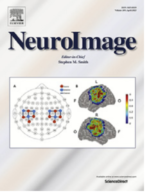
Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



The nested hierarchy of overt, mouthed, and imagined speech activity evident in intracranial recordings

Pedram Z. Soroush[a], Christian Herff[b], Stephanie K. Ries[c], Jerry J. Shih[d], Tanja Schultz[e], Dean J. Krusienski[a,*]

[a] *Virginia Commonwealth University, Richmond, VA, USA*
[b] *Maastricht University, Maastricht, Netherlands*
[c] *San Diego State University, San Diego, CA, USA*
[d] *UCSD Health, San Diego, CA, USA*
[e] *University of Bremen, Bremen, Germany*

ARTICLE INFO

Keywords:
Stereotactic electroencephalography (sEEG)
Brain-computer interface (BCI)
Speech decoding
Imagined speech
Speech activity detection

ABSTRACT

Recent studies have demonstrated that it is possible to decode and synthesize various aspects of acoustic speech directly from intracranial measurements of electrophysiological brain activity. In order to continue progressing toward the development of a practical speech neuroprosthesis for the individuals with speech impairments, better understanding and modeling of imagined speech processes are required. The present study uses intracranial brain recordings from participants that performed a speaking task with trials consisting of overt, mouthed, and imagined speech modes, representing various degrees of decreasing behavioral output. Speech activity detection models are constructed using spatial, spectral, and temporal brain activity features, and the features and model performances are characterized and compared across the three degrees of behavioral output. The results indicate the existence of a hierarchy in which the relevant channels for the lower behavioral output modes form nested subsets of the relevant channels from the higher behavioral output modes. This provides important insights for the elusive goal of developing more effective imagined speech decoding models with respect to the better-established overt speech decoding counterparts.

1. Introduction

Speech is the first and foremost means of human communication. Millions of people worldwide suffer from severe speech disorders due to neurological diseases such as amyotrophic lateral sclerosis (ALS), brain stem stroke, and severe paralysis. A speech neuroprosthesis that decodes speech directly from neural signals could dramatically improve life for these individuals. Intracranial brain-computer interfaces (BCIs) using electrocorticography (ECoG) (Angrick et al., 2019b; 2021; Herff et al., 2019; 2015; Moses et al., 2019; Soroush et al., 2021; Soroush and Shamsollahi, 2018; Sun et al., 2020) or stereotactic electroencephalography (sEEG) (Angrick et al., 2021; Herff et al., 2020; Li et al., 2022; Petrosyan et al., 2022; Soroush et al., 2021; Vadera et al., 2013) have demonstrated success in decoding aspects of speech directly from brain activity. These techniques have superior spatial resolution and bandwidth compared to non-invasive scalp electroencephalography (EEG), and superior temporal resolution compared with functional Magnetic Resonance Imaging (fMRI) (Ball et al., 2009; Brumberg et al., 2016; Schalk and Leuthardt, 2011). sEEG has recently gained wide clinical acceptance for epilepsy surgery planning as it has been found to lead to fewer surgical complications compared to ECoG (Herff et al., 2020; Iida and Otsubo, 2017). Additionally, while ECoG electrodes record localized activity of the cortical surface, sEEG electrodes generally have a much broader spatial distribution, providing access to brain regions including cortex, deeper structures, and both white and grey matter (Li et al., 2021; Revell et al., 2021; Soroush et al., 2021; 2022).

For those who have completely lost the ability to speak, the objective is to synthesize acoustic speech directly from brain activity during *imagined* speech. However, the lack of acoustic or behavioral output during imagined speech presents challenges in designing an effective decoding model (Angrick et al., 2021; Brumberg et al., 2016; Cooney et al., 2018; Perrone-Bertolotti et al., 2014). To cope with this limitation, it is common to utilize behavioral output from *overt* speech or *mouthed* speech (i.e., performing inaudible speaking articulations without vocalization) as a surrogate to study associated brain activity (Angrick et al., 2019a; 2019b; Herff et al., 2019; 2015; Ibayashi et al., 2018; Livezey et al., 2019; Mugler et al., 2014; Ramsey et al., 2018) or to train de-

[*] Corresponding author.
E-mail address: djkrusienski@vcu.edu (D.J. Krusienski).

https://doi.org/10.1016/j.neuroimage.2023.119913.
Received 4 August 2022; Received in revised form 5 January 2023; Accepted 29 January 2023
Available online 31 January 2023.

Table 1

Demographic information of participants and numbers of sEEG channels and number of unique trial sentences performed for each participant. The first and second columns list the gender and age of the participants, respectively. The third column reports the total number of channels recorded during the experiment. The fourth column reports the total number of electrodes excluded from the analysis due to noise or other anomalies, and the fifth column reports the *p*-value of the acoustic contamination index (Roussel et al., 2020). The last column lists the number of unique trial sentences performed.

Participant	Gender	Age	# Recorded	# Excluded (artifacts)	# Acoustic Contamination *p*-value (*P*)	# Trials
P1	Male	25	90	0	0.34	50
P2	Male	60	70	0	0.46	50
P3	Male	32	80	0	0.18	50
P4	Female	42	175	4	0.12	50
P5	Male	21	232	7	0.76	50
P6	Male	22	94	0	0.62	50
P7	Male	31	108	3	0.50	50

coding models for imagined speech applications (Angrick et al., 2021; Anumanchipalli et al., 2019; Martin et al., 2014).

Numerous prior studies have focused on establishing neural speech decoding model performance in various scenarios such as decoding phonemes and words (Martin et al., 2016; Mugler et al., 2014; 2018), brain-to-text (Herff et al., 2015; Makin et al., 2020; Moses et al., 2019; Sun et al., 2020; Willett et al., 2021), and direct speech synthesis from brain activity (Angrick et al., 2019b; 2021; Anumanchipalli et al., 2019; Herff et al., 2019). While several studies have compared brain activity from fMRI or Magnetoencephalography (MEG) across overt, mouthed, and imagined speech modes (Hickok et al., 2003; Okada et al., 2018; Tian and Poeppel, 2013; Tian et al., 2016; Zhang et al., 2020), there has yet to be a systematic comparison of the electrophysiological activity across these speech modes in the context of speech decoding. Rather than employing sophisticated models that attempt to decode higher-level representations of speech and potentially introduce other confounding factors for a comparative analysis of speech modes, a simplified speech activity detection framework can be utilized to better facilitate this comparison based on a lower-level speech representation (Kanas et al., 2014a; 2014b; Koct et al., 2019). Using causal brain activity features as inputs, speech activity detection models simply classify the respective activity as occurring during intervals of *speech* or *non-speech* intent, whether overt or imagined. Recent studies using sEEG have successfully elucidated the relative contributions of spectral features from grey and white matter for speech activity detection (Soroush et al., 2021; 2022), classified phonetic features from activity located in the superior temporal gyrus (Meng et al., 2021), and provided a preliminary demonstration of real-time synthesis of imagined speech activity (Angrick et al., 2021).

The present study utilizes sEEG recordings, spanning cortical and subcortical areas, to compare brain activity during overt, mouthed, and imagined speech with the objective of elucidating the efficacy of various sEEG features and speech surrogates for informing imagined-speech decoding models. Multiple speech activity detection decoding models are developed and applied within and across overt, mouthed, and imagined speech modes to reveal the similarities and differences in the relevant spatial, temporal, and spectral features among the models. The results indicate that relevant channels for speech decoding reside in both cortical and subcortical areas and appear to form a hierarchy in which the relevant channels for the lower behavioral output modes form nested subsets of the relevant channels from the higher behavioral output modes.

2. Methodology

2.1. Participants and electrode locations

sEEG data were collected from 7 native English-speaking participants being monitored as part of treatment for intractable epilepsy at UCSD Health. The demographic information of the participants is provided in Table 1. The study design was approved by the Institutional Review Boards of Virginia Commonwealth University and UCSD Health, and informed consent was obtained for experimentation with human subjects. The locations of sEEG electrodes were determined solely based on the participants' clinical needs. A subset of the implanted electrodes for each participant was determined to be in or adjacent to brain regions associated with speech and language processing. Figure 1 shows the depth electrode locations for the 7 participants, with sEEG electrode (channel) counts provided in Table 1. Anatomical location of the channels, including brain region and localization in white or grey matter, were identified using the FreeSurfer software package and MNE-Python (Fischl, 2012; Rockhill et al., 2022).

2.2. Experimental design

For the experiment, participants were presented with a sentence displayed on a computer monitor and simultaneously narrated via computer speakers. When visually prompted with an icon as shown in Fig. 2a, the participant was instructed to speak the sentence audibly while the acoustic speech and sEEG signals were simultaneously recorded. The participant was subsequently visually prompted with icons indicating to inaudibly articulate speech (i.e., mouth) and imagine speaking without articulating or vocalizing, respectively, for the same sentences. Herein, these three modes are referred to as *overt, mouthed*, and *imagined*, respectively. Each icon prompt and participant response during the task is referred to as a single *trial*.

The participant was asked to perform the associated task immediately upon presentation of the icon within a 4-second interval. This structure was repeated for 50 unique Harvard sentences, which are phonetically-balanced based on conversational English (Rothauser, 1969).

The three tasks were intentionally presented in a consistent sequence according to degree of behavioral output (i.e., overt-mouthed-imagined) rather than block randomized to better facilitate compliance from the patients. It is believed that the behavioral output of the speaking trial better primes the participant for more reliably performing the subsequent mouthed and imagined trials. Moreover, randomizing the trials would require greater attentional resources and more likely lead to response errors and oddball effects (Bénar et al., 2007), particularly with this patient group who are sometimes attentionally compromised due to their condition and the stress of the hospital environment.

The stimuli were presented and synchronized with the sEEG recordings using Presentationë software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, https://www.neurobs.com). The experimental setup and trial sequence structure are depicted in Fig. 2.

2.3. Data collection

The sEEG electrodes (Ad-Tech Medical Instrument Corporation) were referenced to a pair of subdermal needle electrodes in the scalp and digitized at 1,024 Hz (Natus Quantum Amplifier, Natus Medical

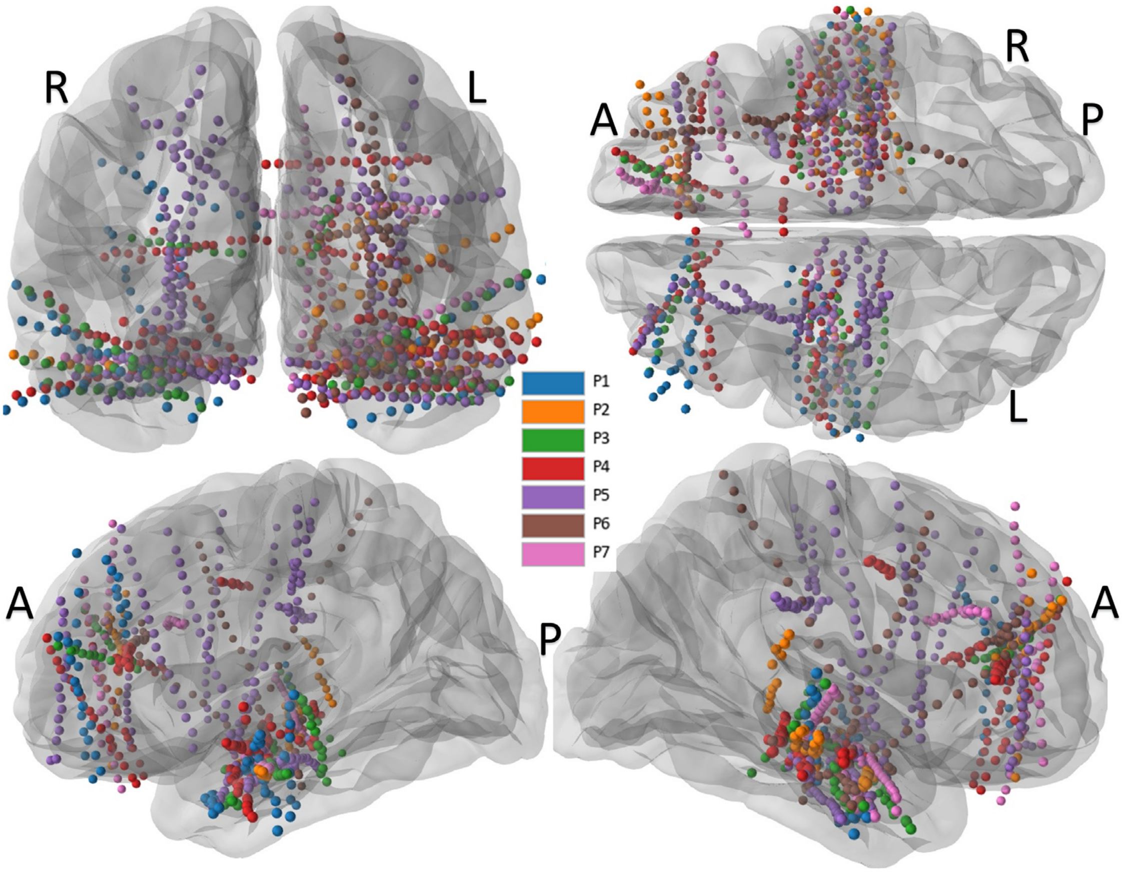
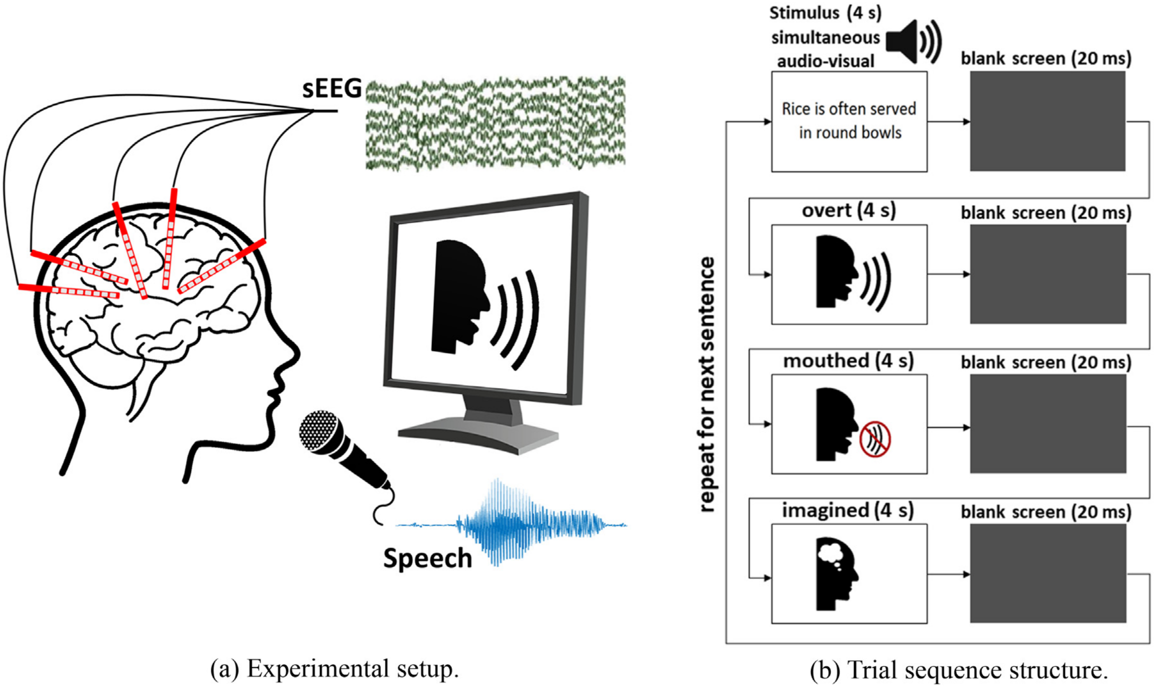


