EEG provides evidence for internal inverse and forward models and organization of speech information during speech perception and production

Abstract

Human speech is a crucial ability central to everyday life. Dual stream models of language processing provide a functional neuroanatomical model for auditory speech processing, however, the neural dynamics underlying speech processing as well as the sensorimotor transformations in these pathways predicted by these models are still poorly understood. Internal models predict a link between sensory and articulatory representations; a forward internal model links motor areas to perceptual areas during speech production (to predict the sensory consequences of speech production), and an inverse internal model from sensory to articulatory representations during speech perception links speech sounds to motor plans. In this study, we used electroencephalography (EEG) with word pairs designed to probe different aspects of speech representations to systematically investigate the nature of neural representations and their predicted dynamic interactions during speech perception and production. We first validated our approach by replicating prior electrocorticography (ECoG) findings regarding the latency and organization of speech information during perception. We then found that the sensorimotor transformation between different neural representations during speech perception and production corresponded to the hypothesized internal model framework. Specifically, we found evidence for an inverse model from perceptual representations in temporal channels to articulatory representations in superior frontal channels during speech perception, and for a predictive forward model from frontal to temporal channels active during speech production. Intriguingly, this forward model appeared to include the anteroventral stream. Additionally, we found evidence that the organization of speech representations in different parts of the brain's speech network is task-dependent and dynamic.

Introduction

Spoken language, which includes speech perception and speech production, is a uniquely human skill central to many aspects of life. Speech production is a highly complex motor skill, engaging the coordinated use of approximately 100 muscles. The production of the intended speech sounds must be precise and reproducible to assure reliable decoding during speech perception. Most amazingly, children usually learn how to speak by simply listening to speech, not by receiving instruction on how to move their articulators[1]. Auditory feedback thus plays a pivotal role in learning to speak, as demonstrated by the difficulty deaf children experience in acquiring this ability.

Learning to correctly produce a speech sound requires an error signal — in this case, information on how much the produced articulation was off from the intended articulation. However, what makes speech production learning so complex *is that the only available error typically is in acoustic (sound) space* – i.e., there is a mismatch between what one *heard* and what one wanted to *produce*. Thus, to improve the articulation, one would need to know in what way the movement of the articulators was incorrect, i.e., an error in *motor* space. Leveraging techniques from robot learning and artificial intelligence, computational neuroscience has provided a solution for this kind of learning problem: forward and inverse modelling[1-4]. Adapted to speech production learning, a *forward* internal model maps a specific articulation to a predicted sound, and an *inverse* model maps a sound or word to the articulator movements needed to produce it. In this internal model framework, the forward model is first trained through external feedback, e.g., through "motor babbling"[5,6] (which involves activating articulators more or less randomly and learning to predict which sounds result). This model can then be used to train an inverse model. A motor articulation produces a sound, which is perceived and compared to the intended sound, producing an error in the auditory domain. This error is then used by the internal models to translate into a motor error, which can be used as an input to train the inverse model[3,4,7,8]. *While this computational theory can readily explain how the brain learns to produce speech, the implementation of the required internal models and the underlying neural signal dynamics within the brain's speech system are less clear.*

Classically, speech has been posited to be localized in two major hubs, in the frontal and temporal regions of the brain[9,10]; however, over the past century, our understanding of the speech system has morphed into a more distributed, dual-stream[11,12] architecture which describes each process in the context of functional networks of brain regions organized along a posterior/dorsal pathway (moving posteriorly from primary auditory cortex along the superior temporal gyrus, STG, via parietal cortex to premotor areas and ultimately the inferior frontal gyrus, IFG) incorporating sensorimotor integration and control[13] and an anteroventral pathway roughly extending anteriorly in the STG from primary auditory cortex to the IFG for speech perception[11,14].

Probing the neural bases of internal models and the transformation between perceptual and articulatory representations requires the ability to measure neural activity with high temporal resolution, given the fast temporal dynamics of speech. Much insight into the organization of speech representations in different brain areas has been obtained from human electrocorticography (ECoG) studies[15-18], with their exquisite spatial and temporal resolution. Specifically, for articulatory representations in sensorimotor cortex, several studies have provided evidence for a somatotopic organization[17,18] (also supported by functional magnetic resonance imaging studies[19]). That is, *changes in place of articulation (e.g., dental vs. labiodental, as in /θ/ vs. /f/) caused the strongest changes in the* *spatial distribution of neural responses in articulatory regions during speech production*. In pronounced contrast, *representations in perceptual areas in the STG were found to be organized by manner of articulation* (e.g., fricative vs. stop, as in /s/ vs. /t/)[15-17].

Yet, a fundamental limitation of ECoG studies is that they usually only cover a subset of brain areas relevant for speech processing, as electrode placement is dictated by clinical considerations. This makes it difficult to study internal models, which by design connect perceptual and articulatory representations located in distant parts of the brain. In contrast, electroencephalography (EEG) has the advantage of whole-head coverage, yet with much lower spatial resolution. This low spatial resolution in turn has traditionally made it difficult to probe speech representations using EEG, as the mere presence of activation does not provide information about the speech specificity of the response, let alone the nature of the underlying representation, e.g., whether it is organized by place of articulation as in sensorimotor cortex, or manner of articulation, as in perceptual areas, nor allow to probe whether the nature of the encoding differs by task (perception vs. production), as suggested by recent studies[17,20].

Testing the hypothesized internal model architecture and the predictions of dynamic interactions between perceptual and motor representations during speech perception and speech production therefore requires an approach with a combination of whole-brain coverage, fine temporal resolution and the ability to probe the specificity and organization of speech representations giving rise to the neural signals, such as sensitivity to articulatory (i.e., place of articulation) vs. perceptual (i.e., manner of articulation) features. In the present study, we address these challenges by employing whole-head electroencephalography (EEG) with a novel stimulus set that enabled us to probe neural dynamics during both speech perception as well as speech production. Using a combination of searchlight-based classification and functional connectivity (Granger causality) analyses, we first validated our technique by showing agreement with prior ECoG studies on auditory speech representations in STC organized by manner of articulation, and then provide evidence for the existence of both an inverse internal model during speech perception as well as a forward internal model during speech production.

Materials and Methods

Participants

A total of 23 right-handed healthy adults (mean age 24.4 years, 13 females) with no hearing loss, no speech loss, and no neurological, movement, or psychiatric disorders were enrolled in the experiment. All participants were native English speakers. The university's Institutional Review Board approved all experimental procedures, and written informed consent was obtained from all subjects before participating in the experiment. One subject was removed from the sample due to failure to complete the experiment and two additional subjects were removed for excessive movement during the EEG recordings, resulting in the final study sample (N=20).

Behavioral stimuli

Ten pairs of pseudowords were designed to differ by manner of articulation (n=5 pairs) or place of articulation (n=5 pairs) – see Table 1. Each stimulus was pre-recorded by a trained speech-language pathologist (Audio-Technica AT4033a microphone, sampling rate 96KHz) and controlled for pitch, frequency, and duration (duration 800 ms +- 50 ms). The effect of these features of articulation (manner vs. place) has been extensively investigated in both perception and production[21,22], finding that differences in manner of articulation are more salient in speech perception[23-25] than differences in place of articulation, and vice-versa for speech production[26-29]. These behavioral findings dovetail well with the aforementioned studies of speech representations in the brain that have reported that perceptual representations are organized by manner of articulation[15-17] whereas articulatory representations are organized by place of articulation[17-19]. We therefore designed our stimuli so that the two word stimuli in each "prodDiff" pair (differing in place of articulation, /θ/ vs. /f/) were expected to cause stronger macroscopic differences in neural activation patterns in articulatory areas, whereas the two words in each "percDiff" pair (differing in manner of articulation, /s/ vs. /t/) were expected to cause stronger macroscopic differences in neural activation patterns in perceptual areas. These two groups of words, 'percDiff' and 'prodDiff', were thus the probes with which we investigated the neural signal dynamics over the whole brain during speech perception and speech production.

To consolidate the novel words prior to the EEG recordings[30-32], each subject was instructed to listen to and also produce each word five times the night before the experiment. To ensure consistent voicing patterns, subjects were instructed to produce the initial phoneme in the /θ/ group of words as in "think", i.e., unvoiced.

Behavioral Pseudoword Stimuli	
Word Pairs	**Differences**
THEEP – FEEP	Place of Articulation “prodDiff”
THIPE – FIPE	
THEEN – FEEN	
THOPE – FOPE	
THUP – FUP	
SAFF – TAFF	Manner of Articulation “percDiff”
SARG – TARG	
SOOG – TOOG	
SILP - TILP	
SEEB – TEEB	

Table 1. Behavioral stimuli in the study. Two sets of word-pairs (consisting of 5 pairs each) were designed to differ maximally across one feature of articulation (either place or manner of articulation) while keeping voicing constant. The word pairs that had differences in place of articulation, but similar manner of articulation (/θ/ vs /f/ initial phonemes) were named “prodDiff” word pairs. The word pairs that had differences in manner of articulation, but similar place of articulation (/s/ vs /t/ initial phonemes) were named “percDiff” word-pairs.

Experimental paradigm

The experimental paradigm was implemented using Psychoolbox 3[33] running on MATLAB R2022b[34]. EEG data acquisition featured a standard[35] setup with a dedicated paradigm PC that controlled the presentation of auditory and visual cues to the participants and a data acquisition PC that recorded the neural imaging data from the EEG system. All recording was completed in a single session, with two types of break periods: the first was an intra-session self-paced blink break which occurred at an interval of every 10 trials. During the blink breaks, the experiment was paused and subjects were allowed to blink freely or rest their eyes for a few seconds. The second type of break was an inter-session break, which occurred at the end of every full run and lasted 1-3 minutes, in order to alleviate boredom and fatigue. The experiment consisted of two parts: Speech perception and speech production. Speech perception was always run first, followed by speech production.

Speech perception

The trial structure of the speech perception part of the experiment consisted of a fixation cross present throughout the duration of the trial and an auditory stimulus that was played at t=1000ms with a 500ms jitter (see Figure 1). Further, 20% of all trials per run were oddball trials, which included an English pseudoword that was distinctly different from the other stimuli (‘MEEV’). Subjects were instructed to press a button whenever they heard this oddball stimulus, and at the end of each run their response accuracy to these oddball stimuli was reviewed to gauge subject alertness during the task. Auditory stimuli were presented using pneumatic stereo earbuds. Proper fitting and usage were ensured, and audio output levels were calibrated for consistency during all auditory tasks.

Each subject participated in a single session of 5 runs. Each run consisted of 125 trials (the 20 words from above, repeated 5 times each, and 25 oddball trials, all arranged pseudorandomly with the additional constraint that no word was ever directly repeated).

Speech production

The trial structure of the speech production part of the experiment consisted of a fixation cross present at the beginning of the trial, which had a duration between 500-1000ms (500ms base + jitter). Next, the to-be-produced stimulus was presented on the screen for a duration of 500ms, and then a blank screen remained for the next 500ms. Finally, a visual cue appeared on the screen in the form of a vertical bar which moved along the screen at a preset velocity. The subjects were trained to use this moving bar as a guide for pacing their overt production – which we used as means of standardizing overt productions between participants. Articulations were recorded using a high-quality hypercardiod condenser microphone (Audio-Technica AT4053B) placed at a distance of 10cm from the subject's mouth.

