Subject-Agnostic Transformer-Based Neural Speech Decoding from Surface and Depth Electrode Signals

Junbo Chen* **(jc7489@nyu.edu)**
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Xupeng Chen* **(xc1490@nyu.edu)**
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Ran Wang (rw1691@nyu.edu)
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Chenqian Le (cl6707@nyu.edu)
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Amirhossein Khalilian-Gourtani (akg404@nyu.edu)
Department of Neurology, New York University
223 East 34th Street, Manhattan, NY 10016

Adeen Flinker (adeen.flinker@nyulangone.org)
Department of Neurology, New York University
223 East 34th Street, Manhattan, NY 10016

Yao Wang† **(yaowang@nyu.edu)**
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

∗ These authors contributed equally
† Corresponding author

Abstract

A growing body of research is aimed at decoding human speech from neural signals captured by intracranial electrodes. Most prior works with high decoding quality can only work with electrodes on a 2D grid (i.e., Electrocorticographic or ECoG array) and data from a single patient. Here we design a deep-learning model that accommodates surface (ECoG) and depth (stereotactic EEG or sEEG) electrodes from multiple participants with large variability in electrode placements. The proposed novel transformer-based model named SwinTW can work with arbitrarily positioned electrodes. We train subject-specific and subject-agnostic models exploiting data from multiple participants. The subject-specific models using only low-density ECoG achieved high decoding performance, outperforming our previous ResNet model [1]. Incorporating additional strip and depth electrodes led to further improvement. For participants with only sEEG electrodes, subject-specific models still enjoy comparable performance.The subject-agnostic models generalized well to unseen participants through a cross-validation study. The proposed SwinTW decoder enables future speech neuroprostheses to utilize any electrode placement that is clinically optimal or feasible for a particular participant, including using depth electrodes, which are more routinely implanted in chronic neurosurgical procedures. Importantly, the generalizability of the multi-patient models suggests the exciting possibility of developing speech neuroprostheses for people with speech disability without relying on their own neural data for training.

Keywords: Neural Speech Decoding; Electrocorticographic (ECoG); stereotactic EEG; Brain-Computer Interface(BCI)

Introduction

Brain-related speech disability, which can be caused by stroke, injury, or tumor [2, 3], can seriously decrease a patient's quality of life. There has been growing interest in developing approaches to directly decode human speech from the neural signals recorded using intracranial electrodes, for future adoption as a Brain-Computer Interface to allow patients with speech disabilities to communicate [4, 5, 6].

Recent studies have explored the use of RNNs, wavenet vocoders, GANs, and HuBERT synthesizers for neural speech decoding, and their effectiveness varies regarding intelligibility and fidelity [7, 8, 9, 10, 11]. We have reported speech decoding with high correlation with actual speech spectrograms for 43 participants with low-density ECoG and 5 participants with hybrid density ECoG [12, 1]. Our decoding pipeline consists of an ECoG Decoder and a Speech Synthesizer, with the ECoG Decoder utilizing ResNet [13] and 3D Swin Transformer [14] architectures. Despite these advances, the application to non-grid electrode configurations remains challenging due to the dependency of prior approaches on spatial convolutions and positional embeddings specific to grid indices [1, 12, 15]. Additionally, sEEG, which uses depth electrodes with minimal cranial disruption, is compatible with DBS methods for potential long-term use in speech neuroprosthetics [16, 17, 18]. It also requires a non-grid-based neural decoder. Previous research on decoding from sEEG data has produced relatively low decoding accuracy[19, 20, 21, 22].

The reliance of fully connected models on specific electrode placements reduces a neural decoder's generalizability across patients, necessitating tailored datasets per subject and limiting scalability [7, 8, 23]. Our research introduces the Swin transformer with temporal windowing (SwinTW), a novel transformer-based Neural Decoder that does not depend on grid structures. By utilizing the anatomical locations of electrodes rather than grid indices, SwinTW surpasses both ResNet and 3D Swin Transformer in performance on grid electrodes, as well as enhancing performance with off-grid electrodes [1]. Significantly, the SwinTW model, trained on data from multiple participants, generalizes effectively to new subjects.

Methodology

The study includes 52 native English-speaking subjects (43 subjects with ECoG electrodes and 9 subjects with only sEEG electrodes) with refractory epilepsy[1]. Participants vocalized 50 target words in response to prompts across five tasks: Auditory Repetition, Auditory Naming, Sentence Completion, Visual Reading, and Picture Naming. This protocol yielded 400 speech production trials per participant with an average duration of 500ms. Electrode configurations included 8x8 ECoG grids with additional strips or depth electrodes as necessary. Preprocessing involved isolating the high gamma band (70-150 Hz). Data exclusion criteria are applied to any channels displaying artifacts or epileptiform activity. For subject-specific models, 350 trials were used for training models, and 50 were reserved for testing across the tasks.

