An Interpretable Deep Learning Model for Speech Activity Detection Using Electrocorticographic Signals

Morgan Stuart, Srdjan Lesaja, Jerry J. Shih, Tanja Schultz, *Fellow, IEEE*, Milos Manic, *Fellow, IEEE*, and Dean J. Krusienski, *Senior Member, IEEE*

***Abstract*—Numerous state-of-the-art solutions for neural speech decoding and synthesis incorporate deep learning into the processing pipeline. These models are typically opaque and can require significant computational resources for training and execution. A deep learning architecture is presented that learns input bandpass filters that capture task-relevant spectral features directly from data. Incorporating such explainable feature extraction into the model furthers the goal of creating end-to-end architectures that enable automated subject-specific parameter tuning while yielding an interpretable result. The model is implemented using intracranial brain data collected during a speech task. Using raw, unprocessed timesamples, the model detects the presence of speech at every timesample in a causal manner, suitable for online application. Model performance is comparable or superior to existing approaches that require substantial signal preprocessing and the learned frequency bands were found to converge to ranges that are supported by previous studies.**

***Index Terms*—Brain-Computer Interfaces (BCIs), deep learning, electroencephalography.**

I. Introduction

BRAIN-COMPUTER Interfaces (BCIs) hold the potential for a direct connection to thoughts and intentions, as well as direct neural control of external devices [1]. Due to superior spatial resolution and spectral bandwidth, invasive BCIs have advantages over non-invasive BCIs for more intricate direct neural control applications. Electrocorticography (ECoG) is an invasive measurement of the electrical potentials generated from the neocortex of the brain [2]. ECoG signals have been shown to successfully control the movement of an upper-limb neuroprosthetic [3] or typing interface [4], as well as decoding speech processes [5].

In the last decade, neural speech decoding systems have made significant progress, including describing brain regions and mechanisms involved in speech, predicting words or phonemes, and translating neural signals to articulatory kinematics, text, or directly to speech waveforms [6], [7], [8], [9], [10], [11], [12]. Recent efforts have progressed to real-time decoding and synthesis of overt and imagined speech [13], [14], [15], [16], [17], [18]. While these studies primarily focus on broadband gamma activity (~70-250 Hz), recent studies have shown that traditional lower-band frequencies (~0-50 Hz) also contain relevant and complementary information for speech decoding [19].

Deep learning has been demonstrated to be an effective method for decoding speech from ECoG signals and its inclusion in the decoding and synthesis pipeline has increased in recent years [12], [16], [20], [21]. Although an end-to-end architecture may eventually be wholly effective with sufficient training data, some current approaches have adopted a modular scheme with several sequential component models, each configured for a specific aspect of the speech decoding process [15], [16], [22].

Regardless of the specific approach, the overarching goal is to decode imagined or attempted speech directly from brain signals to provide an alternate communication channel for those who have lost the ability to speak. Here, the goal is not to maximize a metric for the quality of speech decoding. Instead, the approach is conceived from the perspective of identifying brain activity associated with intervals of intended speech output, with the ultimate objective of reliably detecting activity associated with imagined speech.

The present work introduces a component model, SincIEEG, based on a convolutional neural network (CNN) architecture developed for the task of speech activity detection [23]. The model is designed as a gateway, constantly monitoring brain activity to identify the segments pertinent to speech production. These detected segments can then be sent to downstream models for subsequent speech decoding

Manuscript received 5 November 2021; revised 23 May 2022 and 2 August 2022; accepted 5 September 2022. Date of publication 19 September 2022; date of current version 7 October 2022. This work was supported by NSF under Grant 2011595/1608140 and Grant BMBF 01GQ1602. *(Morgan Stuart and Srdjan Lesaja contributed equally to this work.) (Corresponding author: Srdjan Lesaja.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the institutional review boards of Virginia Commonwealth University under Application No. HM20014564; University of California, San Diego; Old Dominion University; and Mayo Clinic, Florida.

Morgan Stuart and Milos Manic are with the Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284 USA (e-mail: stuartms@vcu.edu; misko@ieee.org).

Srdjan Lesaja and Dean J. Krusienski are with the Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA 23284 USA (e-mail: slesaja@vcu.edu; djkrusienski@vcu.edu).

Jerry J. Shih is with the Neurology Department, UCSD Health, San Diego, CA 92093 USA (e-mail: jerryshih@ucsd.edu).

Tanja Schultz is with the Cognitive Systems Laboratory, University of Bremen, 28359 Bremen, Germany (e-mail: tanja.schultz@uni-bremen.de).

Digital Object Identifier 10.1109/TNSRE.2022.3207624

and synthesis. SincIEEG, unlike a traditional CNN, learns a set of bandpass filter coefficients at its input layer. This provides several advantages over a traditional CNN since the number of required model parameters is significantly reduced by comparison, making it computationally efficient in terms of training and implementation. This compactness allows for flexibility without increasing the optimization problem. Moreover, unlike most traditional CNNs, the SincIEEG model has the distinct advantage of yielding interpretable parameters. The bandpass filters learned by SincIEEG can be visualized and equated to conventional spectral brain features.

The results demonstrate that SincIEEG is capable of detecting the presence or absence of speech during each time interval with a high level of accuracy, and compare the model's performance to a traditional CNN model, as well as non-deep learning methods. In addition, the generalizability of the model architecture is highlighted in terms of providing empirical, interpretable insights about the discriminable bandpass spectral features for any physiological data that can be represented as an aggregate of bandpass activity.

II. Materials and Methods

A. Participants

ECoG data were recorded from 5 participants with pharmacoresistant epilepsy undergoing clinical monitoring for surgical planning. No participants reported hearing deficits. In all cases, a tumor was not the source for the seizures and no lesions were indicated by any electrode used for analysis. All participants gave written informed consent and the study protocol was approved by the institutional review boards of Virginia Commonwealth University; University of California, San Diego; Old Dominion University; and Mayo Clinic, Florida.

Participants were implanted with subdural electrode grids or strips (Ad-Tech Medical Instrument Corporation, 1-cm spacing) based purely on their clinical need. Electrode locations were verified by co-registering preoperative MRI and postoperative computerized tomography scans. For combined visualization, electrode locations were projected to common Talairach space. Electrode locations were rendered using NeuralAct [24], as shown in Figure 1. While brain areas associated with speech are predominantly found on the dominant hemisphere, which is the left hemisphere in the majority of right-hand dominant people, the neural correlates of speech production are not exclusively localized in the left hemisphere [25], [26]. For this reason, both left and right hemisphere cases are evaluated. In total, ECoG activity was recorded from 416 (96 left hemisphere, 320 right hemisphere) subdural electrodes. Of these, electrodes that exhibited unnatural signal anomalies based on visual inspection were excluded from the analysis, leaving 364 electrodes (96 left hemisphere, 268 right hemisphere). For each participant, the number of electrodes implanted, analyzed, and identified as not located over the auditory cortex (non-auditory) are provided in Table I.

