Journal of Neural Engineering

CrossMark

OPEN ACCESS

RECEIVED
11 March 2024

REVISED
24 November 2024

ACCEPTED FOR PUBLICATION
15 January 2025

PUBLISHED
28 January 2025

PAPER

Transformer-based neural speech decoding from surface and depth electrode signals

Junbo Chen[1,5], Xupeng Chen[1,5], Ran Wang[1], Chenqian Le[1], Amirhossein Khalilian-Gourtani[2], Erika Jensen[2], Patricia Dugan[2], Werner Doyle[4], Orrin Devinsky[2], Daniel Friedman[2], Adeen Flinker[2,3,6] and Yao Wang[1,3,6,*]

[1] Electrical and Computer Engineering Department, New York University, 370 Jay Street, Brooklyn, NY 11201, United States of America
[2] Neurology Department, New York University, 223 East 34th Street, Manhattan, NY 10016, United States of America
[3] Biomedical Engineering Department, New York University, 370 Jay Street, Brooklyn, NY 11201, United States of America
[4] Neurosurgery Department, New York University, 550 1st Avenue, Manhattan, NY 10016, United States of America
[5] These authors contributed equally to the work.
[6] These authors jointly supervised the work.
[*] Author to whom any correspondence should be addressed.

E-mail: yaowang@nyu.edu

Keywords: ECoG, neural speech decoding, neural speech prosthesis, speech synthesis, electrocorticographic

Supplementary material for this article is available online

Abstract

Objective. This study investigates speech decoding from neural signals captured by intracranial electrodes. Most prior works can only work with electrodes on a 2D grid (i.e. Electrocorticographic (ECoG) or ECoG array) and data from a single patient. We aim to design a deep-learning model architecture that can accommodate both surface ECoG and depth (stereotactic EEG or sEEG) electrodes. The architecture should allow training on data from multiple participants with large variability in electrode placements. The model should not have subject-specific layers and the trained model should perform well on participants unseen during training. *Approach.* We propose a novel transformer-based model architecture named SwinTW that can work with arbitrarily positioned electrodes by leveraging their 3D locations on the cortex rather than their positions on a 2D grid. We train subject-specific models using data from a single participant and multi-subject models exploiting data from multiple participants. *Main results.* The subject-specific models using only low-density 8×8 ECoG data achieved high decoding Pearson Correlation Coefficient with ground truth spectrogram ($\mathrm{PCC} = 0.817$), over $N = 43$ participants, significantly outperforming our prior convolutional ResNet model and the 3D Swin transformer model. Incorporating additional strip, depth, and grid electrodes available in each participant ($N = 39$) led to further improvement ($\mathrm{PCC} = 0.838$). For participants with only sEEG electrodes ($N = 9$), subject-specific models still enjoy comparable performance with an average $\mathrm{PCC} = 0.798$. A single multi-subject model trained on ECoG data from 15 participants yielded comparable results ($\mathrm{PCC} = 0.837$) as 15 models trained individually for these participants ($\mathrm{PCC} = 0.831$). Furthermore, the multi-subject models achieved high performance on unseen participants, with an average $\mathrm{PCC} = 0.765$ in leave-one-out cross-validation. *Significance.* The proposed SwinTW decoder enables future speech decoding approaches to utilize any electrode placement that is clinically optimal or feasible for a particular participant, including using only depth electrodes, which are more routinely implanted in chronic neurosurgical procedures. The success of the single multi-subject model when tested on participants within the training cohort demonstrates that the model architecture is capable of exploiting data from multiple participants with diverse electrode placements. The architecture's flexibility in training with both single-subject and multi-subject data, as well as grid and non-grid electrodes, ensures its broad applicability. Importantly, the generalizability of the multi-subject models in our study population suggests that a model trained using paired acoustic and neural data from multiple patients can potentially be applied to new patients with speech disability where acoustic-neural training data is not feasible.

1. Introduction

Brain-related speech disability, which can be caused by stroke, injury, or tumor [10, 35, 45], can seriously decrease a patient's quality of life. In the United States, an estimated 2.5 million people suffer from speech disability due to stroke alone [20]. There has been growing interest in using intracranial electrodes to record neural activity during speech production in order to directly decode human speech from these signals, making it possible to design Brain–Computer Interface to allow patients with speech disabilities to communicate [7, 8, 30, 33, 37, 39].

Recent advancements in deep neural networks have been leveraged to push the boundary of speech decoding from ECoG signals. The decoding pipeline proposed in [9, 48] first applies a Neural Decoder (called ECoG Decoder) to predict time-varying speech parameters and then uses a novel Speech Synthesizer to generate speech spectrograms from speech parameters. Using ResNet [15] or 3D Swin Transformer [29] as the Neural Decoder, high speech decoding performance in terms of PCC between the decoded and ground-truth spectrograms has been achieved. In [1], densely connected 3D convolutional neural networks (CNN) were applied to decode speech from ECoG signals. Besides CNN and Transformer, recurrent neural networks (RNN) and long short-term memory (LSTM) networks have also been explored as Neural Decoders [4, 31, 34]. Some approaches produced naturalistic reconstruction leveraging wavenet vocoders [1], generative adversarial networks [47], and unit selection [16], but with limited accuracy. These studies demonstrate that deep neural networks can decode speech information from the complex neural activity recorded by the ECoG signals. A recent study in one patient with Microelectrode array [49] was successful in decoding text from the neural activity with high accuracy using the RNN model. Another recent study in one patient with implanted high-density ECoG electrodes [32] was successful in decoding naturalistic speech with high word decoding accuracy by leveraging the quantized HuBERT features [18] as an intermediate representation space and a pre-trained speech synthesizer which converts the HuBERT features into speech. However, HuBERT features do not carry speaker-specific acoustic information and thus can only be used to generate a generic speaker's voice, requiring a separate model to translate the generic voice to a specific patient's voice.

The deep neural networks in previous speech decoding studies have architecture designs with several limitations. First, architectures that use spatial convolution among electrodes, e.g. [1, 9, 48], are only applicable to grid electrodes like an ECoG array and hence do not work with strip or depth electrodes. Vision transformers' absolute position embeddings and relative positional bias are also based on the 2D or 3D grid index [12, 27–29] and hence are only applicable to grid electrodes [9, 24, 41]. On the other hand, the implantation of depth electrodes (stereotactic EEG or sEEG) has been a more popular neurosurgical approach which does not require the removal of a large skull portion with reports of fewer surgical complications [19, 44]. Further, the approach and electrodes employed in sEEG are similar to those used in deep brain stimulation (DBS), which has demonstrated long-term electrode safety, suggesting the possibility of chronic sEEG for speech neuroprostheses [17]. Multiple sEEG depth probes may be implanted, which can assay a wide range of deeper structures and thus may provide additional information not available from the surface of the cortex. Therefore, decoding speech from sEEG signals would have significant clinical advantages.

Secondly, models that use fully connected computations among the electrodes, e.g. [4, 31, 40], can only be trained for a specific participant, as the weights learned depend on the actual locations of the electrodes in the brain and the ordering of the electrodes. Because electrode placement varies quite widely across patients, fully connected architectures cannot be trained effectively with data from multiple participants unless subject-specific layers are introduced to map the original electrode data from different participants to a common feature domain. Likewise, convolutional [1, 9, 48] models or transformers [9, 24, 41, 48] that leverage grid indices for position embedding cannot generalize well to different participants because they do not specifically consider the locations of the electrodes on the brain. Therefore, studies to date have developed subject-specific models, which suffer from small data challenges as they cannot leverage signals from multiple subjects. Several studies have proposed models that have subject-specific layers along with a shared module that can be trained with data from multiple participants [11, 31, 43]. However, such models still require collecting training data for each participant to refine the subject-specific layer, limiting its practical applicability.

Our study proposes a novel transformer-based Neural Decoder that does not rely on a regular grid structure. We call it the Swin transformer with temporal windowing (SwinTW). Instead of relying on the grid index, the model leverages the anatomical location of electrodes in the standardized brain template to learn the attention between electrodes. The proposed Neural Decoder achieved superior performances than ResNet and 3D Swin Transformer across 43 participants, given the same grid electrodes, reported in [9]. The model demonstrated further performance increase by leveraging the off-grid electrodes that cannot be utilized in the previous studies. Importantly, the model demonstrated promising performance given sEEG electrodes only across 9 participants. Most significantly, the SwinTW model can be effectively trained with data from multiple participants without any subject-specific layers, and the resulting model can generalize well to participants outside the training cohort.