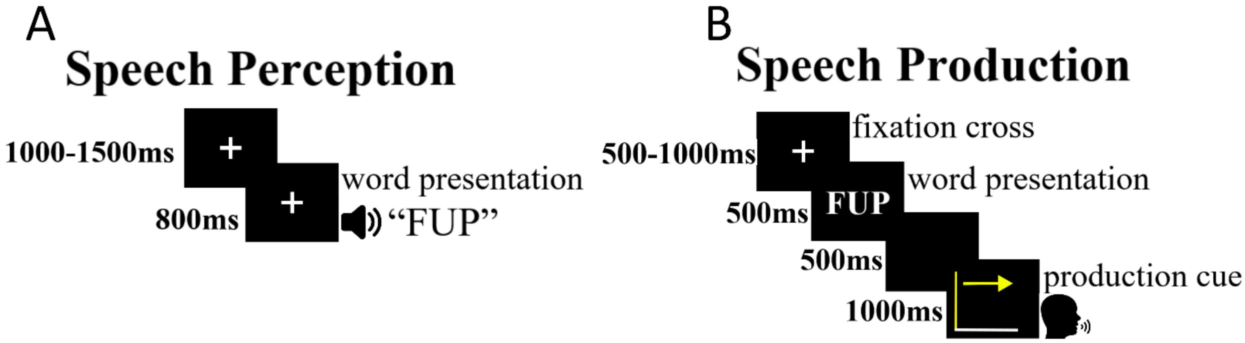


Figure 1. Behavioral paradigm trial structure. (A) Speech perception trial layout. Each trial included two parts: 1) a silent period that lasted between 1000-1500ms, 2) an auditory stimulus presentation which lasted for 800ms. A fixation cross was visible on the screen throughout the entire trial duration. Subjects performed an oddball detection task in the perception trials, see text. (B) Speech production trial structure, which included three parts: 1) a jittered pre-stimulus period with a fixation cross displayed for between 500-1000ms, 2) a visual stimulus presentation period with a fixed duration of 500ms followed by a blank screen for 500ms, 3) a speech production period that began with a visual cue (vertical bar) that moved across the screen at a set velocity to guide production duration.

EEG data acquisition

Electroencephalography (EEG) data were acquired using a BIOSEMI ActiveTwo 64-channel system (Biosemi, Amsterdam, The Netherlands) in an electrically shielded room. Data acquisition included the standard 64 active EEG channels as well as 2 linked-mastoid reference channels with 24-bit resolution and a sampling rate of 2048Hz. Standard BIOSEMI EEG head caps with pre-fixed electrode positions were used.

During setup, electrode impedances were monitored using Biosemi Actiview software, ensuring that no electrode's offset was greater than ±20mV before beginning the recording session. These values were monitored throughout the experimental session and, if necessary, high offsets/unstable electrodes were adjusted during inter-run breaks. Incoming data were low-pass filtered using a fifth order cascaded integrator-comb (CIC) filter response with a -3 dB point at 1/5th of the 2048Hz sampling rate. Event triggers were determined using a photodiode system synced with the experimental paradigm.

EEG data preprocessing

EEG data were preprocessed using EEGLAB[36] and custom MATLAB scripts. For each run, custom MATLAB scripts were used to combine the EEG data with behavioral data prior to importing into the EEGLAB preprocessing pipeline. Event markers were determined using an analogue photodiode channel marked onto the raw BIOSEMI EEG recordings and post-processed during EEGLAB preprocessing. For the speech perception task, each trial had two events of interest: 1) trial start time, and 2) stimulus onset time. For the speech production task, each trial had four events of interest: 1) trial start time, 2) stimulus presentation onset, 3) production cue time, 4) production onset time. The first three events were recorded using a photodiode signal directly fed into the BIOSEMI A-to-D converter via a custom-built Arduino processor, and production onset time was determined using manual annotations of overt production recordings (see Supp. Figs. 1 and 2 and Supplementary Materials). All 64 channels were re-referenced to the averaged of linked mastoids. Separate high pass (0.1Hz) and low pass (30Hz) causal filters were applied to the raw EEG data.

Automatic artifact removal was conducted using EEGLAB's automated artifact rejection method which applies a series of preprocessing steps, including: 1) removal of flatline channels (flatline criterion set to 4 seconds), removal of channels with excessive line noise relative to their signal (line noise criterion set to 4), removal of channels that are poorly correlated with other channels (channel criterion set to 0.8), 5) artifact subspace reconstruction (ASR) (burst criterion set to 20 and burst rejection enabled), and finally removal of time windows with excessive residual artifact content after ASR. Artifact removal was executed with GPU acceleration enabled to enhance processing speed. This procedure ensured the elimination of non-informative channels, channels with abnormal activity, significant line noise contamination, and high-amplitude transient artifacts, thereby improving the overall quality of the EEG data. Bad channels were marked for rejection and subsequently removed. These removed channels were then interpolated using a spherical spline interpolation to estimate the data for the bad channels based on the surrounding good channels. Bad trials were identified using both threshold-based artifact rejection (set to ±120 mV). In addition, for the speech production trials, trials with inter-rater onset ratings differing by more than 40ms were excluded. For the speech perception task, on average the final number of rejected trials was 6.1% of the total number of trials over all 5 runs (n=500). Rejected trials for the speech production task were, on average, higher (17.2%) – likely due to muscle artifacts associated with speech production. The total number of rejected trials per subject is shown in Supp. Fig. 3.

Speech perception data were stimulus-locked to the auditory presentation onset event trigger, with each epoch time range starting 200ms prior to auditory stimulus onset and ending 600ms post-stimulus onset. Speech production data were time-locked to the actual production onset (determined by the average of two expert raters, see below) and the trial epoch was set to 1600ms pre-production onset to 400ms post-production onset. A baseline period was defined from -1600 to -1500ms. This long interval in the speech production case was to include stimulus presentation as well as motor planning (~1000ms prior to the production cue), see Figure 2.

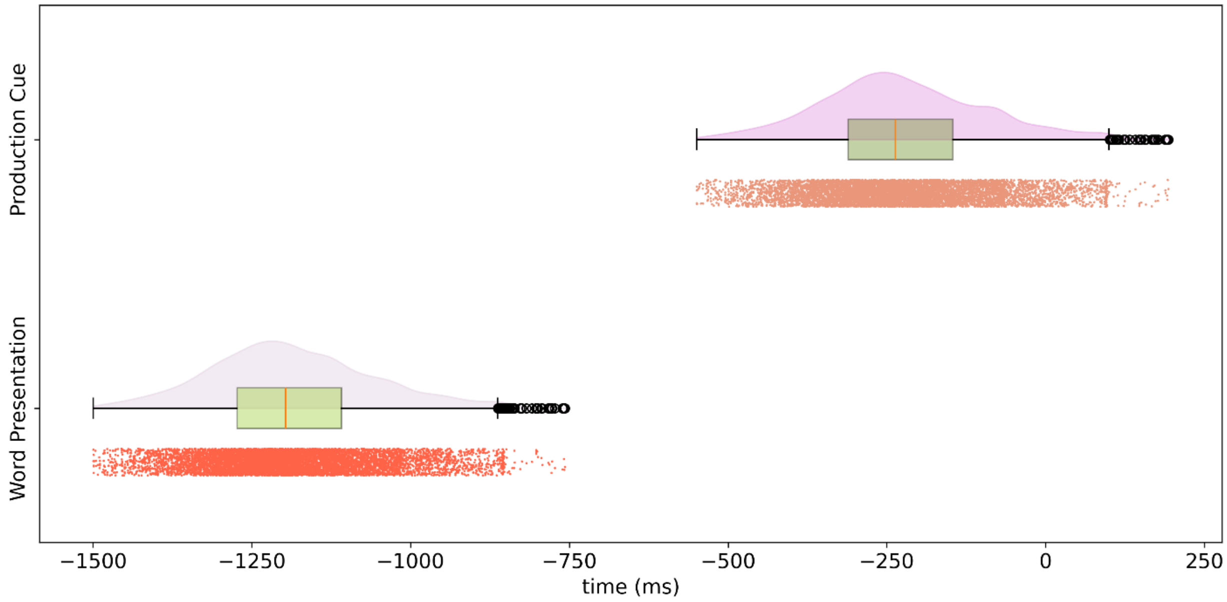


Figure 2. Speech production onset latency distributions. Epochs were time-locked to overt production onset (t=0ms). Production cue and stimulus presentation events had a variable onset with respect to the actual production onset event. The trial-by-trial variability for each event is displayed in the raincloud plot here. The average production cue onset occurred approximately 240ms before production onset, whereas the average stimulus presentation onset occurred by design one second prior to that (see trial structure in Fig. 1B). Activity after t=0ms designate trials in which the subjects' articulation occurred before the production cue displayed onto the screen. Although infrequent, since the time between stimulus presentation and production cue was fixed, subjects did sometimes have anticipatory articulations.

Multivariate analyses of neural responses

A support vector machine (SVM)[37] classifier was used to probe the nature of neural representations engaged during speech perception and production. Classification was performed on multi-channel searchlight responses obtained using sliding window of duration 40ms – results were robust with respect to choices of different sliding window sizes. Following our prior studies that showed an approximate correspondence of estimated cortical sources and overlaying sensor groups [38,39], we chose searchlight sensor groups to roughly overlay left hemisphere brain areas of interest in speech perception and production: "inferior frontal", roughly overlaying inferior frontal cortex, "superior frontal", roughly corresponding to premotor cortex[40], "superior parietal" [11] (we chose not to define a separate "inferior parietal" searchlight, given the close proximity of the relevant areas such as Spt [41] to posterior STG), and "temporal", roughly corresponding to temporal cortex (see Figure 3). Each neural searchlight set was comprised of 3 sensors, and sensor groups were selected so that all searchlights contained disjoint groups of sensors.

Below is an illustration of the sensors that comprise each neural searchlight ROI:

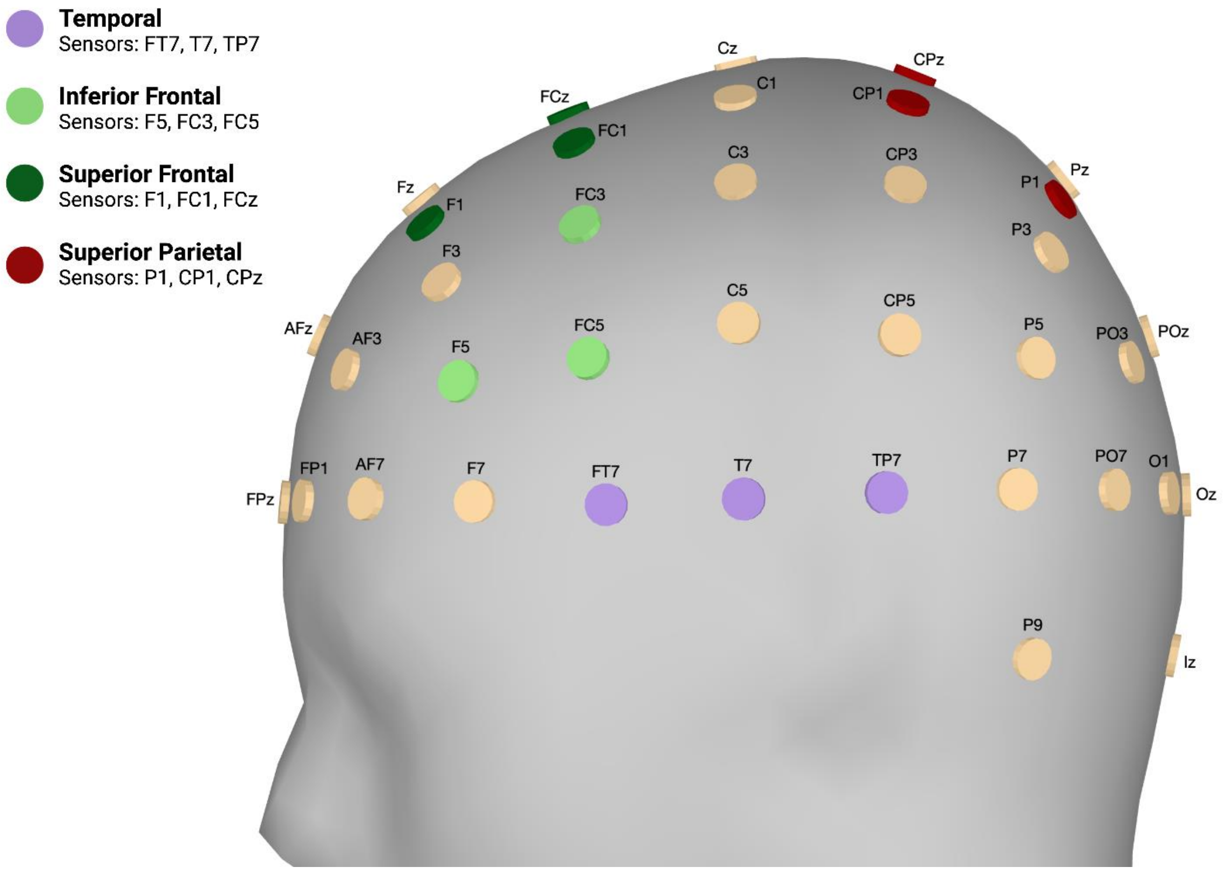


Figure 3. Neural searchlight sensor groups. Four neural searchlights sets were defined, each consisting of 3 sensors. The temporal searchlight included [FT7, T7, TP7], the inferior frontal searchlight included [F5, FC3, FC5], the superior frontal searchlight included [F1, FC1, FCz], and the superior parietal searchlight included [P1, CP1, CPz].