Fig. 1. The combined sEEG depth electrode (channel) locations of the 7 participants from different perspectives using an averaged brain model. A, P, R, and L indicate Anterior, Posterior, Right, and Left sides of brain, respectively.



(a) Experimental setup.

(b) Trial sequence structure.

Fig. 2. (a) sEEG and acoustic speech were simultaneously recorded as the participant performed the task as prompted by icons presented on a monitor. (b) Participants were presented with a sentence displayed on a monitor and simultaneously narrated through speakers. Participants were instructed to speak (overt), mouth, and imagine the sentence in sequence as cued by the respective icons. This was repeated for a bank of 50 sentences.

Inc.). The audio signal, recorded via an external microphone, was digitized at 44,100 Hz. The data from the audible speech portions of the task were used to extract *speech* and *non-speech* segments from the audio recordings.

2.4. Labeling the audio files (Speech vs. Non-speech)

The recorded speech from the overt mode was manually transcribed using the Wavesurfer software package (Sjölander and Beskow, 2000) for a separate analysis, but was found useful to provide precise labeling of the *speech* and *non-speech* segments for the present study. This was accomplished by shifting a 10 ms non-overlapping frame across the audio recording to identify the onset and offset of the spoken sentence, with the resulting timings from the transcription word boundaries being used as the frame label. Each frame was identified as *speech* if at least half of the frame length overlapped with a transcribed word, and as *non-speech* otherwise. For each 4-second interval encompassing the entire sentence utterance, the entire duration between the first onset and last offset was labeled as *speech* and the periods before and after these were labeled as *non-speech*. The frame length was chosen to be 10 ms to better represent brain signals' non-stationary nature and the fast changes of speech activity for eventual closed-loop implementation.

2.5. Labeling the audio-less modes (Mouthed and Imagined)

Due to non-existent speech audio for the mouthed and imagined modes, the average onset timings and durations of the overt speech intervals and the audio data from the corresponding overt mode were used to define respective surrogate *speech* and *non-speech* labels for the mouthed and imagined speech modes (Pei et al., 2011b). For each sentence and mode, the onset and offset of speech activity were estimated based on the average onset and offset timings of the overt mode of the corresponding participant, while the duration of speech interval was determined according to the respective overt audio. The start of each mode trial (i.e., time t_0) to time t_1 was labeled as *non-speech*, from time t_1 to t_2 was labeled as *speech*, and time t_2 to the end of the trial was labeled as *non-speech*.

For each participant and sentence, the average latency between presentation of the vocalization icon and the onset of actual vocalization was computed (t_a). This average latency was used as the transition from *non-speech* to *speech* ($t_1 = t_0 + t_a$). To set the transition from *speech* to *non-speech* (t_2), the duration of the corresponding sentence vocalization (t_s) was used ($t_2 = t_1 + t_s$). The interval from t_2 to the end of the trial was labeled as *non-speech*.

2.6. Data pre-processing

All sEEG data were visually inspected for noisy or anomalous channels for exclusion from the analysis, as reported in Table 1. Additionally, 25 sentence trials from Participant 1 were excluded due to a data mislabeling issue in the recording software. The number of trial sentences for each participant is provided in Table 1. The sEEG data were also analyzed for potential spatio-temporal correlations with the sound produced by the participants or present in the environment, and it was verified that the recordings were not subject to acoustic contamination (Roussel et al., 2020). The p-value of the acoustic contamination index (Roussel et al., 2020) are provided in Table 1. The resulting raw sEEG channels for inclusion in the analysis were re-referenced using the Laplacian method (Li et al., 2018; Mercier et al., 2017).

2.7. Feature extraction

The narrow-band power of each sEEG channel was computed in four conventional frequency bands: theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and broadband gamma (70–170 Hz). In a prior study performed with a subset of the present participants, these frequency bands exhibited better performance for speech activity detection compared to the delta (0–3 Hz) and low-gamma (30–70 Hz) bands (Soroush et al., 2021). The selected frequency bands have also been shown to be informative for other speech-related tasks (Kanas et al., 2014a; Li et al., 2020; Proix et al., 2022).

To extract the features, using the labeled 10-ms frames from the audio signals, the sEEG channels over a specified temporal window around each audio frame were zero-phase filtered over the respective frequency range using a sixth-order Butterworth filter. The window length was chosen to be 210 ms (corresponding to 200 ms before the frame to the end of the frame), based on the expected duration of speech planning revealed by intracranial recordings (Brumberg et al., 2016). However, this is insufficient for reliably estimating the energy of the lower frequency bands as at least 3–4 cycles are needed to convey meaningful information in a particular band. Hence, for theta, alpha, and beta bands, the duration of four cycles of the lowest frequency of the band was used to determine the causal model's window onset, and the window offset was always fixed at the 10-ms frame length. For example, for the 4–8 Hz theta band, the window onset is 1 s (4 cycles × 0.25 s/cycle) before the start of the frame, giving a 1.01 s window length.

An additional 118–122 Hz notch filter was applied to broadband gamma to suppress the second harmonic of the 60 Hz line noise. Finally, the features were computed every 10 ms as the natural logarithm of the signal energy over 210 ms, representing 10 ms overlapping the audio frame and 200 ms prior to the frame to emulate a causal design. Such a causal design aims to decode activity related to speech production rather than perception and can be implemented for real-time feedback for future closed-loop applications.

The features from each included channel were concatenated to form the feature vector (# channels × 21 features × # frequency bands - representing spatial, temporal, and spectral neural signal features, respectively) for the decoding models. A diagram of the feature extraction process is provided in Fig. 3.

2.8. Model training and evaluation

All significance tests were performed using a Benjamini-Hochberg corrected Wilcoxon signed-rank test (Benjamini and Hochberg, 1995; Wilcoxon, 1992). The resulting p-value (p), effect sizes (r), and sample sizes (n) are reported for each respective test (Rosenthal et al., 1994). Throughout the paper, n is either equal to 10 (when the values over the 10-fold cross-validation process were compared for two or more distributions) or 50 (when the values over the 50 trials were compared for two or more distributions).

2.8.1. Logistic regression model

All models are designed using logistic regression with L1 regularization and are specific for each participant and mode. A proximal AdaGrad optimizer with SoftMax function was selected for training the model (Duchi et al., 2011). This model was chosen over more complex machine learning and deep learning models because of the small size of data available (due to the differences between brain coverage of different participants, data from individual participants is processed separately), and that it has been shown to be effective for evaluating individual features and providing a convenient interpretation of the feature contributions (Soroush et al., 2021).

The performance of all models was obtained based on a 10-fold non-shuffled cross-validation analysis process, where each fold contained approximately one tenth of the data. For participants with 50 trials, the trials were randomly arranged into ten folds with each fold containing five trials. For the participant with 25 trials, the trials were randomly arranged into five folds with two trials and five folds with three trials. During the cross-validation process, each fold was used once as test data, and approximately one tenth of the remaining trials (2 and 4 trials in participants with 25 and 50 trials, respectively) were randomly selected as validation data. Trials not selected for validation or testing were used

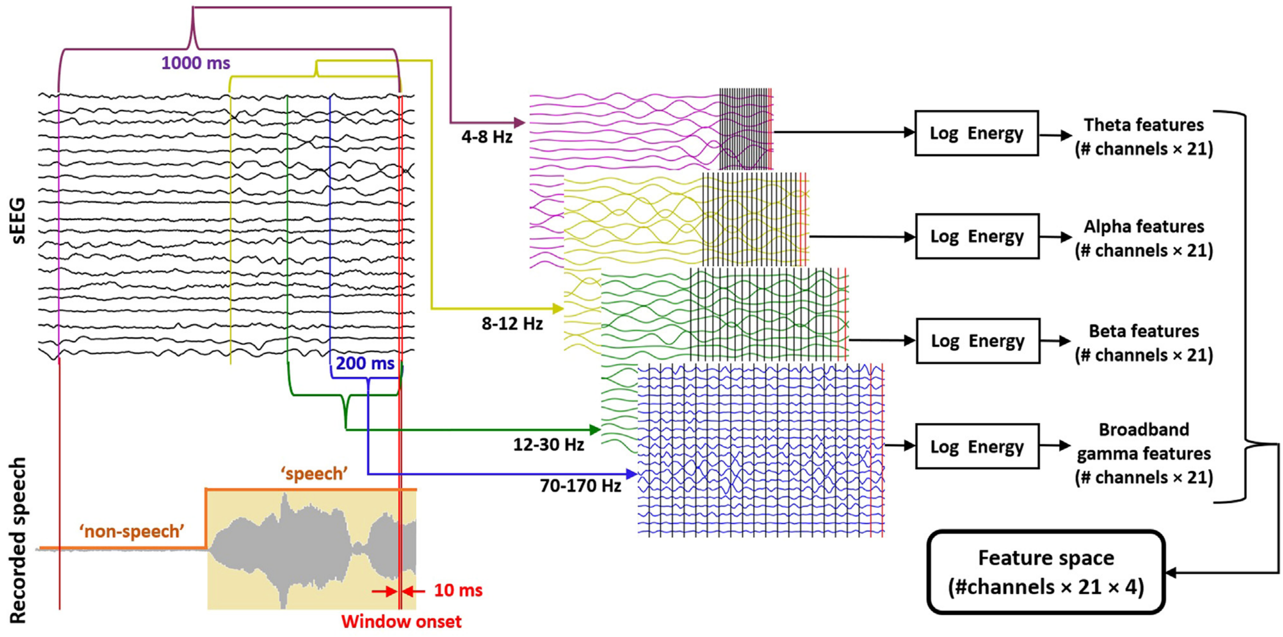

1000 ms
sEEG
200 ms
Recorded speech
'non-speech'
'speech'
10 ms
Window onset
4-8 Hz
8-12 Hz
12-30 Hz
70-170 Hz
Log Energy
Log Energy
Log Energy
Log Energy
Theta features (# channels × 21)
Alpha features (# channels × 21)
Beta features (# channels × 21)
Broadband gamma features (# channels × 21)
Feature space (#channels × 21 × 4)

Fig. 3. Extraction of theta, alpha, beta, and broadband gamma features from 200 ms prior to the audio frame to the frame's offset, resulting in a feature space of # channels × 21 temporal lags × 4 frequency bands. The recorded speech is time-aligned with the sEEG and indicator of the *speech/non-speech* labeling is shown. The filtered bands are presented in different time scales, with the vertical bars indicating the 21 10-ms temporal frames for each band.

as training data. To prevent training bias, the training data were normalized to zero mean and unit variance, and the same normalization parameters were applied to the validation and test data. The validation data were used to optimize the hyperparameters of the training models, while the test data were solely used to obtain the performance of the trained models for each fold.

