: R^2 = 0.09, $F_{2,98}$ = 10.9, $p = 1.3e\text{-}4$, S2: R^2 = 0.13, $F_{2,28}$ = 5.68, $p = 0.024$). Dotted lines indicate the 95% confidence interval of the linear model fit. The presence of a spatial gradience, extracted using high density µECoG electrode arrays, demonstrate a potential neural mechanism for the transition from planning to execution for speech production.


a
Hierarchical speech encoding
Syllable coding
Phoneme coding
b
Neural decoding of speech features
Syllable : 93%
Phoneme : 38.3%
True Class
CVC
VCV
Predicted Class
% Prediction
c
Normalized accuracy
chance
Time from Response onset (s)
d
IFG
rMFG
cMFG
PCG & POG
IPC
Insula
aSTG
pSTG
STS
PAC
Normalized accuracy
Time (s)
e
Onset decoding
Planning
Articulation
Monitoring
f
Peak decoding
Time from Response onset (s)
g
P
A
M
T Syllable - T Phoneme
h
P
A
M
T Syllable - T Phoneme
i
/a/→/b/→/ae/
Planning
Execution
Articulation
Monitoring
S
P

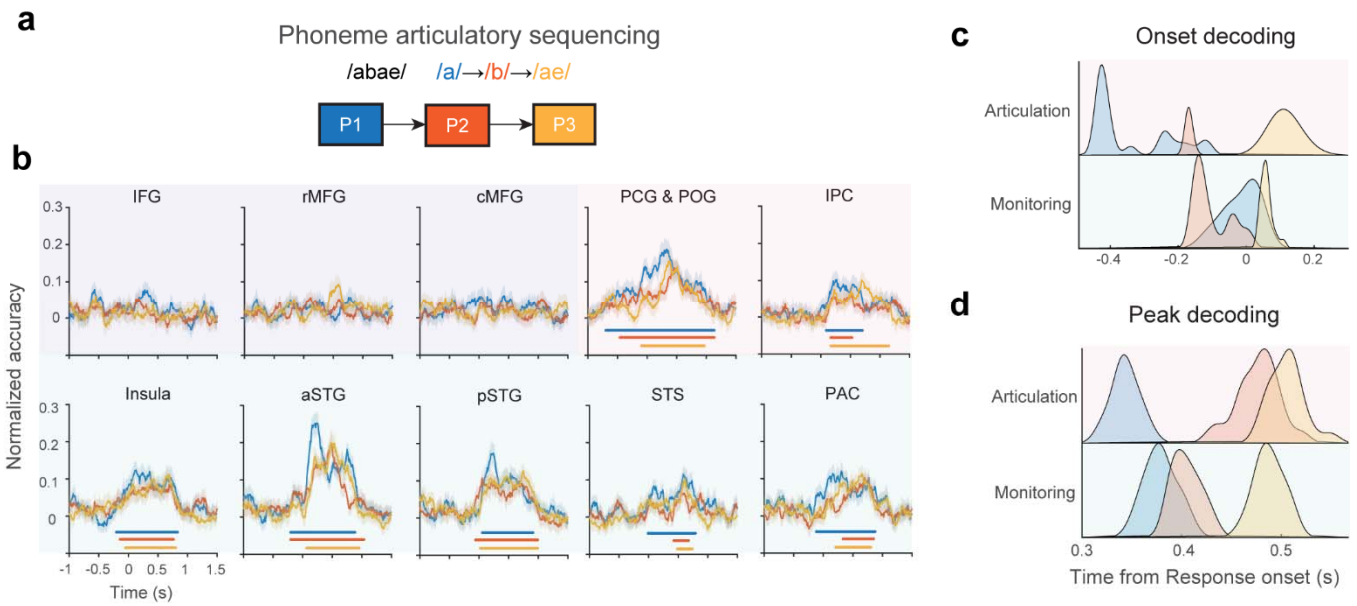


Figure 6. **Phonological units are sequentially articulated during execution. a.** Graphical representation of phonological un sequential execution that depicts phonological units articulated in their correct order during the utterance. **b.** Time-resolve decoding models were trained on neural segments aligned to the utterance to decode phonological units at each position withi the pseudo-word. Each colored trace indicates temporal evolution of positional phonological units in time for an ROI (mean acros 20 decoding instances). Results are shown for all ROIs in the left hemisphere, grouped by planning, articulation, and monitorin networks. Significant phonological unit decoding for all positions was observed only in execution ROIs (articulation an monitoring). Horizontal bars indicate durations of significant decoding (temporal permutation cluster test, $p < 0.05$; one-side permutation test against 250 shuffled decoders). Background shading indicates ROI networks for planning, articulation an monitoring. **c, d.** Colored histograms show distribution times (decoding onsets and peaks) for each positional phonological un within the ROI network. Positional phonological unit codes were decoded in the correct serial order (except onset decoding fo monitoring ROIs, pairwise one-sided Mann-Whitney U test, $p < 0.05$). The presence of ordering demonstrates the serial executio of positional phonological unit codes by the execution network.