Prior to training the SVM classifier, all channel responses were first standardized (i.e. 0 mean and unit variance) using the StandardScaler in Scikit-Learn[42]. The SVM classifier was trained using 10-fold cross validation and a linear kernel with default hyperparameters using Scikit-Learn[42]. Average subject-specific classifier performance was computed as the average across all folds. These analyses were performed across all 20 subjects and averaged across subjects. Standard error of the mean (SEM) was computed using the average SVM performance for each subject for the sliding window centered at each sample time point. To determine whether SVM classifier performance was above chance, the group-level performance at each time point was tested against chance (0.5) using a one-sample one-sided t-test ($p<0.05$).

Granger causality analyses

To probe the flow of information between sensor groups during speech perception and speech production, we computed directed functional connectivity between sensor groups (based on the sensor-averaged response at each time point) using pairwise (bivariate) Granger-Causality (GC) using the BSMART toolbox[43] and custom MATLAB scripts[38,39]. GC was computed separately in different frequency bands, specifically the theta band (4-7Hz) and the beta band (13–20 Hz). To reduce computational demands, all preprocessed data were first down-sampled from 2048Hz to 500Hz. We used a model order of 20ms (as in ref 43) and a forward-looking sliding window of 40ms (results were robust to slight variations in model order and sliding window width).

Lastly, upon completion of GC analysis, a 100ms baseline time period was defined for each speech task; [-200 to -100ms] (relative to auditory stimulus presentation onset) for speech perception and [-1600 to -1500ms] (relative to speech production onset, see Fig. 2, in order to place the baseline period before visual stimulus presentation) for speech production. The average GC value over this 100ms baseline period was used to compute statistical significance, testing for GC increases relative to baseline (one tailed t-test, $p<0.05$).

Results

Evidence for activation of an inverse model during speech perception

Temporal searchlight speech classification results replicate earlier ECoG studies of human STG in terms of organization and timing of speech information

We first investigated the neural responses and signal dynamics during speech perception. We defined four neural searchlight ROIs along the temporal, inferior frontal, superior frontal, and superior parietal regions in sensor space (see Methods). We hypothesized that since neuronal representations in human STG are organized by manner of articulation[15,16], neural response patterns to word pairs whose initial phoneme differed in manner of articulation (our "percDiff" stimuli) would show greater differences in our temporal sensor group than word pairs whose initial phonemes differed in place of articulation (our "prodDiff" pairs). Thus, an SVM classifier trained on the "percDiff" contrast would show higher classification accuracy in the temporal ROI than a classifier trained on the "prodDiff" stimuli. In addition, based on the ECoG literature on the response latency of phoneme-selective representations in the STG[15,16], we expected the "percDiff" stimuli to be discriminable (i.e., show above-chance classification performance) starting around 100ms.

This is exactly what we observed in our temporal searchlight, see Fig.4A. Significant decoding for percDiff contrasts were seen starting at 73ms and reached a peak decoding accuracy at 181ms, in excellent agreement with prior ECoG studies that found peak discriminability around 180ms[16]. This agreement between invasive human ECoG studies and our EEG study is remarkable. In contrast, prodDiff contrasts (place of articulation) were not decodable in temporal channels until much later, with significant decoding only observed starting at 234ms and at a much lower level than for the percDiff (manner) contrast. These results validate our approach and confirm the prior ECoG finding of an organization of speech representations in human temporal cortex by manner of articulation.

Evidence for an inverse internal model from temporal to frontal via parietal areas: Speech information and latencies

An advantage of our EEG approach relative to prior ECoG studies is its whole-head coverage, allowing us to potentially test the hypothesis of an inverse internal model from temporal regions via parietal to frontal regions. As discussed in the Introduction, in the internal model framework, the function of this inverse model is to map speech sounds to the articulations that produce it. Of special interest therefore was the question of the nature of the organization of the speech information, from a perceptual representation around manner of articulation in sensory cortex (represented by our temporal searchlight, see above) to place of articulation in articulatory brain regions in (pre)motor cortex [17] (represented by our superior frontal searchlight). We hypothesized that this transformation was accomplished in parietal cortex[11], but possibly only further downstream, in frontal regions, as suggested by reports of an organization of speech information by manner of articulation in sensorimotor cortex during listening[17].

Indeed, in our superior parietal searchlight (Fig. 4B), in pronounced contrast to the temporal results, we found significant decoding of the place of articulation contrast (prodDiff word pairs) starting at 107ms, compatible with an encoding of speech information that more strongly emphasized place of articulation.

Significant decoding of manner of articulation was only found *later*, starting at 136.7ms and peaking at 185ms, slightly later than in temporal cortex, compatible with the putatively downstream location of the parietal searchlight in the dorsal pathway.

Moving on to the superior frontal searchlight (Fig. 4C), in agreement with the articulatory function of underlying cortex[40], we again found significant decoding for the place of articulation contrast (112ms), at a slightly later latency than in the superior parietal searchlight.

Evidence for an inverse internal model from temporal to frontal via parietal areas: Information flow

Thus, the kind of speech information that was decoded as well as its latency changes support an inverse internal model from temporal to frontal areas via parietal regions. We next turned to analyses of neural information flow, using Granger Causality between the different searchlights, to further test this hypothesis. We specifically focused on changes in Granger Causality relative to baseline in the theta band (4-7 Hz), following an influential earlier study[44] that had shown that feedforward information flow is mediated by the theta band, whereas feedback influences are conveyed through the beta band (13-20Hz). In agreement with that hypothesis, as previously demonstrated in EEG studies of information flow in the visual system[39], we had provided evidence for theta band (but not beta band) modulations from temporal to parietal areas.

Bearing out the hypothesized role of theta band modulations for feedforward information transmission, we found strong and significant interactions among the different searchlights in the perception task in the theta band (in contrast, beta band modulations were weaker, see Supp. Fig. 4). The timing and directionality of these influences further supported the hypothesized inverse model. Specifically, while there was significantly increased Granger Causality from the temporal to the superior frontal searchlight in the theta band (Fig. 4F), as predicted by the internal inverse model theory, there was no significant increase in GC in the opposite direction (supp. Fig. 5A). Furthermore, while there was only a small increase in GC from the frontal to the parietal searchlight following stimulus onset (supp. Fig. 5B), there was a strong increase in GC from the parietal to the frontal sensor group (Fig. 4H), again supporting directionality from sensory to articulatory representations via parietal areas as predicted by the inverse model.

Thus, taken together, these classification and Granger causality data converge to provide strong experimental support for the predicted internal inverse model from perceptual to articulatory representations via parietal cortex[13,45,46].

Support for an anteroventral stream

In addition to the inverse internal model in the dorsal stream, dual stream models also posit an anteroventral stream connecting perceptual temporal and inferior frontal areas [11].We approximated these regions via our temporal and inferior frontal searchlights. As Fig. 4D shows, classification in the inferior frontal searchlight was similar to that in the temporal searchlight, albeit with a later perceptual decoding peak (210 ms), compatible with a downstream location in the hierarchy. Likewise, Fig. 4G shows strong GC from the temporal to the inferior frontal searchlight starting at 70ms following stimulus onset. In contrast, there was no significant increase in GC in the opposite direction (supp. Fig. 5C). These results are compatible with an anteroventral pathway from temporal to inferior frontal cortex.

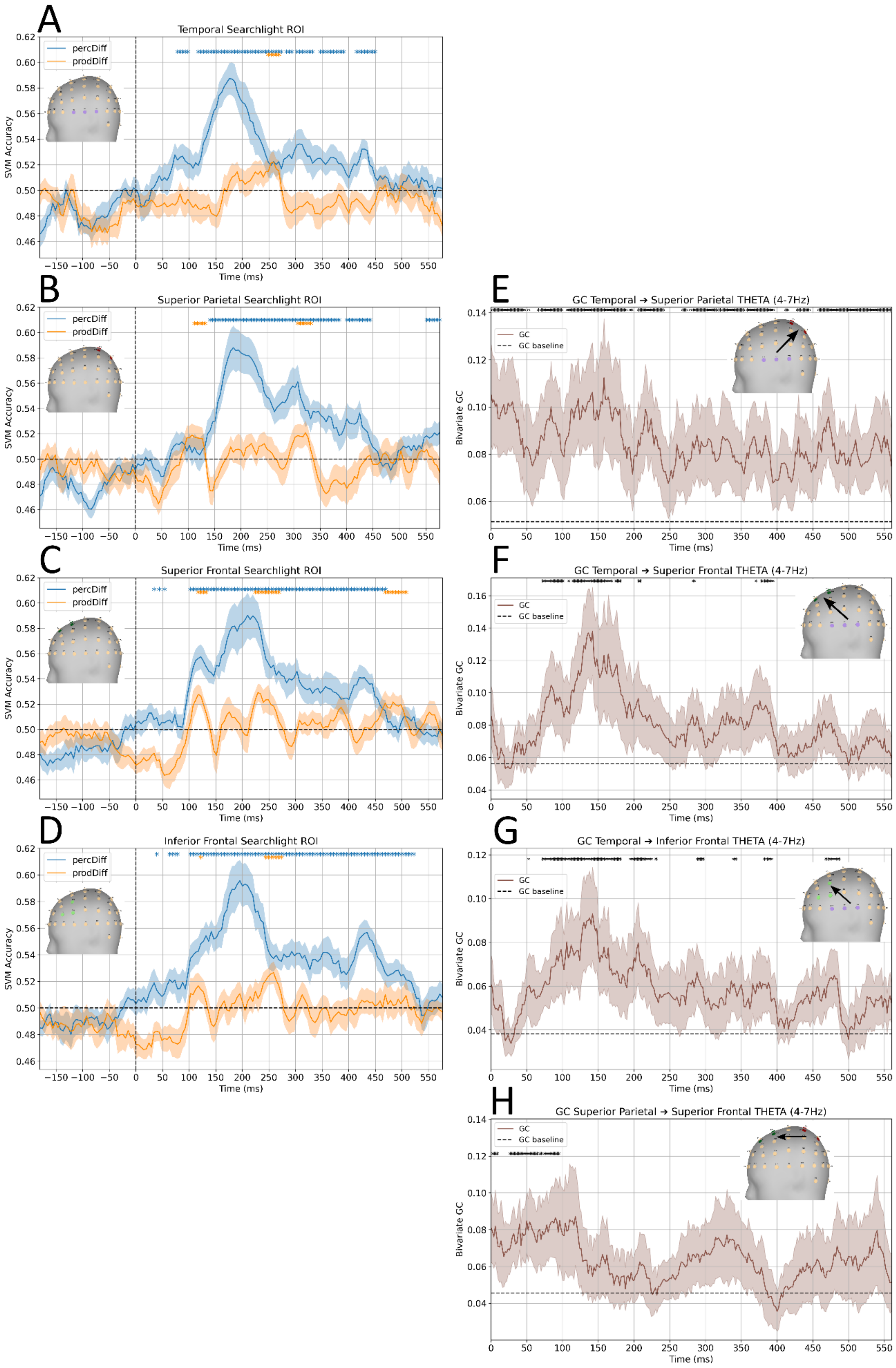

A
Temporal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
Time (ms)
B
Superior Parietal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
Time (ms)
C
Superior Frontal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
Time (ms)
D
Inferior Frontal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
Time (ms)
E
GC Temporal → Superior Parietal THETA (4-7Hz)
GC
GC baseline
Bivariate GC
Time (ms)
F
GC Temporal → Superior Frontal THETA (4-7Hz)
GC
GC baseline
Bivariate GC
Time (ms)
G
GC Temporal → Inferior Frontal THETA (4-7Hz)
GC
GC baseline
Bivariate GC
Time (ms)
H
GC Superior Parietal → Superior Frontal THETA (4-7Hz)
GC
GC baseline
Bivariate GC
Time (ms)

Figure 4. Speech Perception SVM and Granger Causality results. Left column shows support vector machine classifier decoding performance for (A) temporal, (B) superior parietal, (C) superior frontal, and (D) inferior frontal EEG searchlights (see insets showing searchlight channels, compare to Fig. 3). Orange corresponds to decoding of place of articulation contrasts ('prodDiff' word pairs) and blue corresponds to decoding of manner of articulation contrasts ('percDiff' word pairs). Auditory stimulus onset at t=0ms. Above-chance decoding (p<0.05, one-tailed t-test) is shown by asterisks. (E-G) Bivariate Granger Causality (GC) illustrating information flow between sensor groups (see inset). In each figure, the time window begins with auditory stimulus onset (t=0ms). The dashed black line denotes the average baseline GC value (from [-200:-100] ms). Task GC values statistically above baseline are denoted with an asterisk (t-test, p<0.05).

