Ministry of Education and Science of Ukraine
National University of «Kyiv-Mohyla Academy»
Department of Mathematics in the Faculty of Computer Sciences

NATIONAL UNIVERSITY OF
KYIV-MOHYLA ACADEMY

Bachelor Thesis

In the specialty «Computer Science» (122)

DECODING SPEECH FROM ECOG WITH MACHINE TRANSLATION MODELS

Supervised by:

Prof. Nadyia Shvai _____ *(signature)*

Prof. Bo Wang _____ *(signature)*

Authored by a 4^{th} - year student:
Roman Burakov

«_____» _____ 2023

Kyiv - 2023

INDIVIDUAL TASK

Ministry of Education and Science of Ukraine

National University of «Kyiv-Mohyla Academy»

Department of Mathematics in the Faculty of Computer Sciences

APPROVED

Head of the Department of Mathematics,

Professor, Ph.D.

_____*Bogdana Oliynyk*

(signature)

"____"_____ 2023

INDIVIDUAL TASK

For Bachelor Thesis

for the 4^{th} - year student at the Faculty of Computer Sciences *Roman Burakov*

Title: «Decoding Speech from ECoG with Machine Translation Models»

Contents of the work:

1. Calendar plan
2. Contents
3. Introduction
4. Related work
5. Methods
6. Discussion
7. References

Issue date "____"_____ 2023, Supervisor _____

(signature)

Task received _____

(signature)

Concerted plan of work on Bachelor thesis

№	Stage	Deadline
1.	Choosing the graduate thesis project	26.10.2022
2.	Getting familiar with project domain	24.12.2022
3.	Developing research plan	11.01.2023
4.	Overview of the previous literature	11.02.2023
5.	Investigating holes in previous works and solutions to them	24.02.2023
6.	Reproducing previous research	12.03.2023
7.	Conducting experiments with discussed methodology	1.04.2023
8.	Gathering results and comparing to previous research	1.05.2023
9.	Writing the manuscript and preparing the presentation	22.05.2023

Student signature:

Roman Burakov _____ _____
(signature) *(date)*

Supervisor signatures:

Nadiya Shvai _____ _____
(signature) *(date)*

Bo Wang _____ _____
(signature) *(date)*

Contents

Decoding Speech from ECoG with Machine Translation Models

Roman Burakov[1, 2], Nadiya Shvai[1], and Bo Wang[2]

[1]*Faculty of Computer Sciences, National University of «Kyiv-Mohyla Academy»*

[2]*Department of Computer Science, University of Toronto*

Abstract

This paper explores the use and improvement of brain-computer interface (BCI)-based speech neuroprostheses, devices designed to enhance communication for individuals with speech disorders. Focusing on the machine learning aspect, we address the existing challenges associated with these systems, such as the limited vocabulary and simple algorithms of previous research and the individual variances in electrode implantation sites. Our approach reframes the decoding of speech from BCI as a machine translation problem and employs existing language models for semantic knowledge transfer. This research provides an extensive analysis of current neural speech decoding and multilingual neural machine translation methods, adapts the pre-existing M2M100 neural machine translation model for decoding ECoG data into text, and introduces a state-of-the-art model for neural speech decoding that improves upon current methods in semantic text reconstructions.

1 Introduction

Speech neuroprostheses are devices that are used to restore or improve speech function in individuals with disorders of the speech production system. One type of speech neuroprosthesis directly decodes speech from a brain-computer interface (BCI), which is a system that allows a person to communicate or control external devices through their brain activity. BCI-based speech neuroprostheses work by recording and analyzing the brain signals that are associated with speech production. These signals are then used to control a speech synthesizer, which produces the sounds of speech or decodes brain activity into text. The goal of these devices is to enable individuals with speech disorders to communicate more effectively by

bypassing damaged or impaired speech production systems. BCI-based speech neuroprostheses are important because they can greatly improve the quality of life for individuals with severe speech disorders by enabling them to communicate more effectively with others.

Despite the potential benefits of BCI-based speech neuroprostheses, this issue has not been completely solved yet, as it is an extremely challenging task both from the perspective of artificial intelligence and engineering. This work does not directly address the difficulty of engineering efficient BCI systems but rather focuses on the machine learning aspect. BCI systems generally fall within three categories: non-invasive, minimally invasive, and invasive. The former ones use sensors that are placed on the surface of the scalp to read signals from the outer layers of the brain. Although these devices are highly available and relatively low-cost, their signal is less accurate compared to the latter two. The latter two, on the other side, implant electrodes directly into brain tissue, increasing the signal accuracy, while significantly compromising availability.

Over the last several years, a substantial amount of research has been done on developing speech neuroprostheses and decoding frameworks. For example, [1] have developed a speech neuroprosthesis for a patient with Anarthria, that is capable of detecting and classifying a set of 50 singular words; [2] synthesize speech by first predicting articulatory movement from cortical activity, and then synthesizing it into speech signal. However, most previous research either uses very limited vocabulary [1, 3, 4], or algorithms that are too simple to capture the structure and semantics of spoken language [1, 3–5] or both. This is explained by the small amount of available speech production data from BCIs, which does not allow training a complex enough deep learning model from scratch. In addition, because electrodes can be implanted in different subregions of the brain, and their measurements vary from person to person, it is a common practice to train a separate decoder per person. However, it limits the amount of available data even further. To address this issue, [6] implied the transfer learning approach between two participants.

