IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 October 2022, accepted 13 November 2022, date of publication 19 December 2022,
date of current version 29 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3230688

RESEARCH ARTICLE

Self-Supervised Learning of Neural Speech Representations From Unlabeled Intracranial Signals

SRDJAN LESAJA[1,*], MORGAN STUART[2,*], JERRY J. SHIH[3], PEDRAM Z. SOROUSH[1], TANJA SCHULTZ[4], (Fellow, IEEE), MILOS MANIC[2], (Fellow, IEEE), AND DEAN J. KRUSIENSKI[1], (Senior Member, IEEE)

[1]Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA
[2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA
[3]Neurology Department, UCSD Health, San Diego, CA 92103, USA
[4]Cognitive Systems Laboratory, University of Bremen, 28359 Bremen, Germany

Corresponding author: Srdjan Lesaja (slesaja@vcu.edu)

This work was supported in part by NSF under Grant 2011595/1608140, and in part by the Bundesministerium für Bildung und Forschung (BMBF) under Grant 01GQ1602.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Virginia Commonwealth University and UCSD Health IRB.

*Srdjan Lesaja and Morgan Stuart contributed equally to this work.

ABSTRACT Neuroprosthetics have demonstrated the potential to decode speech from intracranial brain signals, and hold promise for one day returning the ability to speak to those who have lost it. However, data in this domain is scarce, highly variable, and costly to label for supervised modeling. In order to address these constraints, we present brain2vec, a transformer-based approach for learning feature representations from intracranial electroencephalogram data. Brain2vec combines a self-supervised learning methodology, neuroanatomical positional embeddings, and the contextual representations of transformers to achieve three novelties: (1) learning from unlabeled intracranial brain signals, (2) learning from multiple participants simultaneously, all while (3) utilizing only raw unprocessed data. To assess our approach, we use a leave-one-participant-out validation procedure to separate brain2vec's feature learning from the holdout participant's speech-related supervised classification tasks. With only two linear layers, we achieve 90% accuracy on a canonical speech detection task, 42% accuracy on a more challenging 4-class speech-related behavior recognition, and 53% accuracy when applied to a 10-class, few-shot word classification task. Combined with the visualizations of unsupervised class separation in the learned features, our results evidence brain2vec's ability to learn highly generalized representations of neural activity without the need for labels or consistent sensor location.

INDEX TERMS Brain–computer interface, intracranial EEG, deep learning, transformers, vector quantization, speech modeling.

I. INTRODUCTION

Speech neuroprostheses are designed to decode and synthesize speech directly from the electrical potentials of the brain. There have been significant advances in neural speech decoding over the past decade using intracranial recordings such as electrocorticography (ECoG) or stereotactic EEG (sEEG). These include describing brain regions and mechanisms involved in speech, predicting words or phonemes, translating neural signals to articulatory kinematics models, text, or directly to speech waveforms [1], [2], [3], [4], [5], [6], [7]. Recent efforts have progressed to real-time synthesis or classification, and decoding of imagined speech [8], [9], [10], [11], [12], [13].

However, due to the nature and limitations of the clinical procedures commonly used to obtain research data,

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Wang.

existing methods for neural speech decoding generally rely on participant-specific models, trained on labeled experiment tasks. Supervised approaches such as these are naturally restrictive, supporting only one particular participant's sensor configuration and task-related behavior. Instead, self-supervised methods with unlabeled data and explicit handling of sensor configuration may allow for much more flexible paradigms in which multiple participants' data can be pooled for learning general purpose features. Furthermore, methods that learn without labels have broader potential applications, including use in closed-loop online systems in which labels are unreliable or non-existent.

The recent introduction of the transformer architecture ushered in a new era for the deep learning field, showing the attention mechanism to be a simple yet powerful tool for natural language processing (NLP) and sequence to sequence models [14]. The self-attention transformer block served as the foundation for BERT [15] and the GPT series [16], which solidified a trend of self-supervised learning (SSL) where models are pretrained on a large, neutral, data corpus before being fine-tuned on a specific task of narrower scope. More recent vision transformers effectively demonstrate that most data can be treated as a sequence, that self-attention performs as well or better than convolutional neural networks, and that computer vision models can benefit from self-supervised pretraining like their NLP counterparts [17]. Transformers have since been shown as a viable or superior method for object detection, video action recognition, point cloud shape classification, and multi-modal models [18], [19], [20], [21], [22], [23].

Recently, several studies have explored training language models directly from audio signals rather than text [24], [25], [26]. The key insight of these methods is that, rather than learning a representation in a latent space with continuous targets, they learn from a discretized set of 'pseudo-speech' units. Thus, these methods essentially use clustering to learn a self-defined lexicon rather than being constrained to map to an externally defined set such as words, phonemes, or characters. This approach is particularly appealing to speech neuroprosthetic development because it is analogous to the way speech is processed by humans, assigning discrete conceptual meaning to physiological inputs from a persisting audio source, which are also concepts underlying speech production.

In this work, we present *brain2vec*, a sensor-level feature learning methodology that builds on recent progress by utilizing self-supervised pretraining, vector quantization, and spatio-temporal positional encoding for use in speech neuroprosthetics. We adapt semi-supervised NLP techniques to allow pooling of data across participants by re-referencing electrode locations of different participants to a common brain atlas before training.

The proposed framework is used to pretrain a sensor-level feature extraction model on unlabeled data from multiple participants. For evaluation, the pretrained model is used to extract features for an unseen participant's speech-related classification tasks. Importantly, the pretrained model's parameters are not updated to accommodate the new participant's data or sensor configuration, forcing the fine-tuning classifier to rely only on the features learned from pooled participant data. We also perform exploratory dimensionality reduction and visualization of the learned features to illustrate class separation for the downstream classification tasks.

Our results demonstrate that brain2vec is capable of encoding rich speech representations which can be used for classifying an array of disparate speech-related downstream tasks. These results show promise for a future in which ''off-the-shelf'' pretrained speech neuroprosthetics models can be used to improve a user's livelihood without the need for extensive data collection and labeling.

TABLE 1. Number of implanted electrodes for each participant.

Participant	# Electrodes
1	90
2	70
3	80
4	175
6	94
7	108

II. STEREOTACTIC EEG DATA

To assess our method, we utilize data collected from seven participants, with time-aligned labels of speech behavior from an experimental protocol. This section describes our data and how it was collected.

A. PARTICIPANTS

sEEG data were collected from 7 native English-speaking participants being monitored as part of treatment for intractable epilepsy at University of California San Diego Health. The locations of sEEG electrodes were determined solely based on the participants' clinical needs. The number of implanted electrodes for each participant are provided in Table 1. The study was approved by Virginia Commonwealth University and UCSD Health IRB.

B. ACQUISITION CONFIGURATION

Data from the sEEG electrodes (Ad-Tech Medical Instrument Corporation) were recorded with a Natus Quantum Amplifier (Natus Medical Inc.) and referenced to a pair of subdermal needle electrodes in the scalp. The amplifier signals were digitized at 1,024 Hz. An external microphone recorded the audio signal, and was digitized at 44,100 Hz. The digitized intracranial signals and microphone audio, along with the experiment cues, were synchronized with the Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

C. DATA COLLECTION PROTOCOL

The experimental protocol is designed to investigate overt and imagined speech processes in the brain by having participants

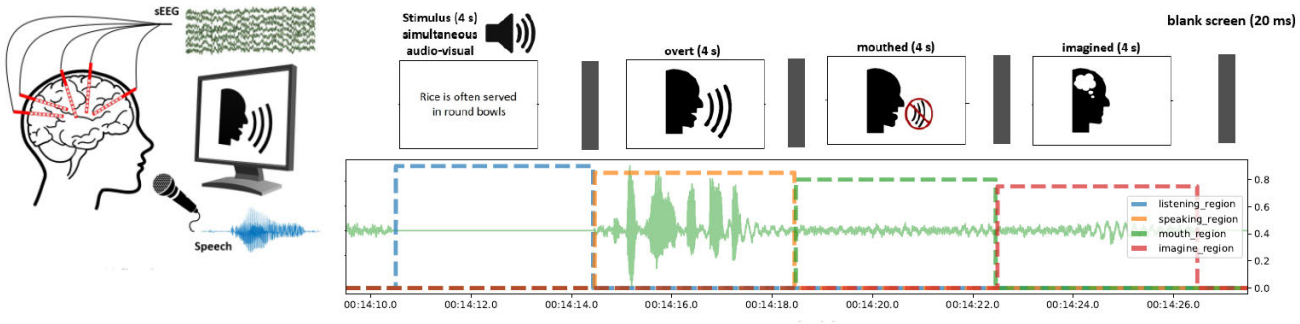


FIGURE 1. Diagram of the harvard sentences experiment protocol. Detailed in section II-C.

repeat a sequence of sentences, each in a series of three different speaking modes.