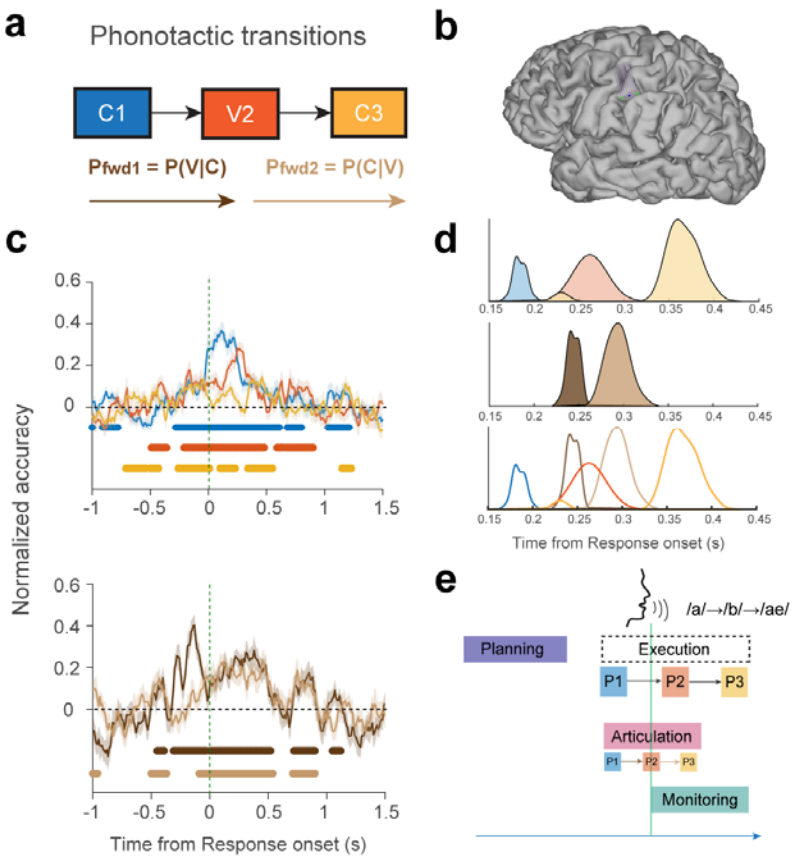
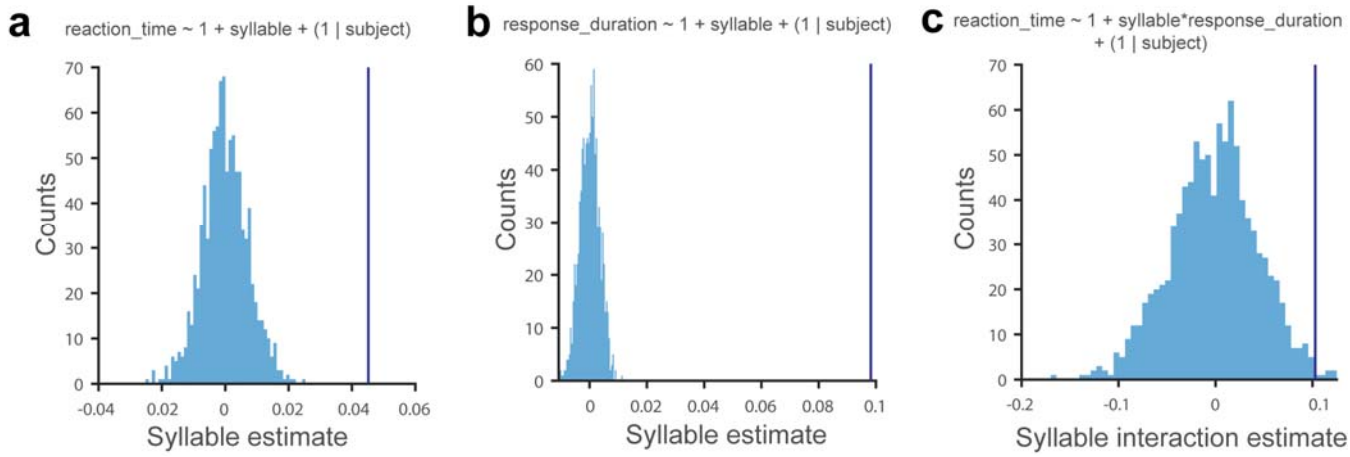
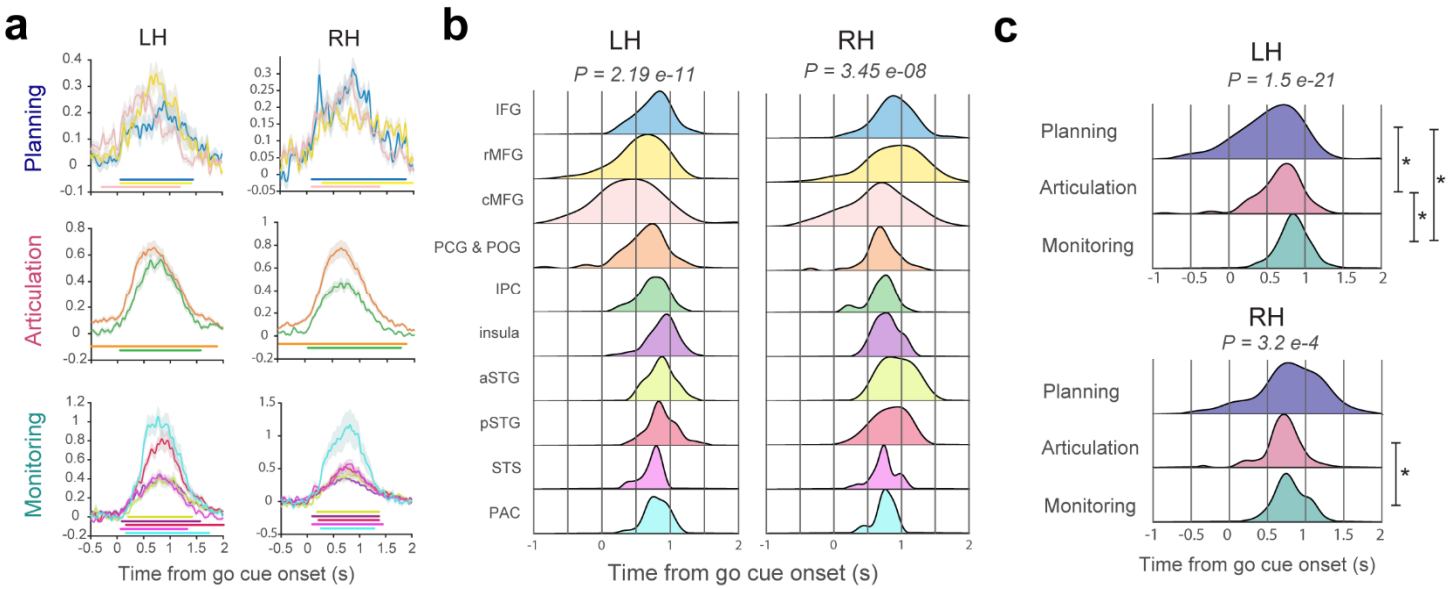