Similarly to [6], we frame the problem of decoding speech from BCI as a machine

translation problem, and evaluate our method on the same dataset. Although the underlying idea is close, we aim to reuse already existing language models for this task to transfer the knowledge of language semantics. For example, [7] has shown that it is possible to encode languages as vectors in latent space, and use a common encoder for all languages. This approach significantly improves the quality of translation for underrepresented languages. We want to train our encoders per person with the same architecture and decode these representations with a shared decoder initialized from [7]. We demonstrate that this approach enables efficient translation of cortical activity into text, as well as sustains the language semantics from machine translation models.

Specifically, the contributions of this paper include:

1. A comprehensive analysis of the existing work in neural speech decoding, and multilingual neural machine translation.

2. Adaptation of the pretrained M2M100 neural machine translation model to decoding ECoG data into text.

3. Creation of new state-of-the-art model for neural speech decoding, significantly improving over existing methods in terms of semantic text reconstructions.

2 Related Work

2.1 Speech neuroprosthetics

Neuroprosthetics is a diverse field of study that aims to replace a dysfunctional part of the nervous system with a bionic counterpart. In the subfield of brain neuroprosthetics, these bionic interfaces are also commonly referred to as Brain-Computer Interfaces (BCIs). These systems can have a vast impact on the lives of people with different impairments. For example, BCIs can be used to predict muscular activity, providing people with limb disabilities (e.g. many war veterans) a new way to perform daily actions. Another example, which is a primary object of this study, are speech neuroprosthetics. These systems can enable new means

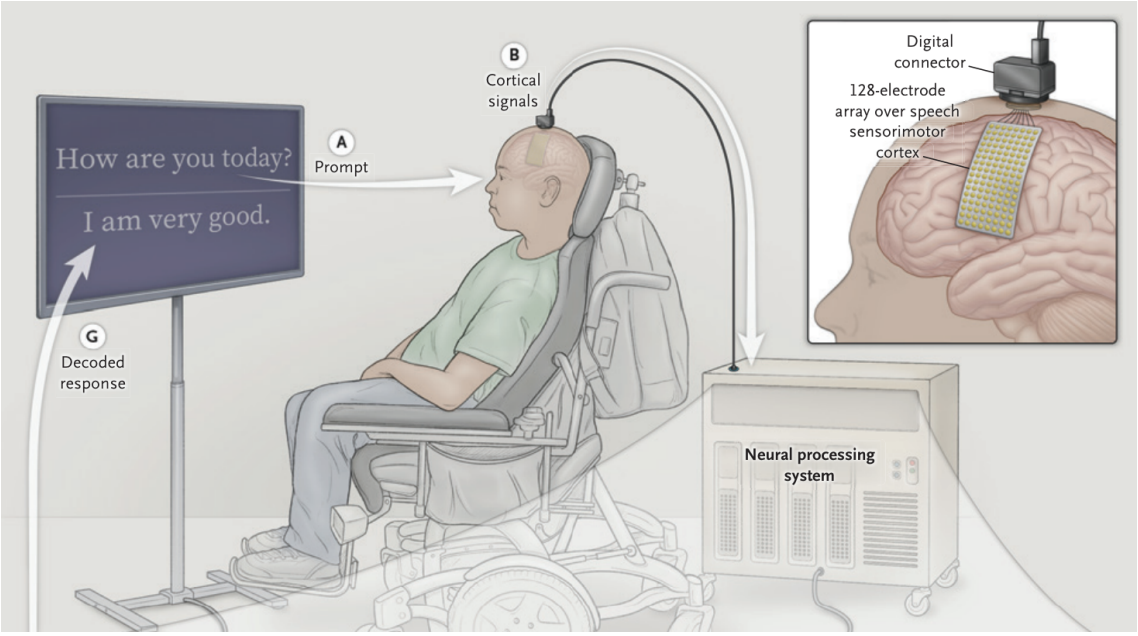

Figure 1: Schematic overview of the direct speech BCI. The illustration is a cropped version of Figure 1 from [1].

of communication for people with severe speech production impairments.

Recent studies have shown that it is possible to engineer a BCI that predicts imagined words from cerebral cortical activity, hence providing real-time communication for people with related disabilities. One of the most successful real-life implementations of such a system is [1]. This study has collected 22 hours of cortical activity of a person with anarthria attempting to pronounce individual words from a vocabulary of 50 words. This data allowed them to create a neural decoding model that reconstructed spoken sentences in real time. The post-hoc accuracy of this system was 98% for classifying an attempt for pronouncing the word, and 47.1% accuracy in word decoding. The resulting model was deployed in a closed-loop BCI system, as shown in Figure 1.

2.2 General neural decoding

Neural decoding entails decoding information from neural signals, which plays a critical role in systems neuroscience and brain-computer interface research [8]. Recent advancements in neuroimaging techniques and deep learning approaches have enabled various brain decoding systems, such as mapping functional Magnetic Resonance Imaging (fMRI) signals to images in an attempt to reconstruct

visual stimulus. Contemporary machine learning approaches have proven this possible with high semantic fidelity.

Research into neural decoding is vital for understanding various cognitive processes, as well as performing tasks related to them. Decoding fMRI signals based on visual stimulation is evidently helpful in understanding the brain's visual function mechanisms, but these fMRI signals can also be used to decode a person's thoughts, memories, and emotions [9]. Performing such research creates new opportunities to study the brain and harness its power for brain-computer interfaces.