B. Task

Participants were instructed to read aloud single words presented in sequence on a computer screen while their brain activity and voice were simultaneously recorded. The words were selected from a bank of 431 unique words, split into 4 sets of 115-116 words. The bank of words are primarily monosyllabic and comprised of the Modified Rhyme Test [27], supplemented with additional words to better reflect the phoneme distribution of American English [28]. While this experimental paradigm was originally designed to examine neural correlates of American English phonemes [7], the data are being used in the present analysis exclusively for speech activity detection without consideration of phonetic aspects.

The experiment begins with a fixation cross at the center of the screen. The cross is then replaced by a word that stays on the screen for 2.5 seconds. The word is then replaced with the cross for 0.5 seconds, before the next word is presented. Words are chosen randomly from the set of 115 words for each session and each session contained a different subset of words. Participants completed between 2 and 4 sessions, depending on willingness and ability to complete the sessions.

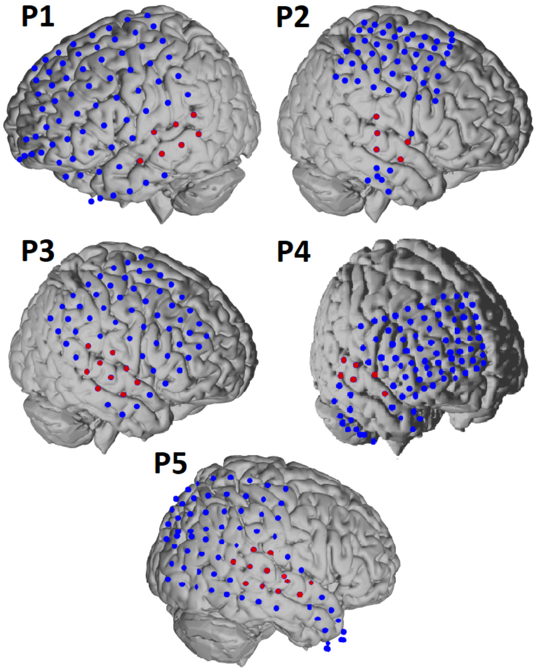

Fig. 1. Electrode locations for all 5 participants. Electrodes identified in the auditory cortex region are highlighted in red.

TABLE I
Electrodes by Participant

Participant	Implanted	Analyzed	Non-Auditory
1	96	96	89
2	64	51	49
3	64	55	48
4	96	77	73
5	96	85	75
Total	416	364	334

C. Data Acquisition

ECoG and audio data were concurrently recorded during the task. ECoG data were bandpass filtered between

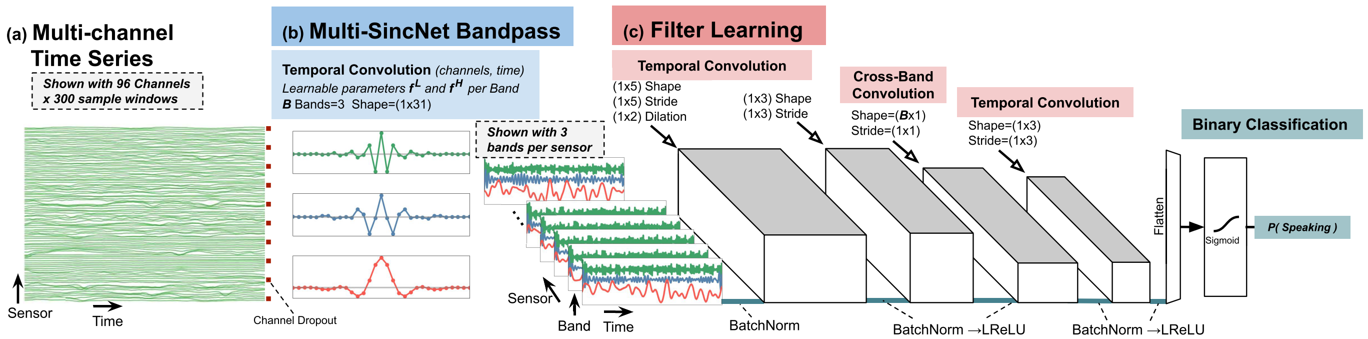

Fig. 2. The **SincIEEG** deep learning architecture: a classification model composed of a Multi-SincNet input layer and multiple subsequent convolutional layers. **(a)** SincIEEG takes raw multi-channel ECoG time series data as input, with channel dropout for improved regularization. **(b)** Multi-SincNet learns bandpass filter parameters to decompose the input signal - illustrated here with three pass-bands. **(c)** The filtered signals are normalized with respect to the band dimension using spatial normalization before convolutional layers learn kernels across time and pass-bands. All hidden layers use batch normalization for regularization and *Leaky Rectified Linear Units* for activation. The model predicts the likelihood of speaking using a *Sigmoid* activation at its output layer.

0.5 and 500 Hz, notch filtered at 60 Hz and recorded using g.USB amplifiers (g.tec Medical Engineering). The data were recorded at a sampling rate of 1200 Hz and subsequently decimated to 600 Hz.

The time series and its frequency spectra were visually inspected for anomalies. Channels having uncharacteristic frequency spectra, substantial artifacts, and/or saturated amplitudes, were excluded from the analysis. In total, 364 (96 left hemisphere, 268 right hemisphere) electrodes were used for analysis.

This basic preprocessing is standard for ECoG acquisition and the data decimation can be equivalently achieved by using a lower sampling rate at the time of data acquisition. Thus, the data used as input to the SincIEEG network effectively represent the raw ECoG timesamples.

Audio data were recorded in parallel using a Blue Microphones Snowball iCE USB microphone connected to the research computer, sampled at 48 kHz. All data recording and stimulus presentation were facilitated by BCI2000 software [29].

D. Speech Labeling

Speech labels used for training the model were made in reference to the stimulus cue of the word being presented in the experiment. Every time-sample from 0.5 seconds after the word presentation cue to 1.5 seconds after the cue were labeled as ‘speaking’. Every time-sample from 2.0 seconds after the word presentation cue to 3.0 seconds after the cue were labeled as ‘not-speaking’. The other segments, from the cue to 0.5 seconds after, and from 1.5 to 2.0 seconds after, were purposefully left unlabeled.