Our neural speech decoding framework adopts a two-step training approach, outlined in [1] and illustrated in Fig.1. Initially, a Speech Encoder extracts temporal speech parameters from input spectrograms, followed by a Speech Synthesizer that reconstructs these parameters into spectrograms. Subsequently, a Neural Decoder predicts these speech parameters from neural signals, synthesizing them into speech spectrograms.

We introduce the Swin Transformer with Temporal Windowing (SwinTW) as a novel Neural Decoder capable of processing signals from arbitrarily positioned electrodes. This advancement allows SwinTW to handle diverse electrode configurations, effectively decoding speech from neural signals across various setups as shown in Fig.1. SwinTW significantly departs from traditional grid-based decoders by temporally partitioning electrode signals into tokens. 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 patch embedding layer is applied to each patch to generate $\frac{T}{W} \times N$ tokens with the latent dimension of C. Instead of grid-based spatial positional biases [24], SwinTW utilizes the anatomical locations of the electrodes, including their coordinates on the standardized Montreal Neurological Institute (MNI) brain map and their brain regions to assign relative positional biases for the electrodes. This approach enables SwinTW to learn embeddings for different brain regions, improving neural decoding accuracy across electrode placements and brain anatomies.

SwinTW supports neural speech decoding regardless of electrode placement. This model can be trained on data from multiple subjects and applied across different individuals, as shown in Fig.1. It exploits neural signals and electrodes' positional information to generate speech parameters. During training, a reference loss is calculated by comparing these parameters to those derived from the corresponding spectrograms. These parameters are then used by patient-specific speech synthesizers (with parameters learned only from audio signals) to produce speech spectrograms. The frame-

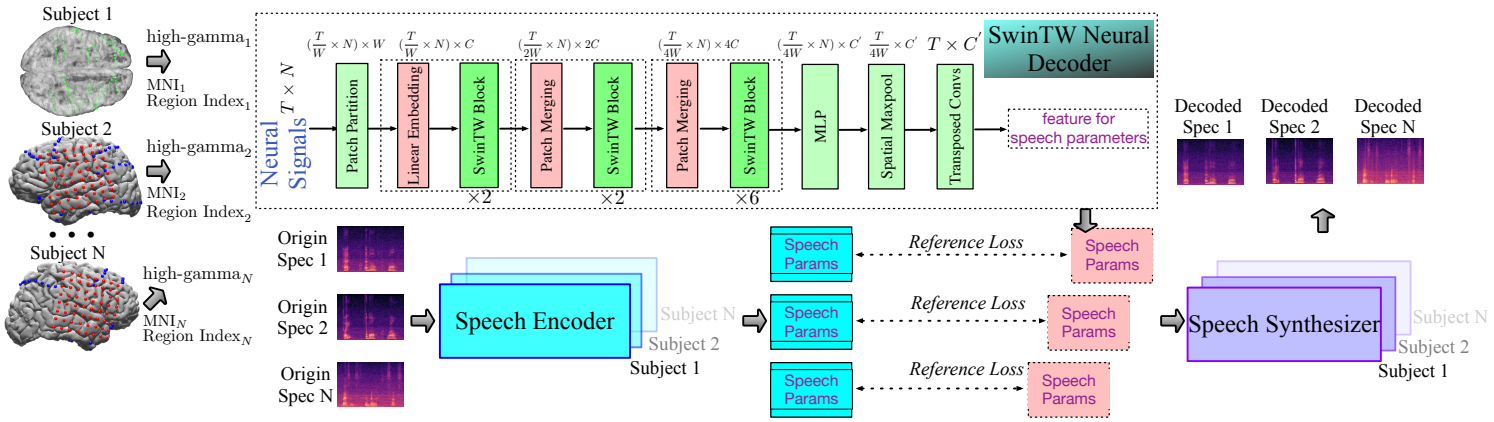

Figure 1: Neural Speech Decoding with SwinTW. Step 1 involves Speech-to-Speech Training using a subject-specific Speech Encoder to generate speech parameters from original spectrograms. Step 2 employs the SwinTW Neural Decoder which utilizes transformer blocks with spatial-temporal attention and temporal windowing for feature extraction. It integrates each participant's neural data and electrode locations (MNI coordinates and ROI index) to predict speech parameters, which are then processed by a participant-specific Speech Synthesizer to reconstruct the speech spectrogram.

work also includes learning embeddings for different brain regions and relative attention biases between two electrodes based on their MNI coordinates and region embeddings, thereby increasing the decoder's adaptability to diverse electrode placements.