Before beginning experiment trails, the experiment paradigm, as well as experimental icons and cues, are explained to the participant. They are instructed to perform the associated tasks immediately upon cue presentation - within a 4-second interval during which the task cue is displayed. A trial begins with a short sentence displayed on a computer monitor while simultaneously narrated through computer speakers. All sentence audio was less than 4 seconds in length, but regardless of the length, the associated text remains on the screen for 4 seconds. Following a 20 ms blank screen, the participant is cued with an icon to vocalize the sentence (i.e., overt mode), and this cue remains on the screen for 4 seconds. Following a 20 ms blank screen, the participant is cued for 4 seconds via icon to articulate the sentence as if they were speaking, but without vocalizing (i.e., mouthing mode). Finally, after a 20 ms blank screen, the participant is cued for 4 seconds by icon to imagine speaking the sentence without articulating or vocalizing (i.e., imagined mode). Then following a 20 ms blank screen, the next sentence trial begins. This protocol is illustrated in Figure 1.

The paradigm is repeated each time for a set of 50 unique Harvard sentences, designed to be phonetically-balanced conversational English [27]. All participants completed the entire set of 50 sentence trials; however, only 25 sentence trials from Participant 1 are evaluated due to a software issue that corrupted the labeling of the other 25 sentence trials.

D. VOLUMETRIC MORPHING OF ELECTRODE LOCATIONS TO A COMMON BRAIN ATLAS

Compared to single audio data streams commonly used for NLP and language modeling domains, neural recordings are commonly acquired from tens to hundreds of electrode channels. Additionally, not only is the location of these channels relative to one another important for modeling neural processes, but the absolute channel locations in the brain are also important.

The 3D electrode coordinates reconstructed from CT and MRI imaging data can not be directly compared across participants due to anatomical brain differences. For this reason, each participants' electrode locations were converted from their native brain space coordinates to corresponding locations on the MNI305 common brain atlas [28], [29]. The mapping was done using the Freesurfer software package [30] and MNE-Python python package [31], where further information on the details of the affine transformation procedure can be found [30], [32].

While the MNI brain was selected because it is a widely used common atlas, the critical step is converting the electrodes to a common coordinate space, then any established common atlas can be implemented. This remapping allows sensing locations to be related across participants or even sensor modalities (e.g. ECoG, scalp EEG, etc.), and allows our modeling methodology to leverage the additional spatial information when learning from many participants.

Figure 2 shows the locations of all participant electrodes on the common brain atlas. Each electrode is represented using a 3-dimensional vector indicating its location on the common brain atlas. These coordinates are given in the Right-Anterior-Superior (RAS) frame, with positive values in the 3 dimensions referring to right vs. left, anterior vs. posterior, and superior vs inferior, respectively. The coordinate units are in meters, and take on a range of values [−0.076 m, 0.079 m] across all dimensions. The origin is located at the Anterior Commissure, and the negative y-axis passes through the Posterior Commissure.

III. SELF-SUPERVISED PRETRAINING METHODOLOGY

Our primary contribution is a model architecture and pretraining methodology for learning generalized feature representations of brain activity, using only unlabeled sensor data pooled from an arbitrary number of participants. We refer to this approach as brain2vec,[1] and this section describes the underlying model, loss functions, and optimization procedure. We later show in Section IV that representations learned by brain2vec can be used to train classifiers on an array of labeled downstream tasks. Importantly, the brain2vec pretraining methodology enables fine-tuning on any number of sensors, including new configurations on unseen users.

[1] https://github.com/Morgan243/brain2vec

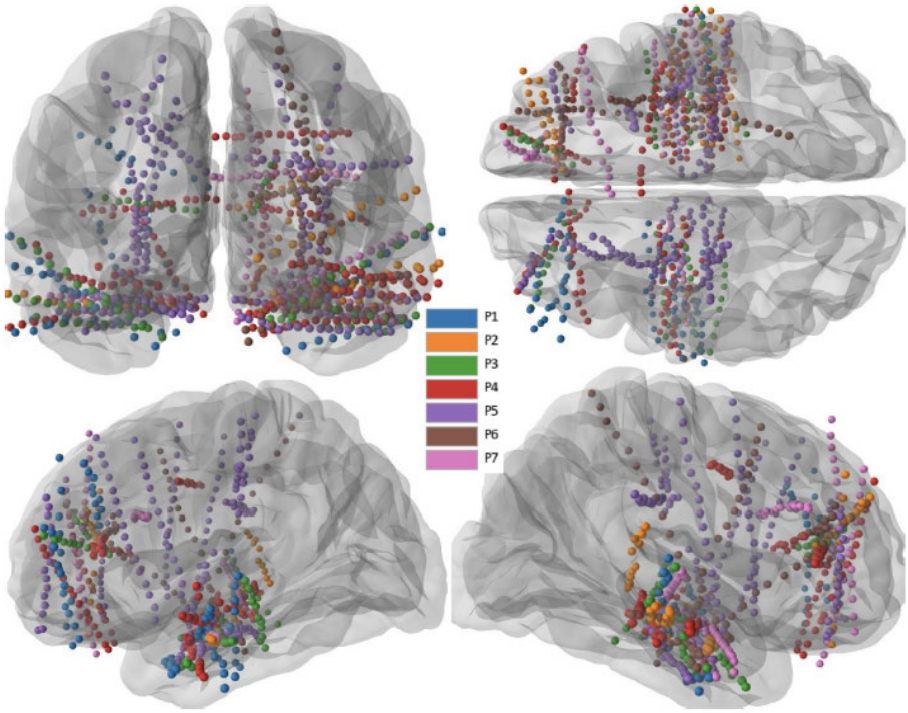


FIGURE 2. Common atlas electrode locations for the 7 participants.

The model consists of a sensor-level feature encoder, implemented as a convolutional neural network (CNN). The feature encoder's outputs are then passed to a transformer network that learns a latent context vector representation of the input sEEG signal. During the pretraining phase, the model is tasked with reconstructing masked regions of the input signal's latent representations, using self-supervised techniques pioneered by language models [15], [16], [24], [33]. The training is aided by a vector quantization module that discretizes the targets, thus guiding the network to learn hidden units. RAS coordinates are used to learn a spatio-temporal embedding that is added to the input of the context model. The resulting sensor-level model can then be used for feature extraction in a task-specific fine-tuning procedure.

A. MODEL ARCHITECTURE

The brain2vec architecture is based on the wav2vec2 audio modeling architecture [24], but with significant modifications to support the modality of intracranial sensor data, including changes to the feature encoder CNN, positional embedding paradigm, codebook configuration, and context network size. In this section, we first overview the input data and the key processing steps across the model's components. Further details on how brain2vec differs from wav2vec2 are described in each subsection.