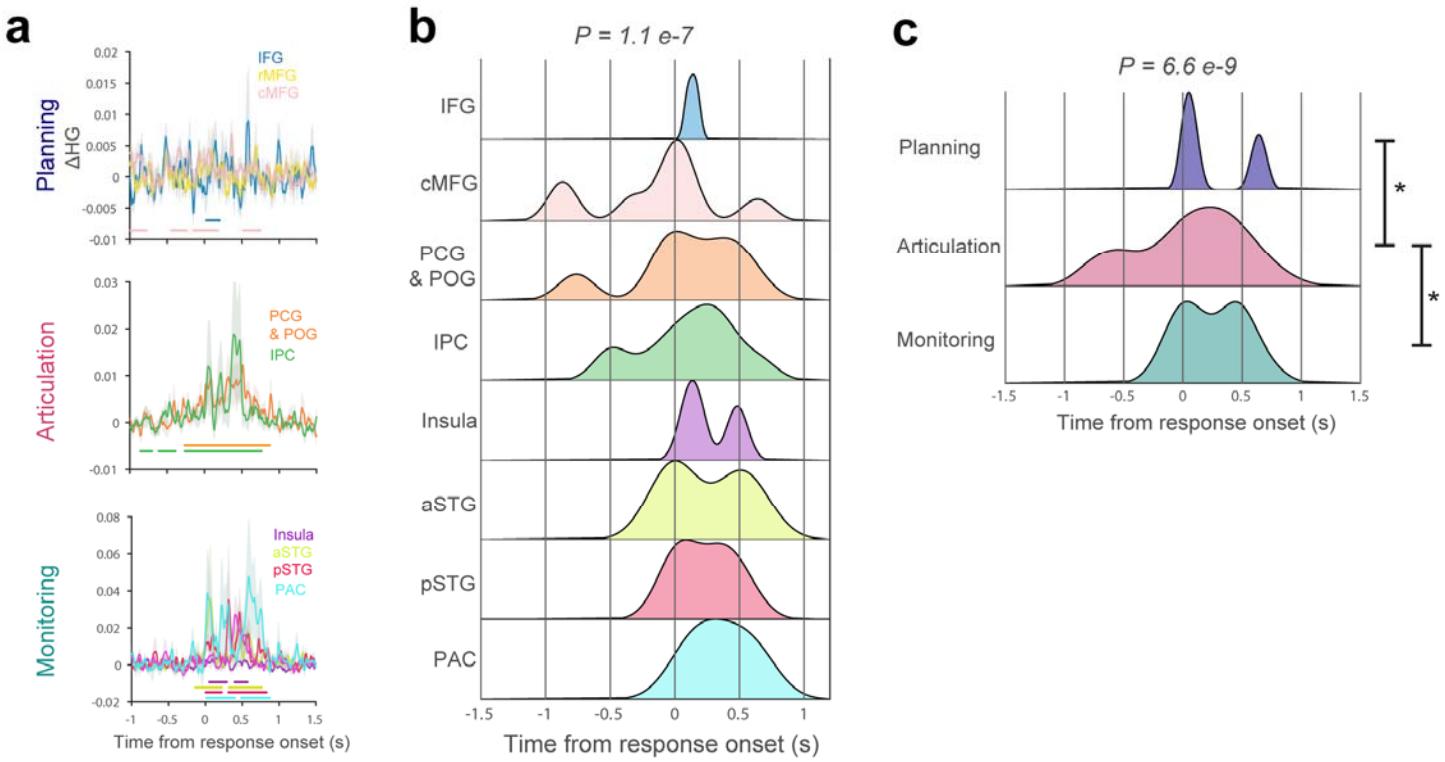
Figure 7. **Continuous phonotactic transitions are present in motor cortex during articulation. a.** Graphical schemati demonstrating that an accurate phonological unit sequence execution involve underlying transitionary statistics or phonotactics Forward phonotactic transition from C1 to V2 is denoted by P_{fwd1}, and from V2 to C3 is denoted by P_{fwd2}. **b.** µECoG electrode arra was implanted over the sensorimotor cortex in an awake patient. The electrode arrays had 128 electrodes with 1.33 mm inter electrode distance (purple). **c.** Time-resoled decoding revealed accurate phonological unit sequencing for CVC trials (C1: blue V2: orange, C3: yellow). Temporal decoding models trained to predict the phonotactic transitions (3-way: low, medium, and hig phonotactic transitions) demonstrated significant decoding for P_{fwd1} (dark brown) and P_{fwd2} (light brown). Horizontal bars indicat durations of significant decoding (temporal permutation cluster test, $p < 0.05$; one-sided permutation test against 250 shuffle decoders). **d.** Colored histograms indicate distribution times of peak decoding onsets for positional phonological units an phonotactic transitions, revealing significant temporal ordering of phonological units and phonotactic transitions (pairwise one sided Mann-Whitney U test, $p < 0.05$, $T_{C1} < T_{V2}$, $p = 1.6e-08$; $T_{V2} < T_{C3}$, $p = 4.2e-07$, $T_{Pfwd1} < T_{Pfwd2}$, $p = 6.6e-08$). Further, th forward transitions were significantly embedded between the phonological unit positions (pairwise one-sided Mann-Whitney U test, $p < 0.05$, $T_{C1} < T_{Pfwd1}$, $p = 3.7e-08$; $T_{Pfwd1} < T_{V2}$, $p = 6.6e-08$, $T_{V2} < T_{Pfwd2}$, $p = 6.2e-08$, $T_{Pfwd2} < T_{C3}$, $p = 7e-07$). **e.** Th significant embedding demonstrates the existence of both discrete and continuous motor coding during the execution c phonological unit sequences for speech production.



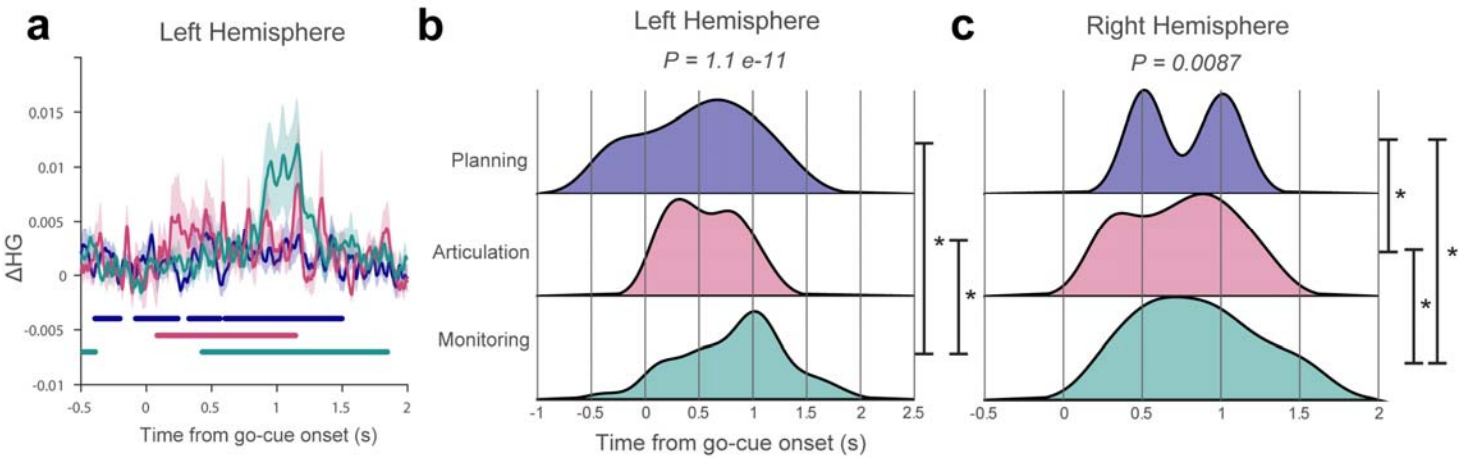
Supplementary Figure 1. **Behavior analysis to quantify syllable frame effects on speech planning. a & b.** Linear mixed effects (LME) modeling to quantify the effects of syllable frame on reaction time and duration. Syllable effect (CVC or VCV trials dark blue) is significantly predicting both the reaction time (LME: $t_{(8698)} = 3.7$, $\beta = 0.045$, $p = 2.2$ e-04, 95% CI = 0.02 to 0.07) and duration (LME: $t_{(8696)} = 10.7$, $\beta = 0.098$, $p = 2.1$e-26, 95% CI = 0.08 to 0.1) of the utterance versus shuffle distribution (n = 1000 light blue histogram). **c.** LME modeling quantifies the effect of syllable frames on predicting reaction time with response duration as an additional predictor. The interaction term between the syllable frame and the response duration is significant in predicting the reaction time (LME: $t_{(9084)} = 2.75$, $\beta = 0.103$, $p = 0.006$, 95% CI = 0.03 to 0.18), indicating that the response duration depends on the identity of syllable frame, highlighting the significant effect of the syllable frame by association. The significant behavior coding provides evidence that speech motor planning reflects the composition of syllable frames.