2. Method

2.1. Speech decoding framework

Our neural decoding framework is trained by following a 2-step approach proposed in our previous study [9], shown in figure 1. In the first step of Speech-to-Speech training, a Speech Encoder is used to extract speech parameters at every time frame (e.g. pitch, formant frequencies, loudness) from the input speech spectrogram, and a differentiable Speech Synthesizer is designed to reconstruct the spectrogram from the speech parameters. The Speech Encoder and the Speech Synthesizer are trained to match the reconstructed spectrogram with the ground truth. In the second step of Neural-to-Speech training, the Neural Decoder is trained to predict the time-varying speech parameters from neural signals using the speech parameters generated by the Speech Encoder as the guidance. The predicted speech parameters from the Neural Decoder are fed to the trained Speech Synthesizer from step 1 to generate the predicted speech spectrogram, which is then converted to the predicted speech waveform.

Following the design from our previous study [9, 48], the Speech Encoder extracts 18 speech parameters at each time step from the original speech spectrogram, which is then fed to the Speech Synthesizer to reconstruct the original speech spectrogram. The Speech Encoder adopts a simple network architecture with MLP (Multilayer Perceptron) and temporal convolution. The differentiable speech synthesizer enables the end-to-end training of the speech-to-speech auto-encoding task (see right of figure 1). Details about the Speech Encoder and Speech Synthesizer and their pretraining using speech signal only can be found in [9].

For Neural-to-Speech training, the Neural Decoder first maps neural activity from all input electrodes to a latent feature, which is then used to predict the 18 speech parameters for each time frame, supervised by the speech parameters generated by the Speech Encoder. Then, the speech parameters predicted by the Neural Decoder will be fed into the Speech Synthesizer to generate the predicted spectrogram, which is then converted to the ECoG-decoded speech signal.

At the inference time, only the trained Neural Decoder and the Speech Synthesizer are needed (figure 1(b)).

2.2. Neural decoder based on temporal swin transformer

In our study, we propose a novel architecture for decoding speech parameters from electrode signals that do not require electrodes to be on a 2D grid. We name the proposed Neural Decoder a SwinTW, inspired by the Swin Transformer [27, 28]. In the vanilla Vision Transformer (ViT) for an image [12], the self-attention layer computes global attention among all tokens (with each token corresponding to an image patch). This global attention causes the absence of the inductive bias of locality and heavy quadratic computational complexity to the input image size. The Swin Transformer solves the problems by grouping tokens into local windows and computing local attention within each window at each self-attention layer. To allow inter-window information exchange, the Swin Transformer shifts the window partition between every two windowed self-attention layers, which prevents different windows from being segregated (details can be found in [27, 28]). However, since the Swin Transformer was designed for 2D images (later extended to 3D videos [29]), its architecture assumes that the input is in the formats of 2D or 3D grids. Our previous transformer-based Neural Decoder used 3D Swin [9] which is inspired by [29], where each 3D window includes nearby 2 × 2 electrodes in two adjacent time steps. In our proposed SwinTW, we made several modifications to allow speech decoding based on electrodes in any topological layout. We still have spatial and temporal attention, but windowing is only applied in the temporal direction to constrain temporal attention. Instead of using electrode location on the 2D grid for spatial positioning information in [9], we use the anatomic location of each grid on the cortex (MNI coordinate and brain region index). The architecture of the SwinTW is shown in figure 1.

Temporal patch partition: In the Swin Transformer [27–29] or ViT [12], the input images or videos are partitioned into 2D or 3D patches, and each patch is then mapped to a token with a patch embedding layer. This patch partition requires ordering all the electrodes into a 2D grid and makes the trained model not invariant to the electrode order. To solve this problem, our proposed SwinTW generates tokens from each electrode individually and

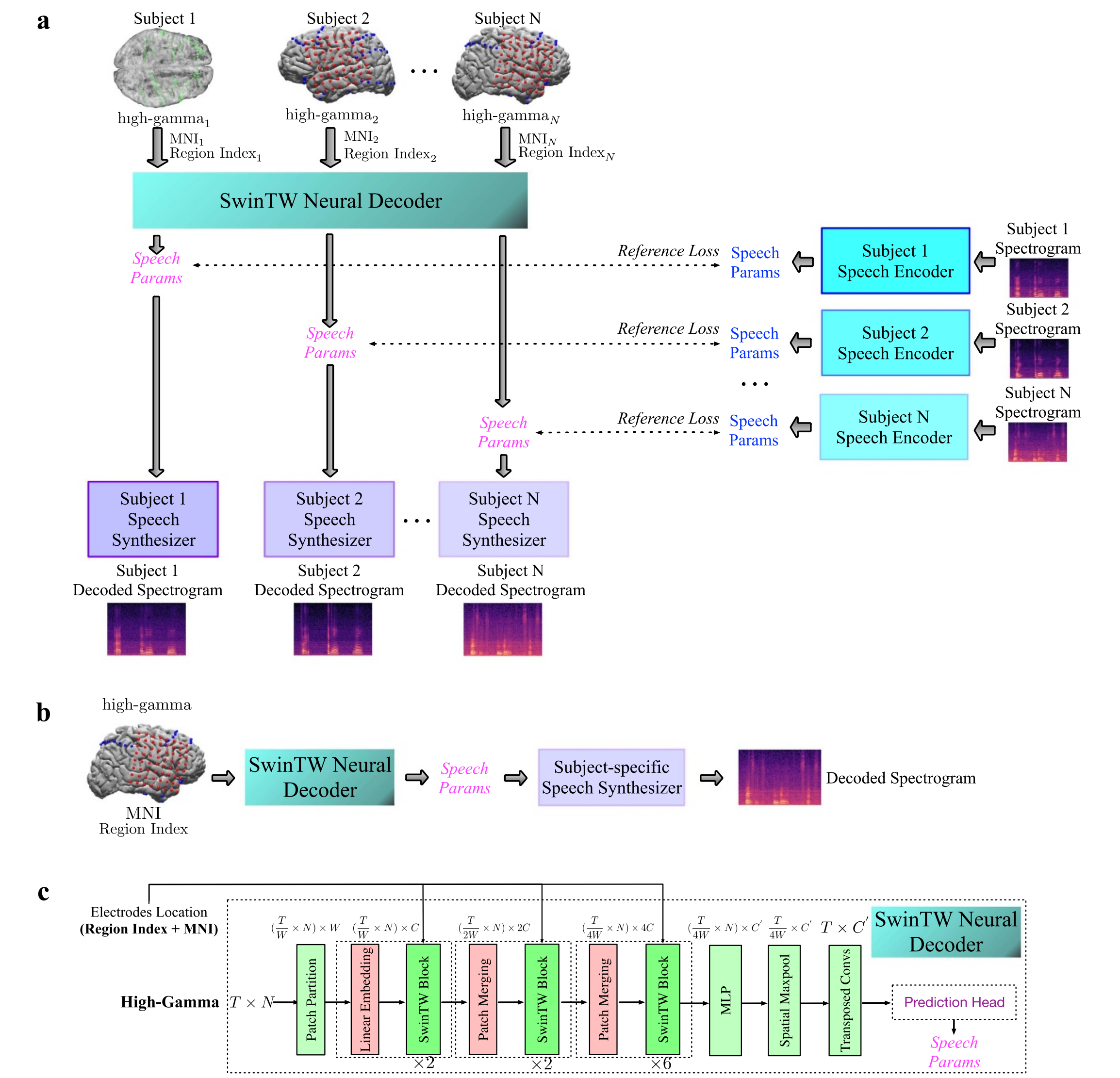

Figure 1. (a) Multiple-subject Neural Decoder training pipeline. Each participant's neural signal and electrodes' location information (MNI coordinates and ROI index) are fed to a shared SwinTW Neural Decoder to predict speech parameters. The predicted speech parameters are supervised by the speech parameters generated by the subject-specific Speech Encoder from the ground-truth speech spectrogram. Each participant's predicted speech parameters are fed into the corresponding subject-specific Speech Synthesizer to generate a speech spectrogram. Once the shared SwinTW decoder is trained, it can be used to decode the speech from the neural signal of any participant. Note that the same training pipeline can be used to train a subject-specific model using data from a single participant. (b) The inference pipeline: The trained SwinTW decoder takes the neural signals and location information of the electrodes of a participant as the input and generates speech parameters. The speech synthesizer (pre-trained using the speech signal of the participant) then converts the generated speech parameters to the decoded spectrogram. (c) The SwinTW decoder architecture uses three stages of transformer blocks with spatial-temporal attention with temporal windowing and patch merging to extract features. Transposed temporal convolution is then used to upsample the features back to the temporal dimension of the input. A prediction head module finally generates speech parameters from the upsampled features.

only partitions the temporal dimension. As shown in figure 1, given an ECoG signal with the shape of $T \times N$ (T: number of frames, N: number of electrodes), for each electrode, the SwinTW partitions the temporal sequence of neural activity into $\frac{T}{W}$ patches with patch size W. The temporal patch partition generates $\frac{T}{W} \times N$ patches in total, and a linear patch embedding layer is applied to each patch to generate $\frac{T}{W} \times N$ tokens with the latent dimension of C.