2.3 Decoding speech from cortical activity

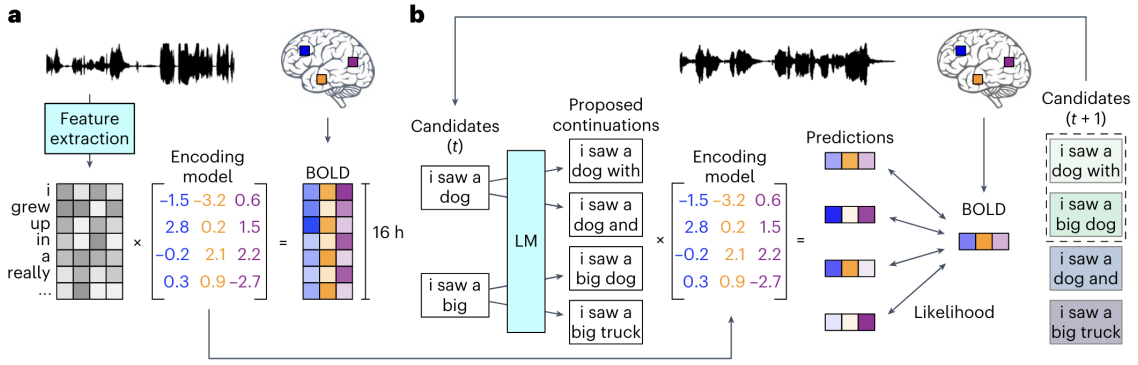

Figure 2: Schematic overview of decoding perceived speech from fMRI. The illustration is a cropped version of Figure 1 from [10].

In addition to many advancements in decoding visual stimuli from fMRI signals, many works have focused on decoding speech from non-invasive recoщdings. For example, [10] recorded 16 hours of fMRI signals while the subject was listening to narrated stories. They then use this data to train models that predict spoken sentences. Specifically, they train a neural encoder for predicting fMRI response from a given next word in a sentence. They then compare the predicted response for every word in the dictionary and select the word that produced the closest value to the real recorded fMRI signals. By doing that iteratively for each word in the sentence, they create a semantic reconstruction of the heard sentence. However, the best obtained Word Error Rate (WER) of 92% and corresponding BERTScore of 0.81 suggest that decoding perceived speech from fMRI signals is

an extremely challenging task. The high-level architecture of their approach is shown in Figure 2.

Because of indirect measurements and low to non-existent temporal resolution of fMRI, most studies on decoding speech from brain activity use other recording methods, such as surface electroencephalogram (sEEG) [11], magnetoencephalography (MEG) [12] or, electrocorticography (ECoG), [1–3, 5, 6, 13, 14]. Being an invasive recording method, ECoG provides an extremely high frequency and high-fidelity signal from narrowly localized brain regions. On the other hand, the invasiveness of ECoG implies smaller sample sizes, which is a significant limitation for developing decoding models. However, the advantages of ECoG make it the most prevalent type of neural recording for speech decoding.

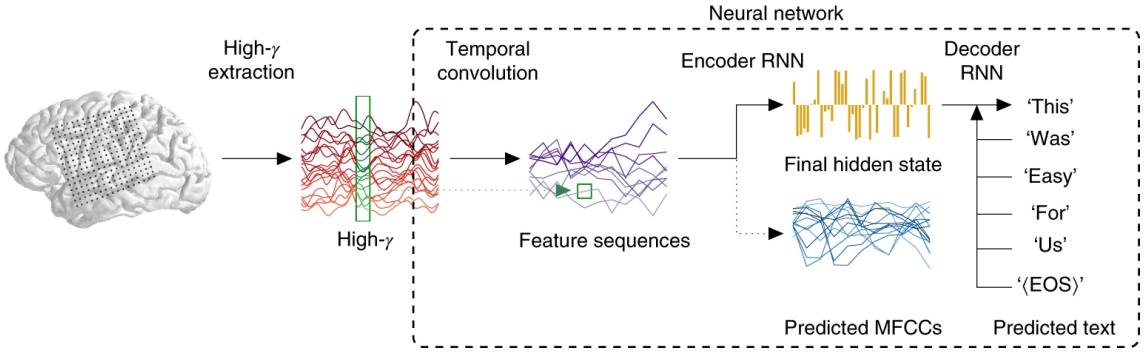

Figure 3: Schematic overview of the decoding pipeline used in [6]. The illustration is a cropped version of Figure 1 from [6].

Many works [1,3,5,13] use ECoG data to train simple machine learning algorithms, such as LDA classifiers or Viterbi decoders. These methods work well with limited data but require additional processing steps to predict spoken sentences instead of singular words. This also significantly limits the capabilities for capturing language semantics. Modern language models require hundreds of millions to billions of parameters, paired with trillions of training text tokens to learn the underlying structure and semantics of natural language [15]. Because of extremely difficult data acquisition, and hence diminutive sample sizes, this is infeasible for ECoG-based speech decoding with simple models.

To address the need of learning better underlying language structure, [6] train a sequence-to-sequence encoder-decoder framework to decode ECoG into text. The schematic overview of their pipeline is depicted on Figure 3. Specifically, they for-

mulate a neural speech decoding task as a neural machine translation task. They first downsample the high-resolution ECoG data with temporal convolutions and then use a bi-directional LSTM encoder to create a latent vector representing input brain data. This vector is then used by a similar LSTM decoder to decode it back into text. This approach provides high-quality neural decoding with WER as low as 3% on a test set that consists of novel repetitions of sentences from the train set. Although achieving high accuracy in predicting words, obtained reconstructions can be significantly different from the original sentence semantically. This is because their method does not use any sort of language modeling.