Shuffling was not performed on any test or validation data. However, the training data were shuffled to better facilitate training of the model. While the feature extraction process was developed to emulate a causal design for the classification models, due to the overlap between the neural features associated with consecutive audio frames (i.e., neural features associated with each audio frame overlap with the previous and subsequent twenty audio frames), randomly assigning audio frames to different folds could potentially cause the train and test folds to contain trials with overlapping neural features, resulting in a leak of information from test data to training data. Hence, for the cross-validation process, only the training data (partitioned from the independent validation and test data) were shuffled.

Due to the difference between the amount of data for each class in some of the models, and to have consistency, the performance of all models was evaluated using balanced accuracy. For all models, the balanced accuracy was evaluated as the average of the recalls of the classes, which ranges from 0 (i.e., worst possible performance) to 1 (i.e., best possible performance). To establish the chance-level classification, a randomization test was performed where all labels were randomly shuffled and the 10-fold cross-validation process was repeated for 1000 separate randomizations of the labels.

2.8.2. Single-channel models and channel selection

Single-channel Within-Mode (WM) and Cross-Mode (CM) decoding models were created to compare the spatial, temporal, and spectral representations of the speech-related activity at each channel with respect to decoding performance within and across modes. For the WM models, the decoding performance for each mode was evaluated using 10-fold non-shuffled cross-validation models. Additionally, the feature weights of the 10 decoding models (one per fold) were averaged to compare the relative contribution of each feature to the models. These weights can provide a convenient interpretation of individual feature contributions based on the non-zero classifier weights and represent the spectral and temporal contributions to speech activity detection (Soroush et al., 2021; 2022).

The single-channel WM models were subsequently used to select channels with relatively superior performance in comparison with the rest of the channels. For each mode and fold in the cross-validation process, the mean plus one standard deviation of the balanced accuracy of all channels was determined as the threshold for the fold to form a distribution of thresholds over the ten fold of the cross-validation process. Additionally, for each channel, the distribution of balanced accuracies over the 10 folds of the cross-validation process was computed.

Single-channel CM models were trained on the data from one mode (train mode) and respectively tested on the other two modes (test mode). The hyperparameters of the decoding model were selected based on the single-channel WM model of the train-mode. The purpose of these models is to identify channels that exhibit similar relevant neural features across various modes versus channels that have dominant features that are unique to specific modes.

2.8.3. Multi-channel models

Multi-channel models were created and evaluated to explore the relative contributions of the spectro-temporal features of the channels in a combined model, as well as to compare the relative performance of these more potent models. Using the features of all channels, for each participant and mode, multi-channel WM models were evaluated using 10-fold non-shuffled cross-validation process. Multi-channel CM models were also created for which a decoding model was trained on the entire data of the train-mode and tested on the entire data from a different test-mode, herein labeled as (train mode)-to-(test mode). The hyperparameters of the decoding model were selected based on the multi-channel WM model of the train-mode.

3. Results

The three modes can be compared with respect to degree of behavioral output, with overt having the highest, mouthed having intermediate, and imagined having no behavioral output. In the subsequent paired comparisons, the mode in the pair having the higher behavioral output will be denoted as the ***HBO mode*** and the mode with lower behavioral output will be denoted as the ***LBO mode***.

3.1. Single-channel models

Figure 4 illustrates violin plots of the distributions of averaged balanced accuracy of the 10-fold cross-validation single-channel WM models trained on the data from each participant and each mode. To compare to chance-level performance, permutation tests were performed by randomly shuffling the labels and performing the 10-fold cross-validation process 1000 times. The black dots represent the channels that performed significantly above the thresholds ($p < 0.01,\ r > 0.5,\ n = 10$) and were thus selected as mode-relevant channels. For all participants and modes, the thresholds were significantly above the chance-level ($p < 0.01,\ r > 0.8,\ n = 10$). Additionally, over all participants, channels that performed significantly better than chance-level ($p < 0.05,\ r > 0.5,\ n = 10$) in all three modes performed significantly better in overt compared to mouthed and imagined modes, and in mouthed compared to imagined ($p < 0.01,\ r > 0.5,\ n = 10$).

3.1.1. Spatial characterization

Figure 5 shows the individualized averaged balanced accuracy of the cross-validated single-channel WM models on the data from a representative participant (Participant 1) for all three modes. This is equivalent to an expansion of the violin plots from Participant 1 in Fig. 4. Since chance-level was a subset of and nearly identical to chance-level for all modes in Fig. 4, the average of the three modes for each channel over the 10 folds of cross-validation is used for reference. Fig. 5 shows that channels along each shaft, if relevant for the mode and significantly above the chance-level ($p < 0.05,\ r > 0.5,\ n = 10$), generally follow a similar trend across the modes, indicating similar activity across these brain regions.

By comparing each mode pair, channels were observed with performance significantly better than chance-level in both modes ($p < 0.05,\ r > 0.5,\ n = 10$), while having a significantly better performance in the HBO mode than the LBO mode ($p < 0.05,\ r > 0.5,\ n = 10$). For instance, channels in the superior temporal gyrus of both hemispheres (channels 5–10 on the blue shaft and 8–10 on the light red shaft) show roughly similar above-chance performance for both mouthed and imagined, and significantly better performance for overt ($p < 0.05,\ r > 0.5,\ n = 10$). However, there were also brain regions where single-channel models exhibited significantly better performance in the HBO mode than in the LBO mode ($p < 0.05,\ r > 0.5,\ n = 10$), while showing no significant difference from chance-level performance in the LBO mode ($p > 0.05,\ r < 0.1,\ n = 10$). This is observed for regions in the left middle temporal gyrus (channels 9 and 10 on the cyan shaft) and left inferior temporal gyrus (channels 4, 5, 9, and 10 on the light green shaft), where single-channel models performed significantly better for overt than mouthed or imagined ($p < 0.05,\ r > 0.7,\ n = 10$), while showing no significant difference from chance-level performance for mouthed and imagined ($p > 0.05,\ r < 0.1,\ n = 10$). Despite a few channels exhibiting slightly higher average performances for imagined than overt and mouthed (i.e., channels 3, 4, and 5 on the red shaft), these are not statistically significant and no channels exhibited significantly better performance in mouthed or imagined than overt.

The averaged balanced accuracy values of the single-channel WM models, normalized for each participant over all three modes, are shown spatially in Fig. 6a. It is observed that most selected channels are clustered in groups of two or more adjacent channels along the same shaft, suggesting these channel groups each reside in a common functional region.

The overt panel of Fig. 6a shows that relevant channels are predominantly located around the border between the parietal and temporal lobes, including the middle frontal gyrus, superior temporal gyrus and sulcus, and Sylvian parieto-temporal regions. Additionally, some relevant channels were localized to regions adjacent to the auditory cortex. However, these channels were also selected in one or both of mouthed or imagined, which did not involve auditory feedback or perception. This is consistent with previous studies showing activation in the auditory cortices during imagined speech or imagined hearing, regardless of the presence of an auditory stimulus (Martin et al., 2016; Orpella et al., 2022; Pei et al., 2011b; Rampinini et al., 2017; Tian and Poeppel, 2010; 2012; Zhang et al., 2020).

The imagined panel of Fig. 6a also indicates that there are brain regions, such as parts of the left frontal lobe, with channels relevant to imagined that also exhibited activity relevant to overt and mouthed. However, these channels generally exhibited lower performance compared to the top performing channels of each of these modes. A similar result is observed between overt and mouthed when comparing brain regions, such as the right frontal lobe across modes.

In contrast, it can be seen that numerous relevant channels for overt, primarily located in the right and left temporal lobes, were not selected for mouthed or imagined. This is also observed between mouthed and imagined. However, there are several channels located in the right and left temporal lobes, such as channels in the superior temporal gyrus of both hemispheres of Participant 1, that perform well in two or more modes.

To characterize the neural activity with respect to relative laminar depth, the radial distance between the closest selected channel to the center of the MNI brain model (Evans et al., 1993) and the furthest selected channel from the center was divided into ten uniformly-spaced levels, forming spherical shells. For each level, the average of the balanced accuracies of the channels located in that level was calculated. Fig. 6b shows the accuracy and number of electrodes corresponding to each relative depth level and mode. For each mode, the average chance-level classification, which were equivalent to chance-levels of Fig. 4, are indicated with magenta dashed lines. The overt results in Fig. 6 show that the best performing channels are comparatively few and near the cortical surface. However, there are a greater number of channels selected at multiple intermediate depths that also exhibit reasonable performance. These observations are also generally consistent for mouthed and imagined.

3.1.2. Comparison across modes

The results from the single-channel models for the WM and CM paradigms were used to assess and spatially visualize the shared relevance of channels across modes. Channels were organized in three groups according to shared relevance for the nested behavioral output hierarchy of (1) overt, (2) overt-mouthed, and (3) overt-mouthed-imagined. The channels uniquely relevant to overt were identified by corresponding single-channel models performing significantly above the chance-level for the overt WM paradigm and not significantly above chance-level for either overt-to-mouthed or overt-to-imagined. The channels relevant to both overt and mouthed were identified by corresponding single-channel models performing significantly above chance-level for both the overt and mouthed WM paradigms, both overt and mouthed CM paradigms, but not significantly above chance-level for either overt-to-imagined or mouthed-to-imagined. The channels relevant to all three modes were identified by corresponding single-channel models performing significantly above the chance-level for all three WM paradigms and all combinations of CM paradigms. Channels were selected based on the respective performance differences in the groups using a threshold of $p < 0.05$. It should be noted that this grouping process resulted in every channel selected in Fig. 4 being assigned to exactly one group.

Fig. 7a shows the brain regions of the three channel groups across all participants on an average brain model. The green channels are relevant to all three modes, whereas the orange channels are relevant to only overt and mouthed, and the blue channels are relevant to only overt. Channels selected in the overt group show relatively higher performance in the overt panel of Fig. 6a. These channels reside in a wide range of cortical and sub-cortical brain regions, including the superior and middle temporal gyrus and superior frontal gyrus, which is in line with previous studies (Arya et al., 2019; Kohler et al., 2021; Leuthardt et al., 2011; Martin et al., 2014; 2016; Okada et al., 2018; Orpella et al., 2022; Pei et al., 2011a; Zhang et al., 2020).

Channels selected in the overt-mouthed group show relatively higher performance in the mouthed panel of Fig. 6a. Fewer channels were selected in this group in comparison with the other two groups, which could be due to the narrow coverage of the motor cortices for these participants. The brain regions exhibiting relevant activity for this group included motor cortex and adjacent, right superior frontal gyrus, and right and left inferior temporal gyrus, which is in line with previous studies (Okada et al., 2018; Pulvermüller et al., 2006; Zhang et al., 2020). Channels selected in the overt-mouthed-imagined group show relatively higher performance in the imagined panel of Fig. 6a and are predominantly located on or near the Broca's area (e.g., left inferior and middle frontal gyrus) and the auditory cortices (e.g., middle and inferior temporal gyrus and sulcus and Sylvian parieto-temporal region), which is in line with previous studies (Geva et al., 2011; Kohler et al., 2021; Leuthardt et al., 2011; Martin et al., 2016; Okada et al., 2018; Orpella et al., 2022; Pei et al., 2011a; 2011b; Zhang et al., 2020).

The brain regions associated with all three groups span grey and white matter in both brain hemispheres. This is in line with recent studies that have reported speech-related activity in both hemispheres and both deeper and superficial brain areas during both overt and imagined (Alexandrou et al., 2017; Cogan et al., 2014; Geva et al., 2011; Kohler et al., 2021; Orpella et al., 2022; Soroush et al., 2021; Tourville et al., 2008).

It should be noted that channels in the overt group may also contain neural features shared with the other modes, but these features are relatively less dominant than those unique to overt, as the models trained on overt do not perform significantly above the chance-level on the mouthed or imagined data. This also applies to the overt-mouthed group, which may have neural features common in all three modes that are likewise relatively less dominant than those unique to overt and mouthed. Moreover, no channels were identified that were uniquely relevant to imagined, compared to mouthed or overt, or uniquely relevant to both imagined and mouthed, compared to overt. Thus the inverse nested hierarchy of (1) imagined, (2) imagined-mouthed, and (3) imagined-mouthed-overt was not observed. This is in line with prior work that did not identify any brain regions that were significantly more active for imagined compared to mouthed modes (Okada et al., 2018). This HBO-LBO hierarchy was also consistent when grouping channels according to their relevance for only mouthed and imagined modes, further confirming the nested behavioral output hierarchy of (1) overt, (2) overt-mouthed, and (3) overt-mouthed-imagined.

