A Study of Decoding Silent Speech from Brain Signals Using Deep Learning

Ishita Saraf
School of Computer Science and Engineering

Professor Guan Cuntai
Dr Tushar Chouhan
School of Computer Science and Engineering

Abstract **-** Neurological impairments such as strokes, paralysis, epilepsy, and amyotrophic lateral sclerosis (ALS) can result in loss of motor communication abilities. Modern researchers have been trying to develop a system that can decode speech directly from such patient's unaffected neural activity.

Two broad methods for recording neural activity for the purpose of speech decoding have been studied in the literature. The first is electrocorticography (ECoG), an invasive method. The second approach uses electroencephalography (EEG), a non-invasive method to record brain activity from the patient's scalp. Although more convenient for participants, EEG has much lower signal-to-noise ratio. Hence, we decided to explore non-invasive EEG recordings for speech decoding using deep learning.

In our research, we have studied and implemented novel, state-of-the-art deep learning methods to decode speech from EEG. One of the methods, proposed by Saha et al., 2019 aims to capture the "information exchange between different parts of the brain" during speech production, in the form of channel-cross-covariance (CCV) matrix and use it to train a four-stage hierarchical model comprising a Convolutional Neural Network (CNN), a Long Short-Term Memory (LSTM) network, an unsupervised deep autoencoder (DAE) and an Extreme Gradient Boost classification layer (XG Boost). In this work, we review the literature and draw on the possible strengths and working principles of the different methods reported in the literature. We also discuss some challenges that the current methods and datasets face. In doing so, we aim to highlight the limitations in these works that may provide motivation for future studies to further improve the research in this area.

Keywords – Deep learning, Speech decoding, EEG, Brain Computer Interface.

1. INTRODUCTION

Sudden loss of verbal communication abilities due to neurological disorders make an individual's life devastating. Currently, such patients rely on primitive spelling-based communication tools by using their residual head or eye movements. With technology advancement and increased availability, patients have increasingly started making use of Brain Computer interfaces (BCIs) for communication purposes. A brain-computer interface (BCI) is a computer-based system that acquires brain signals, analyzes them, and translates them into commands that are relayed to an output device to carry out a desired action. [7] Patients can control cursors by just their thoughts and no physical movement at all, but current BCIs still function on a spelling-based approach. Although greatly helpful, most users struggle to transmit more than 10 words per minute through these devices as compared to 150 words per minute for natural speech. Hence availability of technology that can translate neural activity into speech would be life-changing for such patients. We aim to help cross the hurdle of spelling-based communication to develop a system that can decode speech directly from the user's neural activity that remains intact even after such tragedies. Extensive research is ongoing in this field to make speech neuroprosthetics a reality. In our study, we have familiarized ourselves with state-of-the-art deep learning methods to decode speech from brainwaves and attempted to reproduce their results on the KARA ONE dataset (University of Toronto, Department of Computer Science).

We have studied speech decoding on broadly two categories of brainwaves – the first is Electrocorticography (ECoG) and the second is Electroencephalography (EEG). ECoG is a type of electrophysiological monitoring that uses electrodes placed directly on the exposed surface of the brain to record electrical activity from the cerebral cortex. [8] An invasive method, ECoG yields brain signals of high resolution and effectively captures high-frequency gamma waves, crucial for speech decoding. Despite having high signal-to-noise ratios, ECoG is only used in severe cases due to the complex nature of surgery required to place the electrodes on the surface of the brain of the patient. In contrast, conventional EEG electrodes monitor this brain activity from outside the skull. Electrodes are placed on the subject's scalp to

record the electrogram of neural activities. Being non-invasive in nature, it is the more feasible choice for users, but the brain waves being recorded suffer from attenuation from the skull and scalp, resulting in a much poorer signal-to-noise ratio and temporal and spatial resolution. Hence, to effectively extract speech from EEG waves, we must process the raw waves using methods such as artifact removal, band-pass filtering, Laplacian filtering, and independent component analysis.
In this report, we will review latest deep learning models for speech decoding from both ECoG and EEG waves, outline our findings on replicating one of the models and discuss the shortcomings and challenges faced while implementing the network.