This labeling scheme was chosen based on the stimulus presentation cue, as opposed to direct energy detection in the audio signal, to develop a more robust model that does not directly rely upon the acoustic signal. This was done to emulate the scenario where the user is unable to speak, and thus precise labels for the presence or absence of speech would be unavailable. Instead, the proposed labeling indicates the time segments where speech is most expected, which can be generalized to imagined speech.

III. Model Design and Optimization

The SincIEEG model is a Multi-SincNet based convolutional deep learning architecture adapted for real-time detection of human speech from ECoG input signals. Proposed in [30] for hand-pose classification from myoelectric sensor readings, and based off the work from [23], the Multi-SincNet architecture learns the coefficients of a set of parallel finite impulse response (FIR) bandpass filters, applied across the input channels. Subsequent convolutional layers learn kernels that aggregate across time and bandpass frequency dimensions. A final global view, established by a fully connected layer and sigmoid activation, classifies either ‘speaking’ or ‘not-speaking’ from labeled data. Figure 2 illustrates the SincIEEG model and its layer configurations. This section details the architecture and training strategy to produce models for validation described in Section IV.

In overview, the inputs to the model are 500 ms windows of raw IEEG data (300 time samples) with a stride of 2 ms (1 time sample). Each 500 ms window represents one training sample for the model, described in Section II-D. A model was trained for each participant, using all of the quality electrodes available. Electrodes over the auditory cortex were excluded for a model validation check, detailed in Section IV-C.2. A K-fold training methodology was used and is detailed further in Section III-E.

This architecture was developed and implemented using Pytorch [31] deep learning Python library. Other critical software libraries used for development and discovery include matplotlib [32], numpy [33], pandas [34], [35], seaborn [36], and SciPy [37].

A. Multi-SincNet Input Convolution

The first layer in the SincIEEG model is a Multi-SincNet layer, an extension to the the Kaldi speech framework’s [38] SincNet, which applies a SincNet to each of the incoming sensor channels. A SincNet layer learns a configurable number of bandpass filters, parameterized through two cutoff frequencies, f_L and f_H. The Multi-SincNet layer can therefore be used to decompose a collection input signals into a fixed set of learned bands.

In equations 1 and 2, multiple filters are conceptualized as vectors of low and high cutoffs, F_L and F_H respectively, identifying regions of the input's spectrum that the model uses for classification. These vectors are a parameterization of a SincNet layer, which is shared in the experiments across all sensors $s \in S$.

$$F_L = \{f_0^L, f_1^L, \ldots, f_{i=B-1}^L\} \in \mathbb{R}^+ \tag{1}$$

$$F_L = \{f_L^0, f_L^1, \ldots, f_L^{i=B-1}\} \in \mathbb{R}^+ \tag{2}$$

$$K : (f_L, f_H, f_s) \mapsto \mathbb{R}^W \tag{3}$$

$$\text{SincNet}(F_L, F_H) = \{K(F_L(i), F_H(i))\} \tag{4}$$

$$\text{Multi-SincNet} = \text{SincNet}_{\mathrm{F_L,F_H}}(s) \quad s \in S \tag{5}$$

Sharing bandpass filters across each sensor reduces parameters, improves model latency, and regularizes the treatment of sensor data.

Each FIR filter, k is implemented as a set of kernel coefficients and applied through convolution with the input signal X.

$$X \otimes k_{(f_L,f_H)} = \sum_{j=0}^{M-1} \sum_{i=0}^{N-1} X[i] * k_{(f_L,f_H)}[j-i] \tag{6}$$

where X is the input signal and k_{f_L,f_H} is the vector of kernel coefficients that allows frequencies in $[f_L, f_H]$ to remain in the signal. Additional details on the calculation of k coefficients and how they compare to learned kernels can be found in [23].

Filters are initialized to uniformly sub-divide the majority of the available spectrum (i.e., 0-300 Hz) with a 3 Hz region of overlap between adjacent bands. The original Kaldi implementation initializes bands starting at a low-cutoff of 30 Hz, but this minimum starting frequency is reduced to 10 Hz for the present analysis to help encourage use of lower frequencies that may be relevant for this application [19]. The Kaldi SincNet implementation also includes a minimum frequency and minimum bandwidth constraint, which are configured to be 1 Hz and 3 Hz, respectively. Kaldi enforces these minimums by increasing the absolute value of the learned low-cutoffs and bandwidths by their respective minimums. Future work should explore the impact of different potential initialization schemes.

B. Activation

Rectified linear units (ReLU), defined as $y = max(0, x)$, provide a linear gradient for all input $x \in \mathbb{R}^+$ and 0 gradient for $x \leq 0$. With zero-centered bandpass outputs, a large portion of values will not have a gradient with ReLU activation. Instead, the Leaky ReLU activation (LReLU) provides a small gradient for $x \leq 0$, while still being non-linear and computationally simple. The LReLU activation is defined in equation 7, where the default $\alpha = 0.01$ is used for for all experiments.

$$Leaky\ ReLU(x) = \max(0, x) + \alpha * \min(0, x) \tag{7}$$

Using LReLU on zero-centered data still greatly diminishes negative inputs. However, the learned affine parameters within the batch normalization layers can learn to offset any inputs into regions with higher variance.

C. Batch Normalization

The amplitude of the output from the Multi-SincNet filters scale directly with the amplitude of the input signal. Between-sensor relative magnitudes are important to maintain, so scaling at the sensor dimension of intermediate data is avoided in the early layers.

Brain dynamics are not evenly distributed in the frequency domain, however, and will tend to have higher amplitudes at lower frequencies. This means the additional bandpass dimensions may be distributed at different scales, making it difficult to learn shared kernels in subsequent convolution layers. Furthermore, the scale of the intermediate values may shift as the cutoff frequencies of the learned bandpass filters are optimized.