Temporal window attention: In Swin transformer [27–29], tokens are partitioned into windows, where each window contains a local subset of adjacent tokens, and attention is calculated only among tokens within the same window. In conventional 3D Swin, the windowing is applied spatially and temporally, making the model only suitable for electrodes arranged in a 2D grid. In SwinTW, to remove this grid input constraint, the model only partitions tokens into local windows in the temporal dimension and allows spatial attention across all electrodes (this can be thought of as using a spatial window size that includes all electrodes). Given $N = N_t \times N_s$ tokens (N: total number of tokens, N_t: number of tokens in the temporal dimension, N_s: number of tokens in the spatial dimension, equal to number of electrodes) and window size W_t, the N tokens are partitioned into $\frac{N_t}{W_t}$ windows and attention is calculated among $W_t \times N_s$ tokens within each window.

Temporal patch merging: The Swin Transformer leverages patch merging to achieve inductive bias of locality and hierarchical feature maps. However, merging nearby patches in the spatial dimension is not feasible when the electrodes are not arranged in a grid. Therefore, instead of using the spatiotemporal patch merging in the 3D Swin Transformer [29], the SwinTW conducts temporal patch merging for each electrode individually. For each electrode, every two consecutive tokens in the temporal dimension with feature dimension C will be concatenated as a $2C$ dimensional latent and get mapped to a $2C$ dimensional merged token.

Grid-free positional embedding: The SwinTW follows Swin Transformers [27] to exploit positional information through relative positional bias. However, instead of using the 2D or 3D grid index difference as the relative position like the Swin Transformer, our SwinTW defines the relative positional bias based on each token's anatomical location and time-frame index. The positional bias is defined as below:

$$\text{Attention}\left(Q,K,V\right) = \text{Softmax}\left(SIM\left(Q,K\right)\right)V \quad (1)$$

$$\text{SIM}\left(q_i,k_j\right) = \frac{q_i\,k_j}{|q_i||k_j|}/\tau + B_{i,j} \quad (2)$$

$$B_{i,j} = MLP\left(x_i,y_i,z_i,t_i,x_j,y_j,z_j,t_j,x_i-x_j,y_i-y_j,\right.$$
$$\left.z_i-z_j,t_i-t_j\right) + r_i\cdot r_j \quad (3)$$

Given $Q,K,V \in R^{N\times C}$ (Q,K,V are query, key and value generated from each token, N is number of tokens and C is the latent dimension), shown in equation (1), the softmax of SIM(Q,K) for all pairs of token in the window is used to aggregated V (values of tokens within the window) to get the output token values. We define query-key similarity following the scaled cosine attention of SwinV2 [27], defined in equation (2). τ is a learnable parameter not shared among attention heads and layers. $B_{i,j}$ is the relative positional bias between token i and token j. In SwinTW, $B_{i,j}$ consists of two terms: MNI-based positional bias and region-based bias. We project each subject's electrodes to a standardized Montreal Neurological Institute (MNI) brain anatomical map and collect each electrode's x,y,z location in the MNI coordinate. For each token pair, the MNI coordinates of the corresponding electrodes and time-frame index, along with their differences, will be mapped to the MNI-based positional bias with a 2-layer MLP, which is shown in the first term of equation (3). We also parcellate the standardized brain into regions of interest (ROIs) and learn a dictionary of embeddings for all ROIs, with r_i denoting the embedding features for region i. Given N_r ROIs and N_h attention head, the learnable dictionary has N_h sets of $N_r \times C_r$ region embeddings (C_r is the region embedding dimension). The region embeddings $r_i, \forall i$ are learned during the training. For a pair of tokens, the dot product of the embeddings of their corresponding electrodes' ROIs will be added to the positional bias, shown in the second term of equation (3). The dot product is used instead of cosine similarity to allow the model to assign high attention to certain regions by letting them have large embedding values.

The architecture of SwinTW is shown in figure 1. The input ECoG signal with a size of $T \times N$ is partitioned into $(\frac{T}{W} \times N)$ patches, each with a patch size of $W \times 1$. A linear patch embedding layer maps each patch to a C dimensional token. The SwinTW has three stages with 2, 2, and 6 layers. Swin Transformer Block (consists of a windowed multihead self-attention layer and an MLP) is applied in each layer, detailed in [28], and we replace the spatial-temporal windowing with temporal-only windowing. Following the Swin Transformer, the temporal window partition is shifted for every two consecutive layers to allow inter-window information exchange, detailed in [28]. SwinTW performs temporal patch merging after the first and second stages, each stage decreasing the token number by half and doubling the latent dimension. After stage 3, an MLP is applied to decrease the $4C$ latent dimension to C'. Spatial max pooling across the electrodes is then applied to convert $(\frac{T}{4W} \times N) \times C'$ feature maps to $\frac{T}{4W} \times C'$. Transposed temporal convolutions are then employed to upsample $\frac{T}{4W} \times C'$ to $T \times C'$, where T is the frame number of the input neural signal. As shown in 1, the $T \times C'$ latent from SwinTW next goes through the prediction head consisting of temporal convolutions (kernel-size = 3) and MLP to predict the 18 speech parameters at every frame.

In our study, we set $C = 96$ and $C' = 32$. Patch-size $W = 4$ and window size $W_t = 4$ is applied to partition temporal dimension. In our 3 stages SwinTW with 2, 2, and 6 layers, the self-attention layers in the 3 stages have 3, 6, and 12 attention heads, respectively. The MLP in SwinTW has 3 layers (384→196→96→32) with layer norm [5] and LeakyRELU activation in between. The transposed convolution for temporal upsampling contains 4 1D transposed convolutional layers with stride = 2 and kernel-size = 3, padding = 1. These parameter choices are determined through empirical trials and errors.

2.3. Training of subject-specific neural decoders

The training procedures for both the Speech Encoder and Speech Synthesizer follow the methods described in our previous work [9]. Therefore, we omit the details in this section. Following [9], we use two types of supervision to guide the training of the Neural Decoder that predicts speech parameters from neural signals. Firstly, we train the decoder to generate speech parameters that match the parameters generated by the speech encoder. Besides, the ground truth speech spectrograms act as additional supervision for the decoder, as the predicted speech parameters are converted to spectrograms by the speech synthesizer. The fact that our Speech Synthesizer is differentiable enables us to use the spectrogram reconstruction loss for end-to-end training. The reference loss $L_{\text{reference}}$ for the speech parameters is defined as:

$$L_{\text{reference}} = \sum_{i,t} \lambda_i ||\hat{C}_i^t - C_i^t||_2^2,\ \ i \in \times \left[{f^t}_0, {f^t}_1, \ldots, {f^t}_6, a_1^t, \ldots, a_6^t, {f^t}_u, b_u^t, a_u^t, \alpha^t, L^t\right] \tag{4}$$