Another common challenge in brain decoding comes from the variability in human brain activations. Not only there is a significant difference in brain response across different subjects, but activations of the same subject can vary from day to day. This poses an issue for the generalization of decoding models. To address this issue, [6] implies transfer learning from subject to subject by training a new encoder per subject, but sharing decoder weights. They show that in rare cases such an approach greatly benefits the resulting reconstructions.

2.4 Multilingual neural machine translation

Multilingual neural machine translation task is a challenging yet extremely important problem. Having a single system that can adapt to hundreds of languages can open new means of communication between different demographic groups. The challenge of this task is efficiently translating to and from underrepresented languages. This is difficult because while common languages such as English or French have trillions of texts available, less spoken languages such as Tagalog can have only a few thousand texts in open access. This challenge is addressed in [7], where they represent languages as trainable tokens and use a transformer-based [16] encoder-decoder architecture to enable multilingual translation. By creating a shared language-invariant latent space, this approach simultaneously significantly improves the general multilingual translation and translation from and to underrepresented languages. The schematic of their method is depicted in Figure 4. The intuition of this work is that brain activity can be treated as a severely underrepresented new language and that by training a new encoder from

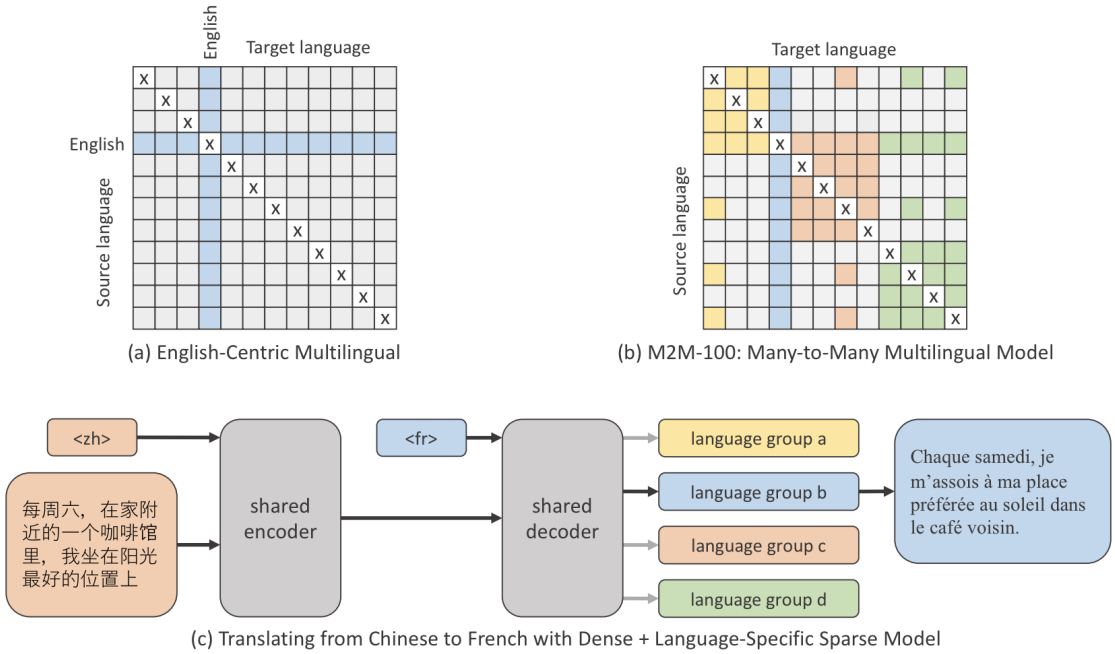

Figure 4: Summary of M2M100 dataset and multilingual model. English-Centric data (a) only contains training data to and from English, whereas M2M100 multilingual setting (b) contains data directly through various different directions. The proposed model, M2M100, combines dense and sparse language-specific parameters to translate directly between languages (c). The illustration and description are an adaptation of Figure 1 from [7].

the brain to shared latent space, we can create accurate semantic reconstructions of spoken sentences.

3 Methods

3.1 Speech production dataset

We use the dataset by [6]. It was collected from four subjects that had 16x16 ECoG grids implanted over peri-Sylvian cortices for diagnostic purposes. The anatomical reconstructions of these grids are shown in Figure 5. Over the course of multiple sessions, each subject was asked to read aloud multiple repeats of sentences in English from a total set of 30-50 unique sentences. Sentences were sampled from two benchmarks: a) picture descriptions (30 sentences, 125 unique words) and b) MOCHA-TIMIT (460 sentences, 1800 unique words). The test set