Therefore, in order to balance influence when learning kernels applied across bands, and to scale hidden outputs to activation regions, a spatial batch normalization [39] is applied at the band dimension in the three hidden outputs following the Multi-SincNet input layer. Re-scaling each band independently maintains within-band relative dynamics that can be learned using shared weights.

$$\mu_f = \frac{1}{BST} \sum_{b=0}^{B-1} \sum_{s=0}^{S-1} \sum_{t=0}^{T-1} X[b, s, f, t] \tag{8}$$

$$\sigma_f = \frac{1}{BST} \sum_{b=0}^{B-1} \sum_{s=0}^{S-1} \sum_{t=0}^{T-1} (X[b, s, f, t] - \mu_f)^2 \tag{9}$$

$$y = \frac{X - \mu_f}{\sqrt{\sigma_f + \epsilon}} * \gamma + \beta \tag{10}$$

$$forf \in F$$

where B is the batch size, S is the set of sensors, F is the set of bandpass regions, and T is the number of input samples. Learned affine parameters β and γ allow the model to adjust the center and scale away from the origin and unit variance. Following cross-band convolution, spatial normalization is applied across sensors - computing μ_s and σ_s analogous to μ_f and σ_f. At this point in the architecture, distributions across sensors are well-normalized and suitable for batch normalization's regularizing effect, reducing internal covariate drift.

D. Monte Carlo Dropout

Sensor systems with many highly responsive input channels may have spurious errors or drift, and sometimes must be removed in pre-processing. Additionally, for general tasks such as speech activity detection from an ECoG array, some important brain regions may have multiple sensors covering them, resulting in high co-linearity across channels. To regularize co-linearity across sensors, channel dropout [40] is applied on the input to the model during training. Channel dropout on the sensors zeros all signal values for a sensor with an independent Bernoulli random number parameterized by probability p. It is common to avoid using dropout when using batch normalization since the noise caused by the dropout will skew the mean and variance statistics used in normalization towards zero.

However, for SincIEEG, the data modality is already centered at zero, and the practical application motivates robustness to sensor dropout.

E. Optimization Procedure

All deep learning models in this work, both the SincIEEG described above and CNN model described in Section IV-C.4, use stochastic gradient descent from gradients produced by error back-propagation. The Adam optimizer [41] is employed with the learning rate fixed to $\alpha = 0.001$ for all experiments. Binary cross-entropy loss between the target label and the model's output is used as the objective criteria.

Models are evaluated through multiple refits using a K-Fold procedure across a participant's sessions. A single holdout session is used for evaluation in each fold and the remaining sessions are used for training. Some participants had three sessions, providing two training sessions per fold, while others had only two sessions overall and provided one session per training fold. The training data is randomly split into a 25% cross-validation portion for monitoring model performance during training. After each epoch of training, a model under optimization is applied to the cross-validation data and scored. For the SincIEEG and CNN experiments, the best model on the cross-validation is maintained and stored after 100 epochs of training.

Experiments without auditory sensors and other supplementary architecture exploration used early stopping. For these experiments, if the cross-validation performance did not improve for 10 epochs during training, then the best model at that point was stored and the training procedure ended. The early stopping procedure generally produced models with similar performance to their 100 epoch counterparts. Other configurations that were explored using this truncated procedure include variations of activation function, batch normalization, number of learned kernels, and other modifications to convolution configuration. Performance was robust for most configurations and these preliminary experiments focused on reducing model complexity.

IV. Model Validation

ECoG data acquired from participants performing the speech task were used to further validate the model. The models are validated both quantitatively for predictive performance, as well as qualitatively for convergence of the spectral band filters to physiologically plausible ranges.

A. Prediction Accuracy

The prediction accuracy is simply computed as the proportion of windows correctly classified as 'speaking' or 'not-speaking'. Visualizations that overlay the stimulus cue, curated labels, speech audio signal, and the model's predicted likelihood of speech are presented. Aligning recorded speech with model predictions across multiple training windows enables an examination of the model's predictions with both the labeled regions and recorded speech data. The model's ability to predict speech occurring outside the labeled region help to validate the model's generalization capabilities. Ultimately, this visualization provides an indication as to how the model would perform in practice. For instance, frequent oscillations in the predicted likelihood may achieve reasonable accuracy but ultimately be unreliable for use in a classification pipeline.

B. Spectral Band Convergence

A key aspect of this model's utility is its ability to learn spectral bands that minimize the loss function of the network. When the band parameters are combined with the loss and cross validation loss for each training batch, a visualization of the band convergence over time can be obtained. This visualization can serve several purposes. For the present analysis it serves as an additional method of model vetting and interpretation, to establish the frequency bands the model identified as empirically predictive. For other analyses, it could serve as an exploratory tool to investigate whether frequency information is central to the phenomenon.

C. Comparison Models and Benchmarks

1) Randomization Tests: In order to compare the model performance to random chance, model prediction was assessed when trained on randomly labeled segments. The labeling scheme maintained a proportional amount of speaking/not-speaking labels, and thus the chance accuracy should be 50%. To confirm this, the train and test paradigms were kept identical, except that before training, a labeled segment was randomly assigned a 'speaking' or 'not-speaking' label. The hyperparameters chosen for model configuration were 1-Band with a dropout of P = 0.5.

2) Auditory Cortex Electrode Removal: To verify that classification performance was not merely being driven by auditory feedback, electrodes in the auditory cortex region were manually identified based on anatomical landmarks and removed from the analysis (see Figure 1). An abbreviated evaluation of SincIEEG was performed to confirm that the classification performance was not significantly degraded by the exclusion of the auditory electrodes. Optimization time of these additional models was reduced by using early stopping as described in Section III-E. Additional testing verified that early stopping does not unfavorably bias the resulting model performance.

3) LDA and SVM Benchmarks: To explore whether the frequency bands that the SincIEEG model identified would confer some benefit over using the entire broadband spectrum, the performance using the bands that 3-band SincIEEG learned for each participant was compared to the performance using broadband activity from 0.5-170 Hz frequencies. The 3-band version was chosen to compare because it is more distinct from broadband than the 5-band version which generally occupies a greater proportion of the spectrum. A Linear Discriminant Analysis (LDA) and a linear Support Vector Machine (SVM) were implemented as performance benchmarks. Because these comparatively simple classifiers are not capable of attaining reasonable performance using raw ECoG timesamples, a preprocessing method derived from [13] was implemented that generates a band power aggregate measure over a 500 ms