Figure 7b shows a Venn diagram of the nested hierarchical channel groups from Fig. 7a, with relative areas proportional to the number of channels in each group, indicated by the numeric values. To summarize, all 119 significant channels are relevant to overt, a subset of 90 are relevant to mouthed, and a nested subset of 78 are relevant to imagined. Figure 7c shows the average performance of WM paradigm of these channel groups, in a mutually-exclusive fashion (i.e., the colored channels comprise the respective groups, not the nested subsets). For each mutually-exclusive grouping, the results are compared across the three modes indicated by the legend using 10-fold non-shuffled cross-validation process for each respective mode. While channels in all groups performed significantly above chance-level ($p < 0.0001$, $r > 0.8$, $n = 10$), the channels in overt group performed significantly better in the overt mode than channels in overt or the other two groups in any of the three modes ($p < 0.01$, $r > 0.7$, $n = 10$). No significant difference was observed between the performance of the channels in the overt group in mouthed and imagined modes ($p > 0.05$, $r < 0.2$, $n = 10$). No significant difference was observed between the performance of the channels in the overt-mouthed group in overt and mouthed modes ($p > 0.05$, $r < 0.1$, $n = 10$). The performance of the channels in the overt-mouthed group in both overt and mouthed modes was significantly better than the same channels in the imagined mode, the channels in the overt group in mouthed and imagined modes, and the channels in the overt-mouthed-imagined group in overt, mouthed, and imagined modes ($p < 0.05$, $0.8 > r > 0.3$, $n = 10$). Lastly, no significant difference was observed between the performance of the channels in the overt-mouthed-imagined group in overt, mouthed, and imagined modes ($p > 0.05$, $r < 0.1$, $n = 10$), but each individually was significantly larger than the channels in the overt group in mouthed and imagined modes and the channels of the overt-mouthed group in the imagined mode ($p < 0.05$, $0.6 > r > 0.3$, $n = 10$).

To compare the relevance of channels within and across different modes, performance of each selected channel from Fig. 7a was compared across HBO-LBO mode pairs (i.e., overt-mouthed, overt-imagined, and mouthed-imagined). For each mode-pair and group, all channels selected in at least one mode in the pair based on the respective single-channel WM models were analyzed. Figure 8 shows a scatter plot of the mode-pair performance of each channel across all participants. Channels are marked according to mode-pair and colorized by nesting grouping. As a visualization aid, bivariate Gaussian distributions were fit to the channels of each nested group. The 90% probability contours of the respective distributions are shown as ellipses in Fig. 8. Chance-level classification, which were a subset of and nearly identical to the chance distributions in Fig. 4, are depicted by the magenta dashed lines for HBO and LBO modes, respectively. The average of 10-fold cross-validation performance of the individual channels was generated for each group and the Pearson correlation coefficient was calculated for the performance of each channel in the mode pairs. While a significantly positive correlation ($r = 0.71$ and $p < 0.0001$) was observed between the LBO and HBO axes of overt-mouthed-imagined group, no significant correlation was observed between the LBO and HBO axes of the other two groups ($p > 0.05$).

3.1.3. *Spectro-temporal characterization*

Figure 9 illustrates the absolute value of normalized feature weights of the single-channel WM models for each mode, averaged over the selected channels of the channel groups from Fig. 7a. While these weights do not directly represent the spectro-temporal cognitive patterns associated with the decoding models, they do convey the relative contributions of spectro-temporal features to the models. As expected from previous studies, features temporally closer to the frame being decoded have a greater contribution to the models (Soroush et al., 2021; 2022). A strikingly similar pattern is observed between the weights of the overt-mouthed-imagined group across the three modes. Such similarities are also observed for the overt-mouthed group in the overt and mouthed modes. While, as expected, broadband gamma was a prominent feature, it was observed that the lower frequency bands also provide important contributions to the models. This also supports previous studies that have shown alpha band to be promising for distinguishing movement from rest (Li et al., 2021) and *speech* from *non-speech* (Soroush et al., 2021; 2022).

3.2. *Multi-channel models: Within-mode and cross-mode*

Fig. 10 shows the averaged balanced accuracy in the multi-channel WM and CM models across participants. To indicate the significance of the classification results, permutation tests were performed by randomly shuffling the labels and performing the 10-fold cross-validation process 1000 times. Since the chance-level distributions were nearly identical for all models, a single distribution of all random permutation results of all models was generated, with a mean indicated by the magenta dashed line. The table in Fig. 10 indicates the significance level of comparison tests between the mode pairs for the multi-channel WM and CM models.

All multi-channel WM and CM models performed significantly better than chance-level ($p < 0.05$, $r > 0.5$, $n = 10$ and 50 for WM and CM models, respectively), except for the overt-to-mouthed and overt-to-imagined models of Participant 3. Overt-to-overt models performed significantly better than all other models ($p < 0.0001$, $r > 0.7$, $n_1 = 10$

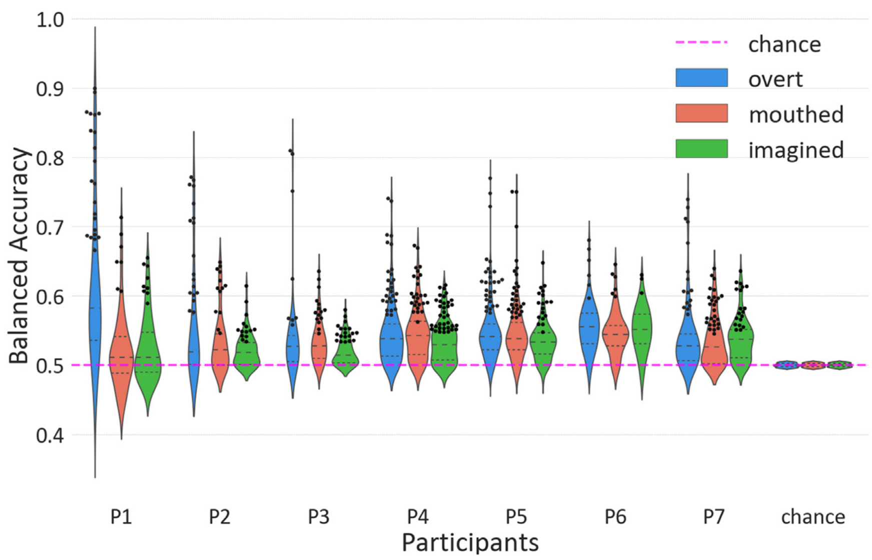


Fig. 4. Distributions of classification performance (averaged balanced accuracy over 10-fold non-shuffled cross-validation models) of decoding models for all channels in the single-channel WM models for each participant and mode. The chance distributions obtained by randomly permuting the class labels are shown for each speech mode, with the magenta dashed line indicating the average chance-level classification results over all participants, modes, and channels. The black dots represent the selected channels for each group. The dashed line within each violin indicate the median and the dotted lines indicate the first (Q1) and third (Q3) quartiles.

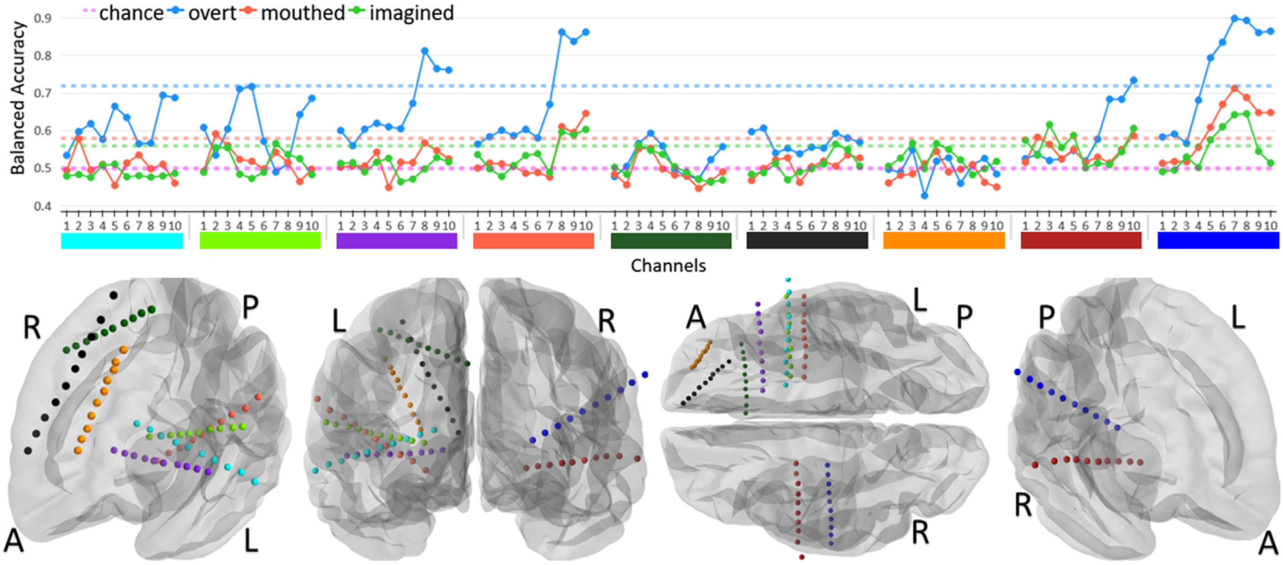


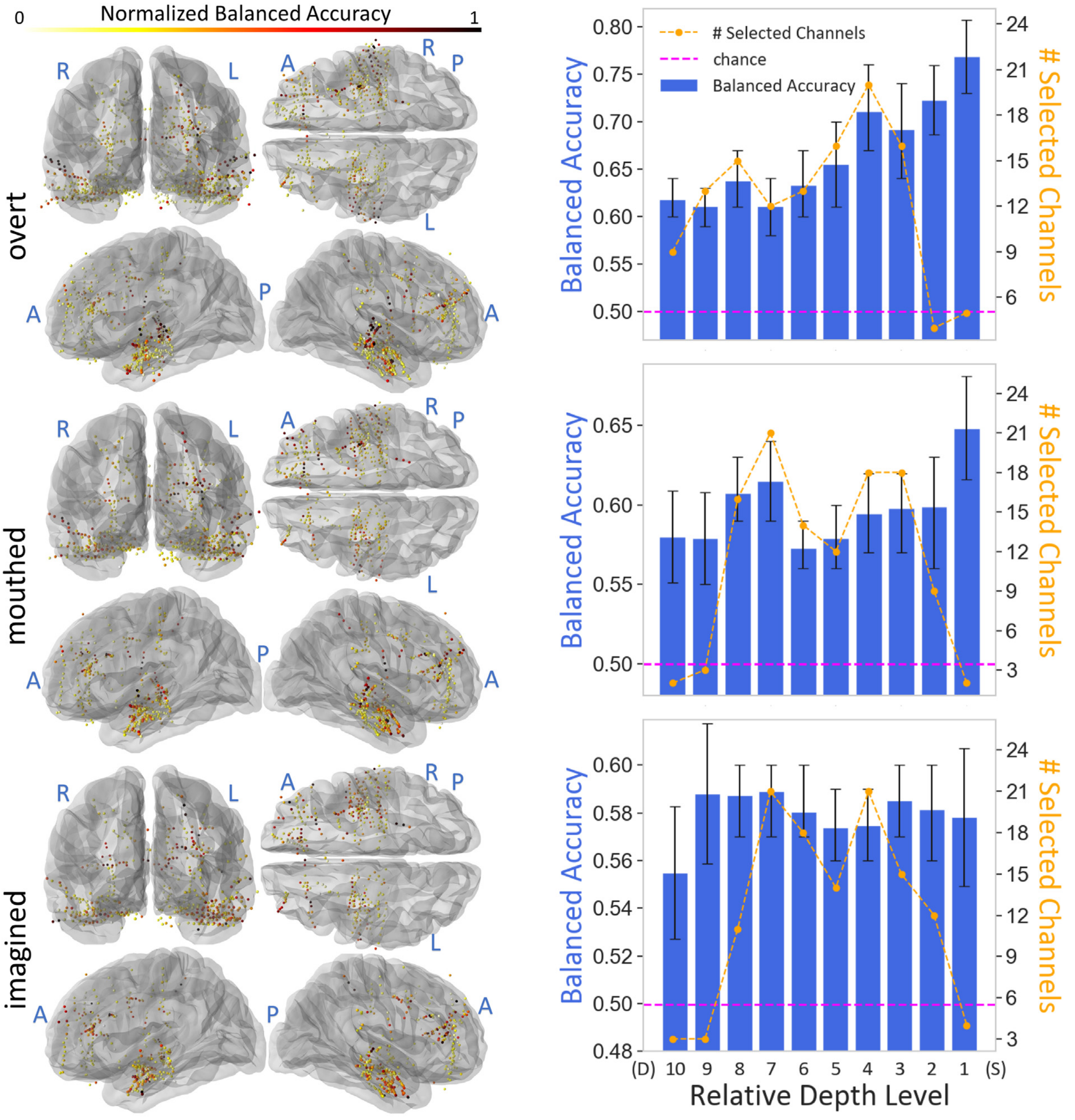
Fig. 5. Averaged balanced accuracy of single-channel WM decoding models for all channels for a representative participant (Participant 1). Channels are grouped by shaft, with 1 representing the deepest channel and 10 representing the most superficial channel. The blue, red, and green horizontal dashed lines show the selection thresholds for each mode, averaged across respective folds. The magenta dashed line represents the chance-level classification results of each channel, averaged over the 10 folds of the cross-validation process. The color-coded bars below each shaft plot correspond to the colored electrodes in the caudal and ventral views of both hemispheres and frontal views of left and right hemispheres at the bottom of the figure. A, P, R, and L indicate Anterior, Posterior, Right, and Left sides of brain, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and $n_2 = 10$ or 50), which may be attributed to the more precise labeling of the overt trials. Models trained on mouthed performed significantly better than all other models ($p < 0.05$, $r > 0.5$, $n_1 = 10$ or 50 and $n_2 = 10$ or 50) on both mouthed (mouthed-to-mouthed models) and overt (mouthed-to-overt models). However, no significant difference was observed between the performance of mouthed-to-mouthed and mouthed-to-overt models ($p > 0.05$, $r < 0.1$, $n = 10$, 10, and 50, respectively). Overt-to-mouthed models performed significantly better than overt-to-imagined models ($p < 0.001$, $r > 0.8$, $n = 50$).