Evidence for a forward internal model during speech production

We next analyzed neural responses in the trials in which individuals were asked to produce the prodDiff and percDiff words. As described in Methods, we epoched the data in each trial to encompass the delay period between the visual word cue presentation up to and beyond production onset. Specifically, the SVM decoding window was extended to -1600ms pre-articulation onset, so as to also encapsulate the stimulus presentation event onset (see Fig. 2). The EEG trials were epoched to the actual overt production onset time (t=0ms). Due to the nature of response time variability between the production cue and the actual overt production onset, there was a range of onset times for the word presentation and production cue events, respectively (Fig. 2). The stimulus presentation event occurred approximately between 1500ms to 850ms prior to production onset. The production cue event onsets ranged between -500ms to +100ms relative to production onset (positive values denote anticipatory articulations which occurred prior to the production cue event appearing on the screen). This large epoch window thus included reading the word, mental rehearsal, motor planning, and, finally, articulation.

Evidence for a "phonological loop" linked to production planning

Intriguingly, while there was no significant decoding of the visually presented word pairs when locked to stimulus onset (supp. Fig 6), when locked to production onset, we were able to decode the prodDiff contrast more than a second prior to speech production onset, with the most consistent classification results for the temporal searchlight (Fig. 5A). This is evidence for the neural substrate of the "phonological loop" [47], as subjects were "keeping in mind" the to-be-produced word. Interestingly, in contrast to the perception results, the temporal searchlight now, in production, showed more significant classification for the place of articulation contrast, i.e., a more "articulatory" representation than during the perception trials, offering an intriguing counterpart to reports that activations in motor areas during *perception* were more similar to perceptual representations in STG and an organization by manner of articulation than the somatotopic representation in motor areas found during production [17].

Evidence for a forward internal model

The internal model theory includes a forward model that predicts the sensory consequences of intended articulations, from frontal premotor to auditory cortex [11,48]. A notable feature of this prediction is that it occurs *before* the actual utterance, in order to enable fast corrective motor control. Indeed, it has been shown that in primate auditory cortex, self-vocalization-induced suppression of neuronal responses began several hundred milliseconds before the onset of the vocalization[49].

These predictions were borne out in our data: We found GC increases from the frontal searchlights (both inferior and superior, Fig. 5E&F) to the temporal searchlight, starting around 450ms *before* production onset and peaking around 250ms before production onset. Intriguingly, shortly thereafter, but still around 200ms before production onset, there was increased decoding for both percDiff and prodDiff contrasts (Fig. 5B&C), indicating the presence of predictive speech information, compatible with the forward model concept. Notably, GC in the opposite direction, from the temporal to the frontal searchlights was flat in the time period before production onset (Supp. Fig. 7.).

In stark contrast to the perception data that found strong modulations of Granger Causality in the theta band between the different searchlights, compatible with the feedforward nature of the hypothesized dorsal stream inverse model[44], GC modulations in the production trials were mostly found in the beta band, with little modulations in the theta band (Supp. Fig. 8), suggesting a "feedback" nature for the forward model from premotor to sensory areas[44].

Somewhat surprisingly, our data did not support involvement of a fronto-parietal-to-temporal route in the forward model: While there was a modest increase in GC from the superior parietal to the temporal searchlight before production onset (Fig. 5G), GC from the superior frontal to the superior parietal searchlight (Fig. 5H) in fact *decreased* to baseline levels 300ms before production onset and rebounded only after speech production onset. In addition, the increase in speech information in the superior parietal searchlight (prodDiff beginning at t=-85ms, Fig. 5D) only occurred only *after* the increase in the temporal searchlight (prodDiff beginning at t=-180ms, Fig. 5A).

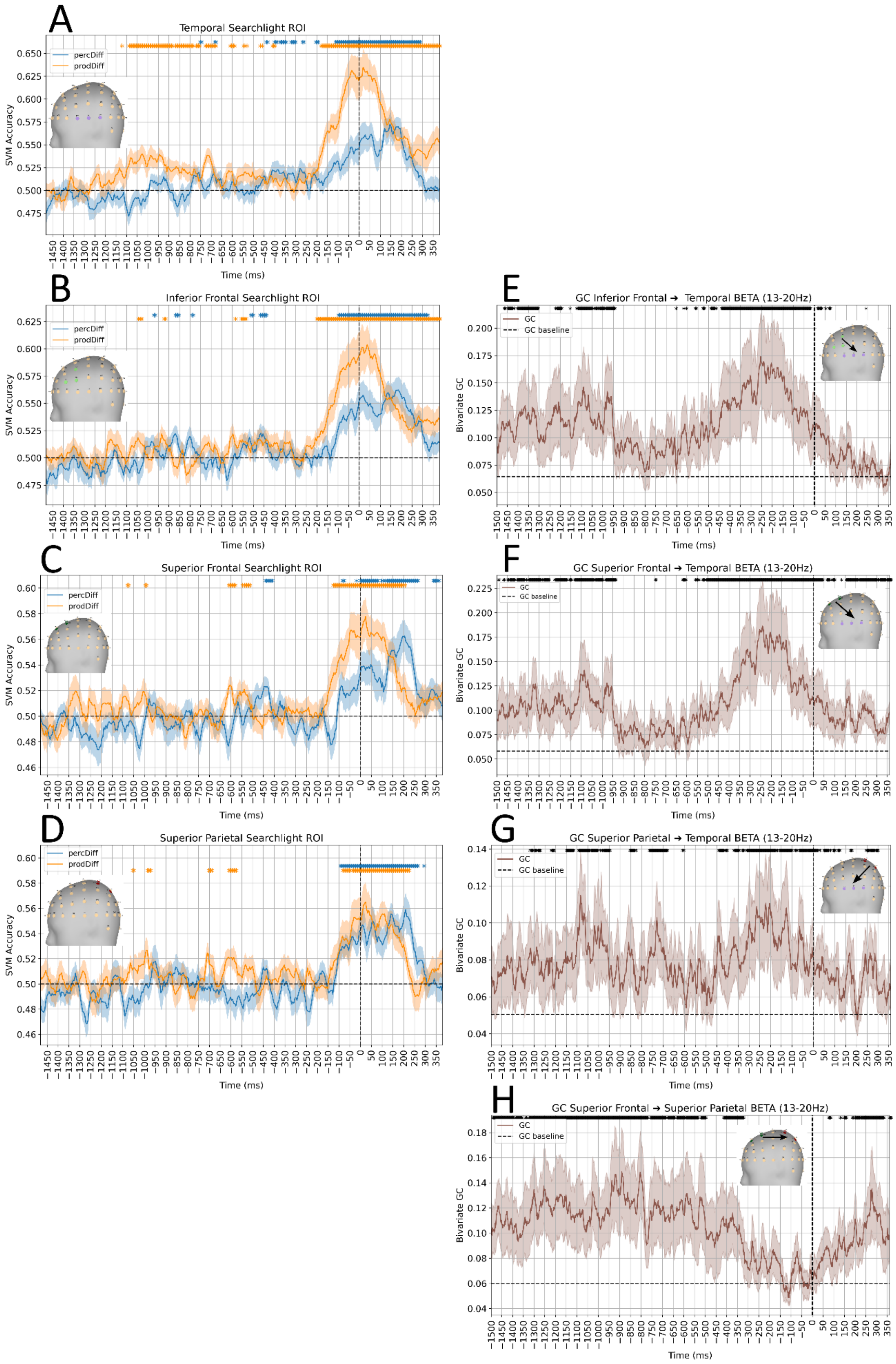

A
Temporal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
0.475
0.500
0.525
0.550
0.575
0.600
0.625
0.650
Time (ms)
B
Inferior Frontal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
0.475
0.500
0.525
0.550
0.575
0.600
0.625
Time (ms)
C
Superior Frontal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
0.46
0.48
0.50
0.52
0.54
0.56
0.58
0.60
Time (ms)
D
Superior Parietal Searchlight ROI
percDiff
prodDiff
SVM Accuracy
0.46
0.48
0.50
0.52
0.54
0.56
0.58
0.60
Time (ms)
E
GC Inferior Frontal ➔ Temporal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Time (ms)
F
GC Superior Frontal ➔ Temporal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0.225
Time (ms)
G
GC Superior Parietal ➔ Temporal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.04
0.06
0.08
0.10
0.12
0.14
Time (ms)
H
GC Superior Frontal ➔ Superior Parietal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
Time (ms)
−1500
−1450
−1400
−1350
−1300
−1250
−1200
−1150
−1100
−1050
−1000
−950
−900
−850
−800
−750
−700
−650
−600
−550
−500
−450
−400
−350
−300
−250
−200
−150
−100
−50
0
50
100
150
200
250
300
350

Figure 5. Speech Production SVM & GC results. Left column shows support vector machine classifier decoding performance for (A) temporal, (B) inferior frontal, (C) superior frontal, and (D) superior parietal. EEG neural searchlight ROIs are the same (see insets showing searchlight channels, compare to Fig. 3). Orange corresponds to decoding of articulatory contrasts ('prodDiff' word pairs differing in place of articulation) and blue corresponds to decoding of perceptual contrasts ('percDiff' word pairs differing in manner of articulation). Time t=0ms corresponds to overt production onset. The dashed black line denotes the average baseline GC value (from [-1500:-1400] ms). Task GC values statistically above baseline are denoted with an asterisk (one-tailed t-test, p<0.05). Similarly, SVM classification performances above chance (0.5) are denoted with an asterisk (one sided t-test, p<0.05).

Discussion

Dual stream models of language[11] [12] provide a neuroanatomical framework for the functional networks of brain regions involved in auditory speech processing. A key feature of the architecture is that the information flow between nodes in the speech system is not just feedforward, but dynamic and varies between speech perception and speech production. Specifically, internal models, which are essential for speech production control and learning, provide a compelling computational framework for mapping the dynamic sensorimotor transformations between the perceptual and articulatory neural representations engaged during either speech perception or speech production. However, testing hypotheses put forth by the internal model framework requires whole-head, high temporal resolution neuroimaging techniques as well as high specificity for detecting the engagement of the target neural representation of interest. To this end, we designed sets of pseudowords optimized to uncover organizations of speech information: one set of words varying by manner of articulation, designed to probe perceptual representations[15-17] and another set varying by place of articulation to probe articulatory representations, previously found to be organized by somatotopy[17,18]. Our whole-brain EEG approach allowed us to examine network dynamics across key frontal, temporal, and parietal sensor regions. With this approach, despite the lower spatial resolution of EEG compared to ECoG used in those previous studies, we were able to decode perceptual representations, organized by manner of articulation, within the same latency periods as those reported in previous ECoG studies[15,16]. We then used the same approach to investigate neural signals during speech perception and production and the information flow across sensor regions, providing evidence for the hypothesized forward and inverse internal models.

Inverse model: from perceptual (sound) to articulatory (motor) representations

As predicted for the inverse internal model linking perception to articulatory representations, we found that speech perception triggers an information flow from perceptual representations in temporal channels to more articulatory representations in parietal and ultimately superior frontal channels. These findings dovetail well (and extend them by revealing the underlying neural signal dynamics) with prior fMRI studies which found evidence of an inverse model during speech perception, in particular of categorical phoneme representations in premotor cortex that were functionally connected to perceptual representations in posterior STG[46].