Brain2vec's input is an unnormalized 0.5 second segment from a single sEEG channel. The input window is first downsampled to 512 Hz and standardized to a zero mean and unit variance within the half-second window. The segment is then passed through a CNN-based feature encoder that generates the latent representations. These latent representations are then passed to both the Quantization Module, where they are discretized into a codebook vector for the objective function, as well as to the context network. The context network is a standard transformer encoder architecture, producing context representations from the codebook distribution. Before entering the context network, regions of context representations across time are masked from the context network by replacing the context representation with a learned mask embedding. Then, spatio-temporal positional information is embedded in the latent representations before being passed to the context model. The masked context representations are learned by having to correctly choose their corresponding quantized latent representation from a set of distractors.

The decision to use a 0.5 s window was driven primarily by prior work, and the intuition that the majority of pertinent information for decoding speech from neural signals will be encapsulated in the neural activity immediately preceding the produced speech. In [34], a speech re-synthesis task was shown to be largely dependent on only 400 ms of neural data centered at the corresponding 400 ms audio signal to be reconstructed, despite the preceding and trailing 400 ms of neural data being included in the predictive model.

1) FEATURE ENCODER NETWORK

The feature encoder network is used to reduce the dimensionality of the input signal before being passed to the Quantization Module and Context Network. The encoder is therefore a

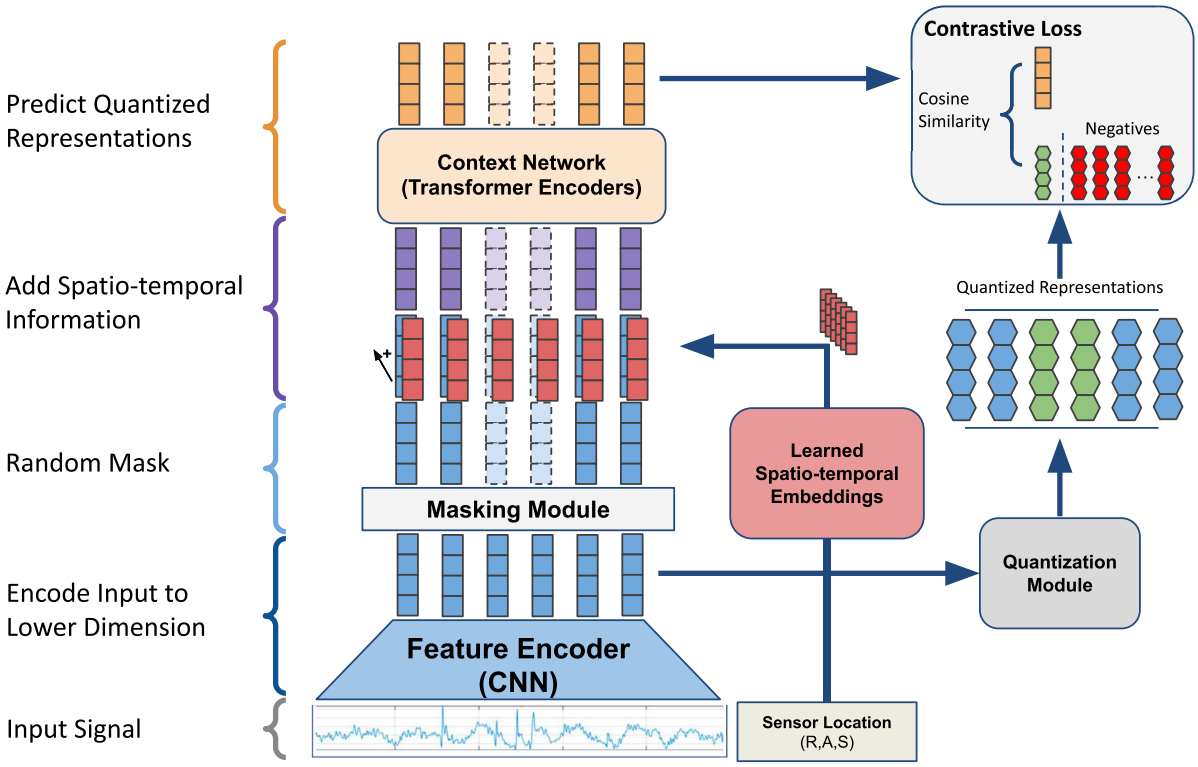


FIGURE 3. Brain2vec pretraining architecture that learns sensor-level representations. A 0.5 s window of normalized sEEG for a single electrode signal is passed to a CNN feature encoder producing latent representations (blue). Spatio-temporal embeddings are created using the 3D RAS coordinates of the electrode (red). The latent representations from the feature encoder are sent to the quantization module. The latent representations are then passed to the masking module, and then the positional embedding is added to the masked latent representations (purple). The embedded latent representations are then passed to the context network, which is a set of transformer blocks, which finally produce the context representations. The reconstructed context representations corresponding to the masked latent representations are then compared to the quantized vectors using cosine similarity in a contrastive loss paradigm. Further details of each component are in Section III.

1-D CNN, operating on the fixed length, single-channel, 0.5 s of 512 Hz input sEEG data. The network has 5 convolution layers, each consisting of a 1-D convolution, dropout regularization with probability $p = 0.25$, layer normalization [35], and a GELU activation function. The first convolutional layer learns 128 filters with a width of 7 samples. The next two layers reduce to 64 filters with a smaller 3 sample kernel. The final two layers further reduce dimensionality to 32 filters with a kernel width of 3. All layers use no padding and a stride of 2 to reduce dimensionality. The resulting feature encoding architecture encodes a 0.5 second window of sEEG into 6 sequential steps of 32 channel data (32×6).

2) POSITIONAL EMBEDDING

The original wav2vec2 architecture utilized a grouped convolution relative positional embedding scheme to include temporal position information to the network. Unlike the single-channel audio used in the original design, there is a need to encode the brain signals according to their spatial locations. In order to include not only temporal but also spatial channel information, a positional embedding scheme was implemented that incorporates the electrode RAS coordinates.

The positional embedding used in brain2vec is produced from a learned transformation of the RAS coordinates described in Section II-D. The first linear layer of the transformation receives the electrode's 3-element RAS coordinates and transforms the input to 32 hidden units. Another 32-unit hidden layer then further transforms the features, before a final output layer produces a 32×6 -dimensional embedding vector. A "Leaky" Rectified Linear Unit (ReLU) with negative slope equal to 0.01 is used as the non-linear transform after each linear layer. We use a leaky ReLU, rather than a standard ReLU, to better handle negative values of the RAS coordinates, while still being computationally simple. The resulting embedding vector is added to the latent representation vector before being passed to the context network.

3) QUANTIZATION MODULE

The vectors are quantized using a combination of the product quantization [36] and Gumbel Softmax [37] techniques. Product quantization involves creating a set of discrete vectors by defining a number of codebooks G, each with a set of codewords W. Quantization vectors are made by concatenating codewords sampled from each codebook. Thereby a maximum number of quantization vectors is given by W^G.

We assign the hyperparameters $G = 2$ and $W = 40$ for a maximum possible 1,600 vocabulary size.

Gumbel Softmax enables one-hot encoding of the quantization vectors in a fully differentiable way. A vector of $G * W = 80$ logits are produced for a latent representation vector which after Gumbel Softmax produce one-hot encoding of a word within a group. The quantization vectors are learned via a linear layer, ReLU, and another linear layer which outputs the logits. A diversity loss term, discussed in more detail in the training section, encourages diverse use of the codebook and codewords. This prevents collapse of the codebook, such that it uses only one or few codewords. Details on the exploration of the effect of modulating number of groups and words on a performance of a vector quantized approach are examined in [38].

4) MASKING PROCEDURE

All the latent representations are quantized before the masking step in order to serve as targets for the objective function. The same latent representations from the feature encoder that are passed to the quantization module are also masked before being fed into the context network.

This masking is the basis of the self-supervised learning of the model and is implemented according to [24]. Due to our shorter sequence dimension of only 6 elements, masking is simplified to choosing two consecutive time steps at random.

Each masked latent representation is replaced by the same learnable masking token vector. Overall this results in 1/3 of latent representation vectors masked for the context network. An example of this masking is provided in Figure 4.