The performance of the mouthed-to-imagined models was only significantly better than the overt-to-imagined models ($p < 0.01$, $r > 0.6$, $n = 50$), and the performances of both of these two models were significantly worse than all other multi-channel WM and CM models ($p < 0.05$, $r > 0.5$, $n_1 = 50$ and $n_2 = 10$ or 50). While no significant difference was observed between the performances of the imagined-to-overt, imagined-to-mouthed, and overt-to-mouthed models ($p > 0.05$, $r < 0.1$, $n = 50$), the performances of all of these three models were significantly worse than imagined-to-imagined models ($p < 0.05$, $r > 0.6$, $n_1 = 50$ and $n_2 = 10$).

Mouthed-to-overt and imagined-to-overt models performed significantly better than the overt-to-mouthed and overt-to-imagined models, respectively ($p < 0.01$, $r > 0.5$, $n = 50$). Imagined-to-mouthed models performed significantly better than the mouthed-to-imagined models ($p < 0.05$, $r > 0.5$, $n = 50$).

It should be noted that, for each participant and mode, the channels selected as capturing relevant neural features in Section 3.1 provided the largest contributions to the multi-channel models. The model weights were examined for each participant and mode, and it was observed that the channels selected in overt, overt and overt-mouthed, and overt-mouthed-imagined groups in Section 3.1.2, respectively, exhibited the largest contributions to the multi-channel models. In contrast, the other channels yielded minor or no contributions to the models (i.e., model weights near or equal to zero). For reference, without additional



(a)

(b)

Fig. 6. (a) Channels from the single-channel WM analysis having averaged balanced accuracy significantly above chance-level for all participants on an averaged brain model ($p < 0.05$, $r > 0.5$, $n = 10$). The electrodes are colorized based on the averaged balanced accuracy values, which were normalized to 0–1 for each participant over all three modes. A, P, R, and L indicate Anterior, Posterior, Right, and Left sides of brain, respectively. (b) Average balanced accuracy across participants and number of selected channels of the single-channel WM decoding models, grouped by relative electrode depth. Selected channels are grouped into ten uniform depth levels based on the center of the Montreal Neurological Institute (MNI) average brain model (Evans et al., 1993) from the deepest (D) to the most superficial (S) along the respective electrode shafts. The error bars indicate the 95% confidence intervals of balanced accuracy over channels. For each mode, the magenta dashed line indicates the average chance-level classification.

model optimization, the performances of the overt-to-overt models are comparable to a prior speech activity detection study using ECoG, where the models for a single participant ranged from 95.3–98.8% in detection accuracy (Kanas et al., 2014a).

3.3. *Speech activity detection proportions*

Figure 11 a shows the distributions of proportions of speech activity detection during each trial for each of multi-channel WM and CM models, compared to the actual proportion of speech based on the true or approximated labels over all participants. The mean of the detected speech proportions for all multi-channel WM models was significantly larger than the mean of the speech proportions based on the actual labels ($p < 0.001$, $r > 0.7$, $n = 50$). For the WM models, the proportions of detected speech for overt-to-overt was significantly smaller than both mouthed-to-mouthed and imagined-to-imagined ($p < 0.05$, $r > 0.4$, $n = 50$); however, no significant difference was found between mouthed-to-mouthed and imagined-to-imagined ($p > 0.05$, $r < 0.1$, $n = 50$).

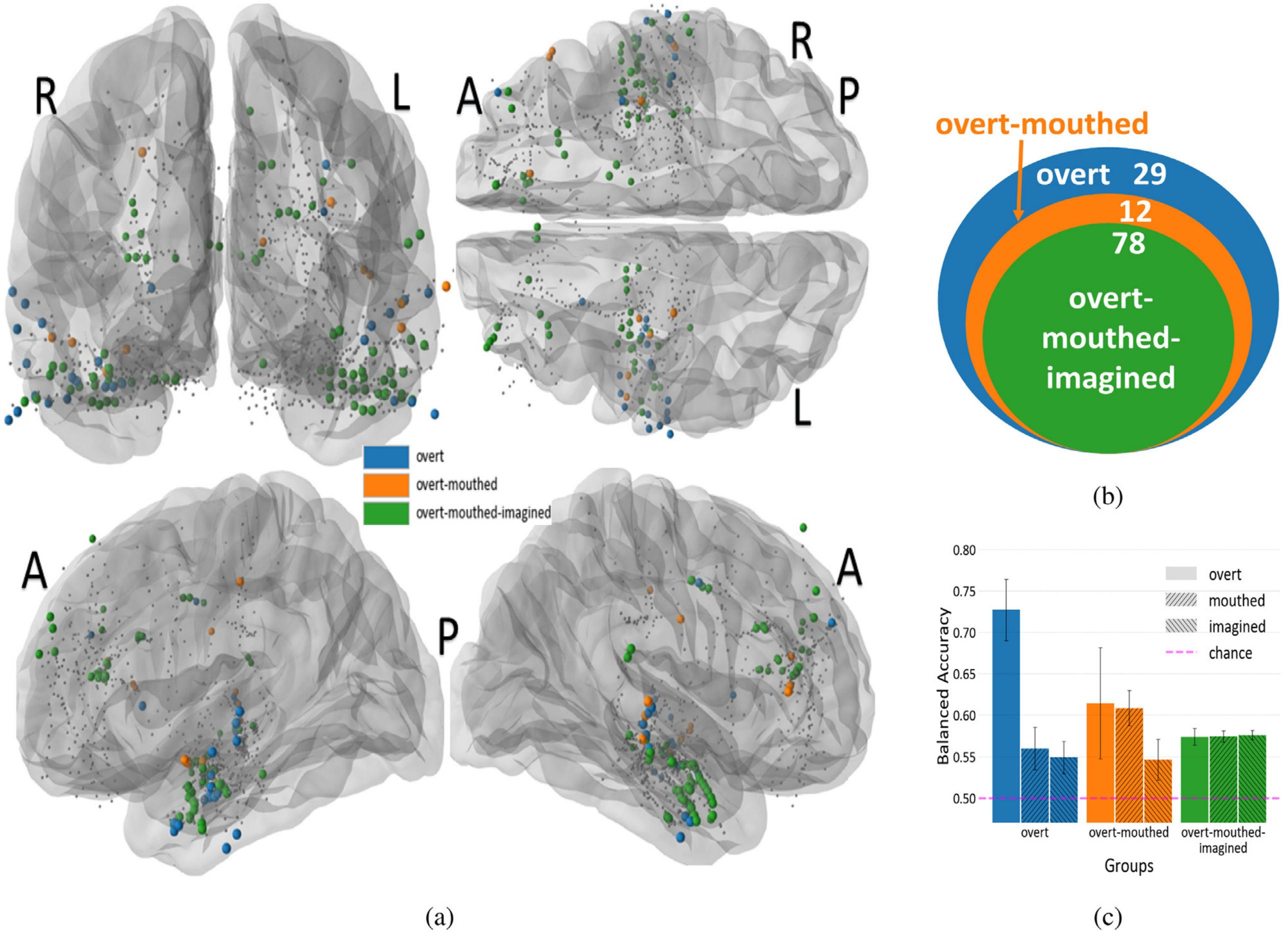


Fig. 7. Representations of the hierarchical channel nesting. (a) Channels for all participants on an averaged brain model, color-coded by mode relevance. A, P, R, and L indicate Anterior, Posterior, Right, and Left sides of brain, respectively. *Overt-mouthed-imagined* represents the subset of selected channels with common neural activity across all three modes. *Overt-mouthed* represents the mutually exclusive subset of channels with neural processes only relevant to overt and mouthed but not imagined. *Overt* represents the mutually exclusive subset of channels with neural processes only relevant to overt. Channels not satisfying the relevance criteria for any mode are shown as smaller grey points. (b) Venn diagram representing nested hierarchical mode groupings including the number of channels identified in each group, selected from more than 800 channels across all participants. (c) Average balanced accuracy across participants of the single-channel WM decoding models in hierarchical mode groupings. For each grouping, the results are compared across the three modes indicated by the legend using 10-fold non-shuffled cross-validation process for each respective mode. The error bars indicate the 95% confidence intervals. The magenta dashed line indicates average chance-level classification, which were a subset of and nearly identical to chance-level of Fig. 4.

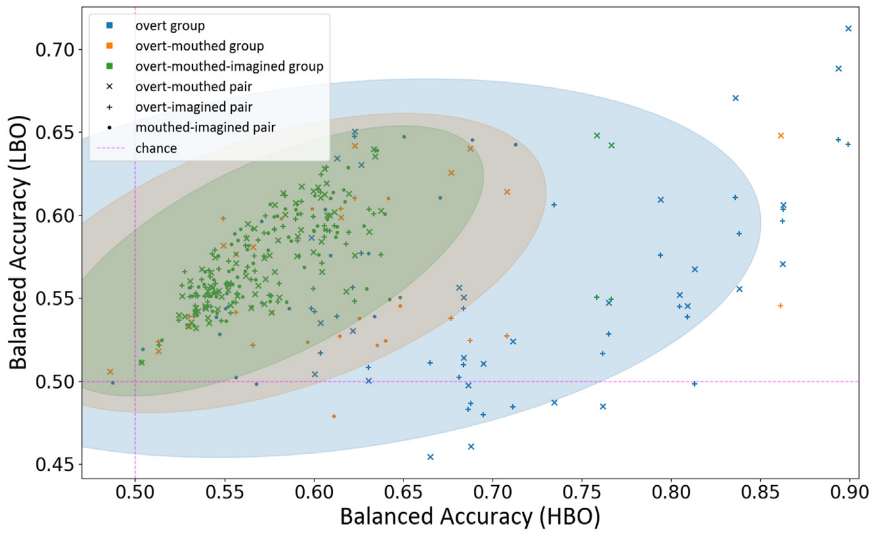


Fig. 8. Scatter plot of the mode-pair performance of each selected channel of each group from Fig. 7a across all participants. Channels are marked according to mode-pair and colorized by nesting grouping. As a visualization aid, bivariate Gaussian distributions were fit to the channels of each nested group, and the 90% probability contours are shown as ellipses. Each sample represents the averaged balanced accuracy over the 10-fold cross-validation of two modes of one channel, with the horizontal axis indicating HBO mode and the vertical axis indicating the LBO mode. The magenta dashed lines indicate the average chance-level classification for HBO and LBO modes, respectively.