The involvement of motor regions in speech perception has been a decades-long topic of interest, with numerous studies providing evidence for this phenomenon[50,51]. For instance, Wilson et al. used fMRI to show that listening to speech activates motor areas involved in speech production[40]. The question of the organization of speech information encoded in motor areas during perception has been more controversial: We found evidence for both manner and place of articulation representations in motor regions during speech perception. This aligns with previous fMRI studies which found that these features could be decoded from dorsal stream areas during listening[46,52-55]. Cheung et al. found that somatosensory cortex primarily represented speech by manner, and not place, during listening (and vice versa; place, but not manner, during speech production)[17] and a prior fMRI study likewise could not decode place of articulation information during speech perception [55]. Somewhat resolving the discrepancies, our results reveal dynamic changes in the organization of speech information in channels over dorsal stream regions during speech perception, compatible with the sensorimotor transformation function of the dorsal auditory stream. The presence of both manner and place neural codes suggests a mapping between perceptual and articulatory representations all along the dorsal stream, including parietal as well as right in (pre)motor areas. This is somewhat reminiscent of the presence of gaze-related information along the dorsal visual stream, thought to relate to transforming retino-centric visual information to an ego-centric frame that can support actions such as reaching.

Forward model: from articulatory plans (motor/place) to sensory (sound/manner) predictions

The concept of a forward model has been central to several influential theories. Rauschecker & Scott's model refers to this internal model as the *efference copy* that is generated during motor planning[11], and the State Feedback Control model posits that a forward model generates predictions of the sensory consequences of overt articulations[56].

During speech production, our findings reveal the forward model at work, connecting motor commands to anticipated sensory outcomes. Indeed, our Granger causality results show information flow from frontal to temporal searchlights and the SVM results show the presence of both manner and place speech information in auditory areas several hundred milliseconds prior to speech production. This is in agreement with NHP electrophysiological studies[49], that showed that neurons in auditory cortex are modulated by vocal production 220ms before production onset.

The fact that we did not observe any directed information flow from superior frontal to superior parietal may suggest that there is instead a direct frontal-to-temporal pathway engaged via the dorsal route (via the arcuate fasciculus/superior longitudinal fasciculus[57]). But our results are also compatible with the existence of a forward model in the anteroventral stream[58], especially given the Granger causality results between the inferior frontal and temporal sensor groups. Such a forward model in the anteroventral stream could, for instance, facilitate robust and reliable speech recognition under noisy conditions[58]. Predictive processes may therefore be more distributed throughout the speech processing system than previously thought. Further, the rapid activation of these underlying neural representations suggests a highly efficient mapping from motor plans to predicted sensory consequences. This not only aligns with what has previously been proposed by other models[8], but also reemphasizes the necessity for such real-time feed-forward mapping. This interplay between perception and production representations in auditory and motor regions suggests that the brain maintains the ability to flexibly recruit neural resources depending on task demands. In summary, our results, which agree with previous cognitive[59] and neuroanatomical speech models[11], provide evidence of a forward model which generates predictions of the sensory consequences of speech motor commands. Further evidence is garnered from investigating auditory representations of learned sound sequences in motor regions of the macaque brain[60].

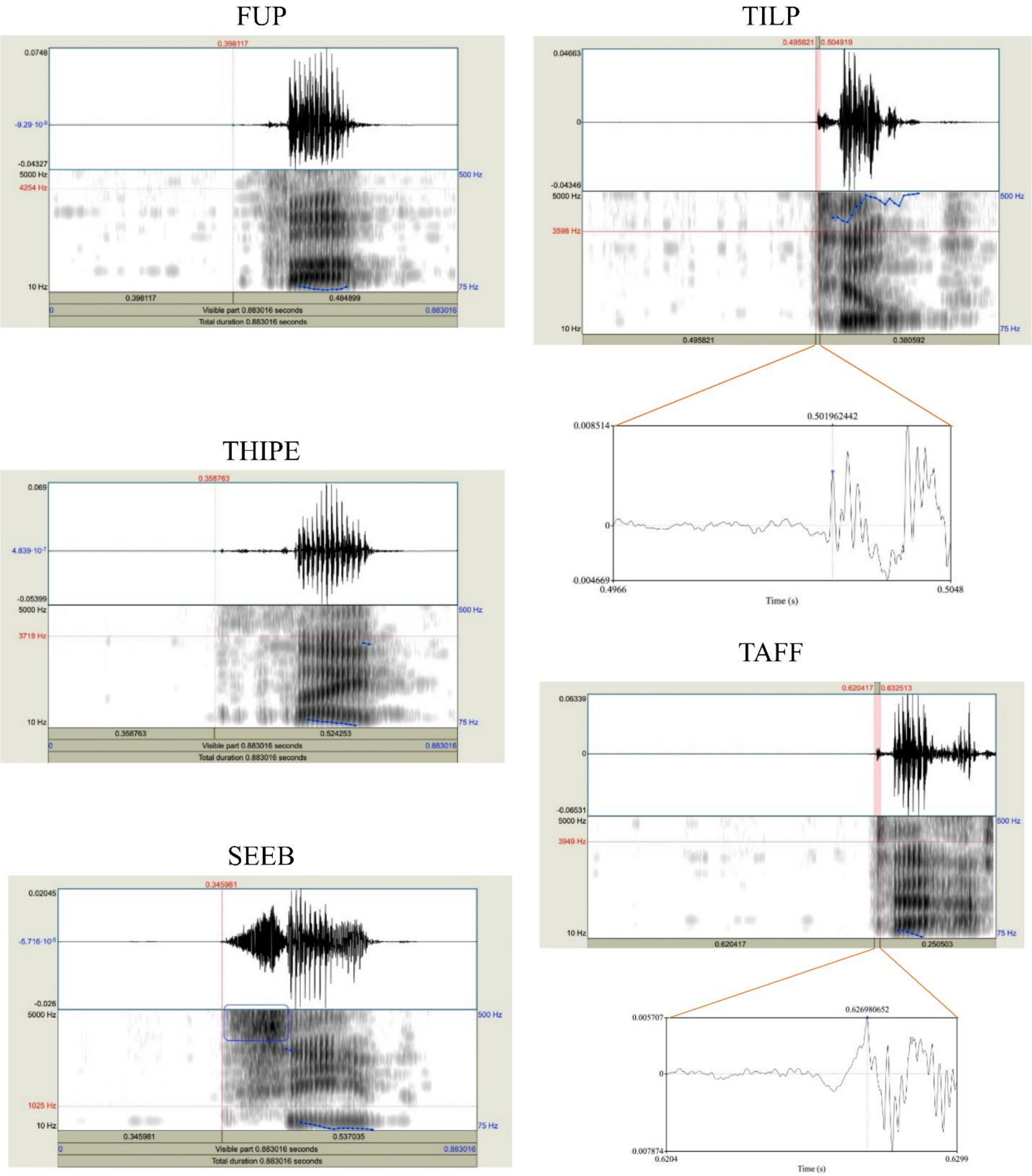
Our results also revealed robust decoding for place (but not manner) of articulation contrast approximately one second prior to production onset in the temporal ROI. This showed evidence of the "phonological loop", a key component of working memory models wherein verbal information is temporarily stored and rehearsed[47]. This suggests that participants actively maintained the to-be produced words in a "phonological store" and rehearsed these words prior to producing them[61]. Our SVM decoding results revealed that the neural substrate of this phonological loop in the temporal searchlight was primarily organized by place of articulation, thereby having a more "articulatory" representation than during speech perception.

Opportunities for future studies

Our study demonstrated the ability to probe the organization of speech information as well as information flow across the large-scale speech network in the brain using EEG. The chief limitation of the current study is its limited spatial resolution, inherent in the low number of sensors used in EEG. It will be interesting to translate our approach to imaging modalities with higher spatial resolution such as MEG, which would help to resolve specific areas such as the role of area Spt[62], e.g., in the phonological loop, or in particular a more fine-grained resolution of the sensorimotor transformations in parietal and frontal cortex. In addition, it will be interesting in future studies with higher spatial resolution to probe the existence of two forward models, via the anteroventral as well as the dorsal stream.

Our study also opens the door to future studies in language acquisition (native and L2), in particular those that investigate the link between perception and production learning. For instance, given the tight link between speech perception and speech production learning predicted by the internal inverse model, our results predict that in development, speech selectivity in motor areas during speech perception derives from (but is not identical to) perceptual speech representations in temporal areas, and that perceptual selectivity predicts production learning, as suggested by recent studies[63].

Supplementary Materials


FUP

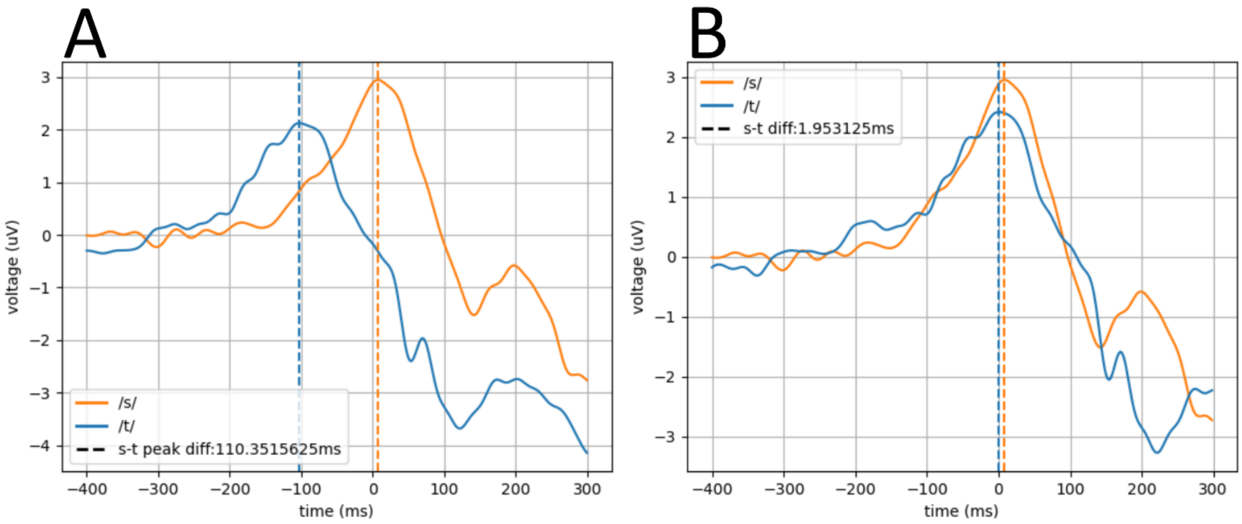
TILP

THIPE

TAFF

SEEB

Supplementary Figure 1. Illustrative examples for speech production onset detection using auditory recordings. Using the PRAAT software package, all trials were first batch filtered between 310-15000Hz. Next, each trial was independently annotated by two expert raters (red vertical lines, shaded region shows inter-rate difference) to determine the production onset time.