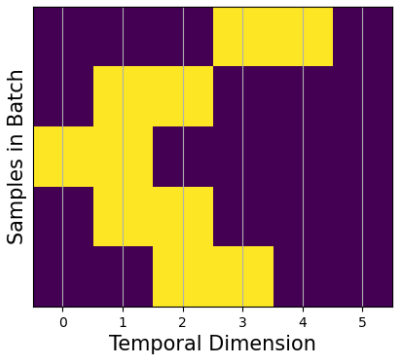


FIGURE 4. Illustration of a random mask on hypothetical batch of 5 samples. A model would be required to identify the correct encoding for each of the yellow regions depicted in the figure.

5) CONTEXT TRANSFORMER NETWORK

The context network is a transformer that follows the same architecture as the encoding side [14], also employed by BERT [15], which provides the in-depth details of the Transformer architecture. The proposed context network consists of 6 transformer block layers, each with four attention heads, 2048 feed forward units, and dropout regularization with $P = 0.25$. The output of each layer is the same dimension as the latent representations fed into the network.

B. PRETRAINING

During pretraining, brain2vec learns speech activity representations from intracranial signals based on an objective function that requires it to correctly identify the true quantized latent representation vector from a set of distractors using the corresponding context representation vector. By using discrete targets rather than continuous vector space targets, the network is influenced towards a parsimonious set of 'hidden unit' clusters which represent the underlying speech activity.

1) LOSS FUNCTIONS

The objective in the pretraining phase is achieved by balancing three loss terms. The first is the contrastive loss function. Given a context representation vector c_t for a masked time step t, the model must choose the correct quantized vector $q_t = QM(z_t)$, which represents the quantization of the latent representation z_t at timestep t, from a set of quantized vectors $q \in Q$ which include itself and K distractors uniformly sampled from other masked timesteps. The loss is calculated by first computing the cosine similarity between context representation vector c_t and quantized vectors Q. The similarity logits are then normalized before taking the negative log of the result for the true vector q_t. All experiments presented in this work use $k = 100$ during pretraining.

$$L_c = -log \frac{exp(cosinesim(c_t, q_t)/\kappa}{\sum_{q \in Q} exp(cosinesim(c_t, q)/\kappa)}$$

This contrastive loss is combined with a diversity loss term. The diversity loss L_d is used to ensure that the use of codewords and codebooks is diverse. The equal use of W codewords from G codebooks is encouraged by maximizing the entropy of averaged softmax distribution over the codewords for each codebook $\overline{p}_g$

$$L_d = -\frac{1}{GW} \sum_{g=1}^{G} \sum_{W=1}^{W} \overline{p}_{g,w} log \overline{p}_{g,w}$$

Finally, a feature penalization term L_z is included as the L2-norm of the feature encoder's output. This encourages smaller features and reduces variance.

$$L_z = \sqrt{\sum_{i=1}^{i=N} |z_t(i)|^2}$$

The final objective function weighs the diversity loss L_d with α, and the L2-norm L_z with λ. Both α and λ can be treated as model hyperparameters during pretraining to help ensure the model converges. All experiments presented in this work use $\alpha = 1$ and $\lambda = 10^{-4}$ during pretraining.

$$L = L_c + \alpha L_d + \lambda L_z$$

2) OPTIMIZATION PROCEDURE

Models are pretrained using stochastic gradient descent, with batches of 1,024 sensor windows over 100 epochs. A random

20% of training samples, stratified at the participant-sentence level, are set aside for cross validation at the end of each epoch during training. The final model is taken from the epoch with the lowest loss L on the cross validation samples. A learning rate of 0.001 and betas of $(0.5, 0.999)$ were used with the Adam optimizer [39]. The learning rate is reduced by a factor of 0.1 every 10 epochs without improvement on a validation set drawn from the training set.

IV. EVALUATION ON CLASSIFICATION TASKS

To assess the viability of brain2vec, and the generalizability of its learned representations, the features extracted through the feature encoder and context network are applied to three distinct but related downstream classification tasks. These tasks were chosen to be relevant to different aspects of speech decoding; however, they vary in complexity and the components of speech being classified.

For all three classification tasks, 0.5 seconds of sEEG data from all available electrodes is considered, with labels for the half-second window assigned in a task-specific manner. In all cases, classification performance is evaluated using balanced accuracy.

The first classification task is *Speech Activity Detection*. This task is the binary classification of whether a participant is speaking or not-speaking during the half-second window. The second task is *Speech Behavior Recognition*, a multi-class problem of predicting which of 4 speech-related behaviors is being performed: listening, speaking, mouthing, or imagining. The third task is *Word Classification*, where the model must classify which word from a reduced set is being spoken during the window.

A. LEAVE-ONE-PARTICIPANT-OUT PRETRAINING

The scarcity of well-labeled intracranial brain data is an important motivation for this work, and with only seven participants, our evaluation must also confront these challenges. We design a leave-one-participant-out pretraining evaluation method, in which six participants of our seven are used for pretraining and a single participant's data is held out for fine-tuning a downstream classifier.

For each participant, that participant's data is excluded and all remaining participants' data is pooled into an unlabeled training dataset. Thus, a unique pretrained model is generated for each participant, one that has never seen a sample from the patient before fine-tuning. This paradigm minimizes data leakage in context feature learning, and ensures the model is not simply memorizing inputs. Additionally, it is intended to simulate the ultimate intended scenario for which a pretrained model based on a larger data corpus is used as the initial model for a new user and subsequently fine-tuned. Herein, a pretrained brain2vec model refers to such a participant-specific, leave-one-out model. All models employ the same architecture and only differ with respect to the training data.

B. DOWNSTREAM CLASSIFICATION

The utility of learned features is assessed by optimizing parsimonious supervised classification models using only the features extracted from brain2vec. The parameters of the brain2vec model are frozen, and not updated, to better assess practical applications where new data and available training time are both small. We refer to these procedures interchangeably as fine-tuning or downstream classification.

All three downstream classification tasks follow a similar structure in terms of architecture. Each 0.5 second window of sEEG data is labeled for each of the three tasks, respectively, as described in subsequent sections. To train the downstream tasks, the weights of the entire pretrained model are fixed. For every 0.5 second window of labeled sEEG data, every electrode belonging to a participant is passed through the pretrained model in sequence. Every electrode generates the context vector representation of the sEEG input. These representations are flattened and concatenated. This vector, containing the context representations of all electrodes of a participant for a 0.5 window, is then provided to one 16-unit linear layer and a final output linear layer which learns to map to the task-specific classes. The activation function is a leaky ReLU with a negative slope of 0.01. We use dropout with $P = 0.75$ and batch normalization to help regularize the classification optimization.

During fine-tuning, only the additional linear layers and normalization layers are updated. Fine-tuning is performed separately for each participant. That is, a classifier is trained for each participant on their set of electrodes and corresponding labels.

1) SPEECH ACTIVITY DETECTION

For speech activity detection, the audio data is labeled using an energy threshold to generate binary speech/non-speech labels for each segment. Only task segments from the speaking region are processed for speaking labels, but non-speaking labels are taken from any low-energy windows in any task region. The sentence narration audio was removed to prevent false positives in this automatic labeling process. Windows of 0.5 s sEEG data corresponding to overt speech are assigned a *speaking* label. An approximately equivalent quantity of windows with audio below the threshold were assigned a label of *non-speaking*.

2) SPEECH-RELATED BEHAVIOR RECOGNITION

The behavior recognition task labels each 0.5 s sEEG window according to one of four speech-related behaviors; *listening*, *speaking*, *mouthing*, or *imagining*. The resulting 4-class classification problem challenges the model to disambiguate highly related activities. The experiment protocol codes the regions with associated experimental cues that are visualized in Figure I. Labels are assigned to the sEEG data according to these task intervals. Each interval is 4 s in length; however, the initial 0.5 s and the final 1.0 s of the 4-s interval is not labeled to better ensure that the labeled data is representing the speech-related behavior within the interval.