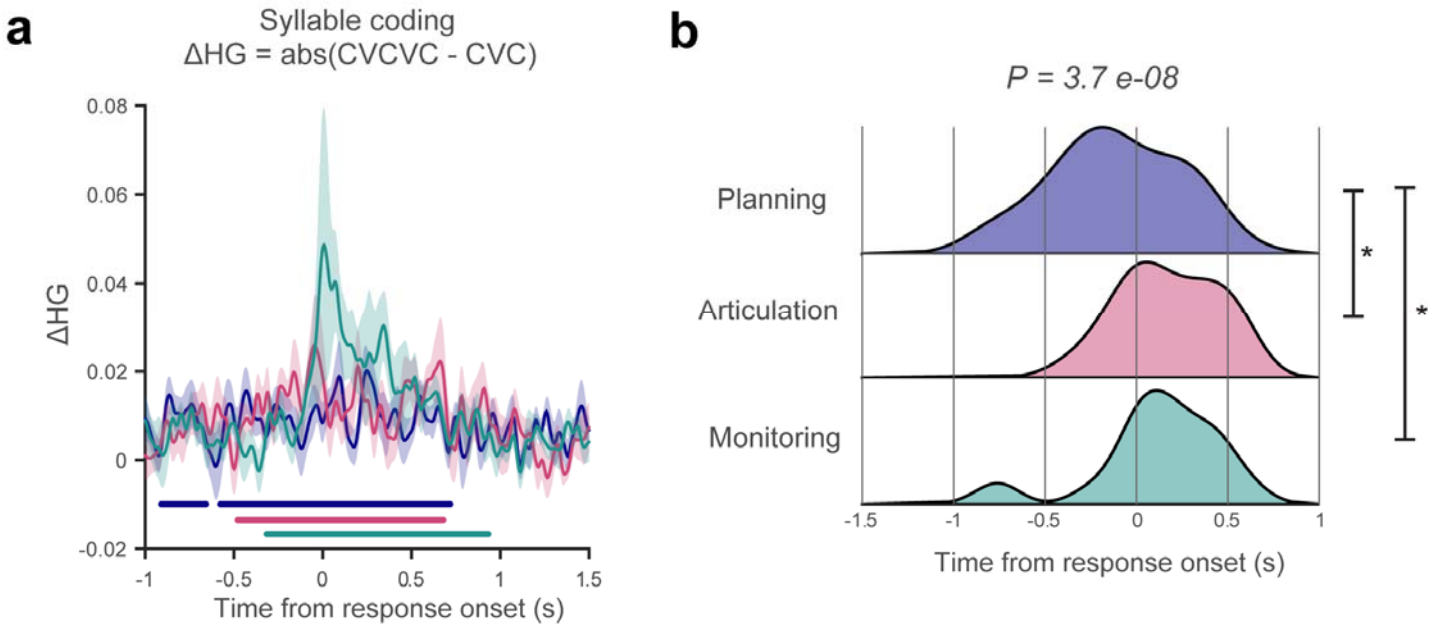
Supplementary Figure 2. **Cortical activations from intracranial recordings during speech production aligned to the G Cue. A.** Mean HG activation of ROIs aligned to go cue onset for both left and right hemispheres, averaged within patient. Shade area denotes standard error. Colored horizontal bars indicate time-periods of significant activation above baseline (one-side permutation test with cluster correction, *p<0.05*). **b.** Electrode distributions of peak HG time-points per ROI for each hemispher with respect to go cue onset. Significant differences in peak activations were observed in both left and right hemisphere suggesting a temporal progression of neural activations across ROIs around the speech utterance (one-way ANOVAs: le $F_{9,717}$=8.03, *p = 2.2e-11*, right: $F_{9,720}$=6.03, *p = 3.6e-08*). **c.** Electrode distributions of peak HG time-points per speec network for each hemisphere. This temporal progression was also observed via significant differences in peak activations in bo left and right hemispheres (one-way ANOVAs: left: $F_{2,644}$=52, *p=1.5e-21*, right: $F_{2,727}$=8.13, *p=3.2e-4*). Post-hoc tests indicate pairwise significance between the speech networks in left hemisphere ($T_{Planning} < T_{Articulation}$, $t_{(376)}$ = 3.13, *p* = 0.001 $T_{Articulation} < T_{Monitoring}$, $t_{(504)}$ = 7.6, *p* = 1.5e-13, $T_{Planning} < T_{Monitoring}$, $t_{(410)}$ = 9.7, *p* = 3.9e-20). Similar significant differences we observed in right hemisphere only for differences between articulation and monitoring networks ($T_{Planning} < T_{Articulation}$, $t_{(529)}$ = -1.1, = 0.27 – not significant (NS); $T_{Articulation} < T_{Monitoring}$, $t_{(526)}$ = 3.3, *p* = 0.0011, $T_{Planning} < T_{Monitoring}$, $t_{(395)}$ = 1.1, *p* = 0.29 - NS). Note th the negative t-statistic with planning network denote late neural activations, indicating the monitoring role of right-hemispheric pr frontal regions. These analyses demonstrate the left-hemispheric temporal progression of neural responses for plannin articulation, and monitoring networks for speech production. Note that some planning activation is seen prior to the go cu