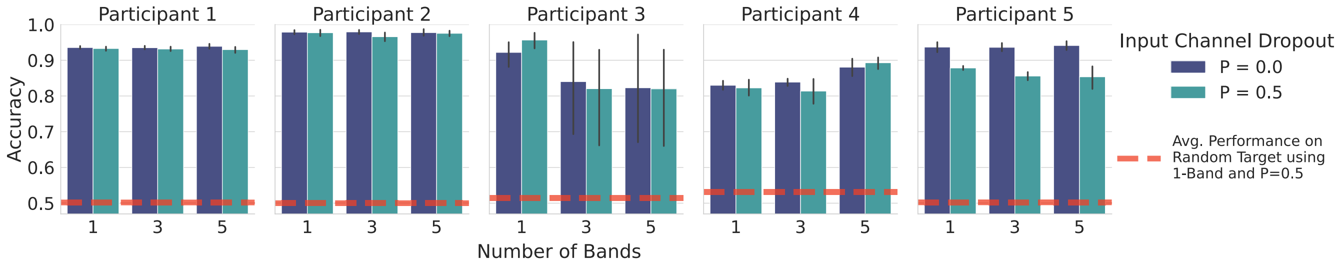

Fig. 3. Mean and variance of accuracy for all repetitions' test folds, for each participant model configuration.

window that updates every 50 ms. The labels were accordingly downsampled to 20 Hz. For each label, the preceding 500 ms of the corresponding preprocessed ECoG signals were used to compute the input features. The resulting feature array was flattened into a vector for training the LDA and SVM models. This process was performed for both the broadband and 3-band SincIEEG versions.

4) Standard CNN: To establish how SincIEEG performs compared to a traditional deep learning method, a standard CNN was implemented and evaluated based on [42]. For this CNN, the first convolutional layers aggregate across time with kernels and stride of five samples, and a dilation of two samples to further downsample. The next layer maintains the kernel's size and stride, but returns to default dilation of one. The remaining two convolutional layers learn 3x3 kernels with unit stride and dilation until a final dense layer outputs to a sigmoid activation. A total of 16 filters were learned in each convolutional layer. The standard convolutional network model is an important alternative to SincIEEG as it uses the same convolution operation but is not directly interpretable. The training and testing paradigms remained unchanged, only the model architecture was exchanged.

V. Results

A. Prediction Accuracy

The average SincIEEG model accuracy across all participants was 94.1% (s.e. 3.5%), and all but one participant achieved an accuracy above 90%. Figure 3 shows the accuracy of all model configurations per participant with each configuration repeated three times. Results from Participants 1 and 2 were very consistent regardless of hyperparameter, while Participant 3 showed significant variability in the 3- and 5- band versions, and Participant 5 performed better without dropout. These differences are most likely mediated by electrode number and placement. However, the ability of the model to achieve good performance on such a variety of electrode locations is a testament to its robustness, and the advantages of a participant-specific feature set.

As described in Section II-D, target labels were created from the timings of experiment cues, rather than the participant's speech. Therefore, to better gauge speech detection performance for practical speech detection applications, predictions were qualitatively assessed by visual inspection into one of three categories: *Full Success*, *Partial Success*, and *Failure*.

A word trial was considered a *Full Success* if the prediction captured the entirety of the spoken word prior to onset and maintained until speech had ceased. Subplots (a), (d), and (g) in Figure 4 are examples of *Full Success* trials. Regions of false positive predictions encompassing a correctly identified speaking region were still categorized as a *Full Success* since false positives are envisioned to be less critical than false negatives for future applications to imagined speech.

TABLE II
PREDICTION SUCCESS OVER TRIALS

Participant	Full Success	Partial Success	Failure
1	93 (81%)	11 (10%)	11 (10%)
2	98 (85%)	10 (9%)	7 (6%)
3	36 (31%)	53 (46%)	26 (23%)
4	43 (37%)	51 (44%)	21 (18%)
5	64 (56%)	37 (32%)	14 (12%)

A trial was considered a *Partial Success* if it captured the majority of the word but clipped either the beginning or the end. Subplots (b), (e), and (h) in Figure 4 are examples of *Partial Success* trials. A trial was considered a *Failure* if the word was missed entirely, if the model prediction was erratic or inconsistent, or if a portion of the word was missed from an otherwise well-placed detection. Subplots (c), (f), and (i) in Figure 4 are examples of *Failure* trials.

For each participant's best model configuration, the model with the best cross-validation performance was selected and its test-set predictions were assessed using the criteria described above.

Table II shows the proportion of words assigned to each category for a 115 word test set for each participant for the respective best model configuration. Participant 1 and 2 models were able to very consistently predict speech before speech onset, suggesting that the model and electrode location combination may capture aspects of speech planning. Participant 3 and 4 models had a majority of partial successes. These trials largely exhibited clipping the beginning portion of words, suggesting that the model may be capturing aspects of speech production rather than speech planning.

B. Spectral Band Convergence

Figure 5 shows a representative example of spectral bands converging over training epochs. While there was a significant amount of variability in the plots across participants and configurations, there are several consistent observations. First, there is a distinct and consistent difference in the