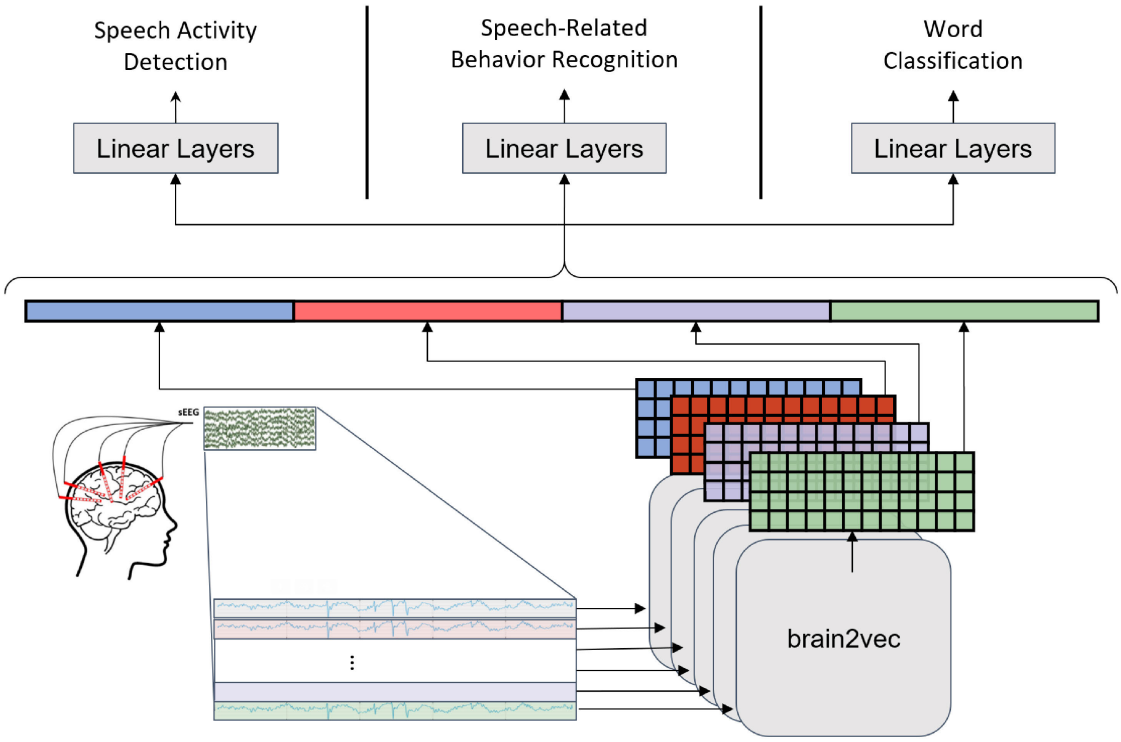


FIGURE 5. Diagram of the downstream task training procedure. Given a participant's sEEG signals, a 0.5 s window across all electrodes is considered. The window for each single electrode, and its corresponding RAS coordinates, are passed to a brain2vec model, producing context representations for each electrode. These representations are flattened, concatenated, then passed through a 16-unit linear layer before finally being passed through the N-class classification output linear layer. The value of the N-class is dependent on the task being optimized.

3) WORD CLASSIFICATION

The word classification task requires the fine-tuning model to classify a word from a restricted set. The data collection protocol does not repeat sentences, but across all sentences, there are a set of words that are repeated and are not stop words. Stop words are the most common words such as articles, prepositions, or pronouns, which are commonly excluded when training natural language schemes. Ten such non-stop words are selected arbitrarily for the present analysis.

Forced word alignment was performed on the audio data to identify word start and stop times. These word start-stop times were used to label the corresponding sEEG segments with the associated word.

The training set consists of the sEEG windows corresponding to all 10 selected non-stop words from their first appearance. For the test set, the model is given an sEEG window from 5 of the 10 words, taken from the second appearance of the word. The remaining second appearances of each word are used for cross-validation during training. For example, if the bolded training word was taken from the sentence *The fish turned on the bent* ***hook***, then the word would be tested on sEEG segments corresponding to the subsequent sentence *He was caught,* ***hook****, line, and sinker.* In this way, the word classification task is challenged with previously unseen data. The selection of which word's second occurrence is included in the cross-validation versus the test set is randomized for each participant's trial.

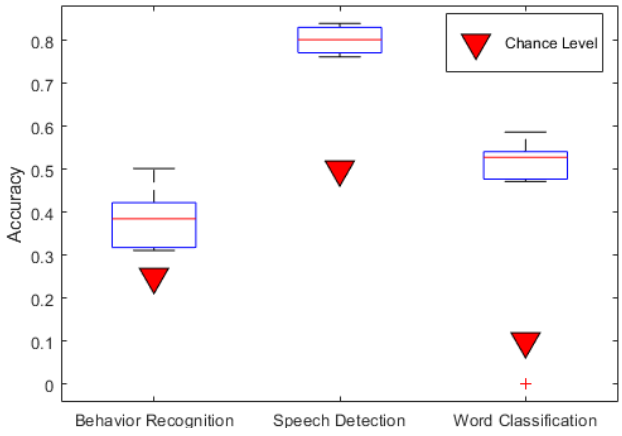


FIGURE 6. Box plot of accuracy across participants for the 3 downstream task. Red triangles represent the chance accuracy for each task.

V. RESULTS

The performance of brain2vec is evaluated by comparing the balanced accuracy for each of the respective classification tasks. Figure 6 and Table 2 show the balanced accuracies of the three tasks for each participant, the overall average

accuracy, and the chance accuracy of the classification task. In order to verify chance accuracy, the downstream tasks were trained on randomly assigned labels, and these results are included in the table.

Compared to the Speech Activity Detection and the Word Classification task, Speech-Related Behavior Recognition had higher inter-participant variability, and was overall closer to chance accuracy for the task.

The Speech Activity Detection task's average balanced accuracy is 89.8% and it achieves the smallest variance among the tasks. All participants were significantly above chance accuracy of 50%, and the worst performer attained 82.7% accuracy. For comparison, in a recent speech activity detection study using the same Harvard Sentence dataset, logistic regression models as well as CNN models achieved an average accuracy of 82-84% [40]. Several other studies using intracranial signals reported results ranging between 80% - 94% accuracy [10], [41]. All of these studies used fully supervised learning methods.

Word Classification yielded the most promising performance of the three tasks. With only one training example of each word from the repeated word set, the average participant accuracy was 52.9% when tested on repeated words. Moreover, the hold-out words were from entirely different sentences with distinct broader contexts. As mentioned in Section II-C, Participant 1 did not complete all 50 sentences during the data collection experiment. They did not have the samples required to be evaluated on the Word Classification task, and thus are excluded from this portion of the evaluation experiments.

A notable observation seen in Figure 6 is that, while there were some exceptions, there was a tendency for participants to perform consistently in comparison to other participants across the three tasks. For example, participants 4 and 6 performed in the top half for all tasks, while participants 3 and 7 performed in the bottom half.

Figure 7 shows the cross-validation loss of during pretraining for all participants. It can be observed that the models converge to generally similar losses, that is, there do not appear to be order-of-magnitude differences. This is expected, as each model shares approximately 6/7 of the electrode data corpus. Nevertheless, it is confirmation of that there is some measure of consistency in the convergence process.

The confusion matrices of downstream classification tasks are shown in Figure 8. The Behavior Recognition task shows that *imagining* was confused more often with *listening* and *mouthing* than with *speaking*. Further, *speaking* was confused most often with *mouthing*. This observation may indicate a closer mechanistic relationship exists between imagined speech and listening or mouthing, than does between imagined and overt speaking [42], [43], [44].