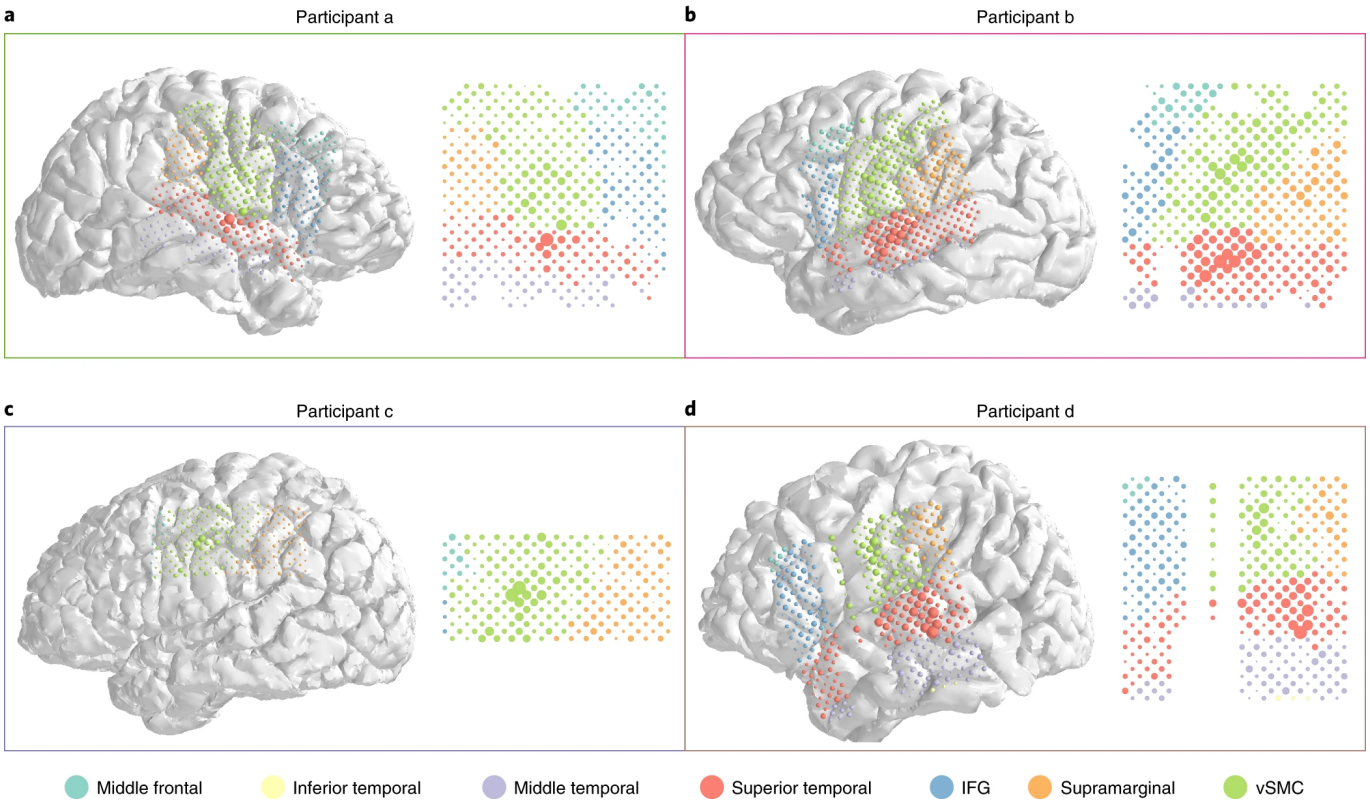

Figure 5: Anatomical reconstruction of the ECoG grids for all four participants (a/b/c/d) in the dataset. Some values are missing, because part of the implanted electrodes were unreceptive. The illustration is a cropped version of Figure 5 from [6].

consists only of sentences that were repeated more than three times. Hence, this provides one repetition for testing, and at least two for training. Such filtration practically restricts the test set to approximately 50 unique sentences, with 250 unique words.

The temporal resolution of the final processed ECoG data is sampled at 200Hz, and spatial resolution was effectively upsampled two-fold by using bipolar referencing. Bipolar referencing leverages the fact that sets of singular electrons generate an electric field, thus allowing for a higher resolution than the number of electrodes. Specifically, from the activity of each electrode, new spatial points were generated by subtracting the activities of its neighbors below and to the right. Because some of the electrodes were corrupted, the amount of channels after bipolar referencing is not trivial but was retrieved via experimentation and further confirmed via a private inquiry to [6]. For a more thorough explanation of data processing, please refer to [6]. Some details on data are depicted in Table

subject	# of channels	# of samples	avg. ECoG length	avg. # of words
a	448	974	577	7.46
b	429	910	526	6.58
c	232	1427	791	7.65
d	371	1791	434	7.74

Table 1: Per subject statistics from the dataset, denoting (from left to right): subject id; the number of working channels in the implanted ECoG grid after the bipolar referencing; the number of sentences captured; the average amount of sampled data points per spoken sentence; the average sentence length.

1.