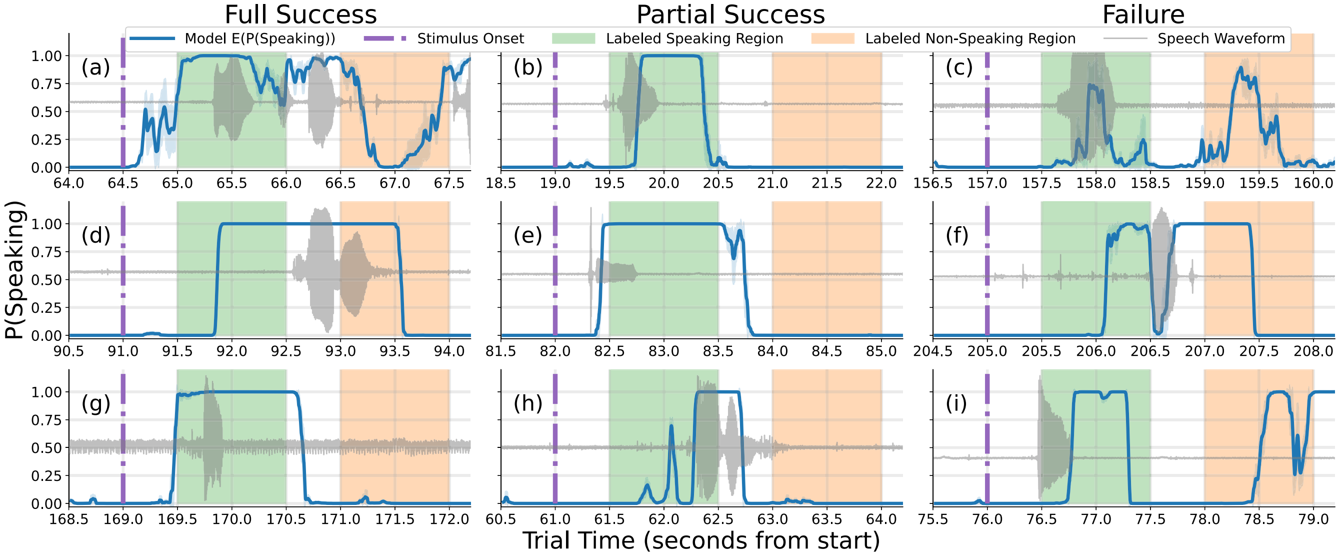

Fig. 4. SincIEEG model predictions of 9 representative words, grouped into 3 categories detailed in Section V.A. (a)–(i) Panels show representative word trials from each category. The grey trace is the audio waveform from the microphone and represents the participants utterances during the word trial. The blue trace, and associated shading, represent the moving average and standard deviation of the model-derived 'speaking' likelihood over the previous 15 samples. The green shaded area represents the region labeled 'speaking', and the orange shaded area represents the region labeled 'not-speaking'. Top row: Participants 5, 4, 5. Middle row: Participants 1, 3, 2. Bottom row: Participants 3, 1, 2.

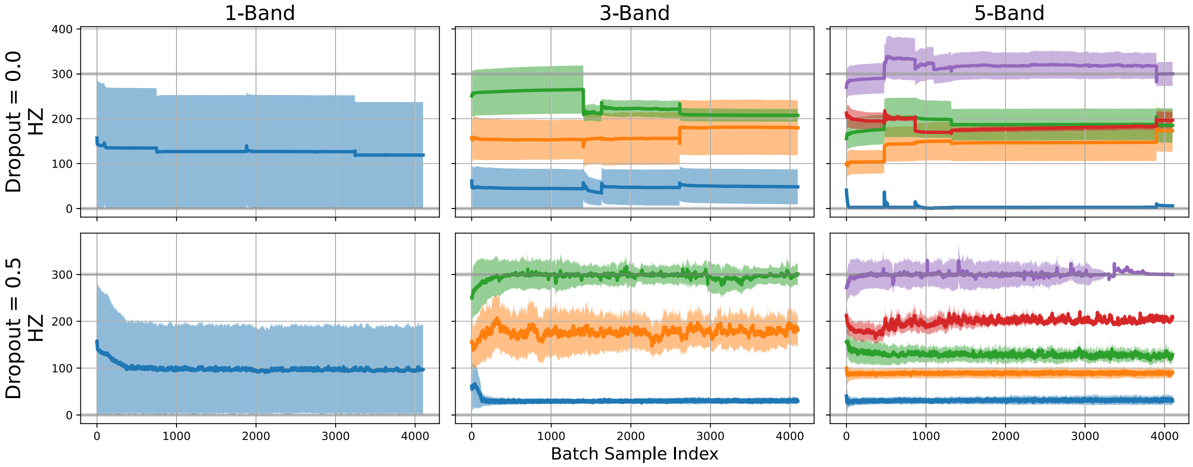

Fig. 5. Spectral band convergence of the 1-, 3-, and 5-band SincIEEG networks for Participant 2. The bold lines are the center of the band, and the shaded regions in the corresponding color are the band bounds. The top row is without dropout, and the bottom row is with dropout.

band evolutions during training when dropout is included in the model. With dropout, bands tended to converge more smoothly, rather than exhibiting large jumps in value as observed without dropout. With shared parameters, zeroing a sensor channel eliminates its influence and subsequently allows other sensors of varying magnitudes to drive parameter updates. Furthermore, zeroed sensors bias downstream normalization layer statistics towards zero. It is posited that these aspects result in the higher variance stochastic search of frequencies illustrated in Figure 5.

The final bands learned for each participant, aggregated across sessions and model configurations, are shown in Figure 6, with the bands aggregated across participants shown in Figure 7. For better visualization, only SincIEEG models with performance in the top 50% for each participant are included in the figures. The bands are superimposed on a single frequency spectrum as a density plot at high transparency. Each band is plotted in a different color, with more saturated hues representing frequencies common across more participants and model configurations than less saturated hues. This provides a compact conceptualization of the final converged frequencies across models.

For the 1-band case, the general tendency is for the band to be broad. However, the aggregated data shows that the bands commonly overlapped around 25-75 Hz, implying the lower frequency band may be more predictive than high gamma for the task, as supported by [19].

The 3-band case indicates one lower-frequency band in a narrow range from 20-40 Hz, a broader middle band roughly spanning 120-200 Hz, and a high frequency band converging above 250 Hz. The 5-band case shows similar bands at the low and high ends of the spectrum, with intermediate

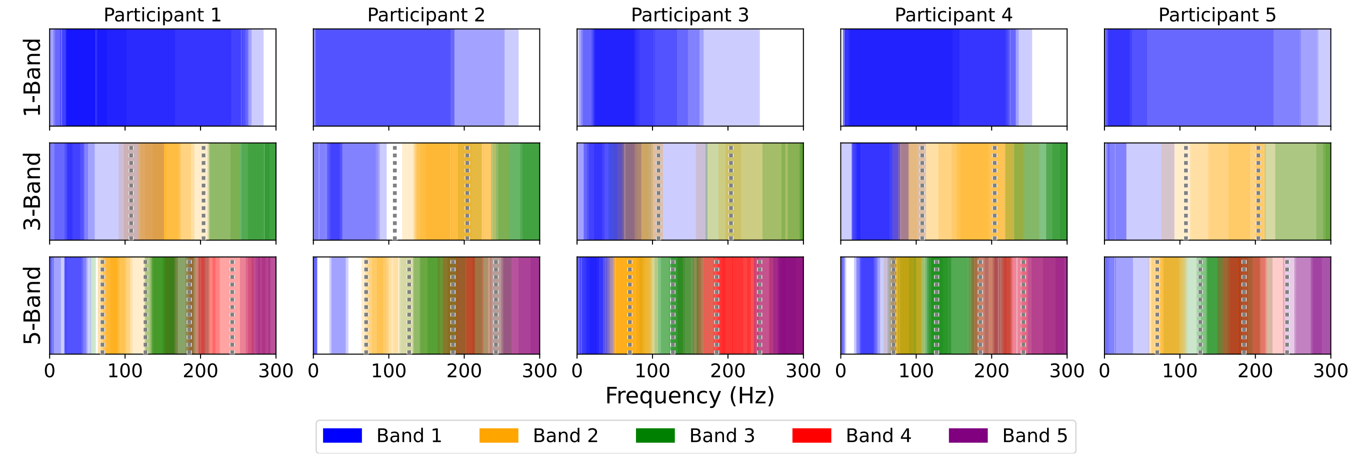

Fig. 6. Learned frequency bands for each participant and 1-, 3-, or 5-band configurations. The selected bands are superimposed on a single frequency spectrum as a density plot at high transparency. Each band is plotted in a different hue: blue, yellow, green, red, and purple. More saturated hues represent frequencies common across a greater number of model configurations than less saturated hues. Vertical dashed lines correspond to the initial cut-off frequencies of adjacent bands prior to convergence. More details on the band initialization procedure can be found in Section III-A.