2. LITERATURE REVIEW

Deep learning models are based on neural network architectures, designed specific to the learning goal. Extensive research has been ongoing in this field in the past few years. There are several novel models designed by researchers, specifically for the task of speech decoding from brainwaves, which have contributed significantly to the goal of developing a speech decoder.

Graves & Schmidhuber, 2005 found that recurrent neural networks (RNNs) are better than convolutional neural networks (CNNs) for the task of framewise phoneme classification i.e., mapping a sequence of speech frames to a sequence of corresponding phoneme labels, as they treat the inputs are temporal rather than spatial. Moreover, Long Short Term Memory (LSTM) networks are even better than recurrent neural networks as it solves the problem of exploding or vanishing gradients. They noticed that contextual information is crucial for speech processing. When using an LSTM for speech processing, outputs are mostly based on previous context as inputs are processed in a unidirectional chronological order. However, in speech production, future context is just as important as past context. To introduce future context, they suggested bidirectional LSTM (bLSTM), an architecture that exploits past as well as future context well for speech processing. In a bLSTM, each training sequence is presented forwards and backwards to the network. Hence the bLTSM network has complete contextual information about every point in the sequence. Graves & Schmidhuber, 2005 stated that " bidirectional networks outperform unidirectional ones, and Long Short Term Memory (LSTM) is much faster and also more accurate than both standard Recurrent Neural Nets (RNNs) and time-windowed Multilayer Perceptrons (MLPs)."

Makin et al., 2021 treated speech decoding as a task of machine translation, exploiting an encoder-decoder model based on LSTM networks. After training their model on ECoG data, they achieved promising accuracies, with WER even close to zero (perfect decoding) for one of their four participants.

Anumanchipalli et al., 2019 suggested a two-stage model to decode acoustic parameters in the form of Mel-frequency cepstral coefficients (MFCCs), from high density ECoG signals recorded from parts of the brain that are most active during speech production. According to Anumanchipalli et al., 2019, speech is a "highly efficient form of communication produced from a fluid stream of overlapping, multi-articulator vocal tract movements". Hence, a biomimetic approach, focusing on articulatory movements of the vocal tract can be a promising way of building a decoder network, and most intuitive for users to learn using. For this reason, Anumanchipalli et al., 2019, introduced an intermediate stage in their decoder model for articulatory kinematic representations. The first stage of the decoder, comprising a bidirectional LSTM network, infers articulatory kinematic features from ECoG recordings. The second stage, again constituting of a bidirectional LSTM network, uses these intermediate vocal tract kinematic trajectories to decode acoustic speech features such as MFCCs and sub-band voicing strengths. As Anumanchipalli et al., 2019, did not have directly recorded vocal tract movement data corresponding to each trial, they formulated a statistical method to estimate articulatory kinematic features such as movement of tongue, jaws, lips and manner of articulation from audio data. They used data from MOCHA-TIMIT and MNGU0 corpora (containing audio and corresponding vocal tract movement EMA recordings) along with some additional complementary speech features like pitch and voicing, to train a stacked encoder decoder model to obtain articulatory kinematic features for various utterances. Hence, using this intermediate model, Anumanchipalli et al., 2019 could obtain articulatory kinematics in addition to ECoG recordings and audio features for various utterances, for multiple participants, across experiment trials. They used these data to train their two-stage bidirectional LSTM based network as shown in Fig. 2. Their speech decoder produced promising results. The decoded audio spectrograms retained the salient energy and spectro-temporal patterns as present in the original audio spectrogram and correctly reconstructed the silence in between sentences. They even conducted single word identification and sentence level transcription tasks for the reconstructed audio from the decoded spectrographs. The tasks produced results consistent with natural speech perception. Measuring the mel-cepstral distortion and Pearson's correlation between the original and decoded acoustic features proved that the network

decoded speech correctly above the level expected by chance (Pearson's correlation > 0.6). This research corroborates the theory that electrophysiological data are correlated to speech can be used for decoding speech.