where $\hat{C}_i^t$ and C_i^t are speech parameters generated by the Neural Decoder and the Speech Encoder (as ground truth), respectively. We assign each speech parameter with individual weight λ_i through testing the performances on three hybrid-density participants with different parameter choices, and the values are detailed in [9]. For spectrogram-based supervision, we use modified multi-scale spectral loss L_{MSS}, Short-Time Objective Intelligibility (STOI) loss L_{STOI}, and supervision loss $L_{\text{supervision}}$. L_{MSS} is inspired by [13]. It supervises speech reconstruction by measuring the distance between the ground truth spectrogram and the reconstructed spectrogram in both linear and mel-frequency scales. L_{STOI} measures the intelligibility of reconstructed speech based on the STOI+ metric [14]. Higher STOI+ indicates better intelligibility, the L_{STOI} is defined as the negative of STOI+: $L_{\text{STOI}} = -\text{STOI+}$. Besides, additional supervision $L_{\text{supervision}}$ is applied to improve the prediction accuracy for the pitch ${f^t}_0$ and formant frequencies ${f^t}_{i=1,2,3,4}$. The $L_{\text{supervision}}$ calculates the L2 distance between each predicted frequency and the corresponding frequency extracted by the Praat method [6]. The overall loss for training the Neural Decoder is

$$L = L_{\text{MSS}} + \lambda_1 L_{\text{STOI}} + \lambda_2 L_{\text{supervision}} + \lambda_3 L_{\text{reference}} \tag{5}$$

where λ_1 λ_2 and λ_3 are set to 1.2, 0.1, and 1.0, following [9] through testing the performances on three hybrid-density participants with different parameter choices.

Adam optimizer [22] with learning-rate $= 5 \times 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used to train the Neural Decoder. As mentioned in section 3.1, following [9], randomly selected 50 out of 400 trials are used as the test set for each subject, and the remaining data are used for training.

2.4. Multi-subject neural decoder training

The proposed SwinTW allows the Neural Decoder to take input with any electrode layout as long as we know each electrode's MNI coordinate and region index. Therefore, this architecture allows the Neural Decoder to be trained using data from multiple participants and then used for inference on any participant. Figure 1 demonstrates the multi-subject Neural Decoder training pipeline. Given data from multiple participants, a shared SwinTW-based Neural Decoder generates speech parameters based on each participant's electrode signals and electrode locations (electrodes' MNI coordinates and region index). Reference loss is calculated between the predicted speech parameters and the speech parameters generated by the subject-specific Speech Encoder. Each subject's predicted speech parameters are fed into the corresponding subject-specific speech synthesizer to generate a speech spectrogram. The neural signals and electrodes' locations are fed into the Neural Decoder to generate speech parameters during inference. The participant's speech synthesizer then generates a speech spectrogram from the predicted speech parameters. Note that the embeddings for different ROIs are also learned as part of the Neural Decoder training. When we train a decoder using participants with left and right hemisphere electrodes, separate region embeddings are learned for the left and right brain hemispheres.

2.5. Evaluation metrics

Following [1, 4, 9, 50], we used three metrics to evaluate the speech decoding performance:

(1) Pearson correlation coefficient (PCC) measures the normalized correlation between the decoded spectrogram and the actual spectrogram and is a widely used metric to evaluate the accuracy of the decoded spectrogram.

(2) STOI+ [14] is another metric that measures the similarity between decoded and original speech. STOI+ has been reported to have a monotonic relationship with speech intelligibility. The STOI+ value ranges from −1 to 1, and higher STOI+ indicates better intelligibility.

(3) Mel-cepstral distortion (MCD)[25] measures the differences between 25 acoustic features generated from the decoded speech and the original speech. A lower MCD is better. MCD is calculated as follows:

$$\text{MCD} = \frac{10}{\ln(10)} \sqrt{\sum_{0<d<25} \left(\text{mc}_d - \hat{\text{mc}}_d\right)^2}$$

where mc_d and $\hat{\text{mc}}_d$ denote the dth feature generated from the original and decoded speech.

3. Results

3.1. Neural data collection and preprocessing

The study includes 52 native English-speaking subjects (43 subjects with ECoG electrodes, 20 males, 23 females; 9 subjects with only sEEG electrodes, 3 males, 6 females) with refractory epilepsy (a disease involving seizures caused by abnormal electrical brain activity). Details about speech and ECoG signals collection can be found in [9]. In brief, at each trial, a subject was requested to speak a specific target word in response to an audio or visual stimulus while their neural activity signals were recorded. Each subject was asked to complete 5 different tasks: (1) Auditory Repetition (repeating the word that the care provider has spoken), (2) Auditory Naming (naming the word based on the definition that the care provider has spoken), (3) Sentence Completion (naming the last word to complete a sentence that the care provider has spoken) (4) Visual Reading (reading the written word shown by the care provider) (5) Picture Naming (naming the word based on a colored drawing shown by the care provider). Each task included the same 50 target words [42], each appearing once in the Auditory Naming and Sentence Completion and twice in each of the other tasks, leading to 400 trials of ECoG signal recording, and the average duration of word production among all trials was 500 ms.

All electrodes were implanted to capture clinically relevant brain regions, detailed in [9]. There were 43 subjects who had 8 × 8 ECoG electrodes with 10 mm spacing capturing signals over the perisylvian cortex (male left hemisphere: 14 subjects; female left hemisphere: 13 subjects; male right hemisphere: 6 subjects; female right hemisphere: 10 subjects). Besides the 8 × 8 grid electrodes, some subjects had additional electrode strips outside the 8 × 8 grid and/or depth electrodes implanted under the brain's surface. We also included 9 subjects with only sEEG electrodes (male = 3, female = 6). The experiments were approved by the Institutional Review Board of NYU Grossman School of Medicine, and written and oral consent was collected from each participant. All implanted electrodes were the clinical standard of care and FDA-approved. The high gamma component (70–150 Hz) was extracted from the raw electrode signal, with electrodes exhibiting artifacts or interictal/epileptiform activity excluded by setting their signal to 0. The preprocessing details can be found in [9]. This study also applies a Savitzky–Golay filter [38] with a 3rd-order polynomial and window size of 11 to further denoise the high-gamma signal in the temporal dimension. Among the 400 trials of ECoG signals recorded from the five-word production tasks, 350 trials were used for model training, and 50 trials were held out for testing (10 randomly selected trials were reserved for testing for each task).

3.2. Subject-specific models: speech decoding with electrodes on One ECoG grid

To compare our proposed grid-free SwinTW with the Neural Decoders based on ResNet and 3D Swin transformer in our previous study [9], firstly, we evaluated the SwinTW trained with 64 ECoG electrodes for each subject individually. Figure 2 compares the decoding performance of the SwinTW decoder with the 3D ResNet and 3D Swin decoders. For each subject, we compute the average PCC, STOI+, and MCD among all the test trials. Each dot in a box plot is the mean metric for one subject. As illustrated in figure 2, the SwinTW (mean PCC = 0.825, STOI+ = 0.309, MCD = 2.341) outperforms ResNet (PCC = 0.804, STOI+ = 0.264, MCD = 2.374) and 3D Swin transformers (PCC = 0.785, STOI+ = 0.216, MCD = 2.425) in terms of PCC, STOI+, and MCD. Statistical tests using the Wilcoxon two-sided signed-rank test further show that these improvements are significant with P <0.001. Additionally, the performance of the three models tested on shuffled data (by randomly shuffling the input neural signals temporally during the entire recording session) is also reported as a control in the supplementary figure. S1. It is evident that the decoding performance with non-shuffled data is significantly better. Note that SwinTW differs from 3D Swin primarily in how the spatial positions of two electrodes affect the spatial attention bias between the two electrodes. With 3D Swin, the relative position between the two electrodes on the 2D grid determines the attention bias, whereas, with SwinTW, the attention bias depends on the MNI coordinates and ROI embeddings of these electrodes. Our results suggest that using the MNI coordinates and ROI information can lead to better decoding performance while making the model applicable to non-grid electrodes.

We also assessed the decoded speech intelligibility using the state-of-the-art ASR tool Whisper [36] and report the results in figure S5. The results demonstrate that SwinTW leads to improved speech intelligibility over ResNet and 3D Swin Transformer decoder.