TABLE III
Model Accuracy Comparison

Participant	SincIEEG	SincIEEG-Non-Auditory	CNN	SincIEEG 3-Band LDA	SincIEEG 3-Band SVM	Broadband LDA	Broadband SVM
1	0.939	0.930	0.941	0.748	0.807	0.735	0.726
2	0.979	0.977	0.983	0.900	0.888	0.832	0.827
3	0.957	0.862	0.932	0.876	0.849	0.811	0.794
4	0.893	0.827	0.885	0.743	0.773	0.728	0.713
5	0.941	0.883	0.941	0.710	0.714	0.695	0.692
Mean	0.942	0.896	0.936	0.796	0.806	0.760	0.751

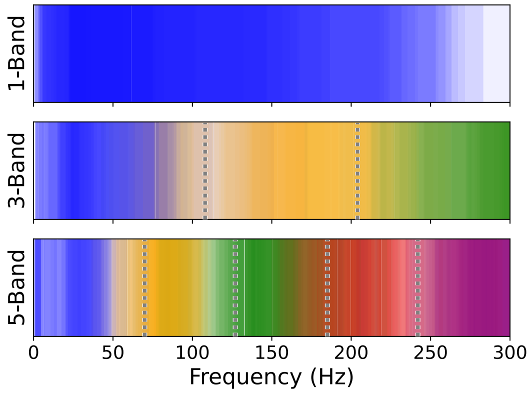

Fig. 7. Learned frequency bands for the top-50% of model configurations across participants for each 1-, 3-, or 5-band configuration, as described in Figure 6. For improved visualization, the figure only includes the top-50% of model configurations of each the participants' sessions.

bands centered at approximately 75 Hz, 150 Hz, and 200 Hz, respectively.

A benefit of the interpretability of learning frequency bands is that the results can be directly compared to known physiologically-relevant bands. Kanas et. al. examined 8 Hz wide frequency bands from 0 to 248 Hz, and produced a histogram ranking bins by contribution to speech detection [43]. It is a multi-modal distribution, with two larger peaks, one spanning 0-40 Hz and one 180-200 Hz, with two smaller, broader peaks in the intermediate frequencies.

The 3- and 5-band plots mirror this trend. In the 3-band version, the lower frequency band at 40 Hz and the middle band covering the 150-200 Hz range coincide quite closely with the peaks in the Kanas et. al. histogram. The 5-band version is even more compelling, with the first band again centering on 40 Hz, the two middle bands covering areas around 100 Hz and in the middle hundreds, and the fourth band centering directly at 200 Hz.

C. Comparison and Benchmarks

Table III shows the performance of all validation measures in comparison to SincIEEG. The SincIEEG and SincIEEG Non-Auditory results are the mean test fold accuracy for each participants' best performing model configuration, effectively the highest bar for each participant in Figure 3. Excluding the auditory cortex electrodes did not significantly impact model performance. The causal formulation of the model, and accurate capture of speech onset within the predicted speech window, provides a strong indication that perception of speech was not a driver of the model classification accuracy. The CNN architecture performance is overall on par with SincIEEG. This shows that the interpretable and parsimonious architecture of the SincNet does not compromise model performance.

The bands identified by the 3-band SincIEEG for each participant were compared to a broadband approach and classified with LDA and SVM. For both classifiers across participants, using learned bands instead of the broadband showed an improvement in classification accuracy. This implies that SincIEEG provides unique and relevant features due to the participant-specific, empirical, and/or parsimonious nature of the learned SincIEEG bands.

It should be noted that, regardless of whether using learned bands or broadband, the LDA and SVM classifiers with the preprocessed ECoG signals did not achieve better results than SincIEEG. Additionally, SincIEEG was able to achieve better results with greater time-domain resolution than the methods using the preprocessed features.

VI. Discussion

This work introduces SincIEEG, a deep learning model with an interpretable architecture. SincIEEG is capable of detecting overt speech using unprocessed ECoG recordings based on a diversity of electrode coverage. SincIEEG meets or exceeds the performance of other ECoG speech detectors, with several additional advantages.

In prior work on using ECoG for speech activity detection, Kanas et. al achieved maximum accuracies of 92% [22], and 98.8% with non deep learning classifiers [43]. Other studies used the detection model as part of a larger speech decoding analysis and so did not report specific results on speech detection performance [15], [16]. In comparison to SincIEEG, which uses unprocessed ECoG recordings, these approaches require appreciable signal preprocessing prior to speech detection. Since the feature extraction is inherent in SincIEEG, any latency introduced via explicit, potentially suboptimal, data-independent preprocessing is mitigated in the processing pipeline - which is critical for real-time implementation.

The architecture of SincIEEG is CNN-based, like that of the foundational work of EEGNet, which showed the viability of CNN's for several tasks using non-invasive EEG signals [44]. The EEGNet architecture was subsequently extended for application in a movement task to intracranial signals, including the addition of a spatial component [45]. This approach is also capable of determining data-driven frequency features, albeit in a manner distinct from SincIEEG. While it is demonstrated that SincIEEG is capable of speech activity detection from ECoG signals, the original implementation was used for acoustic speech detection [23], and it has also been applied to EMG signals [30]. Using a related approach for seizure detection using non-invasive EEG, Fukumori et. al. showed that a data-driven approach was superior to static filter banks [46]. Such models that learn the task-relevant spectral bands can be applied to other domains where frequency analysis is central. This is mainly due to the utility of learning bandpass filters, and the flexibility of the scope on which different filters can be learned.

In terms of interpretability, visualization of the learned bands provides a unique modality for studying the relevant spectral features. One consistent observation is that, across all 1-, 3-, or 5-band models and all participants, a low frequency component was always included. This supports prior work that suggests lower frequency features can play a key role in speech detection in addition to broadband gamma [7], [43]. While the present analysis did not attempt to specifically identify the subset of electrodes related to speech production processes, due to the consistent performance results regardless of the hemisphere of the implant, it is expected that the contributions are largely from the ventral primary motor cortex as shown in prior work [6], [11], [13], [47].