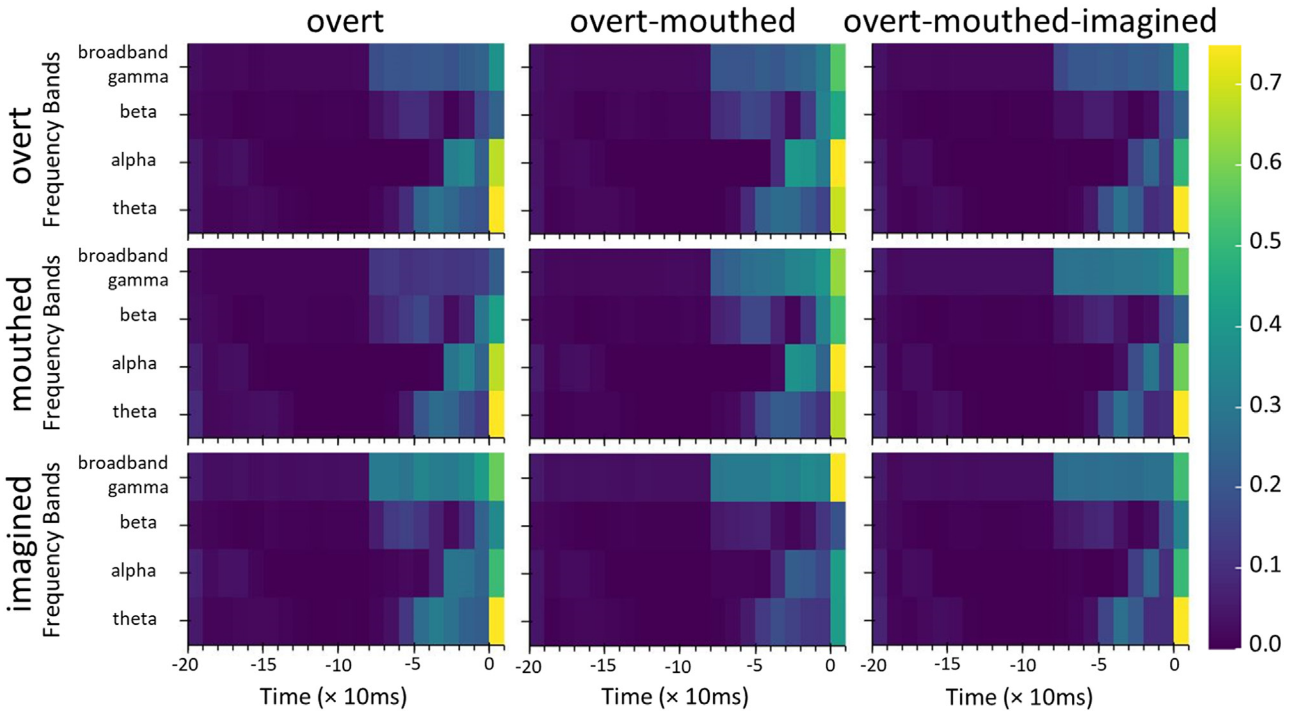


Fig. 9. Average of absolute value of normalized decoding model weights across 10-folds of the channel groups from Fig. 7a. Zero on the horizontal axis indicates the start of the audio frame.

Figure 11 b and c show the histograms of the distributions of all multi-channel WM models' speech-onset and speech-offset, respectively. The black vertical lines shown on the histograms indicate the actual speech-onset and speech-offset of the overt mode, and the blue, green, and red, triangles mark the means of the overt-to-overt, mouthed-to-mouthed, and imagined-to-imagined models' detection distributions, respectively. For all WM models, the detected speech windows ended significantly later than when the actual or estimated speech ended ($p < 0.0001$, $r > 0.7$, $n = 50$), while for all mouthed-to-mouthed and imagined-to-imagined WM models, the detected speech windows started significantly earlier than when the actual or estimated speech started ($p < 0.01$, $r > 0.6$, $n = 50$). For all overt-to-overt WM models, the detected speech windows started significantly later than the detected speech windows of mouthed-to-mouthed and imagined-to-imagined WM models ($p < 0.01$, $r > 0.6$, $n = 50$), while no significant difference was observed between these window starts and when the actual speech started ($p > 0.05$, $r < 0.1$, $n = 50$).

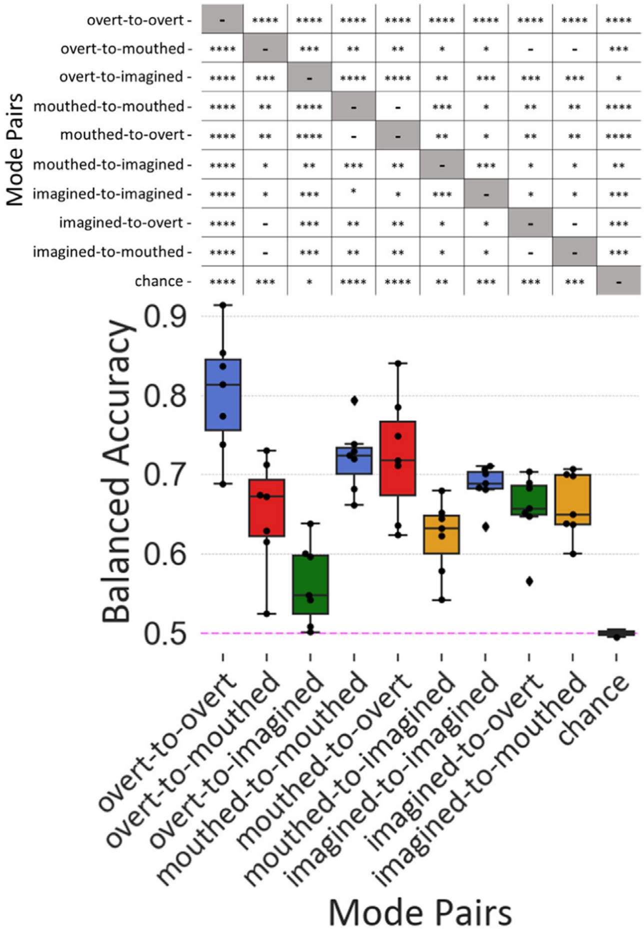
For all multi-channel CM models, except the overt-to-mouthed and overt-to-imagined models, the mean of the detected speech proportions was significantly larger than the mean of the speech proportions based on the actual or estimated labels ($p < 0.001$, $r > 0.7$, $n = 50$). For the overt-to-imagined models, the mean of the detected speech proportions was significantly lower than the mean of the speech proportions based on the estimated labels ($p < 0.001$, $r > 0.7$, $n = 50$). This can be related to the relatively lower performance of these two models as depicted in Fig. 10. No significant difference was observed between the mean of the detected speech proportions of overt-to-mouthed models and the mean of the speech proportions based on the estimated labels ($p > 0.05$, $r < 0.1$, $n = 50$).

4. Discussion

This study used sEEG data collected during overt, mouthed, and imagined speaking conditions to identify common neural features and relationships across these conditions using a speech activity detection paradigm. The relevant features were found to occur near speech-onset, across all frequency bands examined as shown in Fig. 9, which is in line with previous studies (Soroush et al., 2021; 2022) which analyzed data from a subset of the participants from the present study.

4.1. Nested behavioral hierarchy: Single channel models

Recent studies in neurolinguistics have offered evidence for the existence of a nested hierarchy in the brain activity associated with different speech modes, formed from highest behavioral output to lowest behavioral output (Cooney et al., 2018; Hickok et al., 2003; Li et al., 2020; MacKay, 1992; Oppenheim and Dell, 2010; Perrone-Bertolotti et al., 2014; Zhang et al., 2020). Facial micromovements during imagined speech, commonly assumed to be a byproduct of short-circuited motor signals, induced activity in language-associated brain areas (e.g., Broca's and Wernicke's areas) during both overt and imagined speech, and similar motor-to-sensory transformation (starting from frontal and continuing to parietal and temporal lobes) in both overt and imagined speech are among the evidence supporting this hypothesis (Bookheimer et al., 1995; Hickok et al., 2003; Huang et al., 2002; Orpella et al., 2022; Palmer et al., 2001; Perrone-Bertolotti et al., 2014; Tian and Poeppel, 2013; Tian et al., 2016; Zhang et al., 2020). It has also been posited that hierarchical forward predictions, generated by motor commands for comparison of auditory output and its consequences, occur during speech production tasks with and without audible output (i.e., overt, mouthed, and imagined) (Heinks-Maldonado et al., 2006; Hickok et al., 2011; Okada et al., 2018; Pickering and Garrod, 2013). Moreover, the presence of articulatory and acoustic information in motor and auditory cortices, respectively, during imagined (no motor or auditory output), mouthed (no auditory output), and overt modes further supports this hypothesis (Zhang et al., 2020). In an fMRI study of twenty-four participants silently articulating (i.e., mouthing) or imagining to speak a sequence, greater activations were observed in premotor cortex, insula, and auditory cortex during mouthed compared to imagined speech, suggesting forward predictions arise from additional levels of the perceptual/motor hierarchy that are involved in monitoring the intended speech output (Okada et al., 2018). It is hypothesized



overt-to-overt -	-	****	****	****	****	****	****	****	****	****
overt-to-mouthed -	****	-	***	**	**	*	*	-	-	***
overt-to-imagined -	****	***	-	****	****	**	***	***	***	*
mouthed-to-mouthed -	****	**	****	-	-	***	*	**	**	****
mouthed-to-overt -	****	**	****	-	-	**	*	**	**	****
mouthed-to-imagined -	****	*	**	***	**	-	***	*	*	**
imagined-to-imagined -	****	*	***	*	*	***	-	*	*	***
imagined-to-overt -	****	-	***	**	**	*	*	-	-	***
imagined-to-mouthed -	****	-	***	**	**	*	*	-	-	***
chance -	****	***	*	****	****	**	***	***	***	-

Fig. 10. Box plot of the average balanced accuracy of the multi-channel WM and CM models across participants and corresponding table of significance levels between performances of the mode pairs. Blue boxes represent the WM models. Red boxes represent both train-test combinations of the CM models for overt and mouthed. Green and orange boxes represent the CM models for the combinations of overt/imagined and mouthed/imagined, respectively. The horizontal line within each box shows the median, while the extents of the boxes represent the first (Q1) and third (Q3) quartiles. The whiskers extend from the box to 1.5 times the inter-quartile range (IQR). Each dot represents a data point from an individual participant that lies between the two 1.5IQRs and the outliers are indicated with diamonds. The chance distribution represents chance-level classification performance of all models based on a random permutation of the class labels, with the magenta dashed line indicating the average chance-level classification over all participants and models. -, *, **, ***, and **** indicate a significance level of $p > 0.05$, $p < 0.05$, $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively.

that imagined speech is an abbreviation of overt speech, suggesting that cognitive processes relevant to imagined speech are also involved during overt speech, whereas overt speech involves additional processes beyond imagined speech - likely associated with articulatory planning, articulation, sound production, and possibly aspects of perceptual feedback.

The present study provides evidence that channels relevant to different speech modes generally form nested hierarchical subsets from highest behavioral output to lowest behavioral output. Specifically, channels relevant to imagined were found to be a subset of those relevant in mouthed, while those relevant for mouthed were a subset of those relevant for overt. The subset of relevant channels for mouthed and overt that is mutually exclusive with imagined likely represents activity related to direct control of the speech articulators modulated in both modes. The channels exclusive to overt are presumed to be related to brain activities present exclusively for overt, including perceptual feedback, articulatory planning, articulatory motor executions, and/or sound production. The perceptual feedback likely represents both direct and indirect perceptual activity, e.g., forward prediction (Heinks-Maldonado et al., 2006; Hickok et al., 2011; Pickering and Garrod, 2013). The subset of channels relevant for all modes is hypothesized to represent the common substrate of activity for general speech planning and production.

When examining individual channels across modes, as shown in Fig. 7, the nested nature of channels within the mode hierarchy is apparent. The majority of relevant channels were shared amongst the three modes, while only about ten percent were unique to overt and mouthed but not imagined. The channel subset relevant to imagined speech was found to reside in bilateral frontal and temporal regions, which is consistent with prior ECoG and fMRI studies indicating that overt, mouthed, and imagined speech produce neural activity in both right and left cortical hemispheres (Okada et al., 2018; Pei et al., 2011a). Furthermore, these activations occurred at various bilateral depths as indicated in Fig. 6b. This is consistent with previous studies showing neural features from both grey and white matter contributing to decoding models (Angrick et al., 2021; Kohler et al., 2021; Li et al., 2021; Okada et al., 2018; Soroush et al., 2021; 2022), further demonstrating the relevance of deeper structures and white matter for speech decoding.

Roughly a fourth of the relevant channels were unique to overt, which predominantly resided in more superficial temporal regions, despite pre-screening channels for auditory feedback. Notwithstanding the absence of auditory feedback, neural activations in or around the auditory cortex were observed for mouthed and imagined. This is consistent with prior studies using overt and imagined speech and has been hypothesized to be related to inner speech rehearsal or forward predictions of intended speech and its consequences (Brumberg et al., 2016; Cooney et al., 2018; Hickok et al., 2011; Leuthardt et al., 2012; Li et al., 2020; Okada et al., 2018; Palmer et al., 2001; Pei et al., 2011a; 2011b; Pickering and Garrod, 2013; Price, 2012; Zhang et al., 2020).

Figures 7c and 9 further support the nested behavioral hierarchy by comparing the model performances and respective neural feature weights of the nested channel groups across modes. The overt-mouthed-imagined group exhibits highly consistent performance across the three modes, while performance is degraded for the other groupings when evaluated on the LBO modes. While the relevant spectro-temporal features of the overt-mouthed-imagined group are quite similar across all three modes, the features of the overt-mouthed group for the overt and mouthed are also similar and noticeably different from imagined. Furthermore, differences between the feature of the overt group are observed between overt and the two LBO modes.