3.3. Subject-specific models: speech decoding with additional electrodes

As the SwinTW does not rely on the 2D grid positions of the electrodes, the proposed SwinTW can easily leverage off-grid electrodes to provide additional information for speech decoding. In our study, for each participant with additional electrodes beyond one ECoG grid, we selected additional electrodes with a standard deviation of the signal greater than a subject-specific threshold, determined following the approach described in [21] for identifying active electrodes. We then trained the SwinTW Neural Decoder

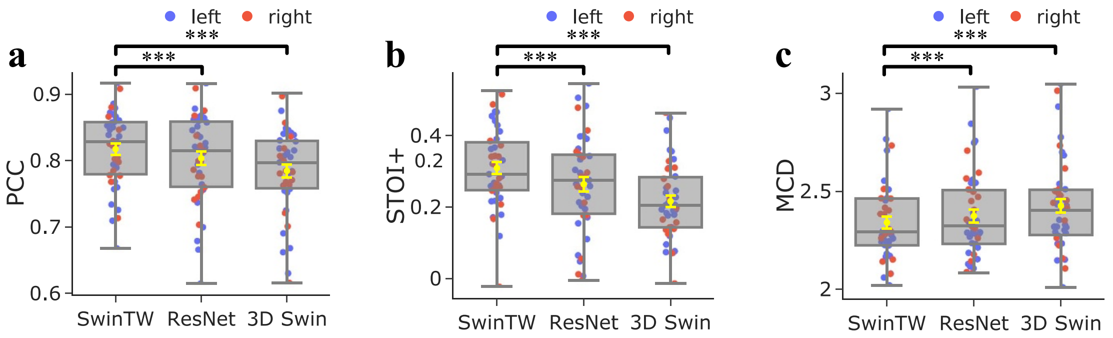

Figure 2. Performance comparison between different models. Subject-specific models when trained and tested on grid electrodes using different Neural Decoder architectures. Comparison of the distributions of decoding PCC (a), STOI+ (b), and MCD (c) over 43 participants. Each dot in a box plot indicates the mean metric for one participant across all testing trials. The yellow error bars denote the mean ± standard error of the mean (SEM) across participants. The SwinTW outperforms the ResNet and 3D Swin Transformer regarding PCC and STOI+. All box plots depict the median (horizontal line inside box), 25th and 75th percentiles (box), 25th or 75th percentiles ±1.5× interquartile range (whiskers) across all participants ($N = 43$), and the yellow error bars denote the mean ± standard error of the mean (SEM) across participants. Distributions were compared with each other as indicated. Black brackets indicate two experiments are compared using the Wilcoxon two-sided signed-rank test. ***: $P < 0.001$, *: $P < 0.05$, ns: $p > 0.05$.

with 64 electrodes from the 8 × 8 grid and the selected additional electrodes for each subject. As 4 participants did not have any additional electrodes that fulfill the threshold requirement, we compared the models based on the remaining 39 participants. Each participant had 1 to 19 strip electrodes, 1 to 21 depth electrodes, 1 to 21 extra grid electrodes, and 1 to 11 electrodes with unknown locations (we set the MNI coordinates of these electrodes as 0 and the region index corresponding to Unknown instead of discarding them). Figure 3 compares the SwinTW Neural Decoder performance using all selected electrodes and with the performance using only electrodes on one ECoG grid. The results demonstrate that additional electrodes can further improve the decoding performance (all electrodes: mean PCC = 0.838, STOI+ = 0.359, MCD = 2.228; grid electrodes: mean PCC = 0.825, STOI+ = 0.318, MCD = 2.341). Wilcoxon's two-sided signed-rank test further shows that these improvements are significant with $P = 0.00025$.

3.4. Subject-specific models: speech decoding with sEEG electrodes only

We also attempted to train the proposed SwinTW model to decode speech production from only SEEG electrodes. Our study included 9 subjects (male = 3, female = 6) with only sEEG electrodes implanted. For each subject, electrodes with a standard deviation of the signal greater than a subject-specific threshold derived following the approach of [21] were included. The number of selected electrodes for each participant ranges from 19 to 178. Figure 3 demonstrates that the SwinTW can achieve promising speech production prediction based on sEEG electrodes only, with the mean of PCC slightly lower than using grid electrodes (0.798 vs 0.825), mean MCD slightly higher (2.396 vs. 2.341), but STOI+ slightly higher (0.341 vs 0.318). Note that the decoding from 64 ECoG grid electrodes is from a different patient cohort consisting of 43 participants. Wilcoxon rank-sum tests comparing these two cohorts demonstrate that the decrease in PCC is significant with $P = 0.035$. However, the differences in STOI+ and MCD are not statistically significant.

3.5. Multi-subject model: evaluation on test trials of participants within the training set

As the proposed SwinTW architecture does not require the electrodes to be arranged in a grid but relies on the electrode position in the brain, it can handle the differences in the electrode placements among different participants and allow a single model to be trained with multiple patient data. To validate this idea, we trained a single SwinTW decoder with 15 randomly selected male participants with ECoG electrodes implanted in either the left or right brain hemisphere (4 on the left and 11 on the right). As detailed in section 2.4, subject-specific speech encoder and speech synthesizer are applied while the Neural Decoder is shared among subjects. We compare the decoding performance of the multi-subject and subject-specific models on the test trials of each of the 15 participants included in the multi-subject model training. As shown in figure 4, the multi-subject SwinTW model (PCC = 0.837, STOI+ = 0.352, MCD = 2.307) showed similar performance with the subject-specific model (PCC = 0.831, STOI+ = 0.334, MCD = 2.313), with no statistically significant differences.

3.6. Multi-subject model: evaluation on participants outside the training set

We also evaluated the multi-subject SwinTW decoder on test trials of the subjects outside the training set. We conducted 5-fold cross-validation separately for

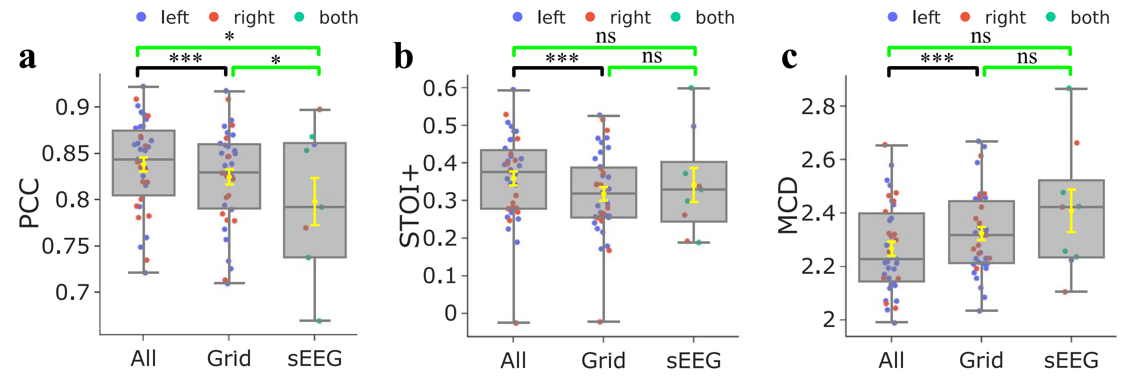

Figure 3. Performance comparison between different modalities with Subject-specific SwinTW model. Decoding PCC (a), STOI+ (b), and MCD (c) shows the comparison between subject-specific SwinTW Neural Decoder performance obtained with all selected electrodes, with only electrodes on one 8 × 8 grid for 39 participants, and with sEEG-only electrodes over 9 participants. Using all electrodes outperforms using grid electrodes or sEEG-only electrodes. Black brackets indicate two experiments are compared using the Wilcoxon two-sided signed-rank test. Green brackets indicate two experiments that are compared using the Wilcoxon rank-sum test, indicated in green. ***: $P < 0.001$, *: $P < 0.05$, ns: $p > 0.05$.