Figures 9, 10, and 11, and respectively show the 3-component t-SNE [45] of the pretrained features for each fine-tuning task. The figures give an indication that the context representations learned by brain2vec are meaningful to each speech domain task. It is observed that, for each task,

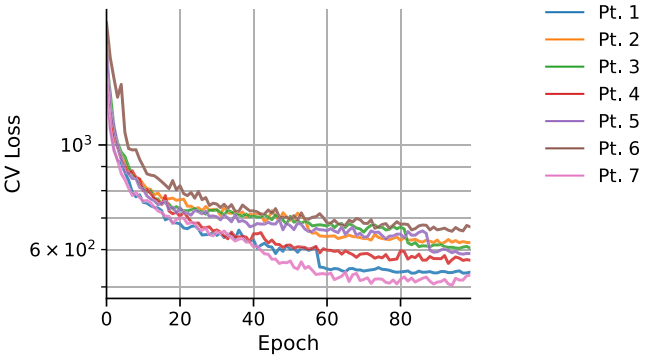


FIGURE 7. Cross validation loss of brain2vec model over pretraining epochs.

TABLE 2. Balanced accuracy of downstream tasks. Participant 1 did not have a complete dataset needed for Word Classification and is therefore omitted.

Participant	Speech-Related Behavior Recognition	Speech Activity Detection	Word Classification
1	33.4%	91.1%	-
2	36.2%	95.0%	54.1%
3	44.3%	82.7%	48.3%
4	49.4%	89.3%	40.9%
5	36.1%	88.9%	55.7%
6	46.4%	89.9%	56.0%
7	49.8%	91.7%	62.6%
Average	42.2%	89.8%	52.9%
Random	27.0%	54.8%	12.4%
Chance Acc.	25%	50%	10%

there are clear regions of separability for each of the classes. Particularly, word classification in Figure 9 shows distinct separation between words. This likely contributes to the impressive performance of the word classification task given comparatively little training data, as the context representations show clear differentiation prior to supervised training.

VI. DISCUSSION

The performance of brain2vec on the three disparate downstream tasks showcases the generalizability of the self-supervised features learned by the procedure. While all tasks achieve better than chance accuracy for all participants, in particular, the speech detection task approaches accuracies on par with other supervised learning methods, and the word classification task exhibits promising results using only a small amount of labeled data.

The main objective of this analysis was to develop and establish the efficacy of the pretraining procedure and model, using the performance on downstream tasks as a measure rather than an end goal. The manner in which the model pretrains inherently makes it difficult to draw conclusions directly from analyzing the context representations, and is further complicated with the addition of the fine-tuning linear layers. Thus, performance on downstream tasks is used to draw indirect evidence of the efficacy of pretrained features.

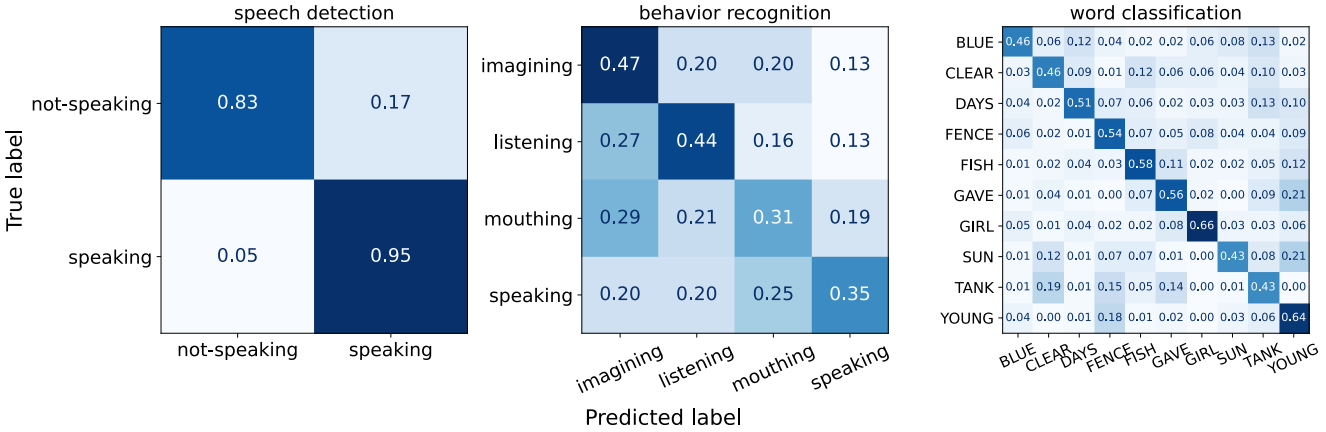


FIGURE 8. **Confusion matrices of fine-tuning classification tasks across all participant test sets. Each row (true label) is normalized independently, giving the portion predicted class labels across all of the true samples evaluated.**

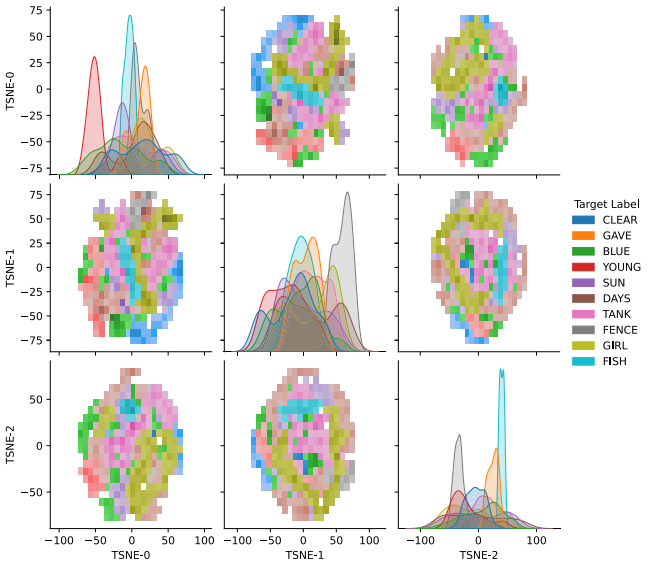


FIGURE 9. **Visualization of 3 t-SNE components from the pretrained features on an unseen users data (Pt. 7), colored by the Word Classification fine-tuning task.**

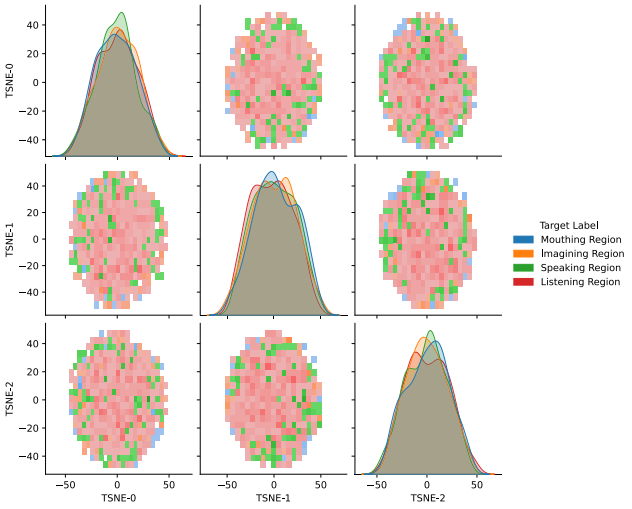


FIGURE 10. **Visualization of 3 t-SNE components from the pretrained features on an unseen users data (Pt. 7), colored by the Behavior Recognition fine-tuning task.**

The classification tasks were purposefully selected to cover disparate speech representations that yield a range of classification challenges. Otherwise, the selected classification tasks are somewhat arbitrary with respect to the common speech representation available in this particular dataset, and the framework is designed to be agnostic to specific speech representations.

Performance on the Speech-Related Behavior Recognition task, while comparatively exhibiting the weakest performance, can also be considered the most challenging of the three classification tasks. The neural circuits for perceiving speech, and producing overt, mouthed, and imagined speech, are highly intertwined [43], [46], [47]. Nevertheless, it is encouraging that the context representations of the model appear to encode some neural correlates of these behaviors.