This hierarchy, with respect to relative decoding performance of relevant channels between mode pairs, is also observed from Fig. 8. Nearly all channels yield an above-chance performance for each mode in the pairs, except for a select group of channels in the lower portion of the plot that perform well for overt but not the other modes. This indicates that nearly all channels with above-chance performance in the LBO mode also performed above chance-level in the HBO mode, while the inverse does not hold. The majority of channels in the overt-mouthed-imagined group and the mouthed-imagined pair of the overt group are clustered within the same range on the LBO and HBO axes (i.e., 0.5-0.65), suggesting that these channels roughly yield comparable performance for both modes. However, the majority of the overt-mouthed and overt-imagined pairs of the overt group reside toward the right side of the plot, with the HBO axis (overt mode) having noticeably larger values, showing that these channels have a more dominant and potentially unique neural activity during overt compared to the other two modes. This relationship is also apparent in the overt group of Fig. 7c. A weaker but similar trend is observed among the overt-mouthed group, with majority of values in the same range on the LBO and HBO axes (i.e., 0.5-0.65) and some located toward the bottom right of the plot, with the

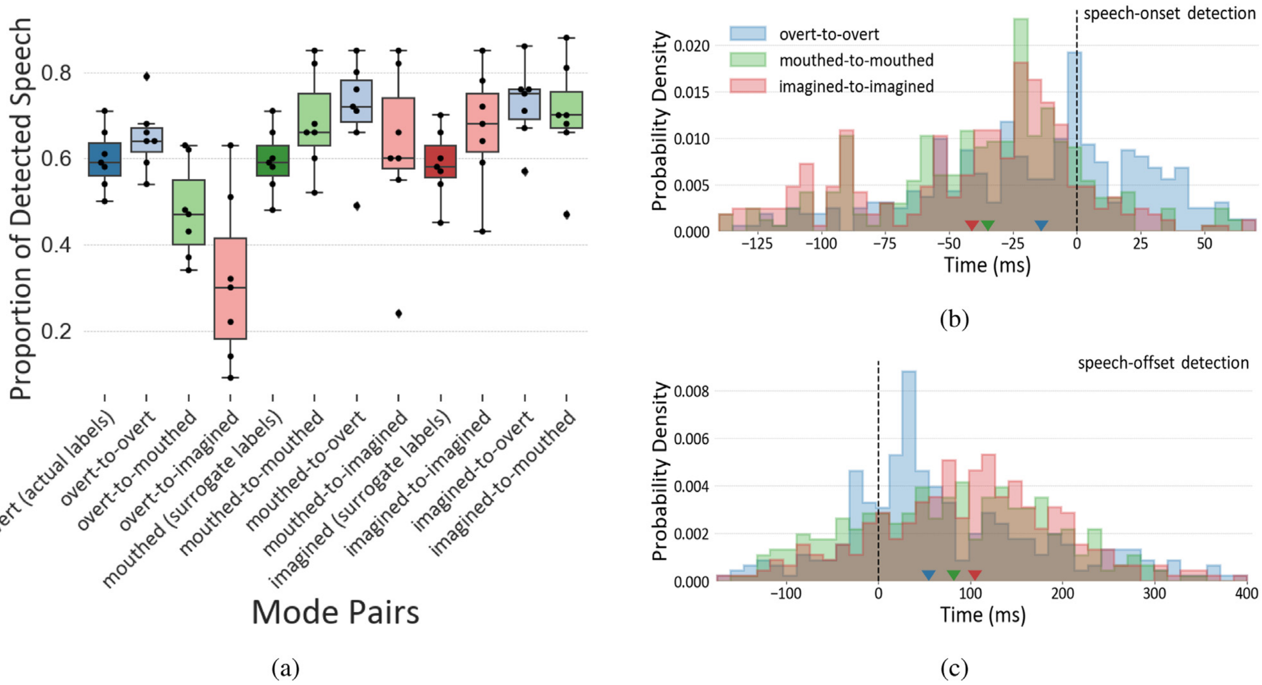


Fig. 11. (a) Average proportions of decoded speech vs. actual/surrogate speech labels for all participants and modes for all multi-channel WM and CM models. The bold-shaded boxes represent the speech proportions based on the actual/surrogate labels. For reference, the colors of lighter-shaded boxes are coordinated with colors of the actual/surrogate labels of the respective test modes. Refer to Fig. 10 for a description of the box plot properties. (b) Histograms of speech-onset detection timings of all multi-channel WM models. The black, vertical line indicates speech-onset based on actual labels of overt mode. (c) Histograms of speech-offset detection timings of all multi-channel WM models. The black, vertical line indicates speech-offset based on actual labels of overt mode. The blue, green, and red, triangles in (b) and (c) mark the mean of overt-to-overt, mouthed-to-mouthed, and imagined-to-imagined distributions, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

HBO axis (overt or mouthed) having noticeably larger values than the LBO axis (imagined). This is also apparent in overt-mouthed group of Fig. 7c.

4.2. Further evidence: Multi-channel models

These findings are relevant for understanding why multi-channel decoding models successfully trained and tested using overt speech tend to be poor at generalizing to imagined speech, as shown in Fig. 10. This figure also shows that there is a consistent decrease in performance when training on an HBO mode and testing on an LBO mode compared to the inverse, and this decrease is more pronounced for larger differences in the behavioral output hierarchy. It is also observed that the imagined models perform consistently well across modes, while the overt-to-imagined performs poorest amongst the combinations. This further suggests that the relevant channels for the imagined models also capture speech processes present for mouthed and overt, whereas the other relevant channels from overt (likely associated with articulation and aspects of perceptual feedback) do not extend to imagined. When interpreting these results, it is important to note that although the relevant LBO channels are available when training the HBO models, they do not appear to be selected or weighted in a way to generalize to the LBO modes. This is presumed due to the neural activity unique to the HBO modes (e.g., motor and auditory) having more prominent contributions to the HBO models, and hence having comparatively larger model weights, while the LBO-relevant channels are weighted near or equal to zero as a result of the L1 regularization.

While the present results offer strong evidence for the nested behavioral output hierarchy, other studies suggest this may be an oversimplification and imagined speech can be more than just an abbreviation of overt speech processes (Cooney et al., 2018; Geva et al., 2011; Li et al., 2020; MacKay, 1992; Oppenheim and Dell, 2010; Perrone-Bertolotti et al., 2014; Scott et al., 2013; Zhang et al., 2020). These prior studies indicate that imagined speech may involve different linguistic processes than those relative to overt speech or may contain unique neural processes (e.g., inhibitory activity) that are not involved in overt speech such as more prominent activity in the middle frontal gyrus, left and right temporal gyrus, left supramarginal gyrus, left superior frontal gyrus, and in various regions of white matter (Cooney et al., 2018; Geva et al., 2011; Li et al., 2020; Okada et al., 2018; Perrone-Bertolotti et al., 2014; Proix et al., 2022; Rampinini et al., 2017; Shuster and Lemieux, 2005; Zhang et al., 2020). In an fMRI study where participants performed overt, mouthed, and imagined trials of sixteen Chinese syllables, increasing monotonically from LBO to HBO modes, similar activity patterns were observed across the three speech modes in different brain regions, including superior temporal gyrus, angular gyrus, and inferior frontal gyrus. While these results indicated substantial overlap in regions activated during the three speech modes, activation unique to one or two modes was also observed in distinct regions (Zhang et al., 2020). Nevertheless, other studies have proposed that imagined speech may be an abbreviation of overt and mouthed speech processes, but further investigation is required to verify the precise mechanisms at the linguistic and motor levels (Okada et al., 2018; Oppenheim and Dell, 2010; Perrone-Bertolotti et al., 2014). The present study is limited by the nature and availability of sEEG recordings from a relatively low number of participants with sparse and inconsistent electrode coverage. While this coverage is not designed or ideal for speech decoding, it does provide important insights regarding previously unexplored neural features for this purpose.

Because this study was designed to specifically investigate speech modes, it is possible that the results may be influenced by neural processes that are not unique to speech production such as participant engagement or type of behavioral task. For example, it is conceivable that a similar nested hierarchy could be revealed for overt, mouthed, and imagined whistling. A separate experimental design is required to test this hypothesis.

4.3. Speech activity detection proportions

While the use of the speech activity detection model provides a very coarse labeling for the actual and surrogate speech, it yielded statistically above-chance performing models across modes, thus providing a solid and simplified basis for exploring and comparing the models and relevant features. Figure 11a shows that the proportions of detected speech tend to be overestimated when training using the surrogate labels. This is likely due to the inherent variability when applying surrogate labels. In contrast, the proportion of detected speech is underestimated when training on the actual (overt) labels and testing on the surrogate labels, and this effect is more pronounced for decreasing behavioral output. The detected proportions are strikingly consistent when training on imagined and testing across other modes. This further suggests the existence of a nested behavioral output hierarchy.

It was observed that the detected windows generally lead the actual speech-onset and lag the speech-offset (Fig. 11b and c), resulting in a higher false positive rate than false negative rate. This is again likely due to the inherent variability of the surrogate labels, but nevertheless may be desirable in practical application where the primary goal is to reliably detect the intention to speak.

5. Conclusion

The main objective of this study was to elucidate neural features associated with imagined speech to inform the development of imagined-speech neuroprostheses. This was achieved by comparing neural features and associated speech activity detection decoding model performance across three speech modes with varying degrees of behavioral output. The results suggest that the relevant channels can be organized in a nested hierarchy according to the degree of behavioral output, with the overt mode encompassing all relevant channels across modes, the relevant channels from the mouthed mode being a subset of overt, and the relevant channels from the imagined mode being a subset of mouthed. This nested hierarchy suggests that there may be a common neural substrate of related speech production processes that progressively extends with increasing behavioral output. These findings also provide important insights toward the design and development of imagined speech decoding models based on available overt speech data. Additionally, through the acquisition of sEEG, relevant neural activity across modes was found beyond the cortex, bilaterally at various depths, in both grey and white matter. This provides further evidence that deeper structures are relevant and may be beneficial in the development of improved speech decoding models. These findings also show that, with proper consideration and treatment, recordings of overt speech can serve as viable surrogates for generating imagined-speech decoding models. Given the limitations of sEEG recordings in terms of coverage and patient accessibility, additional work is needed to further characterize and understand the neural activity relationships across speaking modes. While the speech activity detection model provides a simplified framework for comparison, it is envisioned that these findings can be extended to more sophisticated imagined speech decoding schemes to reveal more nuance to the features and relationships.

Acknowledgments

This work was supported by the National Science Foundation (NSF) [grant numbers 2011595, 1608140] and the German Federal Ministry of Education and Research (BMBF) [grant numbers 01GQ1602, 01GQ2003].

Declaration of Competing Interest

No competing interests declared.

Credit authorship contribution statement

Pedram Z. Soroush: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Christian Herff:** Data curation, Writing – review & editing. **Stephanie K. Ries:** Data curation, Writing – review & editing. **Jerry J. Shih:** Funding acquisition, Data curation, Writing – review & editing. **Tanja Schultz:** Funding acquisition, Data curation, Writing – review & editing. **Dean J. Krusienski:** Conceptualization, Funding acquisition, Data curation, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request. Custom data analysis code is available on GitHub (https://github.com/pedramzs/The-nested-hierarchy-of-overt-mouthed-and-imagined-speech-activity-.git).

References

Alexandrou, A.M., Saarinen, T., Mäkelä, S., Kujala, J., Salmelin, R., 2017. The right hemisphere is highlighted in connected natural speech production and perception. Neuroimage 152, 628–638.

Angrick, M., Herff, C., Johnson, G., Shih, J., Krusienski, D., Schultz, T., 2019. Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. Neurocomputing 342, 145–151.

Angrick, M., Herff, C., Mugler, E., Tate, M.C., Slutzky, M.W., Krusienski, D.J., Schultz, T., 2019. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. J. Neural Eng. 16 (3), 036019.

Angrick, M., Ottenhoff, M.C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., Saal, J., Colon, A.J., Wagner, L., Krusienski, D.J., et al., 2021. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. Commun. Biol. 4 (1), 1–10.

Anumanchipalli, G.K., Chartier, J., Chang, E.F., 2019. Speech synthesis from neural decoding of spoken sentences. Nature 568 (7753), 493–498.

Arya, R., Ervin, B., Dudley, J., Buroker, J., Rozhkov, L., Scholle, C., Horn, P.S., Vannest, J., Byars, A.W., Leach, J.L., et al., 2019. Electrical stimulation mapping of language with stereo-EEG. Epilepsy Behav. 99, 106395.

Ball, T., Kern, M., Mutschler, I., Aertsen, A., Schulze-Bonhage, A., 2009. Signal quality of simultaneously recorded invasive and non-invasive EEG. Neuroimage 46 (3), 708–716.

Bénar, C.-G., Schön, D., Grimault, S., Nazarian, B., Burle, B., Roth, M., Badier, J.-M., Marquis, P., Liegeois-Chauvel, C., Anton, J.-L., 2007. Single-trial analysis of oddball event-related potentials in simultaneous EEG-fMRI. Hum. Brain Mapp. 28 (7), 602–613.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological) 57 (1), 289–300.

Bookheimer, S.Y., Zeffiro, T.A., Blaxton, T., Gaillard, W., Theodore, W., 1995. Regional cerebral blood flow during object naming and word reading. Hum. Brain Mapp. 3 (2), 93–106.

Brumberg, J.S., Krusienski, D.J., Chakrabarti, S., Gunduz, A., Brunner, P., Ritaccio, A.L., Schalk, G., 2016. Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task. PLoS ONE 11 (11), e0166872.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory–motor transformations for speech occur bilaterally. Nature 507 (7490), 94–98.