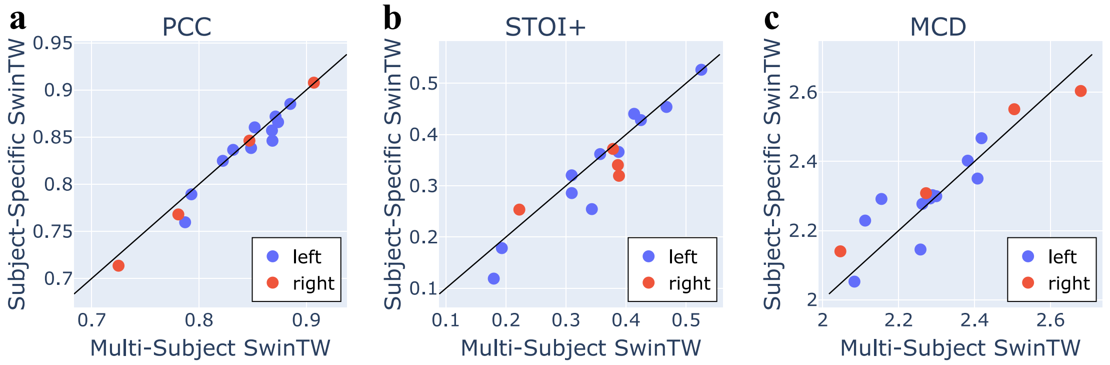

Figure 4. Comparison between a single SwinTW Neural Decoder trained with 8 × 8 ECoG data from training trials of multiple (15) subjects and 15 subject-specific SwinTW models. PCC, STOI+, and MCD were evaluated on test trials from the same 15 participants. Wilcoxon's two-sided signed-rank test is used to compare the two models. P-values of PCC, STOI, and MCD are 0.12, 0.06, 0.27.

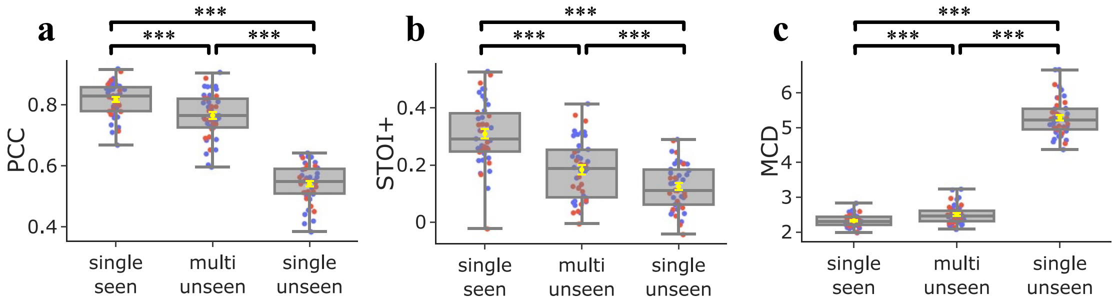

Figure 5. The decoding performance of the trained multi-subject models on ECoG participants outside the training set. Cross-validation was conducted on male and female subjects separately. **single seen** refers to subject-specific model performance on each single subject ($N = 43$), **multi unseen** refers to the multi-subject model tested on unseen subjects, **single unseen** refers to the top five subject-specific models (based on CC) tested on unseen subjects with the same gender. Distributions were compared with each other as indicated using the Wilcoxon two-sided signed-rank test for (a), (b), and (c). ***: $P < 0.001$. .

male ($n = 20$) and female ($n = 23$) participants. Specifically, we partitioned all male (resp. female) participants (with ECoG electrodes implanted in either the left or right brain hemisphere) into five folds. Each time, we used data from four-fold participants to train a SwinTW decoder and evaluate its decoding performance on the remaining one-fold participants. The process is repeated to use every fold as the validation fold once. As shown in figure 5, although the performance achieved for participants outside the training set is lower than the subject-specific models, the decoded speech still has

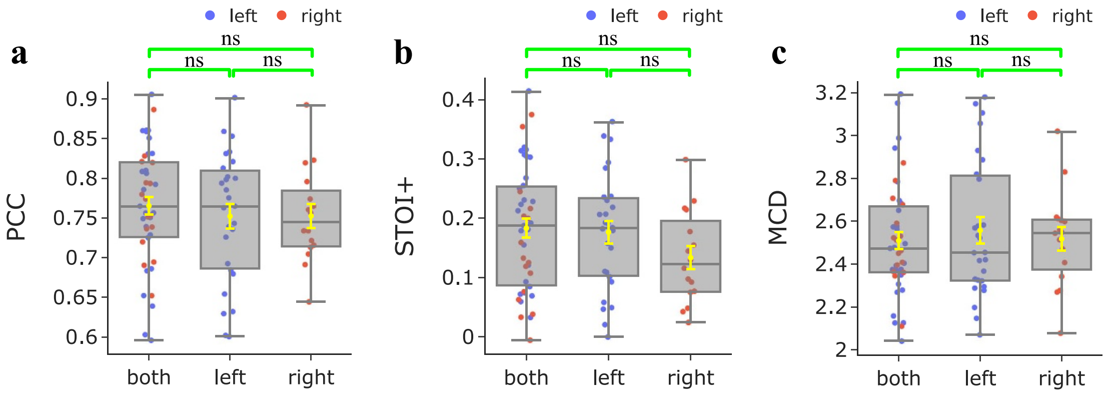

Figure 6. The comparison of speech decoding performance on unseen subjects between SwinTW trained on data from one hemisphere(left or right.) and SwinTW trained on data from both hemispheres. Models were trained separately for males and females. The results demonstrate that, compared with hemisphere-specific models, the SwinTW Neural Decoder trained on both hemispheres can achieve comparable performance when inference on unseen subjects. Box plots as described in figure 2. Distributions were compared with each other as indicated. Green brackets indicate two experiments that are compared using the Wilcoxon rank-sum test. ns: $P > 0.05$.

a high mean PCC of 0.765.We also tested the top-5 single-subject model (with top-5 performance based on CC on subject-specific experiments) on unseen subjects. More specifically, we selected three top-performing subject-specific models for female participants and evaluated each of them on unseen female ($n = 22$) subjects. Similarly, we selected two top-performing subject-specific models for male participants and tested each on unseen male ($n = 19$) subjects. The results indicate that subject-specific models perform significantly worse than multi-subject models on unseen participants. While the performance of **multi unseen** is lower than that of **single seen** (subject-specific models on the testing trials of the same participants used for training), it is still significantly better than the **single unseen** model (subject-specific models tested on other participants). hese results demonstrate that the proposed SwinTW decoder trained with multiple subject data can achieve generalizability for participants who are unseen during model training, while subject-specific models cannot.

To investigate if separate models should be trained for decoding from neural data in the left and right hemispheres, we performed additional experiments, where we trained and evaluated multi-subject models for the two hemispheres separately, each through cross-validation. Among male participants, there were 14 with left hemisphere data and 6 with right hemisphere data. For female participants, we had 13 with left hemisphere data and 10 with right hemisphere data. We used 5-fold cross-validation for training and evaluating each model. As shown in figure 6, compared with hemisphere-specific models, the SwinTW decoder trained using data from both hemispheres achieved comparable performance when tested on unseen subjects. This suggests that a single SwinTW model can effectively extract and synthesize information from both hemispheres for speech decoding. We have also added the audio demos containing sample decoded speech in the demo page.

4. Discussion

This study proposes a new Neural Decoder architecture, SwinTW, that does not have the grid-input assumption and can predict speech parameters from electrodes in any topological layout in the brain. The SwinTW removes the grid-based operations in the 3D Swin Transformer model used in our prior study [9] to make the model applicable for electrodes in any layout. Instead of relying on 2D grid indices to provide positional information about each electrode, the SwinTW relies on each electrode's position in the standardized brain coordinate (i.e. MNI) and the brain region that the electrode resides in to generate relative positional bias for self-attention. The SwinTW was used as the Neural Decoder in the speech decoding pipeline proposed in our previous work [9, 48] and was trained using the 2-step training pipeline in [9, 48]. Our proposed SwinTW Neural Decoder achieved superior performance over the Neural Decoders based on ResNet and 3D Swin Transformer in [9, 48], which can only work with ECoG data. As illustrated in figure 2, over 43 participants with low-density 8 × 8 ECoG electrodes, the SwinTW achieved higher mean PCC, STOI+, and lower MCD (PCC: 0.825, STOI+: 0.309, NCD: 2.341) than the ResNet (PCC:0.804, STOI+: 0.264, MCD: 2.374) and 3D Swin Transformer (PCC: 0.785, STOI+: 0.216, MCD: 2.425) using the same 64 electrodes from the 8 × 8 ECoG grid. We attribute SwinTW's better performance to its utilization of electrodes' locations on the brain cortex (the MNI coordinate and brain region information) rather than the 2D grid index.