The Word Classification task is essentially a few-shot learner, only provided a pair of training examples (i.e. word utterances) of each class before evaluation - one for optimization, and another for validation. In contrast, a study recently showed results ranging from 30-60% on a similar classification task using ECoG signals and a transformer architecture, though in a fully supervised manner [48]. This demonstrates the utility of the self-supervised method: using only unlabeled data, features are learned and guided into hidden, likely subword, units. Then, it is posited, comparatively little data is required to map these features to a word space.

The success of brain2vec is likely due to several factors. The self-supervised training of latent representations with quantized targets, while keeping the learned context representation as continuous, is a gentle influence to learn not fully-discrete codewords, but instead grouped clusters in the continuous space, known as hidden units. In this way, features are guided towards self-determined clusters, while still allowing the model to fully leverage the rich context of continuous-space features. Because of the self-supervised nature, these

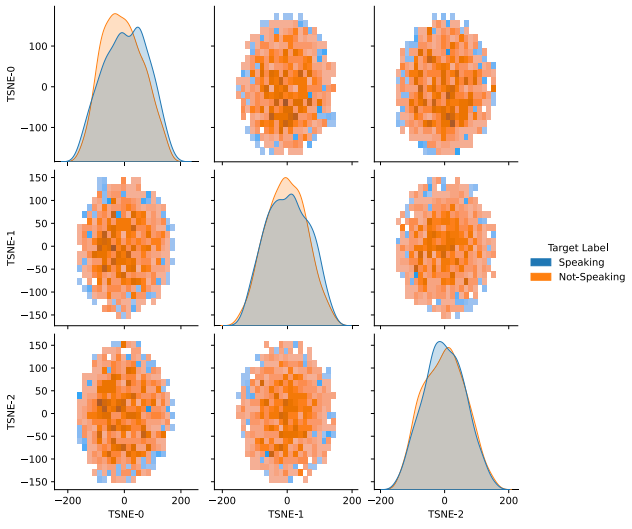


FIGURE 11. Visualization of 3 t-SNE components from the pretrained features on an unseen users data (Pt. 7), colored by the Speech Detection fine-tuning task.

clusters are not matched to any linguistic unit, such as words or phonemes, and instead are self-determined by the network. However, because the training data are strictly from the speech domain, it is likely that the hidden units are converging to neural versions of some, possibly combinations of, linguistic units. This is a potential explanation as to why the Word Classification task was successful using sparse training data.

The projection of RAS electrode coordinates to a common brain atlas allowed for the pooling of data from multiple participants to provide informative absolute brain location data of electrodes to the model. With a sufficient data corpus and electrode coverage, this type of self-supervised model has the potential to train a brain signal regression given neighboring signal data.

During model development, several issues were observed that adversely impacted training success. The objective term weights, *and* λ, required exploration with small experiments to find appropriate configurations that avoided codebook collapse - wherein the model used few codewords or the codewords would have little variance overall.

Under some conditions, brain2vec would fail to converge and be maintained at a high CV loss, but this could not be consistently replicated and never occurred with the configuration presented in this work. We found large improvements in consistency after implementing appropriate weight initialization. Convolution and linear layers were initialized from $\mathcal{N}(0., 0.02)$, BatchNorm parameters from $\mathcal{N}(1., 0.02)$ with a bias of zero, and LayerNorm parameters are initialized with 1.0 and zero bias. This implies a sensitivity to initial conditions and hints at further improvement through more sophisticated initialization schemes and complex learning rate paradigms as explored in other language model methods [26], [38]. This is likely an attribute of the model architecture rather than the particular data.

The number of transformer blocks, the latent representation vector dimension, and other factors that determined model complexity, impacted performance on downstream tasks. This is likely a balance with the amount of available data. Language models using transformer architectures often have a 'large' model variant with 24 transformer blocks [24], [25], [26]; however, these models are typically pretrained using on the order of 60,000 hours of data, whereas the proposed approach was effective using slightly over 1 hour of data for pretraining.

Additional sEEG training data would allow for a deeper model with more transformer blocks, a longer input sequence, or a larger embedding dimension, which might in turn provide greater context and learn richer representations of multiple speech and speech-related processes. The downstream tasks explored here are constrained by the nature of the speech data available. With enough data, and a sufficient depth of network, it is conceivable for brain2vec to serve as the backbone of an even more generalized model; one capable of discriminating overt or imagined speech intention, then decoding the speech from the same initial feature set.

As this work is largely an initial proof-of-concept, there are many possibilities to extend and optimize this framework. Here, a linear output layer was implemented for simplicity and comparability; however, more complex decoder paradigms, such as a GPT transformer stack may be better suited to more complex downstream tasks. The recent and growing corpus of publicly available data sets [49] can be leveraged to pool data from participants across experiments, and potentially across sensing paradigms, as long as the dataset includes electrode coordinates for the positional embedding.

VII. CONCLUSION

This work developed and evaluated brain2vec, a transformer-based self-supervised model that learns speech-related hidden unit representations from unlabeled sensor-level sEEG data. The outputs of brain2vec after pretraining are used to fine-tune a classifier on labeled data from three disparate downstream speech classification tasks. All tasks perform above chance accuracy for all participants, while the speech activity detection and word classification task performance rival competitive supervised learning methods.

REFERENCES

[1] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, pp. 327–332, Mar. 2013.
[2] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "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.
[3] 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.
[4] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Frontiers Neurosci.*, vol. 9, p. 217, Jun. 2015.

[5] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers Hum. Neurosci.*, vol. 9, p. 97, Feb. 2015.

[6] 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.

[7] 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.

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

[9] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, and P. L. Kubben, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Commun. Biol.*, vol. 4, no. 1, pp. 1–10, 2021.

[10] 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, Jul. 2019.

[11] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, P. F. Sun, J. Chartier, M. E. Dougherty, P. M. Liu, G. M. Abrams, A. Tu-Chan, K. Ganguly, and E. F. Chang, "Neuroprosthesis for decoding speech in a paralyzed person with anarthria," *New England J. Med.*, vol. 385, no. 3, pp. 217–227, Jul. 2021.

[12] S. Martin, P. Brunner, I. Iturrate, J. D. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Word pair classification during imagined speech using direct brain recordings," *Sci. Rep.*, vol. 6, no. 1, pp. 1–12, May 2016.

[13] 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.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[18] H. Fan, Y. Yang, and M. Kankanhalli, "Point 4D transformer networks for spatio-temporal modeling in point cloud videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14199–14208.

[19] H. Akbari, W.-H. Chuang, L. Yuan, S.-F. Chang, B. Gong, R. Qian, and Y. Cui, "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24206–24221.

[20] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021, *arXiv:2102.05095*.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.

[22] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," 2021, *arXiv:2103.15691*.

[23] H. Zhao, L. Jiang, J. Jia, P. H. S. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 16259–16268.

[24] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020, *arXiv:2006.11477*.

[25] W.-N. Hsu, B. Bolte, Y.-H. Hubert Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," 2021, *arXiv:2106.07447*.

[26] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," 2021, *arXiv:2108.06209*.

[27] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[28] L. Collins, "3D model-based segmentation of individual brain structures from magnetic resonance imaging data," Ph.D. thesis, Dept. Biomed. Eng., McGill Univ., Montreal, QC, Canada, 1994. [Online]. Available: https://escholarship.mcgill.ca/concern/theses/3t945s766

[29] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *Proc. IEEE Conf. Rec. Nucl. Sci. Symp. Med. Imag. Conf.*, vol. 3, Nov. 1993, pp. 1813–1817.

[30] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, Aug. 2012.

[31] A. Gramfort, M. Luessi, E. Larson, A. D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and S. M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers Neurosci.*, vol. 7, no. 267, pp. 1–13, 2013.

[32] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, 2012.

[33] W.-N. Hsu, Y.-H.-H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6533–6537.

[34] J. Kohler, M. C. Ottenhoff, S. Goulis, M. Angrick, A. J. Colon, L. Wagner, S. Tousseyn, P. L. Kubben, and C. Herff, "Synthesizing speech from intracranial depth electrodes using an encoder–decoder framework," 2021, *arXiv:2111.01457*.