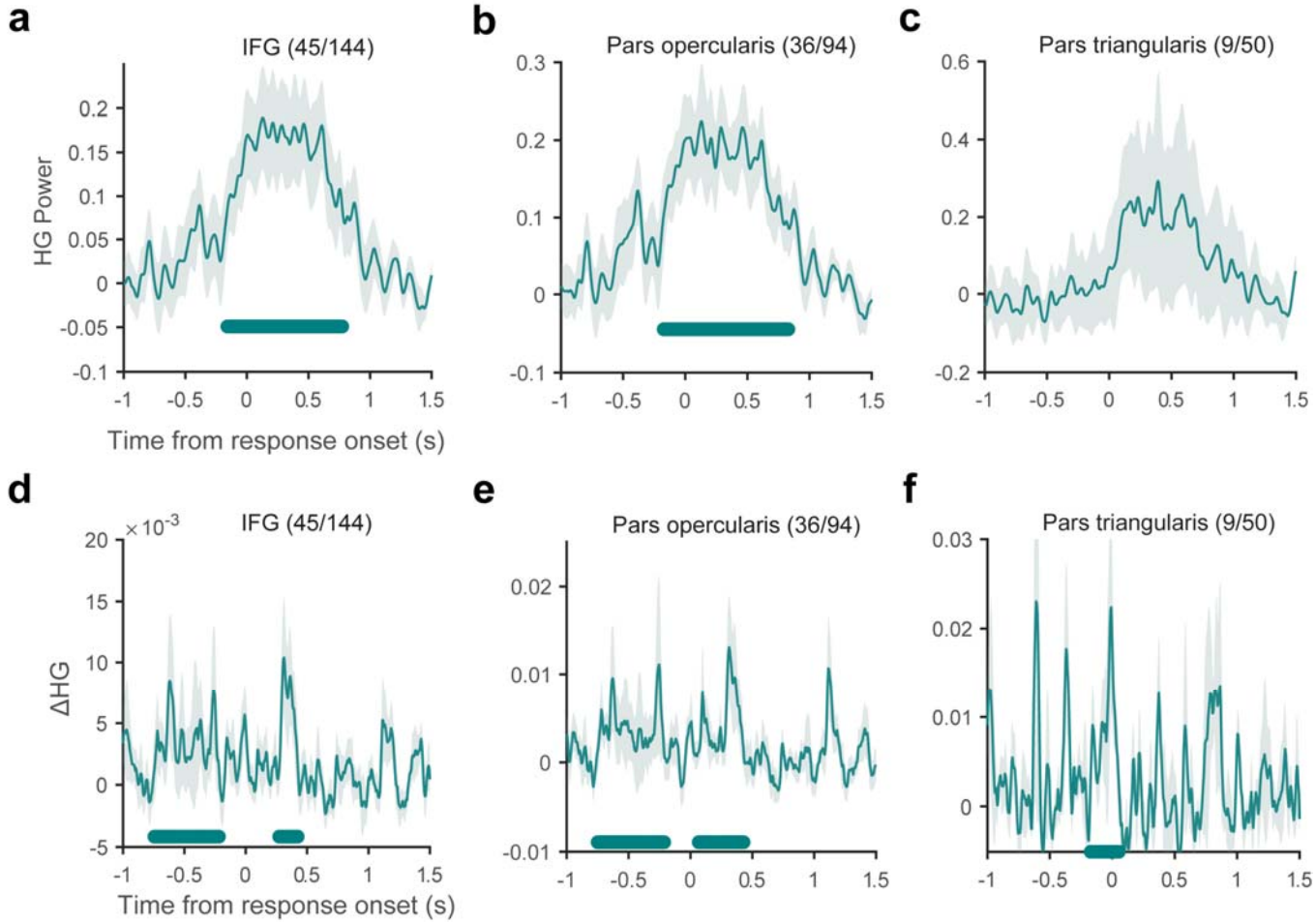
Supplementary Figure 3. **Neural code for syllable frames in the right hemisphere. a.** Contrasting syllabic frames per RO show time points with significant syllable discrimination around the utterance. Plots show mean ΔHG contrasts of VCV versus CVC frame for each ROI in the right hemisphere aligned to the response onset. Shaded area denotes standard error. Horizonta bars indicate durations of significant contrasts (across vs. within frame, one-sided permutation test with cluster correction *p<0.05*). **b.** Histograms show electrode distributions of peak ΔHG contrasts per ROI. Significant differences in peak ΔHG indicate temporal distinction in syllable activations corresponding to syllable frames, around the utterance (one-way ANOVA, $F_{7,442}$ = 6.8 *p = 1.1e-7*). **c.** Histograms show electrode distributions of peak ΔHG contrasts per network in the right hemisphere: planning articulation, and monitoring. Significant differences in peak ΔHG indicate the temporal distinction of syllable activations across speech networks (one-way ANOVA, $F_{2,537}$ = 19.5, *p = 6.5e-9*). Post-hoc t-tests indicate pairwise significance in syllabic frame contrasts (ΔHG) between the networks ($T_{Planning} < T_{Articulation}$, $t_{(406)} = -5.4$, $p = 1.1e-7$; $T_{Articulation} < T_{Monitoring}$, $t_{(415)} = 4.6$, $p = 5.1e-6$) with no significant difference between the planning and monitoring networks, $T_{Planning} < T_{Monitoring}$, $t_{(299)} = 0.16$, $p = 6.1e-4$). Note that the negative t-statistic with planning network denote late processing of syllable frames, indicating the monitoring role of right-hemispheric pre-frontal regions. These analyses demonstrate the temporal progression of the neural coding of the syllabic frame within the execution network (articulation to monitoring) in right-hemispheric speech ROIs. The hemispheric distinction supports the role of right-hemisphere in only executing the planned syllable frames for an utterance.