Unlike ResNet and 3D Swin Transformer, the SwinTW does not rely on 2D grid indices of electrodes and can accommodate both ECoG electrodes, strip and depth electrodes, and additional grid electrodes. Our results demonstrate that leveraging the additional electrodes can improve speech decoding performance, as illustrated in figure 3. Specifically, for 39 subjects with additional active electrodes, the SwinTW utilizing the additional electrodes achieved better mean PCC (0.838), STOI+ (0.359), and MCD (2.228) compared with the SwinTW using grid electrodes only (PCC: 0.825, STOI+: 0.318, MCD: 2.341). The superior results indicate that the neural activity recorded by the additional electrodes contains complementary information for decoding speech.

Our results further demonstrate that the SwinTW can achieve high decoding quality based only on sEEG electrodes. Specifically, as shown in figure 3, for nine subjects with only sEEG electrodes implanted, we achieved mean PCC: 0.798, STOI+: 0.341, and MCD: 2.396. The mean and range of PCC are slightly lower than the decoding performance obtained using ECoG electrodes but significantly higher than previously reported decoding performance from sEEG only, ranging between 0.54 to 0.77 in mean PCC [2, 3, 23, 46]. It is notable that there is no statistical difference between the decoding performance from sEEG vs. from ECoG or all electrodes in terms of STOI+ and MCD, the metrics that are better indicators of the intelligibility of the decoded speech.

We present the distribution of electrode coverage in figure S6, alongside the electrode contribution analysis in figure S8, which quantifies the importance of each electrode across all participants. The methodology employed for the contribution analysis is detailed in [9]. The contribution analysis is conducted using subject-specific models. Despite the variation in electrode coverage across the grid, all-electrode, and sEEG cases, the contribution analysis reveals consistent patterns. Specifically, electrodes in the motor and temporal lobe regions exhibit greater contributions compared to those in other regions. Conversely, electrodes in other and unidentified regions show the lowest contributions, despite the relatively large number of electrodes present in these areas. This may explain why similar decoding performance can be achieved despite variations in electrode coverage. That is, the sEEG is sampling sufficient cortical regions as ECoG, thus leading to similar decoding performance.

Since the SwinTW directly uses the anatomical positions of electrodes rather than their grid indices, it can be trained with data from multiple subjects. As shown in figure 4, when evaluated on the testing trials of the participants in the training cohort and using only data on 8 × 8 grids, the resulting multi-subject model trained with data from 15 participants achieved statistically on par decoding performance (mean PCC: 0.837, STOI+:0.352, MCD: 2.307), compared with the SwinTW trained for each subject individually (mean PCC: 0.831, STOI+: 0.335, and MCD: 2.313). This implies that the SwinTW model structure is able to effectively deal with the significant variability in the electrode placements among patients and make use of electrodes' positions on the cortex. Previously, we have attempted to train ResNet and 3D Swin-based Neural Decoders using ECoG data from multiple participants. We obtained significantly worse decoding performance compared to subject-specific models. That is likely because ResNet and 3D Swin models rely on the electrodes' relative positions in the 2D grid. Because the ECoG grid is placed differently among the participants, the same relative difference in the 2D grid can be associated with very different anatomical positions in different participants, making using the grid index as positional information unsuitable when the data come from multiple participants.

Most significantly, the SwinTW model trained with multiple participants data demonstrated generalizability to participants outside the training cohorts, with high average decoding PCC (mean PCC = 0.765 over 43 unseen participants through a cross-validation study conducted separately for males and females). Figure 5 shows that the speech decoding performance achieved on unseen subjects overlaps significantly with that of the subject-specific model. Testing on unseen subjects presents a significantly more challenging task than evaluating the model on subjects whose data were used during training. Despite the increased difficulty, our SwinTW model still leads to decoding evaluation metrics that are significantly above chance (shown in figure S3). We consider this reasonable performance as a demonstration of the generalizability of our approach, even though the performance on unseen participants is lower than using subject-specific models. Furthermore, a model trained with data from both the left and right hemispheres performs on par with those trained using only the left or right hemisphere on unseen participants (figure 6). These results suggest that the SwinTW training using multiple participants' data can successfully learn how to handle differences among subjects based on electrode signals and the anatomical position of the electrodes. The success of the left and right hemispheres co-training demonstrates the strong learning capacity of the SwinTW. The two-hemisphere co-training also allows the Neural Decoder to fully leverage the whole dataset as we no longer need to train the model separately for each hemisphere.

To summarize, the SwinTW Neural Decoder can predict speech parameters from electrode signals and electrodes' positions on the brain cortex without requiring the electrodes to be arranged in a grid. The SwinTW Neural Decoder, in conjunction with our previously reported Speech Synthesizer, demonstrated superior speech decoding performance compared with our prior works based on ResNet and 3D Swin Transformers when only electrodes on a single ECoG array were used. Besides, the grid-free architecture of the SwinTW allows the model to leverage off-grid electrodes to improve speech decoding further. When using only sEEG data, the decoding performance was comparable with that using ECoG data. As explained in the Introduction, decoding speech from sEEG signals would have significant clinical advantages over using ECoG data. Furthermore, the SwinTW can be trained with data from multiple subjects regardless if the electrodes were implanted in the left or right brain hemispheres. The multi-subject SwinTW performed statistically on par with the subject-specific models for participants within the training cohort. Most importantly, our SwinTW trained with multiple participants' data demonstrated good generalizability to subjects outside the training cohorts, achieving high average decoding PCC.

We are one of the few studies ([11, 31, 43]) demonstrating speech-decoding models trained across multiple participants. However, these other prior works embed subject-specific layers in their model structures, and hence, the models need subject-specific data for training. To our knowledge, we are the first to design a framework that goes beyond subject-specific training without using subject-specific layers. Our result demonstrates the exciting possibility of developing speech prostheses without collecting subject-specific training data: We can train a reliable decoder with data from selected participants and then directly deploy the model to a new participant. Note that although our experiments on the multi-subject model only considered the ECoG grid data, we expect similar trends when using ECoG plus non-grid data or non-grid data only.

Notably, the proposed SwinTW Neural Decoder is not limited to being used with our speech synthesizer. It could potentially be used to decode other latent features, e.g. the HuBERT latent features [18], which can then drive a corresponding synthesizer [26]. The work in [32] successfully decoded speech with high word decoding accuracy by decoding to quantized HuBERT units using an RNN decoder from high-density ECoG signals of a single participant. However, the RNN structure cannot be trained with multi-subject data without introducing subject-specific layers. It will be interesting to explore the potential of training a SwinTW decoder using data from multiple participants with surface and/or depth electrodes, to map the neural signals to the HuBERT units and compare the decoding performance with the subject-specific RNN model or multi-subject RNN models with subject-specific layers.

One limitation of our current study is that the decoding performance for participants outside of the training cohorts is not consistently high. This could be potentially solved by including more participants in the training set when larger datasets become available. Furthermore, the SwinTW model structure can also be extended to include subject-specific layers for improved performance. We would explore training the non-subject-specific layers with a large precollected multi-subject dataset and refine only the subject-specific layer with a small amount of data for any new participants.

Data availability statement

The data of this study are available from the corresponding author upon request. Although all participants consented to share their data for research purposes, not all participants agreed to share their audio publicly. Given the sensitive nature of audio speech data, we will share data with researchers that directly contact the corresponding author and provide documentation that the data will be strictly used for research purposes and will comply with the terms of our study NYU Langone IRB. The data that support the findings of this study are openly available at the following URL/DOI: https://data.mendeley.com/datasets/fp4bv9gtwk/1.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-1912286 and IIS-2309057 (Y W, A F) and National Institute of Health R01NS109367, R01NS115929, R01DC018805 (A F).

Conflict of interests

The authors declare no conflict of interest.