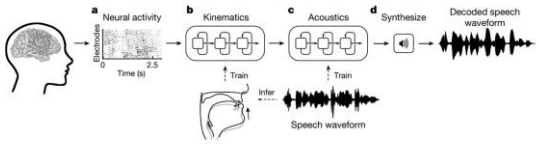

Fig 1: An end to end speech decoding and reconstruction pipeline (Anumanchipalli et al., 2019)

Although these studies are conducted on high resolution ECoG data, obtaining which is difficult due to surgical requirements, it gives us the motivation that similar network architectures and methods can be used for less resolution EEG data for deciphering speech.

We have also reviewed several EEG based speech decoding models. We found the model suggested by Saha et al., 2019 to be intriguing. It consists of a four-stage hierarchical network architecture comprising a CNN, an LSTM, a deep autoencoder and an Extreme Gradient Boost Classifier. We found this model intriguing as it uses the channel cross covariance matrix (CCV) of the EEG channels instead of raw EEG signals to train their model and to obtain encouraging results. We have described and analyzed this EEG based speech classifier pipeline in greater detail in the following section.

3. ANALYSIS OF AN EEG BASED PIPELINE

3.1 DETAILED DESCRIPTION

According to Saha et al., 2019, speech production is a complex task and "higher cognitive processes underlying speech planning and synthesis involve frequent information exchange between different parts of the brain". "Multichannel EEG data is high dimensional multivariate time series data." Training models on high dimensional raw EEG not only needs long training durations and more hardware resources, but it is also seen that conventional deep neural networks, such as CNNs, RNNs and autoencoders, individually, are unable to learn from such complex features in raw EEG data. Hence, instead of directly using raw EEG data as their model's input, they find it beneficial to decrease its dimensionality and capture the information exchange between electrodes by computing the CCV matrix.

Along with using CCV matrix as the model input, and corresponding labels of utterances as target outputs, Saha and colleagues adopt a mixed, hierarchical deep neural network strategy. As shown in Fig. 2, they devise a four-stage decoder model consisting of a CNN, LSTM, deep autoencoder and an XGBoost classifier. At the first stage of the hierarchy, the input is passed parallelly to the CNN and the LSTM. The two-dimensional CNN, comprising two convolutional layers and two fully connected layers stacked together, is trained with the CCV as inputs and corresponding labels as target outputs to decode spatial correlations between electrodes. The LSTM, stacking two fully connected and two recurrent layers, is trained in the same manner as the CNN and is used to extract the hidden temporal features. The outputs from the penultimate layers of the CNN as well as the LSTM are concatenated together to form a single feature vector. This vector "forms a joint spatio-temporal encoding of the cross-covariance matrix". (Saha et al., 2019) At the second stage of the hierarchy, an unsupervised deep autoencoder (DAE) is trained on the vector produced by the first stage, to decrease the dimensionality of the spatiotemporal encodings even more and remove background noise effects. The DAE consists of 3 encoder and 3 decoder layers. The reconstructed vector from the DAE is passed into an Extreme Gradient Boost (XGBoost) classification layer, forming the last stage of the hierarchy. Trained in a supervised manner to predict the final classes, the XGBoost decision tree classifies structured data well. According to Saha et al., 2019, "Since our EEG phonological pairwise classification has an internal structure involving individual phonemes and words, it seems to be a reasonable choice of classifier".

In contrast to ECoG signals, due to the poor resolution of EEG signals, it is difficult to obtain a pairwise signal-phoneme mapping. Hence Saha and colleagues train this network to tackle five binary classification problems. They perform these tasks on experimental data from the KARA ONE database.[5] The dataset consists of EEG recordings and corresponding transcriptions of phonemic and single word prompts as participants image speaking them across several experimental trials. "The five binary tasks include classifying these phonemic/syllabic utterances (*/iy/, /piy/, /tiy/, /diy/, /uw/, /m/, /n/*) as well as words (*pat, pot, knew* and *gnaw*) into the following categories: presence/absence of consonants, phonemic nasal, bilabial, high-front vowels and high-back vowels". (Saha et al., 2019)