Results and Discussion

We first evaluate SwinTW against the previous grid-based Neural Decoders based on ResNet and 3D Swin Transformer architectures [1], trained separately for each of the 43 participants using data from a single 8x8 ECoG grid. SwinTW demonstrated superior performance in terms of the Pearson Correlation Coefficient (PCC) between decoded and actual spectrograms, as shown in Fig.2**a**, underscoring the importance of incorporating MNI coordinates and brain region information of electrodes even when the electrodes are on a 2D grid.

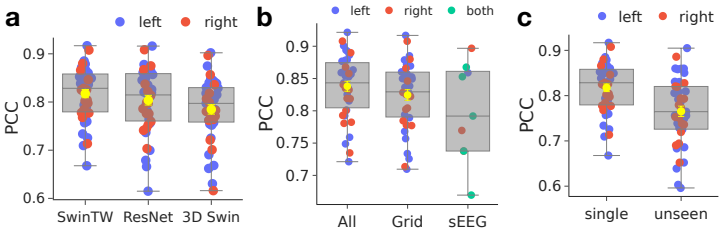

Figure 2: **(a)** Subject-specific models using 8 x 8 grid electrodes across 43 participants; SwinTW has the highest PCC. **(b)** Subject-specific SwinTW Neural Decoder using all electrodes outperforms using only grid electrodes across 39 participants. sEEG-only decoding over 9 participants also indicates the SwinTW can achieve promising speech decoding from sEEG electrode data. **(c)** The decoding performance of the trained multi-subject model on participants outside the training set shows the generalization ability of SwinTW through five-fold cross-validation across 43 participants. Left, right, and both indicate electrode placement on the corresponding brain hemispheres.

Furthermore, for 39 participants with additional electrodes (strip and depth electrodes), active electrodes were selected based on signal variance criteria detailed in [25]. Leveraging these additional electrodes (1 to 19 strip electrodes, 1 to 21 depth electrode per subject) along with the 64 grid electrodes, the subject-specific SwinTW models achieved improved decoding accuracy, as shown in Fig.2**b**. This capability to utilize diverse electrode types without requiring a grid configuration allows for a more versatile application across different participants. We further trained the SwinTW model on sEEG data alone over 9 participants. Electrodes were selected following [25], resulting in an electrode count ranging from 19 to 178 per participant. As shown in Fig.2**c**, the SwinTW model's sEEG-based decoding exhibits promising results, with PCCs marginally reduced compared to using ECoG electrodes.

Since SwinTW utilizes the brain locations of electrodes rather than their positions on a 2D grid, a subject-agnostic model can be trained with data from multiple participants. We assess the generalization ability of the multi-subject SwinTW decoder on unseen participants using a 5-fold cross-validation approach for male (N=20) and female (N=23) participants, each with ECoG electrodes in either left or right brain hemispheres. Participants were divided into five groups, each sequentially serving as the test set while the decoder was trained on the remaining four groups. Despite lower performance on test participants compared to subject-specific models, the SwinTW decoder demonstrated a respectable mean PCC of 0.765, indicating effective generalization to new subjects shown in Fig.2**c**.

This study introduces the Swin Transformer with Temporal Windowing (SwinTW), a novel Neural Decoder that overcomes the grid-input constraints of traditional models like the 3D Swin Transformer and ResNET by using the MNI coordinates and brain regions of electrodes for generating positional biases in a two-step speech decoding pipeline [1, 12]. SwinTW achieves a higher mean PCC than predecessors like ResNet and 3D Swin Transformer on ECoG signals only. SwinTW can effectively exploit neural signals captured by additional electrodes, including strip, depth, and grid electrodes, and obtain improved decoding accuracy with these additional electrodes. Using only sEEG data yields decoding performance comparable to that achieved through ECoG data, conferring substantial clinical benefits as outlined in the Introduction.

SwinTW's design facilitates effective training across multiple subjects. It is, to our knowledge, the first study demonstrating neural speech decoding models trained across multiple participants and generalized well to unseen participants. These results demonstrate SwinTW's capacity to handle diverse electrode setups and its potential for speech decoding applications that do not require subject-specific calibration.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. IIS-2309057 (Y.W., A.F.) and National Institute of Health R01NS109367, R01NS115929, R01DC018805 (A.F.).