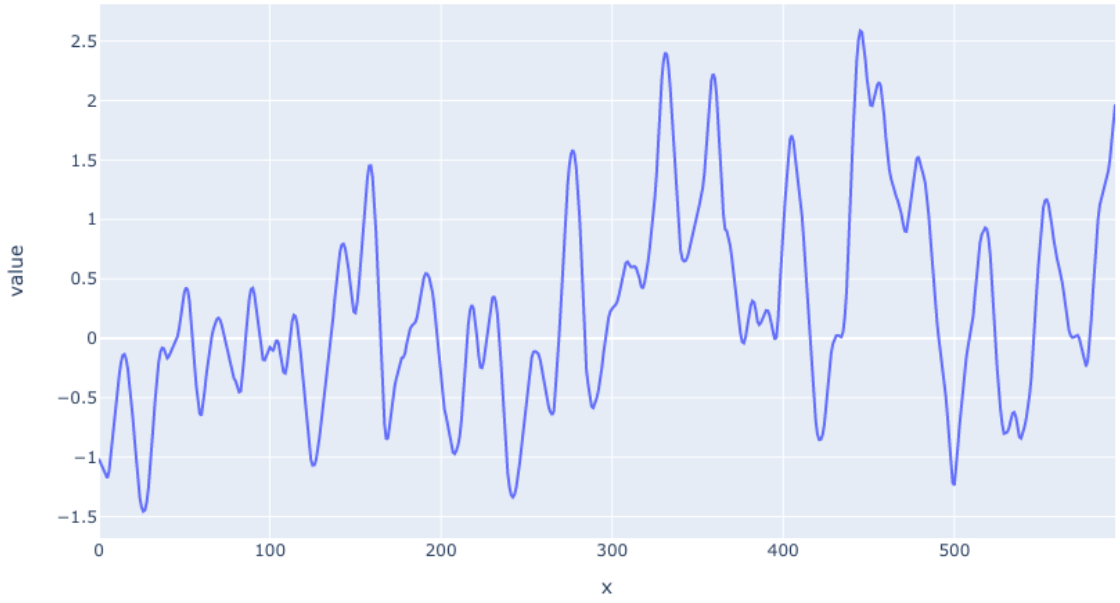

Figure 6: A subsample of data for one of the electrode channels for subject A. x denotes the time axis, and value denotes the value of the corresponding channel. The data has highly seasonal features, and localized trends can be observed.

Overall, the processed dataset contains pairs of ECoG activations and pronounced sentences. On average, sentences are approximately 7 words long, and correspond to about 500 ECoG measurements, providing roughly 2.5 seconds per sentence. Data has high local seasonality, and can generally benefit from proper aggregation and downsampling. An example of one input channel is plotted on Figure 6.

3.2 Evaluation metrics

3.2.1 Word Error Rate

Word Error Rate (WER) is a common metric used in the field of speech recognition, machine translation, and other areas of natural language processing that involve sequence prediction tasks. It is based on Levenshtein distance, which is a string metric for measuring the difference between two sequences.

WER measures the minimum number of operations needed to transform the system output into the reference. These operations can be:

- Substitutions: replacing one word in the hypothesis with one word from the reference.
- Deletions: removing one word from the hypothesis.
- Insertions: inserting one word from the reference into the hypothesis.

Let's denote S as the number of substitutions, D as the number of deletions, I as the number of insertions, and N as the number of words in the reference. With these notations, the WER is computed as:

$$WER = \frac{S + D + I}{N} \tag{1}$$

The WER will be 0 for a perfect match and can otherwise take on any non-negative value. Lower WER corresponds to a closer match to the reference.

3.2.2 BLEU score

The Bilingual Evaluation Understudy (BLEU) score is a widely-used, automated metric for evaluating machine translation systems, introduced in [17]. Its key benefits include its objectivity and scalability, enabling consistent, large-scale evaluations of text translations. Despite its simplicity, the BLEU score has shown a reasonable correlation with the human judgment of translation quality. It is straightforward to implement and can be universally applied to any pair of languages given reference translations. However, it primarily considers n-gram over-

lap and does not account for semantic correctness, grammaticality, or fluency, which can sometimes lead to inaccuracies when assessing the quality of translations.

Let's denote C as the machine-generated text and R as the human-generated reference text. To compute BLEU, we first, for each n-gram length (from 1 to N), calculate the modified precision p_n:

$$p_n = \frac{\sum_{\text{ngram} \in C} \min(\text{Count}_C(\text{ngram}), \text{Count}_R(\text{ngram}))}{\sum_{\text{ngram}' \in C} \text{Count}_C(\text{ngram}')} \tag{2}$$

We then compute the brevity penalty BP:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{3}$$

And lastly, we combine the modified precision scores and the brevity penalty to get the $BLEU$ score:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log(p_n)\right) \tag{4}$$

3.2.3 BERTScore

BERTScore is an evaluation metric used for text generation tasks, including machine translation and text summarization. Unlike simpler metrics like BLEU, BERTScore leverages the BERT [18] language model to capture the semantic similarity between generated and reference texts. It computes precision, recall, and F1 scores based on contextual embeddings from BERT, yielding scores that better align with human judgment. In practice, BERTScore computes the cosine similarity between the contextual embeddings of individual tokens in the candidate and reference texts. These embeddings are extracted from a pre-trained BERT model. This makes BERTScore's evaluations sensitive to the nuanced semantics captured by these embeddings.

Let's denote c_i and j_i as the contextual embeddings for token i in the candidate sentence and token j in the reference sentence respectively, and N as the number of tokens in the candidate sentence, and M is the number of tokens in the reference

sentence. With these notations, to compute BERTScore, we first calculate the precision:

$$P_{\text{BERT}} = \frac{1}{N}\sum_{i=1}^{N}\max_{j=1}^{M}\frac{c_i \cdot r_j}{\|c_i\|_2 \cdot \|r_j\|_2} \tag{5}$$

We then compute the recall:

$$R_{\text{BERT}} = \frac{1}{M}\sum_{j=1}^{M}\max_{i=1}^{N}\frac{c_i \cdot r_j}{\|c_i\|_2 \cdot \|r_j\|_2} \tag{6}$$

And, finally, combine precision and recall into BERTScore, which is analogous to the F1 score in this setting:

$$BERTScore = \frac{2 * P_{\text{BERT}} * R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{7}$$