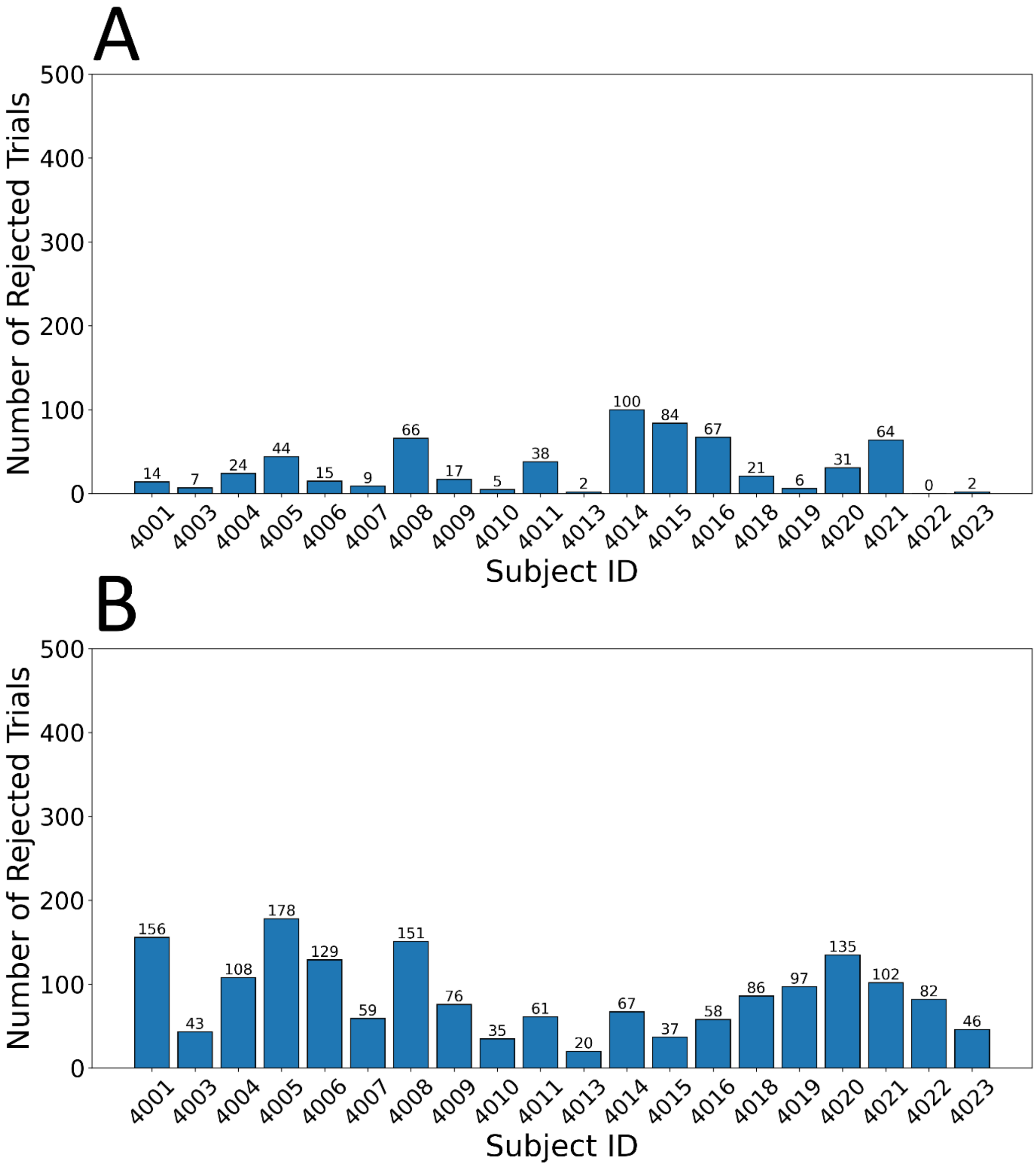


Supplementary Figure 2. Effect of differences in the Acoustic-Articulatory-Interval (AAI) for fricative vs. stop consonants. The non-corrected stop consonant /t/ has a silent period (AAI) followed by a burst of air (acoustic onset) which is difficult to accurately annotate using acoustic trial recordings during the overt speech production task. This AAI discrepancy between /s/ and /t/ is on the order of 100ms[64,65], as also evident from the phoneme-specific grand subject average ERP in (A). To account for this silent period in the articulation of stop consonants (all behavioral stimuli beginning with /t/), we implemented a correction of the manual production onsets, leading to better alignment of the ERPs (B).

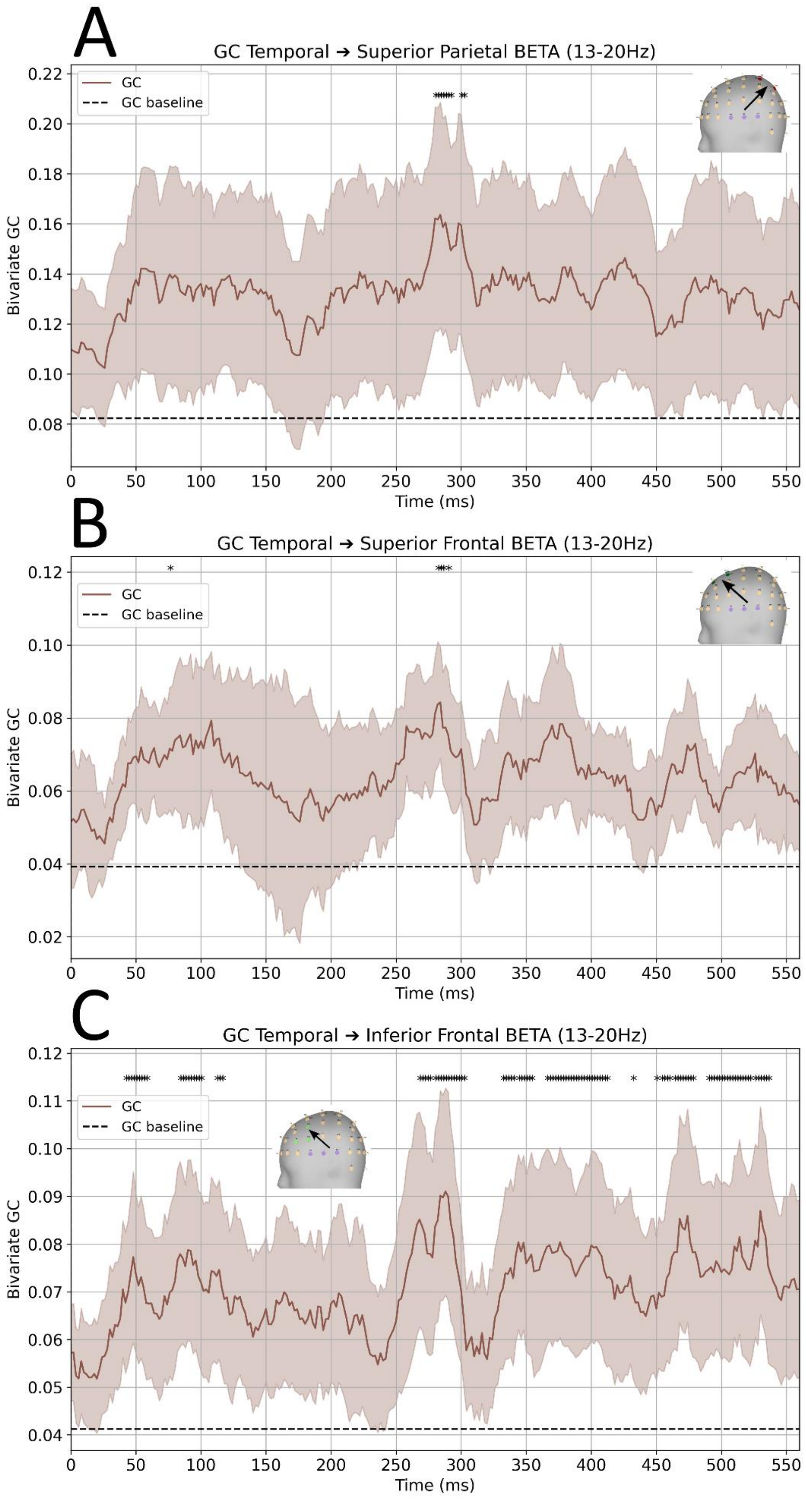
Determining speech production onset times

In order to accurately quantify speech production onset times in the speech production trials, audio recordings in each trial were visualized in PRAAT (Version 6.4.13)[66] and the corresponding spectrograms and time series waveforms were used to determine the onset of articulation (see Supplementary Figure 1 for examples). Recordings for each trial were first filtered between 310-15000Hz, and both time and frequency plots were reviewed independently by two expert raters. Any trials with an inter-rater difference in production onset time greater than 40ms were noted (on average this occurred on less than 10 trials/subject) and the corresponding trial was marked for rejection during preprocessing.

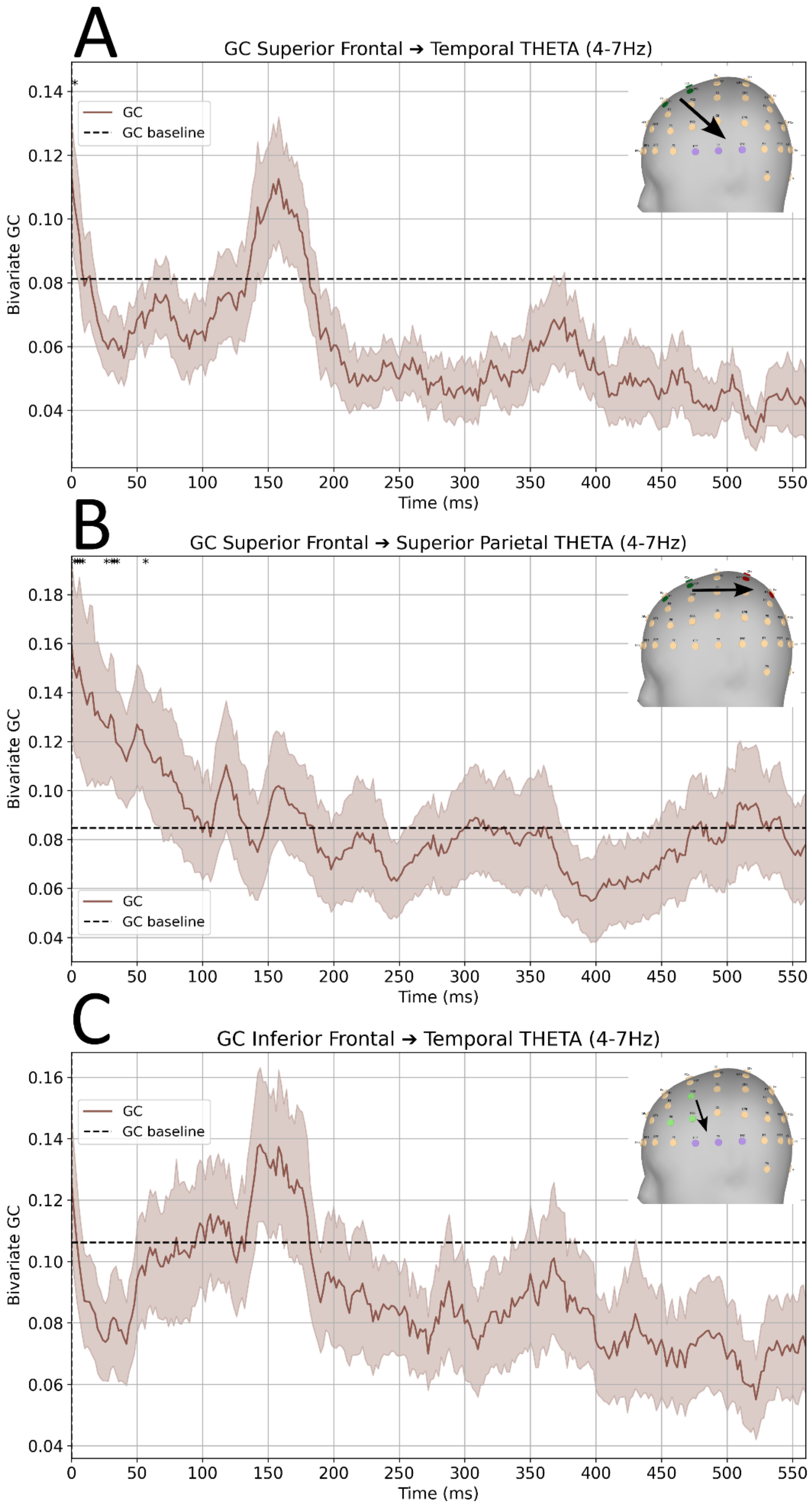
For the /f/ and /θ/ and /s/ phoneme word groups, there is a clear landmark on the spectrograms to determine the onset of articulation. However, for /t/ stops, the utterance included a silent period in which articulatory activation is not captured by the microphone – a phenomenon known as the Acoustic-Articulatory-Interval[67,68]. It has long been established that muscle activity during overt articulation typically precedes the acoustic component by up to 100ms depending on the phoneme type[64,65,69]. This so-called Acoustic-Articulatory-Interval (AAI) difference is particularly relevant for stops, which involve a silent period followed by a release of air; for these phonemes, the articulatory onset during speech production occurs up to 100ms before the acoustic onset[65,68,69] (see Supp. Fig. 2). Indeed, grand subject average ERP signals showed a pronounced difference of about 100ms between fricatives and stops (see Supp. Fig. 2A). To account for the AAI discrepancy and better align trials to the onset of articulatory muscle movement across initial phonemes, the onset times for trials with an initial stop /t/ were therefore corrected by 100ms, leading to a better alignment of ERPs (see Supp. Fig. 2B).