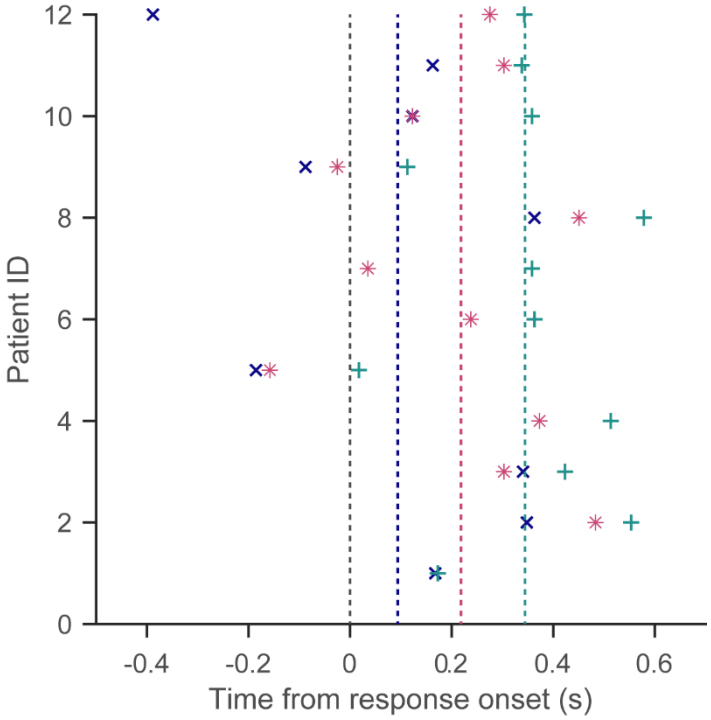
Supplementary Figure 4. **Neural code for syllable frames aligned to go-cue onset. a.** Contrasting the syllabic frames for ea speech network shows time points with significant syllable discrimination aligned to go-cue onset from the left-hemisphere. Plo show mean ΔHG contrasts of VCV versus CVC frame for each speech network. Shaded area denotes standard error. Horizont bars indicate durations of significant contrasts (across vs. within frame, one-sided permutation test with cluster correctio *p<0.05*). **b.** Histograms show left-hemispheric electrode distributions of peak ΔHG contrasts for speech network. Significa differences in peak ΔHG show a temporal distinction in syllable activations corresponding to syllable frames, around the go-cu onset (one-way ANOVA, $F_{2,521}$=26, *p = 1.1e-11*), with some significant coding during the delay, prior to go-cue onset potential indicator of the involvement of verbal working memory). Post-hoc t-tests show pairwise significance in syllabic fran contrasts (ΔHG) between the networks ($T_{Planning} < T_{Articulation}$, $t_{(314)} = 1.2$, $p = 0.23$ – not significant; $T_{Articulation} < T_{Monitoring}$, $t_{(371)} = 6.3$, = 8.1e-10, $T_{Planning} < T_{Monitoring}$, $t_{(347)} = 6.1$, $p = 2.7e-9$). **c.** Histograms indicate right-hemispheric electrode distributions of peak ΔH contrasts per speech. Significant differences in peak ΔHG indicate temporal distinction in syllable activations corresponding syllable frames, around the utterance (one-way ANOVA, $F_{2,495}$=4.8, *p = 0.0087*). Post-hoc t-tests show pairwise significance syllabic frame contrasts (ΔHG) between the articulation and monitoring networks ($T_{Planning} < T_{Articulation}$, $t_{(339)} = -8.1$, $p = 6.6e-1$ $T_{Articulation} < T_{Monitoring}$, $t_{(378)} = 6.1$, $p = 3.1e-9$, $T_{Planning} < T_{Monitoring}$, $t_{(255)} = -3$, $p = 0.003$), supporting the execution role of rig hemisphere. Note that the negative t-statistic with planning network denote late processing of syllable frames, indicating th monitoring role of right-hemispheric pre-frontal regions. These analyses demonstrate the temporal progression of the syllab frame across left-hemispheric speech network is preserved with respect to the go-cue, and the left-hemispheric planning syllable frames start prior to instructional go-cue. These results also support the right hemisphere's involvement in execution po go-cue.



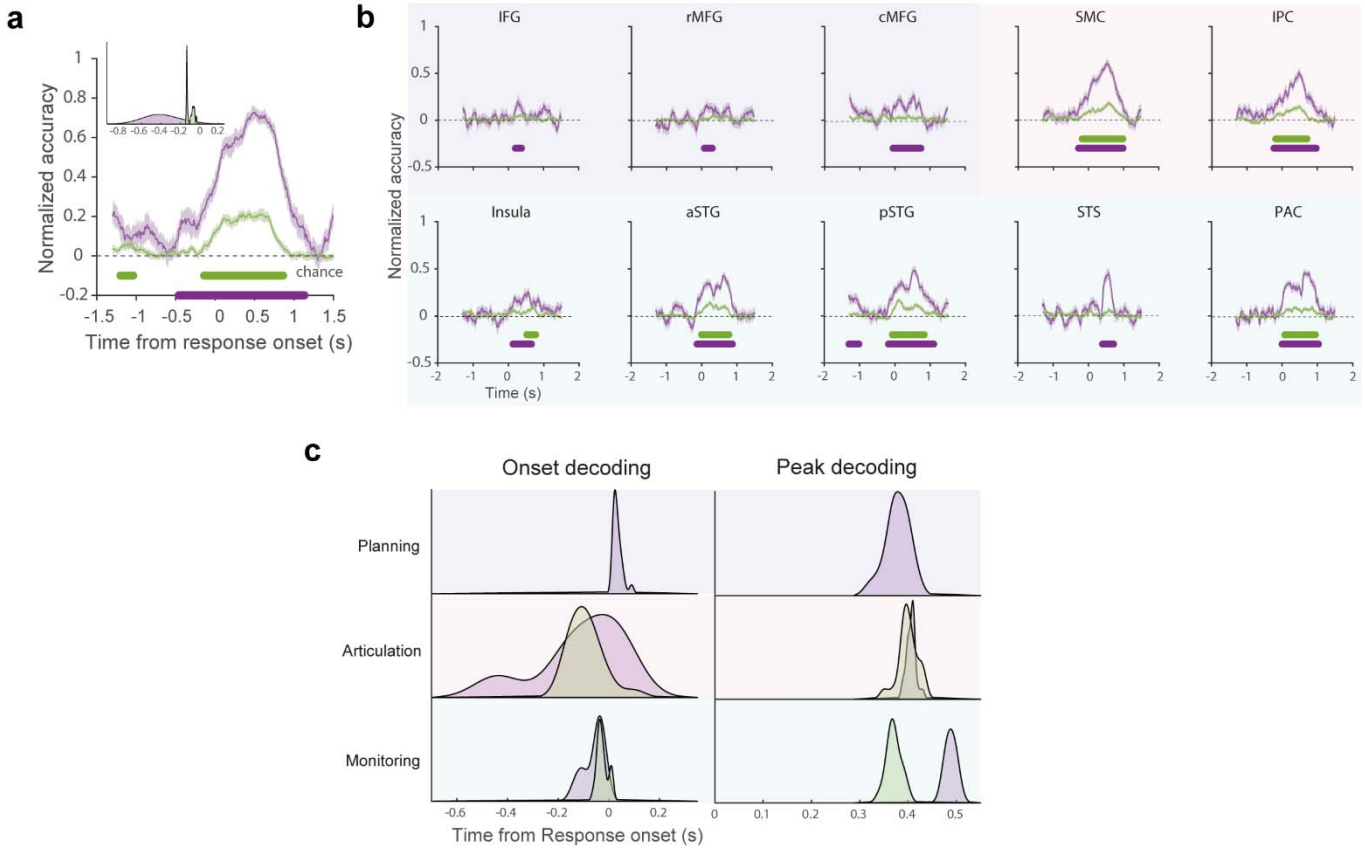
Supplementary Figure 5. **Neural code for syllable frames is specific to number of syllables in the frame. a.** Contrastin syllabic frames (CVCVC vs. CVC) per speech network show time points with significant syllable discrimination in the lef hemisphere aligned to the utterance. In a subset of patients, a control analysis was performed to rule out the effect of difference in initial phoneme (C vs. V). Plots show mean ΔHG contrasts of CVCVC versus CVC frame for each speech network aligned t the response onset. Shaded area denotes standard error. Horizontal bars indicate durations of significant contrasts (across v within frame, one-sided permutation test with cluster correction, *p<0.05*). **b.** Histograms show electrode distributions of peak ΔH contrasts per speech network. Significant differences in peak ΔHG support a temporal progression in syllable activatior corresponding to syllable frames, around the utterance (one-way ANOVA, $F_{2,254}$ = 18.3, *p = 3.7e-8*). Post-hoc t-tests sho pairwise significance in syllabic frame contrasts (ΔHG) between the networks ($T_{Planning}$ < $T_{Articulation,}$ $t_{(155)}$ = 3.97, *p* = 1e-4; $T_{Planning}$ $T_{Monitoring,}$ $t_{(182)}$ = 5.6, *p* = 6.3e-8), except for the differences between articulation and monitoring network ($t_{(169)}$ = 0.88, *p* = 0.3793 These analyses indicate the temporal progression of the syllabic frame across left-hemispheric speech network is specific t number of syllables in the syllable frame, rather than being driven by phoneme specific information (C vs. V) within the syllab