3.3 Decoding speech with machine translation models

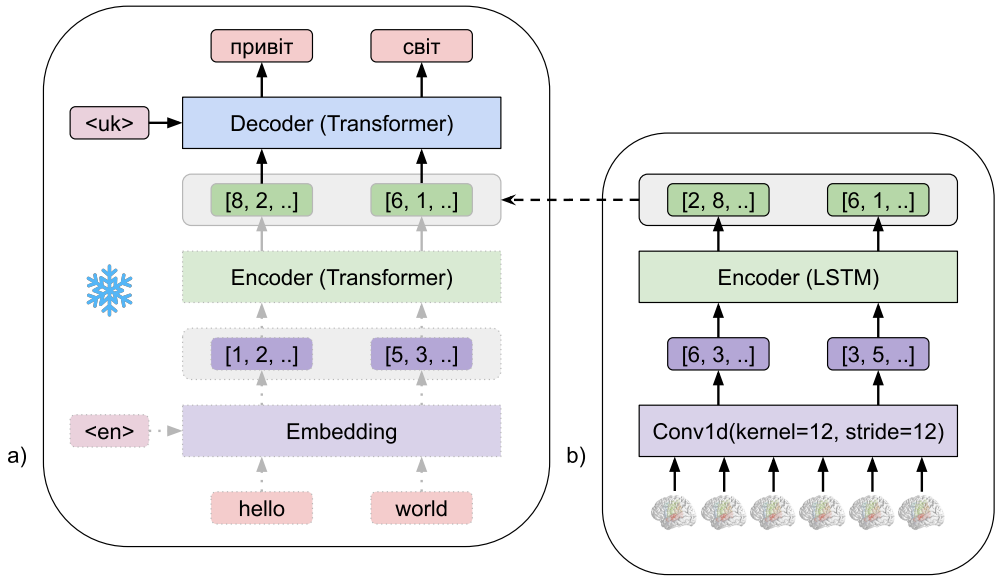

Figure 7: Schematic of the proposed model: a) represents the high-level architecture of the M2M100 model by [7]; b) represents the proposed brain encoding model. We replace the text encoder of (a) with our brain encoder (b) and train the resulting model jointly with the frozen decoder so that only weights in (b) are optimized.

Motivated by recent advancements in neural machine translation, and neural speech decoding with encoder-decoder architecture, we combine these methods.

To do that, we start by initializing our model with a pre-trained M2M100 with 418M parameters. At this point we freeze all of the parameters in M2M100, and never optimize them. Original M2M100 uses an encoder that takes a language code and a sequence of word token embeddings as input. This encoder embeds them into the shared language-invariant latent space, and forwards them to the decoder. Decoder then takes these representations along with output language code as the first forced output token, and autoregressively creates the translation sequence by predicting next word with a simlpe multi-class classification head. The model is trained jointly with Multiclass Cross-Entropy Loss. Specifically, the formula for cross-entropy is:

$$L(\mathbf{y}, \mathbf{p}) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log(p_{ij}) \tag{8}$$

Where:

- $L(\mathbf{y}, \mathbf{p})$ is the cross-entropy loss.
- N is the total number of observations (or instances, samples).
- C is the total number of classes.
- y_{ij} is the binary indication of whether class label j is the correct classification for observation i.
- p_{ij} is the model's predicted probability that observation i falls into class j.

Because our models works with ECoG brain data instead of words, we are not able to use the M2M100 architecture as is. Hence, we create latent representations with our own trainable LSTM encoder. The schematic of the overall approach is shown in Figure 7. Specifically, similarly to [6] we downsample the spatial resolution of ECoG by using temporal convolutions. We then pass the resulting tokens (token dimension equals to the amount of channels after bipolar referencing) into the LSTM netword to create representations of the brain by using the hidden state of this LSTM network (hidden dimension of the LSTM network equals to the latent dimension of M2M100). We then experiment with two ways of propagating these embeddings into M2M100: a) passing them to the encoder, instead of word

embeddings; b) passing them directly to the decoder. We find that b) works best. This final model is trained jointly with (8).

4 Results

To evaluate our results we use the same train/validation/test splits as [6]. We generate optimal validation and test sentences using beam search [19]. This brings slight improvement by exploring a wider space of possible translations compared to naive methods like greedy decoding. Because the final model reseased by [6] was incompatible with our pipeline, we reproduce their work to our best ability, by consulting the original paper and privately inquiring authors about the implementation details. We do not use greedy search when evaluating their network.

subject	model	WER (% ↓)	BLEU (% ↑)	BERTScore (% ↑)
a	Makin et al.	**53.5**	**12.3**	51.2
a	Ours	56.8	9.5	**71.8**
b	Makin et al.	3.4	56.1	85.1
b	Ours	**3.1**	**61.9**	**93.4**
c	Makin et al.	19.3	26.2	77.7
c	Ours	**15.0**	**29.8**	**82.0**
d	Makin et al.	**10.9**	33.7	82.5
d	Ours	11.3	**36.9**	**89.6**

Table 2: Evaluation results for each of the four subjects. Makin et al. is our implementation of [6]. While WER is not always better, we show significant improvement in BERTScore across all subjects, and in BLEU across three out of four subjects.

Table 2 depicts the final comparison metrics of our method to [6]. We observe an overall significant improvement over their method. While difference in WER is not as high, BERTScore is uniformly better for our approach. The same can be said for BLEU, with an exception for subject a. This is explained by the fact that their method aims to predict individual words as closely as possible, while our

method is bottlenecked by M2M100 decoder and language modeling head. This bottleneck forces our model to generate semantic reconstructions rather than just structural, implying that the meaning of our translated sentences is closer to the original sentence, while exact wording may vary. An example of this would be "I love my mum" decoded as "I honour my family". Even though words are different, the semantic meaning is preserved.

Another advantage of our method is that M2M100's decoder dominantly generates coherent and grammatically correct sentences, while the method used in [6] does not account for any grammatical correspondance. We believe this is a significant improvement.