[35] J. L. Ba, J. R. Kiros, and E. G. Hinton, "Layer normalization, 2016, *arXiv:1607.06450*.

[36] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Mar. 2011.

[37] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," 2016, *arXiv:1611.0114*.

[38] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," 2019, *arXiv:1910.05453*.

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

[40] P. Z. Soroush, M. Angrick, J. Shih, T. Schultz, and D. J. Krusienski, "Speech activity detection from stereotactic EEG," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 3402–3407.

[41] 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.

[42] P. Langland-Hassan and A. Vicente, *Inner Speech: New Voices*. New York, NY, USA: Oxford Univ. Press, 2018.

[43] H. S. Gauvin and R. J. Hartsuiker, "Towards a new model of verbal monitoring," *J. Cognition*, vol. 3, no. 1, p. 17, Sep. 2020.

[44] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behav. Brain Sci.*, vol. 22, no. 1, pp. 1–38, Feb. 1999.

[45] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[46] T. Proix, J. D. Saa, A. Christen, S. Martin, B. N. Pasley, R. T. Knight, X. Tian, D. Poeppel, W. K. Doyle, O. Devinsky, L. H. Arnal, P. Mégevand, and A.-L. Giraud, "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features," *Nature Commun.*, vol. 13, no. 1, p. 48, Jan. 2022.

[47] A. Roelofs, "Spoken word planning, comprehending, and self-monitoring: Evaluation of WEAVER++," in *Phonological Encoding and Monitoring in Normal and Pathological Speech*. New York, NY, USA: Psychology Press, 2005, pp. 42–63.

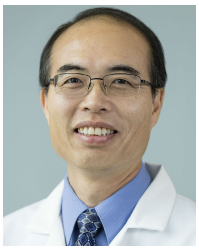
[48] S. Komeiji, K. Shigemi, T. Mitsuhashi, Y. Iimura, H. Suzuki, H. Sugano, K. Shinoda, and T. Tanaka, "Transformer-based estimation of spoken sentences using electrocorticography," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1311–1315.

[49] M. Verwoert, M. C. Ottenhoff, S. Goulis, A. J. Colon, L. Wagner, S. Tousseyn, J. P. Van Dijk, P. L. Kubben, and C. Herff, "Dataset of speech production in intracranial electroencephalography," *Sci. Data*, vol. 9, no. 1, p. 434, Jul. 2022.

SRDJAN LESAJA received the B.S. degree in applied mathematics and the M.S. degree in statistics from the Georgia Institute of Technology, Atlanta, GA, USA, and the Ph.D. degree in biomedical engineering from Virginia Commonwealth University, Richmond, VA, USA. He is currently a Data Scientist working in public health. His research interests include brain–computer interfaces, AI/ML, neuroscience, and biotechnology.

MORGAN STUART is currently pursuing the Ph.D. degree in computer science with Virginia Commonwealth University.

He is currently a Research Data Scientist with the United Network for Organ Sharing. His work experience spans health, finance, and cyber security domains, including applications in embedded systems, distributed computing, and predictive modeling. His research interests include predictive model performance and utility in real-world applications outside the datacenter.



JERRY J. SHIH received the Medical degree from the School of Medicine in Los Angeles, University of California, Los Angeles (UCLA), Los Angeles, CA, in 1984. He completed an Internship in internal medicine at the Cedars-Sinai Medical Center, Los Angeles, and a Residency (Chief Resident) in neurology and a fellowship in clinical neurophysiology and critical care neurology at UCLA Medical Center, Los Angeles.

He is currently a Chief of the Division of Epilepsy and a Professor of neurosciences with the Department of Neurosciences, University of California San Diego, San Diego, CA. He is also the Director of the Comprehensive Epilepsy Center, University of California San Diego Health. He has written or co-written over 100 articles for peer-reviewed journals and book chapters on treating and managing neurological disorders.

Dr. Shih is a fellow of the American Neurological Association, the American Academy of Neurology, and the American Epilepsy Society. He is a Board Certified in neurology with added qualifications in clinical neurophysiology and epilepsy by the American Board of Psychiatry and Neurology. He is also a Board Certified by the American Board of Clinical Neurophysiology. He has served as the Principal Investigator or the Co-Principal Investigator on numerous federal, foundation, and industry research grants and clinical trials. He recently served as a contributing Editor for the peer-reviewed journal Epilepsy Currents. He is an Ad Hoc Reviewer for several journals, including *Annals of Neurology* and *Epilepsia*.

PEDRAM Z. SOROUSH received the B.Sc. degree in electrical engineering (majored in bio electric) from the Sharif University of Technology (SUT), in 2019. He is currently pursuing the Ph.D. degree in biomedical engineering with Virginia Commonwealth University (VCU), where he conducts research as a member of the ASPEN Laboratory. His Ph.D. dissertation title is characterization and decoding of speech-related activity from intracranial signals. His research interests include biomedical signal processing, machine learning, data analysis, brain–computer interfaces, and neural engineering.

TANJA SCHULTZ (Fellow, IEEE) received the Diploma degree and the Ph.D. degree in informatics from the University of Karlsruhe, Germany, in 1995 and 2000, respectively. She joined the Language Technologies Institute at Carnegie Mellon University, Pittsburgh, PA, USA, in 2000, and remained an Adjunct Research Professor, until 2021. From 2007 to 2015, she was a Full Professor of informatics at the Karlsruhe Institute of Technology, Karlsruhe, Germany, before she became a Professor of Cognitive Systems at the University of Bremen, Germany, in April 2015. Since 2007, she has been directing the Cognitive Systems Laboratory, where her research interests include the processing, recognition, and interpretation of human signals for biosignal-adaptive cognitive systems and their applications. She is the Spokesperson of the high-profile area "Minds, Media, Machines" at the University of Bremen and a member of the Board of Directors of the DFG CRC EASE and the Leibniz Science Campus on Digital Public Health. She is a fellow of ISCA, in 2016; the European Academy of Sciences and Arts, in 2017; and the Asia-Pacific Artificial Intelligence Association, in 2021.

MILOS MANIC (Fellow, IEEE) is currently a Professor with the Computer Science Department and the Director of VCU Cybersecurity Center with the Virginia Commonwealth University. He completed over 50 research grants in AI/ML in cyber and energy and intelligent controls. He authored over 200 refereed articles, has given over 40 invited talks around the world, authored over 200 refereed articles in international journals, books, and conferences, holds several U.S. patents and has won 2018 Research and Development 100 Award for Autonomic Intelligent Cyber Sensor (AICS), and one of top 100 science and technology worldwide innovations, in 2018. He holds Joint Appointment with Idaho National Laboratory, as an Inductee of U.S. National Academy of Inventors (class of 2019), and a fellow of Commonwealth Cyber Initiative (specialty in AI and cybersecurity).

He is an IEEE IES President Elect. He was a recipient of the IEEE IES 2019 Anthony J. Hornfeck Service Award, the 2012 J. David Irwin Early Career Award, and the 2017 IEM Best Paper Award. He is an Associate Editor of IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, Open Journal of Industrial Electronics Society, IES President Elect, and a Senior Life AdCom Member. He served as an AE of IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the Founding Chair of IEEE IES Technical Committee on Resilience and Security in Industry, and the General Chair of IEEE IECON 2018 and HSI 2019.

DEAN J. KRUSIENSKI (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from The Pennsylvania State University, University Park, PA, USA. He conducted the Postdoctoral Research in the Brain–Computer Interface Laboratory, Wadsworth Center of the New York State Department of Health. He is currently a Professor and the Graduate Program Director of biomedical engineering with Virginia Commonwealth University (VCU), Richmond, VA, USA, where he also directs the Advanced Signal Processing in Engineering and Neuroscience (ASPEN) Laboratory. His research interests include biomedical signal processing, machine learning, brain–computer interfaces, and neural engineering.