ORCID iDs

Xupeng Chen https://orcid.org/0009-0003-3519-1863
Adeen Flinker https://orcid.org/0000-0003-1247-1283

References

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

[2] Angrick M *et al* 2022 Towards closed-loop speech synthesis from stereotactic EEG a unit selection approach *ICASSP 2022-2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 1296–300

[3] Angrick M *et al* 2021 Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity *Commun. Biol.* **4** 1055

[4] Anumanchipalli G K, Chartier J and Chang E F 2019 Speech synthesis from neural decoding of spoken sentences *Nature* **568** 493–8

[5] Ba J L, Kiros J R and Hinton G E 2016 Layer normalization (arXiv: 1607.06450)

[6] Boersma P and Van Heuven V 2001 Speak and unSpeak with PRAAT (available at: www.fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf)

[7] Brumberg J S, Nieto-Castanon A, Kennedy P R and Guenther F H 2010 Brain–computer interfaces for speech communication *Speech Commun.* **52** 367–79

[8] Chakrabarti S, Sandberg H M, Brumberg J S and Krusienski D J 2015 Progress in speech decoding from the electrocorticogram *Biomed. Eng. Lett.* **5** 10–21

[9] Chen X, Wang R, Khalilian-Gourtani A, Yu L, Dugan P, Friedman D, Doyle W, Devinsky O, Wang Y and Flinker A 2024 A neural speech decoding framework leveraging deep learning and speech synthesis *Nat. Mach. Intell.* **6** 467–80

[10] Chiluba B C 2019 Tackling disability of speech due to stroke: perspectives from stroke caregivers of the university teaching hospital in zambia *Indonesian J. Disability Stud.* **6** 215–22

[11] Défossez A, Caucheteux C, Rapin J, Kabeli O and King J-R 2023 Decoding speech perception from non-invasive brain recordings *Nat. Mach. Intell.* **5** 1097–107

[12] Dosovitskiy A *et al* 2020 An image is worth 16×16 words: transformers for image recognition at scale (arXiv: 2010.11929)

[13] Engel J, Hantrakul L, Gu C and Roberts A 2020 DDSP: differentiable digital signal processing (arXiv: 2001.04643)

[14] Graetzer S and Hopkins C 2021 Intelligibility prediction for speech mixed with white gaussian noise at low signal-to-noise ratios *J. Acoust. Soc. Am.* **149** 1346–62

[15] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8

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

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

[18] Hsu W-N, Bolte B, Tsai Y-H H, Lakhotia K, Salakhutdinov R and Mohamed A 2021 HuBERT: self-supervised speech representation learning by masked prediction of hidden units *IEEE/ACM Trans. Audio, Speech Lang. Process.* **29** 3451–60

[19] Iida K and Otsubo H 2017 Stereoelectroencephalography: indication and efficacy *Neurologia Med. Chirurgica* **57** 375–85

[20] Jacobs M and Ellis C 2023 Aphasianomics: estimating the economic burden of poststroke aphasia in the united states *Aphasiology* **37** 25–38

[21] Khalilian-Gourtani A, Wang R, Chen X, Yu L, Dugan P, Friedman D, Doyle W, Devinsky O, Wang Y and Flinker A 2024 A corollary discharge circuit in human speech *Natl. Acad. Sci.* **121** e2404121121

[22] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv: 1412.6980)

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

[24] Komeiji S, Shigemi K, Mitsuhashi T, Iimura Y, Suzuki H, Sugano H, Shinoda K and Tanaka T 2022 Transformer-based estimation of spoken sentences using electrocorticography *ICASSP 2022-2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 1311–5

[25] Kominek J, Schultz T and Black A W 2008 Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion

[26] Lakhotia K *et al* 2021 On generative spoken language modeling from raw audio *Trans. Assoc. Comput. Linguist.* **9** 1336–54

[27] Liu Z *et al* 2022 Swin transformer v2: scaling up capacity and resolution *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 12009–19

[28] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 Swin transformer: hierarchical vision transformer using shifted windows *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 10012–22

[29] Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S and Hu H 2022 Video swin transformer *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 3202–11

[30] Luo S, Rabbani Q and Crone N E 2022 Brain-computer interface: applications to speech decoding and synthesis to augment communication *Neurotherapeutics* **19** 263–73

[31] Makin J G, Moses D A and Chang E F 2020 Machine translation of cortical activity to text with an encoder–decoder framework *Nat. Neurosci.* **23** 575–82

[32] Metzger S L, *et al* 2023 A high-performance neuroprosthesis for speech decoding and avatar control *Nature* **620** 1037–46

[33] Moses D A, Leonard M K, Makin J G and Chang E F 2019 Real-time decoding of question-and-answer speech dialogue using human cortical activity *Nat. Commun.* **10** 3096

[34] Moses D A *et al* 2021 Neuroprosthesis for decoding speech in a paralyzed person with anarthria *New Engl. J. Med.* **385** 217–27

[35] Nicholas L E and Brookshire R H 1995 Comprehension of spoken narrative discourse by adults with aphasia, right-hemisphere brain damage, or traumatic brain injury *Am. J. Speech-Lang. Pathology* **4** 69–81

[36] Radford A, Kim J W, Xu T, Brockman G, McLeavey C and Sutskever I 2023 Robust speech recognition via large-scale weak supervision *Int. Conf. on Machine Learning* (PMLR) pp 28492–518

[37] Ramsey N F, Salari E, Aarnoutse E J, Vansteensel M J, Bleichner M G and Freudenburg Z 2018 Decoding spoken phonemes from sensorimotor cortex with high-density ECOG grids *Neuroimage* **180** 301–11

[38] Schafer R W 2011 What is a savitzky-golay filter? [lecture notes] *IEEE Signal Process. Mag.* **28** 111–7

[39] Schultz T, Wand M, Hueber T, Krusienski D J, Herff C and Brumberg J S 2017 Biosignal-based spoken communication: a survey *IEEE/ACM Trans. Audio, Speech Lang. Process.* **25** 2257–71

[40] Senda J, Tanaka M, Iijima K, Sugino M, Mori F, Jimbo Y, Iwasaki M and Kotani K 2024 Auditory stimulus reconstruction from ECOG with dnn and self-attention modules *Biomed. Signal Process. Control* **89** 105761

[41] Shigemi K, Komeiji S, Mitsuhashi T, Iimura Y, Suzuki H, Sugano H, Shinoda K, Yatabe K and Tanaka T 2023 Synthesizing speech from ECOG with a combination of transformer-based encoder and neural vocoder *ICASSP 2023-2023 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 1–5

[42] Shum J, Fanda L, Dugan P, Doyle W K, Devinsky O and Flinker A 2020 Neural correlates of sign language production revealed by electrocorticography *Neurology* **95** e2880–9

[43] Sun P, Anumanchipalli G K and Chang E F 2020 Brain2Char: a deep architecture for decoding text from brain recordings *J. Neural Eng.* **17** 066015

[44] Tandon N, Tong B A, Friedman E R, Johnson J A, Von Allmen G, Thomas M S, Hope O A, Kalamangalam G P, Slater J D and Thompson S A 2019 Analysis of morbidity and outcomes associated with use of subdural grids vs stereoelectroencephalography in patients with intractable epilepsy *JAMA Neurol.* **76** 672–81

[45] Thomas R, O'Connor A M and Ashley S 1995 Speech and language disorders in patients with high grade glioma and its influence on prognosis *J. Neuro-Oncol.* **23** 265–70

[46] Verwoert M, Ottenhoff M C, Goulis S, Colon A J, Wagner L, Tousseyn S, van Dijk J P, Kubben P L and Herff C 2022 Dataset of speech production in intracranial electroencephalography *Sci. Data* **9** 434

[47] Wang R, Chen X, Khalilian-Gourtani A, Chen Z, Yu L, Flinker A and Wang Y 2020 Stimulus speech decoding from human cortex with generative adversarial network transfer learning *2020 IEEE 17th Int. Symp. on Biomedical Imaging (ISBI)* (IEEE) pp 390–4

[48] Wang R, Chen X, Khalilian-Gourtani A, Yu L, Dugan P, Friedman D, Doyle W, Devinsky O, Wang Y and Flinker A 2023 Distributed feedforward and feedback cortical processing supports human speech production *Proc. Natl Acad. Sci.* **120** e2300255120

[49] Willett F R *et al* 2023 A high-performance speech neuroprosthesis *Nature* **620** 1031–6

[50] Wu X, Wellington S, Fu Z and Zhang D 2024 Speech decoding from stereo-electroencephalography (seeg) signals using advanced deep learning methods *J. Neural Eng.*. **21** 036055