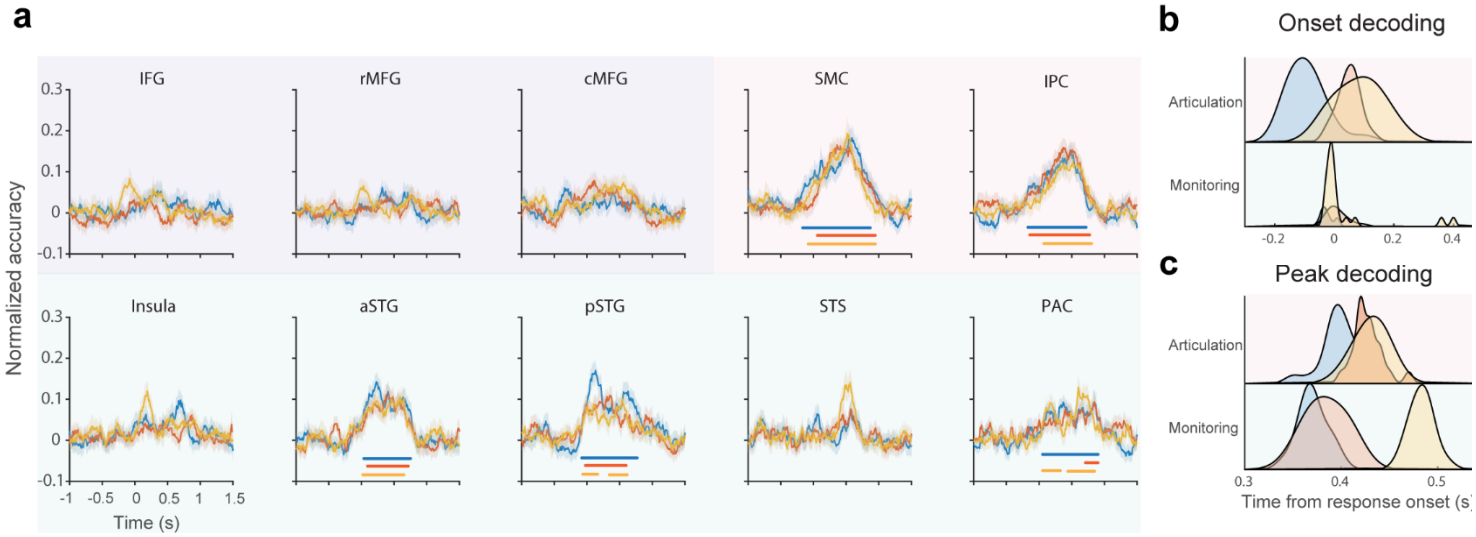
Supplementary Figure 6. **Neural activations in left-hemispheric IFG is specific to pars opercularis. a – c.** Mean HG activations around response onset for left hemisphere inferior frontal gyrus (IFG), pars opercularis, and pars triangularis. Shaded area denotes standard error. Colored horizontal bars indicate time-periods of significant activation above baseline (one-sided permutation test with cluster correction, *p<0.05*). Numbers indicate the total number of electrodes with significant HG activations compared to the total number of electrodes within the ROI across patients. **d – f.** Contrasting syllabic frames (VCV vs. CVC) per speech network shows time points with significant syllable discrimination around the utterance. Plots show mean ΔHG contrasts of VCV versus CVC frame for each speech network aligned to the response onset. Shaded area denotes standard error. Horizontal bars indicate durations of significant contrasts (across vs. within frame, one-sided permutation test with cluster correction, *p<0.05*). These results show that neural activations and neural code for syllable frames (ΔHG) from IFG are mainly confined to pars opercularis.



Supplementary Figure 7. **The temporal progression of syllable neural code is observed on a patient-specific level.** Scatte plot indicates distribution of peak ΔHG time points of planning (x), articulation (*), and monitoring (+) networks for each patier during utterance. Each scatter point denotes the peak ΔHG contrasts of VCV versus CVC frame for each speech network averaged across electrodes within the patient. The dotted vertical lines indicate average onsets for planning, articulation, an monitoring networks across patients, demonstrating the correct order of temporal progression. The analysis indicates th progression of neural activations from planning to execution of speech is present on a patient specific scale.