Beyond interpretability, the flexibility of the SincNet architecture's ability to learn different combinations of relevant frequency bands make it promising for implementing transfer learning to leverage existing data for development and training of generalizable models. Gathering sufficient data and learning robust models for new participants is challenging, particularly for intracranial recordings where available data is limited and the electrode locations are generally sparse and not consistent across participants. In this context, transfer learning can be used to refine the model on a new participant's data after having learned its initial parameters from other participants' data - which can significantly reduce training time and improve model robustness and performance.

Because SincIEEG is capable of learning task-relevant spectral bands across multiple participants independent of precise electrode locations, it has the potential to learn generalized bands for brain regions sampled by the population of electrodes across participants. Furthermore, specific bands can be learned for channel context labels, such as in which brain region an electrode resides. This allows for encoding a spatial component to the transfer learning, initializing different bands dependent on electrode location.

Ultimately, toward the development of a practical speech neuroprosthetic, future work must examine the efficacy of SincIEEG on transfer learning and, moreover, on imagined speech and integration with the subsequent speech decoding pipeline.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE Rev. Biomed. Eng.*, vol. 4, pp. 140–154, 2011.

[3] V. Gilja *et al.*, "Clinical translation of a high-performance neural prosthesis," *Nature Med.*, vol. 21, no. 10, pp. 1142–1145, 2015.

[4] P. Nuyujukian *et al.*, "Cortical control of a tablet computer by people with paralysis," *PLoS ONE*, vol. 13, no. 11, 2018, Art. no. e0204566.

[5] B. N. Pasley *et al.*, "Reconstructing speech from human auditory cortex," *PLoS Biol.*, vol. 10, no. 1, Jan. 2012, Art. no. e1001251.

[6] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, pp. 327–332, Feb. 2013.

[7] E. M. Mugler *et al.*, "Direct classification of all American English phonemes using signals from functional speech motor cortex," *J. Neural Eng.*, vol. 11, no. 3, Jun. 2014, Art. no. 035015.

[8] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, "Progress in speech decoding from the electrocorticogram," *Biomed. Eng. Lett.*, vol. 5, no. 1, pp. 10–21, Mar. 2015.

[9] C. Herff *et al.*, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Frontiers Neurosci.*, vol. 9, p. 217, Jun. 2015.

[10] F. Lotte *et al.*, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers Hum. Neurosci.*, vol. 9, p. 97, Feb. 2015.

[11] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron*, vol. 98, no. 5, pp. 1042–1054, Jun. 2018.

[12] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[13] C. Herff *et al.*, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers Neurosci.*, vol. 13, p. 1267, Nov. 2019.

[14] M. Angrick *et al.*, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Commun. Biol.*, vol. 4, no. 1, pp. 1–10, 2021.

[15] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, Dec. 2019.

[16] D. A. Moses *et al.*, "Neuroprosthesis for decoding speech in a paralyzed person with anarthria," *New England J. Med.*, vol. 385, no. 3, pp. 217–227, 2021.

[17] S. Martin *et al.*, "Word pair classification during imagined speech using direct brain recordings," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 25803.

[18] S. Martin, I. Iturrate, J. D. R. Millán, R. T. Knight, and B. N. Pasley, "Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis," *Frontiers Neurosci.*, vol. 12, p. 422, Jun. 2018.

[19] T. Proix *et al.*, "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features," *Nature Commun.*, vol. 13, pp. 1–14, Jan. 2022.

[20] M. Angrick *et al.*, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 036019.

[21] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature Neurosci.*, vol. 23, no. 4, pp. 575–582, 2020.

[22] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone, "Real-time voice activity detection for ECoG-based speech brain machine interfaces," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 862–865.

[23] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 1021–1028.

[24] J. Kubanek and G. Schalk, "NeuralAct: A tool to visualize electrocortical (ECoG) activity on a three-dimensional model of the cortex," *Neuroinformatics*, vol. 13, no. 2, pp. 167–174, 2015.

[25] M. Patkowski, "Laterality effects in multilinguals during speech production under the concurrent task paradigm: Another test of the age of acquisition hypothesis," *Int. Rev. Appl. Linguistics Lang. Teach.*, vol. 41, no. 3, pp. 175–200, 2003.

[26] C. Code, "Can the right hemisphere speak?" *Brain Lang.*, vol. 57, no. 1, pp. 38–59, Mar. 1997.

[27] A. S. House, C. Williams, M. H. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," *J. Acoust. Soc. Amer.*, vol. 35, no. 11, p. 1899, 1963.

[28] M. A. Mines, B. F. Hanson, and J. E. Shoup, "Frequency of occurrence of phonemes in conversational English," *Lang. Speech*, vol. 21, no. 3, pp. 221–241, 1978.

[29] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–1043, Jun. 2004.

[30] M. Stuart and M. Manic, "Deep learning shared bandpass filters for resource-constrained human activity recognition," *IEEE Access*, vol. 9, pp. 39089–39097, 2021.

[31] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[32] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May/Jun. 2007.

[33] C. R. Harris *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[34] *Pandas-Dev/Pandas: Pandas*, Pandas Development Team, Feb. 2020, [Online]. Available: https://pandas.pydata.org/about/citing.html, doi: 10.5281/ZENODO.3509134.

[35] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python Sci. Conf.*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61.

[36] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.

[37] P. Virtanen *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020.

[38] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6465–6469.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[40] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[42] R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball, "Deep learning with convolutional neural networks for decoding and visualization of EEG pathology," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Dec. 2017, pp. 1–7.

[43] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone, "Joint spatial–spectral feature space clustering for speech activity detection from ECoG signals," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1241–1250, Apr. 2014.

[44] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 056013.

[45] S. M. Peterson, Z. Steine-Hanson, N. Davis, R. P. N. Rao, and B. W. Brunton, "Generalized neural decoders for transfer learning across participants and recording modalities," *J. Neural Eng.*, vol. 18, no. 2, 2021, Art. no. 026014.

[46] K. Fukumori, N. Yoshida, H. Sugano, M. Nakajima, and T. Tanaka, "Epileptic spike detection using neural networks with linear-phase convolutions," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1045–1056, Mar. 2022.

[47] D. Carey, S. Krishnan, M. F. Callaghan, M. I. Sereno, and F. Dick, "Functional and quantitative MRI mapping of somatomotor representations of human supralaryngeal vocal tract," *Cerebral Cortex*, vol. 27, no. 1, pp. 265–278, 2017.