Saha et al., 2019 found their hierarchical model to produce an average accuracy of 77.9%, yielding a considerable 22.5% improvement over older methodologies. [5,6]

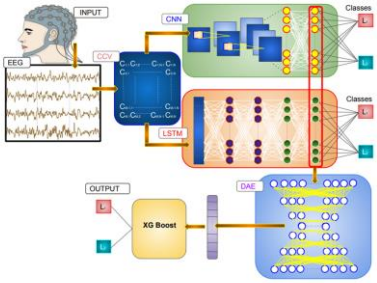

Fig 2: Four-stage hierarchical model for speech classification (Saha et al., 2019)

3.2 LIMITATIONS OF THE EEG-BASED PIPELINE

We replicated this four-stage model and trained it in the same manner as described by Saha et al., 2019. We obtained EEG data for imagined speech with corresponding target labels from the publicly available KARA ONE database. We used python's MNE library to preprocess the raw EEG and used Pytorch libraries for creating and training the CNN, LSTM and deep autoencoder. We trained and run our model on Google Colaboratory's NVIDIA Tesla K80" GPU environment. The results obtained by our implementation does not achieve similar accuracies as reported in findings of Saha et al., 2019. Apart from the overall classification accuracies being lower, we noticed that individually, the CNN and LSTM are greatly biased towards those class labels which have more training data thereby producing higher classification accuracy (~70%) than those reported. Possible implementation differences that are not clarified in the work may have led to the observed differences in the results. For instance, it is not clear how the authors deal with the imbalance of classes in the training data to prevent the model from tilting towards one class. Moreover, the key performance metric used to judge the efficacy of the overall model is classification accuracy, which may not take into account the biased prediction due to imbalanced classes. It is also unclear how the authors segment each trial's data due to which differences may have arisen. We also postulate that the data is insufficient for training such a complex deep learning model. Hence previous works [5,6] which used simpler machine learning algorithms such as Deep Belief Networks (DBN) and Support vector machines (SVM) were able to converge their models effectively even with less data available, to produce satisfactory results for the same five binary classification tasks on the same dataset.

4. CONCLUSION AND FUTURE DIRECTION

As a step towards successfully building a speech decoder from EEG data, we have familiarized ourselves with several state-of the art methods available for speech decoding using both ECoG and EEG data. We have analyzed which networks are the best for accomplishing the task of speech decoding. We have also replicated a novel speech classifier model which gave us insight into the challenges one faces when implementing an EEG based speech decoder/classifier. One must deal with the constraints of insufficient public availability of data as speech decoding from EEG is fairy new topic of research. Future works in this field must also ensure that they clearly specify all steps of their input processing and training as well as use comprehensive performance metrics so that it is reflective of the true performance of the model.

4. ACKNOWLEDGEMENT

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project.

5. REFERENCES

[1] Saha, P., Fels, S., & Abdul-Mageed, M. (2019). Deep learning the EEG manifold for phonological categorization from active thoughts. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2019.8682330

[2] Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568(7753), 493–498. https://doi.org/10.1038/s41586-019-1119-1

[3] Makin, J. G., Moses, D. A., & Chang, E. F. (2021). Speech decoding as machine translation. SpringerBriefs in Electrical and Computer Engineering, 23–33. https://doi.org/10.1007/978-3-030-79287-9_3

[4] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

[5] Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2015.7178118

[6] Pengfei Sun and Jun Qin, "Neural networks based eeg-speech models," arXiv:1612.05369, 2016.

[7] Shih, J. J., Krusienski, D. J., & Wolpaw, J. R. (2012, March). *Brain-computer interfaces in medicine*. Mayo Clinic proceedings. Retrieved June 30, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3497935/#:~:text=A%20brain-computer%20interface%20(BCI,carry%20out%20a%20desired%20action.&text=%E2%96%A0-,In%20principle%2C%20any%20type%20of%20brain%20signal%20could%20be,to%20control%20a%20BCI%20system.

[8] Wikimedia Foundation. (2022, June 30). *Electrocorticography*. Wikipedia. Retrieved June 30, 2022, from https://en.wikipedia.org/wiki/Electrocorticography