5 Discussion

5.1 Conclusion

In conclusion, this paper explores the challenges and potential solutions for speech neuroprostheses, specifically those based on brain-computer interfaces (BCI). It underscores the importance of these devices for improving the quality of life for individuals with severe speech disorders by enabling them to communicate more effectively. Notably, this work delves into the nuances of different BCI systems, including non-invasive, minimally invasive, and invasive methods, each presenting their own unique advantages and disadvantages. The state-of-the-art research on speech neuroprostheses and decoding frameworks is comprehensively reviewed, highlighting existing limitations, such as the need for large amounts of speech production data and the individual-specific nature of decoder training.

Further, the paper introduces a novel approach that treats the problem of decoding speech from BCI as a machine translation problem. Drawing on the idea of encoding languages as vectors in latent space, we propose training individual encoders per person using a shared decoder. By reusing pre-existing language models, we aim to preserve language semantics while translating cortical activity into text. The discussion also highlights the significant potential of multilingual neural machine translation, specifically in the context of decoding underrepre-

sented 'languages' such as brain activity.

We conclude with a call to action for further research and exploration into the intersection of speech neuroprostheses, brain-computer interfaces, machine learning, and multilingual neural machine translation. The development of more efficient and accurate speech neuroprostheses, through the use of machine learning and advanced decoding methods, has the potential to drastically improve the lives of individuals living with speech impairments.

In terms of performance, the our model showed a significant improvement in BERTScore across all test subjects compared to the model by [6]. There are some mixed results in terms of WER and BLEU scores, model by [6] performing better in a few cases but our model performs better in the majority of settings. However, we argue that our model produces translations that are more semantically accurate, as are focuses on preserving the meaning of sentences rather than strictly adhering to structural elements. We also point out that our model generally produces more coherent and grammatically correct sentences, which we believe is a significant advantage over [6].

5.2 Future work

We believe there is a set of improvements that can be done to our method. Some of the most important potential steps include:

- Using autoencoder model, such as [20], instead of M2M100. This can bring better generalization and lower bias to the network. In addition, autoencoders often only use one latent vector instead of encoded tokens to represent sentences. This can significantly simplify and improve the training process.

- Training with distribution penalty, such as KL Divergence loss. This will bring brain encoder's representation even closer to those expected by language models, potentially improving decoding quality.

- Collecting more data. As with any machine learning project, data is crucial. Having access to bigger and higher-quality datasets would significantly increase the generalization capabilities of our models, potentially even allowing

for a single shared encoder across all subjects.

- Using non-invasive recordings. While our method relies on invasive ECoG data because of its exceptional ability to conduct brain signals, more data is available with other recording types, such as sEEG, MEG, or fMRI.

6 Acknowledgements

This project was entirely funded by the Mitacs Globalink Research Award program. The office space and computing resources for conducting experiments and hosting datasets were provided by Vector Institute. We also express our gratitude to Joseph G. Makin for sharing the dataset used in [6], and providing timely support in reproducing their work. This project would also not be possible without [7] releasing pre-trained weights for their models.

References

[1] 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 Journal of Medicine*, vol. 385, no. 3, pp. 217–227, 2021. PMID: 34260835.

[2] G. Anumanchipalli, J. Chartier, and E. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, pp. 493–498, 04 2019.

[3] D. Moses, M. Leonard, J. Makin, and E. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Communications*, vol. 10, 07 2019.

[4] G. Wilson, S. Stavisky, F. Willett, D. Avansino, J. Kelemen, L. Hochberg, J. Henderson, S. Druckmann, and K. Shenoy, "Decoding spoken english from intracortical electrode arrays in dorsal precentral gyrus," *Journal of Neural Engineering*, vol. 17, p. 066007, 11 2020.

[5] C. Herff, L. Diener, M. Angrick, E. Mugler, M. Tate, M. Goldrick, D. Krusienski, M. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in Neuroscience*, vol. 13, p. 1267, 11 2019.

[6] J. Makin, D. Moses, and E. Chang, "Machine translation of cortical activity to text with an encoder-decoder framework," 07 2019.

[7] A. Fan, S. Bhosale, H. Schwenk, M. Zhiyi, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," 10 2020.

[8] J. A. Livezey and J. I. Glaser, "Deep learning approaches for neural decoding across architectures and recording modalities," *Briefings in Bioinformatics*, vol. 22, pp. 1577–1591, 12 2020.

[9] S. Huang, W. Shao, M.-L. Wang, and D.-Q. Zhang, "fmri-based decoding of visual information from human brain activity: A brief review," *International Journal of Automation and Computing*, vol. 18, 01 2021.

[10] J. Tang, A. LeBel, S. Jain, and A. G. Huth, "Semantic reconstruction of continuous language from non-invasive brain recordings," *bioRxiv*, 2022.

[11] M. Angrick, M. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, P. L. Kubben, T. Schultz, and C. Herff, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *bioRxiv*, 2020.

[12] O. Chehab, A. Defossez, J.-C. Loiseau, A. Gramfort, and J.-R. King, "Deep recurrent encoder: A scalable end-to-end network to model brain signals," 2022.

[13] M. Angrick, C. Herff, E. Mugler, M. Tate, M. Slutzky, D. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks:," 11 2018.

[14] S. Metzger, J. Liu, D. Moses, M. Dougherty, M. Seaton, K. Littlejohn, J. Chartier, G. Anumanchipalli, A. Tu-Chan, K. Ganguly, and E. Chang, "Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis," *Nature Communications*, vol. 13, 11 2022.

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

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

[17] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," 10 2002.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[19] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, Association for Computational Linguistics, 2017.

[20] C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao, "Optimus: Organizing sentences via pre-trained modeling of a latent space," 2020.