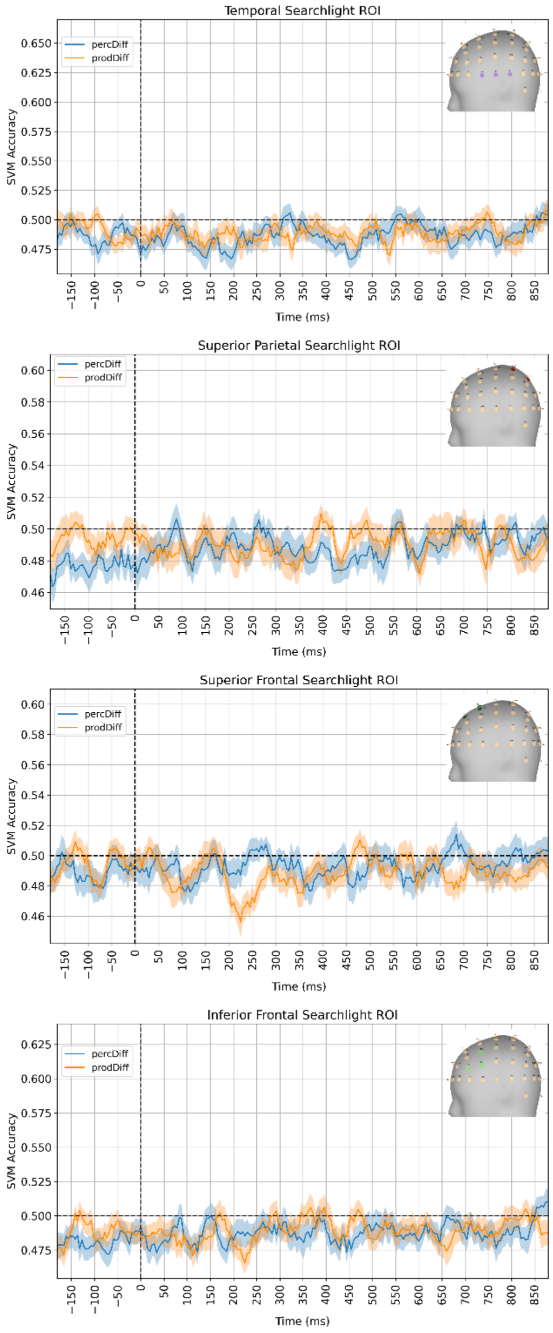
Supplementary Figure 3. Total numbers of rejected trials for each subject after EEG preprocessing. A) for the speech perception task. B) for the speech production task.


A
GC Temporal ➔ Superior Parietal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
Time (ms)
0 50 100 150 200 250 300 350 400 450 500 550
B
GC Temporal ➔ Superior Frontal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.12
0.10
0.08
0.06
0.04
0.02
Time (ms)
0 50 100 150 200 250 300 350 400 450 500 550
C
GC Temporal ➔ Inferior Frontal BETA (13-20Hz)
GC
GC baseline
Bivariate GC
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
Time (ms)
0 50 100 150 200 250 300 350 400 450 500 550

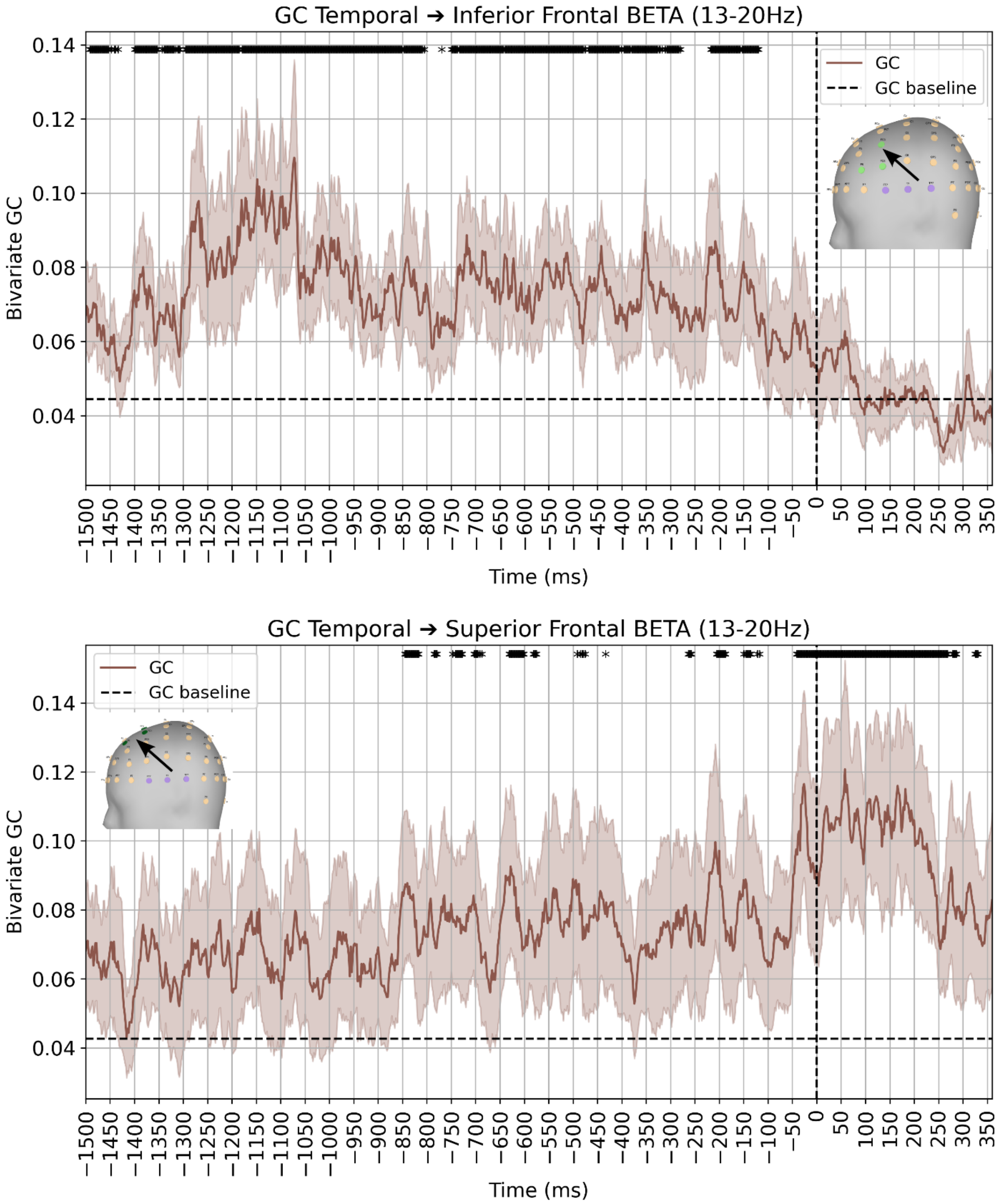
Supplemental Figure 4. Bivariate Granger Causality results in beta frequency band. In each panel the source ROI is temporal and the target is A) superior parietal, B) superior frontal, C) inferior frontal.


A
GC Superior Frontal ➔ Temporal THETA (4-7Hz)
*
GC
GC baseline
Bivariate GC
0.14
0.12
0.10
0.08
0.06
0.04
0
50
100
150
200
250
300
350
400
450
500
550
Time (ms)
B
GC Superior Frontal ➔ Superior Parietal THETA (4-7Hz)
*** *** *
GC
GC baseline
Bivariate GC
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0
50
100
150
200
250
300
350
400
450
500
550
Time (ms)
C
GC Inferior Frontal ➔ Temporal THETA (4-7Hz)
GC
GC baseline
Bivariate GC
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0
50
100
150
200
250
300
350
400
450
500
550
Time (ms)

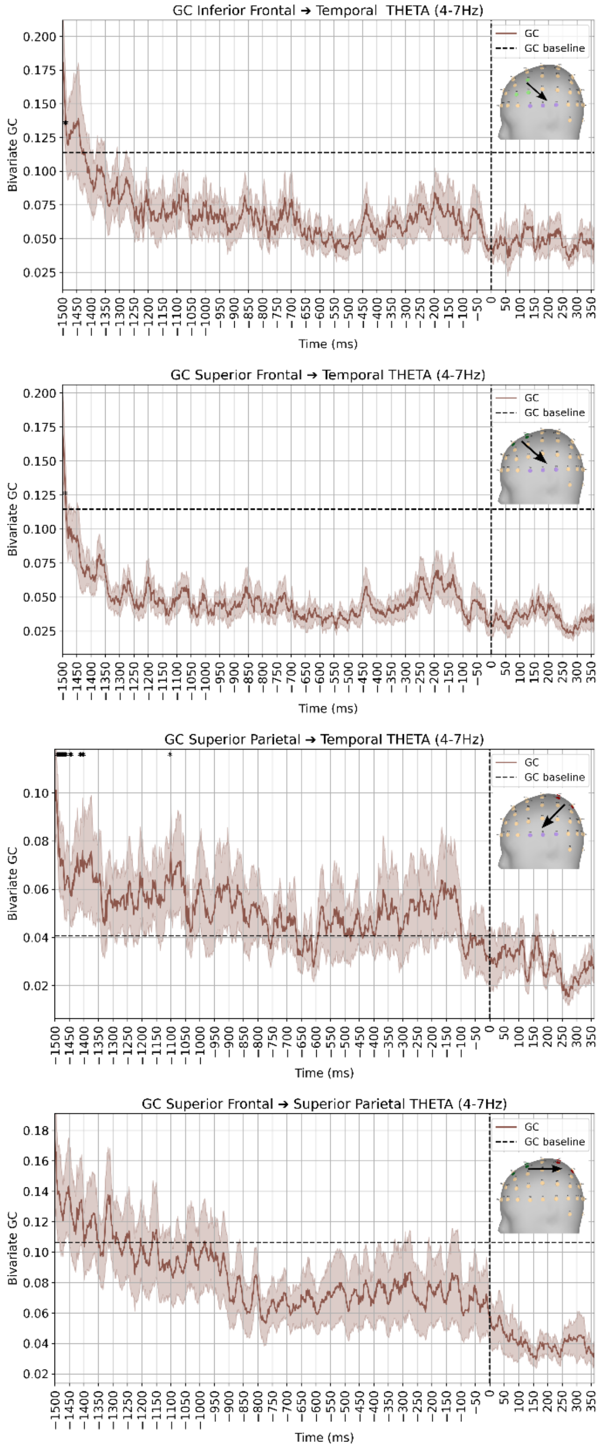
Supplemental Figure 5. Speech Perception bivariate Granger Causality (GC) results in theta frequency band. A) GC from superior frontal to temporal ROI. B) GC from superior frontal to superior parietal ROI. C) GC from inferior frontal to temporal ROI. Asterisks indicate significance GC above baseline (one-tailed t-test, p<0.05).



Supplemental Figure 6. Stimulus-locked SVM classification results during the production trials. For these analyses, the epoch was stimulus-locked to the visual word presentation event. Time t=0ms refers to the onset of the word on the screen, which lasted for 500ms. After this period, a fixation cross was displayed for another 500ms. Notably, no significant above-chance decoding was obtained in any of the searchlights, in pronounced contrast to the production onset-locked results (Fig. 5 in main paper).



Supplemental Figure 7. Speech production bivariate theta band Granger-Causality results – from temporal to frontal searchlights. A) GC from temporal to inferior frontal, B) GC from temporal to superior frontal.



Supplemental Figure 8. Speech production bivariate Granger-Causality results – theta band. Same seed-target pairs as in Fig. 5 E-H.

Works Cited

1 Jordan, M. I. & Rumelhart, D. E. Forward models: Supervised learning with a distal teacher. *Cognitive Science* **16**, 307-354 (1992). https://doi.org/https://doi.org/10.1016/0364-0213(92)90036-T

2 Kawato, M., Furukawa, K. & Suzuki, R. A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics* **57**, 169-185 (1987). https://doi.org/10.1007/BF00364149

3 Wolpert, D. M. & Miall, R. C. Forward Models for Physiological Motor Control. *Neural Netw* **9**, 1265-1279 (1996). https://doi.org/10.1016/s0893-6080(96)00035-4

4 Wolpert, D. M., Doya, K. & Kawato, M. A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci* **358**, 593-602 (2003). https://doi.org/10.1098/rstb.2002.1238.