Supplementary Figure 8. **Right-hemispheric ROIs show significant decoding of syllables and phonemes but do not exhibit a clear hierarchical organization. a.** The temporal evolution of syllables and phoneme activations were obtained using time-resolved neural decoding (200 ms window; 10 ms step size). Solid lines indicate average across 20 instances of decoding iterations, and shaded bars indicate standard errors across instances. Grey traces indicate mean and standard errors of chance decoding which was obtained from randomly shuffling syllable and phoneme labels (250 shuffled decoding models). Horizontal bars indicate durations of significant decoding (temporal permutation cluster test, $p < 0.05$; one-sided permutation test against 250 shuffled decoding models). Colored histograms indicate distributions of decoding onsets of syllables (violet) and phonemes (green), with significant differences between the onsets (one-sided Mann-Whitney U test, $p = 1.9e\text{-}8$). **b.** Temporal decoding models also demonstrate a similar syllable and phoneme decoding for each ROI (except phoneme decodes in planning regions and STS). Dotted lines indicate chance decoding. Horizontal bars indicate durations of significant decoding. Background shades indicate ROI networks for planning, articulation and monitoring. **c.** Colored histograms indicate distributions of onset and peak decoding times for both syllable and phoneme for each ROI network. No significant differences in decoding distributions between syllable and phoneme in speech networks. This analysis demonstrates that although the syllables are represented earlier than phonemes on a hemispheric level, the hierarchical organization is absent on a network level.



Supplementary Figure 9. **Phonological sequence execution in the right hemisphere. a.** Time-resolved decoding model were trained on neural segments aligned to the utterance to decode phonological units at each position within the pseudo-word Each colored trace indicates temporal evolution of positional phonological units in time for an ROI (mean across 20 decodin instances). Results are shown for all ROIs in the right hemisphere, grouped by planning, articulation, and monitoring networks Significant phonological unit decoding for all positions was observed only in execution ROIs (articulation and monitoring – excep insula and STS). Horizontal bars indicate durations of significant decoding (temporal permutation cluster test, $p < 0.05$; one-side permutation test against 250 shuffled decoders). Background shading indicates ROI networks for planning, articulation an monitoring. **b, c.** Colored histograms show distribution times (decoding onsets and peaks) for each positional phonological un within the ROI network. Positional phonological unit codes were decoded in the correct serial order (except onset decoding fo monitoring ROIs, pairwise one-sided Mann-Whitney U test, $p < 0.05$). The presence of ordering demonstrates the serial executio of positional phonological unit codes by the right-hemispheric execution network.

Supplementary Table 1:

Clinical summary of patients: Pre-operative epileptic monitoring

Patient	Diagnosis	Electrode type: Total No. of recording electrodes	Number of electrodes with significant HG activation
D20	Epilepsy	PMT SEEG: 120	7
D22	Epilepsy	PMT SEEG: 100	0
D23	Epilepsy	PMT SEEG: 121	77
D24	Epilepsy	Ad-Tech 48 channel ECoG: 52	12
D25	Epilepsy	PMT SEEG: 168	6
D28	Epilepsy	PMT SEEG: 108	44
D29	Epilepsy	PMT SEEG: 140	60
D31	Epilepsy	PMT SEEG: 160	51
D35	Epilepsy	PMT SEEG: 174	52
D39	Epilepsy	PMT SEEG: 240	107
D40	Epilepsy	PMT SEEG: 188	4
D41	Epilepsy	PMT SEEG: 274	118
D42	Epilepsy	PMT SEEG: 176	27
D45	Epilepsy	PMT SEEG: 202	54
D49	Epilepsy	PMT SEEG: 210	113
D53	Epilepsy	PMT SEEG: 158	11
D54	Epilepsy	PMT SEEG: 200	0
D55	Epilepsy	PMT SEEG: 190	64
D56	Epilepsy	PMT SEEG: 128	54
D57	Epilepsy	PMT SEEG: 178	85
D58	Epilepsy	PMT SEEG: 244	67
D59	Epilepsy	PMT SEEG: 184	63
D60	Epilepsy	PMT SEEG: 252	111
D61	Epilepsy	PMT SEEG: 232	9
D63	Epilepsy	PMT SEEG: 274	77
D64	Epilepsy	PMT SEEG: 246	36
D65	Epilepsy	PMT SEEG: 220	101
D66	Epilepsy	PMT SEEG: 182	44
D67	Epilepsy	Ad-Tech 48 channel ECoG: 72	22
D68	Epilepsy	PMT SEEG: 146	63
D69	Epilepsy	Ad-Tech SEEG: 132	10
D70	Epilepsy	Ad-Tech SEEG: 202	0
D71	Epilepsy	PMT SEEG: 162	49
D72	Epilepsy	PMT SEEG: 214	173
D73	Epilepsy	Dixi SEEG: 200	59
D75	Epilepsy	Dixi SEEG: 215	126
D76	Epilepsy	Dixi SEEG: 213	145
D77	Epilepsy	Dixi SEEG: 157	91
D79	Epilepsy	Dixi SEEG: 256	69
D81	Epilepsy	Dixi SEEG: 222	111
D82	Epilepsy	Dixi SEEG: 263	152
D84	Epilepsy	PMT SEEG: 230	201
D85	Epilepsy	Dixi SEEG: 247	126
D86	Epilepsy	PMT SEEG: 284	49
D88	Epilepsy	PMT SEEG: 184	127
D91	Epilepsy	Dixi SEEG: 332	123
D92	Epilepsy	Dixi SEEG: 218	83
D93	Epilepsy	Dixi SEEG: 210	125
D95	Epilepsy	Dixi SEEG: 294	88
D96	Epilepsy	Dixi SEEG: 234	86
D102	Epilepsy	Dixi SEEG: 227	0
D103	Epilepsy	Dixi SEEG: 247	2

Clinical summary of patients: Intra-operative awake surgery

Patient	Diagnosis	Electrode type/ No. of recording electrodes	Electrode material	Number of electrodes with significant HG activation
S1	Tumor resection	µECoG: 256	Platinum-Iridium	201
S2	Parkinson's	µECoG: 128	Platinum-Iridium	63
S3	Parkinson's	µECoG: 128	Gold	111

Supplementary Table 2: Stimulus labels for speech repetition task

CVC	VCV
/bab/	/abae/
/baek/	/abi/
/bak/	/aeba/
/bup/	/aebi/
/gab/	/aebu/
/gaeb/	/aega/
/gaev/	/aeka/
/gak/	/aepi/
/gav/	/aka/
/gig/	/aku/
/gip/	/ava/
/gub/	/avae/
/kab/	/ibu/
/kaeg/	/ika/
/kub/	/ikae/
/kug/	/ipu/
/paek/	/iva/
/paep/	/ivu/
/paev/	/uba/
/puk/	/uga/
/pup/	/ugae/
/vaek/	/ukae/
/vaeg/	/upi/
/vip/	/upu/
/vug/	/uvae/
/vuk/	/uvi/