References

[1] Xupeng Chen, Ran Wang, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*, pages 1–14, 2024.

[2] Brian Chanda Chiluba. Tackling disability of speech due to stroke: Perspectives from stroke caregivers of the university teaching hospital in zambia. *Indonesian Journal of Disability Studies*, 6(2):215–222, 2019.

[3] Linda E Nicholas and Robert H Brookshire. Comprehension of spoken narrative discourse by adults with aphasia, right-hemisphere brain damage, or traumatic brain injury. *American Journal of Speech-Language Pathology*, 4(3):69–81, 1995.

[4] Shiyu Luo, Qinwan Rabbani, and Nathan E Crone. Brain-computer interface: applications to speech decoding and synthesis to augment communication. *Neurotherapeutics*, 19(1):263–273, 2022.

[5] Jonathan S Brumberg, Alfonso Nieto-Castanon, Philip R Kennedy, and Frank H Guenther. Brain–computer interfaces for speech communication. *Speech communication*, 52(4):367–379, 2010.

[6] David A Moses, Matthew K Leonard, Joseph G Makin, and Edward F Chang. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications*, 10(1):3096, 2019.

[7] Joseph G Makin, David A Moses, and Edward F Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.

[8] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[9] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.

[10] Ran Wang, Xupeng Chen, Amirhossein Khalilian-Gourtani, Zhaoxi Chen, Leyao Yu, Adeen Flinker, and Yao Wang. Stimulus speech decoding from human cortex with generative adversarial network transfer learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 390–394. IEEE, 2020.

[11] Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, pages 1–10, 2023.

[12] Ran Wang, Xupeng Chen, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. Distributed feedforward and feedback cortical processing supports human speech production. *Proceedings of the National Academy of Sciences*, 120(42):e2300255120, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Koji Iida and Hiroshi Otsubo. Stereoelectroencephalography: indication and efficacy. *Neurologia medico-chirurgica*, 57(8):375–385, 2017.

[17] Nitin Tandon, Brian A Tong, Elliott R Friedman, Jessica A Johnson, Gretchen Von Allmen, Melissa S Thomas, Omotola A Hope, Giridhar P Kalamangalam, Jeremy D Slater, and Stephen A Thompson. Analysis of morbidity and outcomes associated with use of subdural grids vs stereoelectroencephalography in patients with intractable epilepsy. *JAMA neurology*, 76(6):672–681, 2019.

[18] Christian Herff, Dean J Krusienski, and Pieter Kubben. The potential of stereotactic-eeg for brain-computer interfaces: current progress and future directions. *Frontiers in neuroscience*, 14:123, 2020.

[19] Miguel Angrick, Maarten Ottenhoff, Lorenz Diener, Darius Ivucic, Gabriel Ivucic, Sophocles Goulis, Albert J Colon, Louis Wagner, Dean J Krusienski, Pieter L Kubben, et al. Towards closed-loop speech synthesis from stereotactic eeg: a unit selection approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1296–1300. IEEE, 2022.

[20] M Angrick, MC Ottenhoff, L Diener, D Ivucic, G Ivucic, S Goulis, J Saal, AJ Colon, L Wagner, DJ Krusienski, et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. commun biol 4 (1): 1055–1055, 2021.

[21] Jonas Kohler, Maarten C Ottenhoff, Sophocles Goulis, Miguel Angrick, Albert J Colon, Louis Wagner, Simon Tousseyn, Pieter L Kubben, and Christian Herff. Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework. *arXiv preprint arXiv:2111.01457*, 2021.

[22] Maxime Verwoert, Maarten C Ottenhoff, Sophocles Goulis, Albert J Colon, Louis Wagner, Simon Tousseyn, Johannes P van Dijk, Pieter L Kubben, and Christian Herff. Dataset of speech production in intracranial electroencephalography. *Scientific data*, 9(1):434, 2022.

[23] Jyun Senda, Mai Tanaka, Keiya Iijima, Masato Sugino, Fumina Mori, Yasuhiko Jimbo, Masaki Iwasaki, and Kiyoshi Kotani. Auditory stimulus reconstruction from ecog with dnn and self-attention modules. *Biomedical Signal Processing and Control*, 89:105761, 2024.

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[25] Amirhossein Khalilian-Gourtani, Ran Wang, Xupeng Chen, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A corollary discharge circuit in human speech. *BioRxiv*, pages 2022–09, 2022.