Cooney, C., Folli, R., Coyle, D., 2018. Neurolinguistics research advancing development of a direct-speech brain-computer interface. IScience 8, 103–125.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12 (7).

Evans, A., Collins, D., Mills, S., Brown, E., Kelly, R., Peters, T., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In: 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, pp. 1813–1817vol.3. doi:10.1109/NSSMIC.1993.373602.

Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.

Geva, S., Jones, P.S., Crinion, J.T., Price, C.J., Baron, J.-C., Warburton, E.A., 2011. The neural correlates of inner speech defined by voxel-based lesion–symptom mapping. Brain 134 (10), 3071–3082.

Heinks-Maldonado, T.H., Nagarajan, S.S., Houde, J.F., 2006. Magnetoencephalographic evidence for a precise forward model in speech production. Neuroreport 17 (13), 1375.

Herff, C., Diener, L., Angrick, M., Mugler, E., Tate, M.C., Goldrick, M.A., Krusienski, D.J., Slutzky, M.W., Schultz, T., 2019. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. Front. Neurosci. 13, 1267.

Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., Schultz, T., 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci. 9, 217.

Herff, C., Krusienski, D.J., Kubben, P., 2020. The potential of stereotactic-EEG for brain–computer interfaces: current progress and future directions. Front. Neurosci. 14, 123.

Hickok, G., Buchsbaum, B., Humphries, C., Muftuler, T., 2003. Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. J. Cogn. Neurosci. 15 (5), 673–682.

Hickok, G., Houde, J., Rong, F., 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69 (3), 407–422.

Huang, J., Carr, T.H., Cao, Y., 2002. Comparing cortical activations for silent and overt speech using event-related fMRI. Hum. Brain Mapp. 15 (1), 39–53.

Ibayashi, K., Kunii, N., Matsuo, T., Ishishita, Y., Shimada, S., Kawai, K., Saito, N., 2018. Decoding speech with integrated hybrid signals recorded from the human ventral motor cortex. Front. Neurosci. 12, 221.

Iida, K., Otsubo, H., 2017. Stereoelectroencephalography: indication and efficacy. Neurol. Med. Chir. 57 (8), 375–385.

Kanas, V.G., Mporas, I., Benz, H.L., Sgarbas, K.N., Bezerianos, A., Crone, N.E., 2014. Joint spatial-spectral feature space clustering for speech activity detection from ECoG signals. IEEE Trans. Biomed. Eng. 61 (4), 1241–1250.

Kanas, V.G., Mporas, I., Benz, H.L., Sgarbas, K.N., Bezerianos, A., Crone, N.E., 2014. Real-time voice activity detection for ECoG-based speech brain machine interfaces. In: 2014 19th International Conference on Digital Signal Processing. IEEE, pp. 862–865.

Koct, M., Juh, J., et al., 2019. Speech activity detection from EEG using a feed-forward neural network. In: 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, pp. 147–152.

Kohler, J., Ottenhoff, M. C., Goulis, S., Angrick, M., Colon, A. J., Wagner, L., Tousseyn, S., Kubben, P. L., Herff, C., 2021. Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework. arXiv preprint arXiv:2111.01457.

Leuthardt, E., Pei, X.-M., Breshears, J., Gaona, C., Sharma, M., Freudenburg, Z., Barbour, D., Schalk, G., 2012. Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task. Front. Hum. Neurosci. 6, 99.

Leuthardt, E.C., Gaona, C., Sharma, M., Szrama, N., Roland, J., Freudenberg, Z., Solis, J., Breshears, J., Schalk, G., 2011. Using the electrocorticographic speech network to control a brain–computer interface in humans. J. Neural Eng. 8 (3), 036004.

Li, G., Jiang, S., Meng, J., Chai, G., Wu, Z., Fan, Z., Hu, J., Sheng, X., Zhang, D., Chen, L., et al., 2022. Assessing differential representation of hand movements in multiple domains using stereo-electroencephalographic recordings. Neuroimage 250, 118969.

Li, G., Jiang, S., Paraskevopoulou, S.E., Chai, G., Wei, Z., Liu, S., Wang, M., Xu, Y., Fan, Z., Wu, Z., et al., 2021. Detection of human white matter activation and evaluation of its function in movement decoding using stereo-electroencephalography (sEEG). J. Neural Eng. 18 (4), 0460c6.

Li, G., Jiang, S., Paraskevopoulou, S.E., Wang, M., Xu, Y., Wu, Z., Chen, L., Zhang, D., Schalk, G., 2018. Optimal referencing for stereo-electroencephalographic (sEEG) recordings. Neuroimage 183, 327–335.

Li, Y., Luo, H., Tian, X., 2020. Mental operations in rhythm: motor-to-sensory transformation mediates imagined singing. PLoS Biol. 18 (10), e3000504.

Livezey, J.A., Bouchard, K.E., Chang, E.F., 2019. Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex. PLoS Comput. Biol. 15 (9), e1007091.

MacKay, D.G., 1992. Constraints on theories of inner speech. Auditory Imagery 121–149.

Makin, J.G., Moses, D.A., Chang, E.F., 2020. Machine translation of cortical activity to text with an encoder–decoder framework. Nat. Neurosci. 23 (4), 575–582.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N.E., Rieger, J., Schalk, G., Knight, R.T., Pasley, B.N., 2014. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng. 7, 14.

Martin, S., Brunner, P., Iturrate, I., Millán, J.d.R., Schalk, G., Knight, R.T., Pasley, B.N., 2016. Word pair classification during imagined speech using direct brain recordings. Sci. Rep. 6 (1), 1–12.

Meng, K., Grayden, D.B., Cook, M.J., Vogrin, S., Goodarzy, F., 2021. Identification of discriminative features for decoding overt and imagined speech using stereotactic electroencephalography. In: 2021 9th International Winter Conference on Brain-Computer Interface (BCI), pp. 1–6. doi:10.1109/BCI51272.2021.9385355.

Mercier, M.R., Bickel, S., Megevand, P., Groppe, D.M., Schroeder, C.E., Mehta, A.D., Lado, F.A., 2017. Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: a glimpse on white matter signal. Neuroimage 147, 219–232.

Moses, D.A., Leonard, M.K., Makin, J.G., Chang, E.F., 2019. Real-time decoding of question-and-answer speech dialogue using human cortical activity. Nat. Commun. 10 (1), 1–14.

Mugler, E.M., Patton, J.L., Flint, R.D., Wright, Z.A., Schuele, S.U., Rosenow, J., Shih, J.J., Krusienski, D.J., Slutzky, M.W., 2014. Direct classification of all American English phonemes using signals from functional speech motor cortex. J. Neural Eng. 11 (3), 035015.

Mugler, E.M., Tate, M.C., Livescu, K., Templer, J.W., Goldrick, M.A., Slutzky, M.W., 2018. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. J. Neurosci. 38 (46), 9803–9813.

Okada, K., Matchin, W., Hickok, G., 2018. Neural evidence for predictive coding in auditory cortex during speech production. Psychon. Bull. Rev. 25 (1), 423–430.

Oppenheim, G.M., Dell, G.S., 2010. Motor movement matters: the flexible abstractness of inner speech. Memory Cognit. 38 (8), 1147–1160.

Orpella, J., Mantegna, F., Assaneo, F., Poeppel, D., 2022. Speech imagery decoding as a window to speech planning and production. bioRxiv.

Palmer, E.D., Rosen, H.J., Ojemann, J.G., Buckner, R.L., Kelley, W.M., Petersen, S.E., 2001. An event-related fMRI study of overt and covert word stem completion. Neuroimage 14 (1), 182–193.

Pei, X., Barbour, D.L., Leuthardt, E.C., Schalk, G., 2011. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J. Neural Eng. 8 (4), 046028.

Pei, X., Leuthardt, E.C., Gaona, C.M., Brunner, P., Wolpaw, J.R., Schalk, G., 2011. Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. Neuroimage 54 (4), 2960–2972.

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M., Loevenbruck, H., 2014. 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.

Petrosyan, A., Voskoboinikov, A., Sukhinin, D., Makarova, A., Skalnaya, A., Arkhipova, N., Sinkin, M., Ossadtchi, A., 2022. Speech decoding from a small set of spatially segregated minimally invasive intracranial EEG electrodes with a compact and interpretable neural network. J. Neural Eng. 19 (6), 066016.

Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behav. Brain Sci. 36 (4), 329–347.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62 (2), 816–847.

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B.N., Knight, R.T., Tian, X., Poeppel, D., Doyle, W.K., Devinsky, O., et al., 2022. Imagined speech can be decoded from low-and cross-frequency intracranial EEG features. Nat. Commun. 13 (1), 1–14.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F.M., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. 103 (20), 7865–7870.

Rampinini, A.C., Handjaras, G., Leo, A., Cecchetti, L., Ricciardi, E., Marotta, G., Pietrini, P., 2017. Functional and spatial segregation within the inferior frontal and superior temporal cortices during listening, articulation imagery, and production of vowels. Sci. Rep. 7 (1), 1–13.

Ramsey, N.F., Salari, E., Aarnoutse, E.J., Vansteensel, M.J., Bleichner, M.G., Freudenburg, Z., 2018. Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. Neuroimage 180, 301–311.

Revell, A.Y., Silva, A.B., Mahesh, D., Armstrong, L., Arnold, T.C., Bernabei, J.M., Stein, J.M., Das, S.R., Shinohara, R.T., Bassett, D.S., et al., 2021. White matter signals reflect information transmission between brain regions during seizures. BioRxiV.

Rockhill, A.P., Larson, E., Stedelin, B., Mantovani, A., Raslan, A.M., Gramfort, A., Swann, N.C., 2022. Intracranial electrode location and analysis in MNE-python. J. Open Source Softw. 7 (70), 3897.

Rosenthal, R., Cooper, H., Hedges, L., et al., 1994. Parametric measures of effect size. Handb. Res. Synth. 621 (2), 231–244.

Rothauser, E., 1969. Ieee recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17, 225–246.

Roussel, P., Le Godais, G., Bocquelet, F., Palma, M., Hongjie, J., Zhang, S., Giraud, A.-L., Mégevand, P., Miller, K., Gehrig, J., et al., 2020. Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. J. Neural Eng. 17 (5), 056028.

Schalk, G., Leuthardt, E.C., 2011. Brain-computer interfaces using electrocorticographic signals. IEEE Rev. Biomed. Eng. 4, 140–154.

Scott, M., Yeung, H.H., Gick, B., Werker, J.F., 2013. Inner speech captures the perception of external speech. J. Acoust. Soc. Am. 133 (4), EL286–EL292.

Shuster, L.I., Lemieux, S.K., 2005. An fMRI investigation of covertly and overtly produced mono-and multisyllabic words. Brain Lang. 93 (1), 20–31.

Sjölander, K., Beskow, J., 2000. Wavesurfer-an open source speech tool. In: Sixth International Conference on Spoken Language Processing.

Soroush, P., Angrick, M., Shih, J., Schultz, T., Krusienski, D., 2021. Speech activity detection from stereotactic EEG. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 3402–3407.

Soroush, P., Herff, C., Ries, S., Shih, J., Schultz, T., Krusienski, D., 2022. Contributions of stereotactic EEG electrodes in grey and white matter to speech activity detection. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, pp. 4789–4792.

Soroush, P.Z., Shamsollahi, M.B., 2018. A non-user-based BCI application for robot control. In: 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE, pp. 36–41.

Sun, P., Anumanchipalli, G.K., Chang, E.F., 2020. Brain2Char: a deep architecture for decoding text from brain recordings. J. Neural Eng. 17 (6), 066015.

Tian, X., Poeppel, D., 2010. Mental imagery of speech and movement implicates the dynamics of internal forward models. Front. Psychol. 1, 166.

Tian, X., Poeppel, D., 2012. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. Front. Hum. Neurosci. 6, 314.

Tian, X., Poeppel, D., 2013. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. J. Cogn. Neurosci. 25 (7), 1020–1036.

Tian, X., Zarate, J.M., Poeppel, D., 2016. Mental imagery of speech implicates two mechanisms of perceptual reactivation. Cortex 77, 1–12.

Tourville, J.A., Reilly, K.J., Guenther, F.H., 2008. Neural mechanisms underlying auditory feedback control of speech. Neuroimage 39 (3), 1429–1443.

Vadera, S., Marathe, A.R., Gonzalez-Martinez, J., Taylor, D.M., 2013. Stereoelectroencephalography for continuous two-dimensional cursor control in a brain-machine interface. Neurosurg. Focus 34 (6), E3.

Wilcoxon, F., 1992. Individual comparisons by ranking methods. In: Breakthroughs in Statistics. Springer, pp. 196–202.

Willett, F.R., Avansino, D.T., Hochberg, L.R., Henderson, J.M., Shenoy, K.V., 2021. High-performance brain-to-text communication via handwriting. Nature 593, 249–254.

Zhang, W., Liu, Y., Wang, X., Tian, X., 2020. The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production. Cogn. Neurosci. 11 (4), 194–204.