5 Oller, D. K. & Eilers, R. E. The Role of Audition in Infant Babbling. *Child Development* **59**, 441-449 (1988). https://doi.org/10.2307/1130323

6 Oller, D. K., Wieman, L. A., Doyle, W. J. & Ross, C. Infant babbling and speech. *Journal of Child Language* **3**, 1-11 (1976). https://doi.org/10.1017/S0305000900001276

7 Kawato, M. Internal models for motor control and trajectory planning. *Curr Opin Neurobiol* **9**, 718-727 (1999). https://doi.org/10.1016/s0959-4388(99)00028-8

8 Tourville, J. A. & Guenther, F. H. The DIVA model: A neural theory of speech acquisition and production. *Lang Cogn Process* **26**, 952-981 (2011). https://doi.org/10.1080/01690960903498424

9 Lichtheim, L. On aphasia. *Brain* **7**, 433-484 (1885).

10 Wernicke, C. The symptom complex of aphasia: A psychological study on an anatomical basis. In R. S. Cohen, & M. W. Wartofsky (Eds.). *Boston studies in the philosophy of science*, 34-97 (1874).

11 Rauschecker, J. P. & Scott, S. K. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience* **12**, 718-724 (2009). https://doi.org/10.1038/nn.2331 PMID - 19471271

12 Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nature Reviews Neuroscience* **8**, 393-402 (2007). https://doi.org/10.1038/nrn2113

13 Rauschecker, J. P. An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hear Res* **271**, 16-25 (2011). https://doi.org/10.1016/j.heares.2010.09.001

14 Kell, A., Yamins, D., Shook, E., Norman-Haignere, S. & McDermott, J. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98** (2018). https://doi.org/10.1016/j.neuron.2018.03.044

15 Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006-1010 (2014). https://doi.org/10.1126/science.1245994

16 Lakretz, Y., Ossmy, O., Friedmann, N., Mukamel, R. & Fried, I. Single-cell activity in human STG during perception of phonemes is organized according to manner of articulation. *Neuroimage* **226** (2021). https://doi.org/10.1016/j.neuroimage.2020.117499

17 Cheung, C., Hamilton, L. S., Johnson, K. & Chang, E. F. The auditory representation of speech sounds in human motor cortex. *eLife* **5** (2016). https://doi.org/10.7554/eLife.12577

18 Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327-332 (2013). https://doi.org/10.1038/nature11911

19 Brown, S. *et al.* The somatotopy of speech: phonation and articulation in the human motor cortex. *Brain Cogn* **70**, 31-41 (2009). https://doi.org/10.1016/j.bandc.2008.12.006

20 Müsch, K., Himberger, K., Tan, K. M., Valiante, T. A. & Honey, C. J. Transformation of speech sequences in human sensorimotor circuits. *Proc Natl Acad Sci U S A* **117**, 3203-3213 (2020). https://doi.org/10.1073/pnas.1910939117

21 Goldstein, L. Categorical features in speech perception and production. *J Acoust Soc Am* **67**, 1336-1348 (1980). https://doi.org/10.1121/1.384079

22 Thorin, J., Sadakata, M., Desain, P. & McQueen, J. M. Perception and production in interaction during non-native speech category learning. *J Acoust Soc Am* **144**, 92 (2018). https://doi.org/10.1121/1.5044415

23 Miller, G. A. & Nicely, P. E. An Analysis of Perceptual Confusions Among Some English Consonants. *J Acoust Soc Am* **27**, 338-352 (1955). https://doi.org/10.1121/1.1907526

24 Benkí, J. R. Analysis of English nonsense syllable recognition in noise. *Phonetica* **60**, 129-157 (2003). https://doi.org/10.1159/000071450

25 Harris, K. S. Cues for the Discrimination of American English Fricatives in Spoken Syllables. *Language and Speech* **1**, 1-7 (1958). https://doi.org/10.1177/002383095800100101

26 Wickelgren, W. A. Phonemic similarity and interference in short-term memory for single letters. *Journal of Experimental Psychology* **71**, 396-404 (1966). https://doi.org/10.1037/h0022998

27 Fromkin, V. A. The Non-Anomalous Nature of Anomalous Utterances. *Language* **47**, 27-52 (1971). https://doi.org/10.2307/412187

28 Buchwald, A. & Miozzo, M. Finding levels of abstraction in speech production: Evidence from sound-production impairment. *Psychological Science* **22**, 1113-1119 (2011). https://doi.org/10.1177/0956797611417723

29 Pouplier, M. & Goldstein, L. Asymmetries in the perception of speech production errors. *Journal of Phonetics* **33**, 47-75 (2005). https://doi.org/https://doi.org/10.1016/j.wocn.2004.04.001

30 Davis, M. H., Di Betta, A. M., Macdonald, M. J. & Gaskell, M. G. Learning and consolidation of novel spoken words. *J Cogn Neurosci* **21**, 803-820 (2009). https://doi.org/10.1162/jocn.2009.21059

31 Bakker, I., Takashima, A., van Hell, J. G., Janzen, G. & McQueen, J. M. Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language* **73**, 116-130 (2014). https://doi.org/https://doi.org/10.1016/j.jml.2014.03.002

32 Dumay, N. & Gaskell, M. G. Sleep-associated changes in the mental representation of spoken words. *Psychol Sci* **18**, 35-39 (2007). https://doi.org/10.1111/j.1467-9280.2007.01845.x

33 Brainard, D. H. The Psychophysics Toolbox. *Spatial Vision* **10**, 433-436 (1997). https://doi.org/https://doi.org/10.1163/156856897X00357

34 MATLAB VER - 9.13.0 (R2022b) (The MathWorks Inc., Natick, Massachusetts, 2022).

35 Nieto, N., Peterson, V., Rufiner, H. L., Kamienkowski, J. E. & Spies, R. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Scientific Data* **9**, 52 (2022). https://doi.org/10.1038/s41597-022-01147-2

36 Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods* **134**, 9-21 (2004).

37 Boser, B. E., Guyon, I. M. & Vapnik, V. N. in *Proceedings of the fifth annual workshop on Computational learning theory* 144–152 (Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 1992).

38 Martin, J. G., Cox, P. H., Scholl, C. A. & Riesenhuber, M. A crash in visual processing: Interference between feedforward and feedback of successive targets limits detection and categorization. *Journal of Vision* **19**, 20-20 (2019). https://doi.org/10.1167/19.12.20

39 Damera, S. R. *et al.* From shape to meaning: Evidence for multiple fast feedforward hierarchies of concept processing in the human brain. *Neuroimage* **221**, 117148 (2020). https://doi.org/10.1016/j.neuroimage.2020.117148

40 Wilson, S. M., Saygin, A. P., Sereno, M. I. & Iacoboni, M. Listening to speech activates motor areas involved in speech production. *Nat Neurosci* **7**, 701-702 (2004). https://doi.org/10.1038/nn1263

41 Hickok, G. The functional neuroanatomy of language. *Phys Life Rev* **6**, 121-143 (2009). https://doi.org/10.1016/j.plrev.2009.06.001 PMID - 20161054

42 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

43 Cui, J., Xu, L., Bressler, S. L., Ding, M. & Liang, H. BSMART: A MATLAB/C toolbox for analysis of multichannel neural time series. *Neural Networks* **21**, 1094-1104 (2008). https://doi.org/10.1016/j.neunet.2008.05.007

44 Bastos, A. M. *et al.* Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* **85**, 390-401 (2015). https://doi.org/10.1016/j.neuron.2014.12.018

45 Guenther, F. H. Cortical interactions underlying the production of speech sounds. *J Commun Disord* **39**, 350-365 (2006). https://doi.org/https://doi.org/10.1016/j.jcomdis.2006.06.013

46 Chevillet, M. A., Jiang, X., Rauschecker, J. P. & Riesenhuber, M. Automatic phoneme category selectivity in the dorsal auditory stream. *Journal of Neuroscience* **33**, 5208-5215 (2013).

47 Baddeley, A. *Working memory*. (Clarendon Press/Oxford University Press, 1986).

48 Hickok, G. Computational neuroanatomy of speech production. *Nature Reviews Neuroscience* **13**, 135-145 (2012). https://doi.org/10.1038/nrn3158 PMID - 22218206

49 Eliades, S. J. & Wang, X. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. *J Neurophysiol* **89**, 2194-2207 (2003). https://doi.org/10.1152/jn.00627.2002

50 Liberman, A. M. & Mattingly, I. G. The motor theory of speech perception revised. *Cognition* **21**, 1-36 (1985).

51 Lotto, A. J., Hickok, G. S. & Holt, L. L. Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences* **13**, 110-114 (2009). https://doi.org/https://doi.org/10.1016/j.tics.2008.11.008

52 Correia, J. M., Jansma, B., Hausfeld, L., Kikkert, S. & Bonte, M. EEG decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. *Front Psychol* **6** (2015). https://doi.org/10.3389/fpsyg.2015.00071

53 Lankinen, K. *et al.* Role of articulatory motor networks in perceptual categorization of speech signals: a 7T fMRI study. *Cereb Cortex* **33**, 11517-11525 (2023). https://doi.org/10.1093/cercor/bhad384

54 Pulvermüller, F., Moseley, R. L., Egorova, N., Shebani, Z. & Boulenger, V. Motor cognition–motor semantics: Action perception theory of cognition and communication. *Neuropsychologia* **55**, 71-84 (2014). https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2013.12.002

55 Arsenault, J. S. & Buchsbaum, B. R. No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. *Psychon B Rev* **23**, 1231-1240 (2016). https://doi.org/10.3758/s13423-015-0988-z

56 Hickok, G., Houde, J. & Rong, F. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* **69**, 407-422 (2011). https://doi.org/10.1016/j.neuron.2011.01.019 PMID - 21315253

57 Geva, S. & Fernyhough, C. A Penny for Your Thoughts: Children's Inner Speech and Its Neuro-Development. *Front Psychol* **10**, 1708 (2019). https://doi.org/10.3389/fpsyg.2019.01708

58 Hickok, G. The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. *J Commun Disord* **45**, 393-402 (2012). https://doi.org/10.1016/j.jcomdis.2012.06.004 PMID - 22766458

59 Houde, J. F. & Nagarajan, S. S. Speech production as state feedback control. *Front Hum Neurosci* **5**, 82 (2011). https://doi.org/10.3389/fnhum.2011.00082

60 Archakov, D. *et al.* Auditory representation of learned sound sequences in motor regions of the macaque brain. *Proceedings of the National Academy of Sciences* **117**, 15242-15252 (2020). https://doi.org/doi:10.1073/pnas.1915610117

61 Jacquemot, C. & Scott, S. K. What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences* **10**, 480-486 (2006). https://doi.org/https://doi.org/10.1016/j.tics.2006.09.002

62 Buchsbaum, B. R., Hickok, G. & Humphries, C. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science* **25**, 663-678 (2001). https://doi.org/10.1016/S0364-0213(01)00048-9

63 Tsao, F. M., Liu, H. M. & Kuhl, P. K. Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Dev* **75**, 1067-1084 (2004). https://doi.org/10.1111/j.1467-8624.2004.00726.x

64 Machač, P. & Skarnitzl, R. *Principles of Phonetic Segmentation*. (2013).

65 Mooshammer, C. *et al.* Bridging planning and execution: Temporal planning of syllables. *J Phon* **40**, 374-389 (2012). https://doi.org/10.1016/j.wocn.2012.02.002

66 Praat: doing phonetics by computer [Computer program]. v. Version 6.1.42 (2021).

67 Kawamoto, A. H., Liu, Q., Mura, K. & Sanchez, A. Articulatory preparation in the delayed naming task. *Journal of Memory and Language* **58**, 347-365 (2008). https://doi.org/https://doi.org/10.1016/j.jml.2007.06.002

68 Rastle, K., Davis, M., Marslen-Wilson, W. & Tyler, L. Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes - LANG COGNITIVE PROCESS* **15**, 507-537 (2000). https://doi.org/10.1080/01690960050119689

69 Jouen, A. L., Lancheros, M. & Laganaro, M. Microstate ERP Analyses to Pinpoint the Articulatory Onset in Speech Production. *Brain Topogr* **34**, 29-40 (2021). https://doi.org/10.1007/s10548-020-00803-3