Neural Network Approach To Speech Classification Using MEG

by

Minkang Suk

Supervisor: Frank Rudzicz

April 2019

Abstract

In this work, we train a deep neural network with raw magnetoencephalography (MEG) data to perform classification between a monosyllabic speech-elicitation task, a multi-syllabic speech-elicitation task, a verb-generation task, and a mouth-open speech-elicitation task. Simple neural networks have been used in previous work to classify data after significant data pre-processing into features based on deep domain knowledge. The model proposed in this work is trained using relatively untouched raw MEG, and we show that this model can learn to mimic some of the feature-engineering normally done by experts. This model may be able to be extended to help better understand research the relationship of the brain with speech using raw MEG without the bias of expert knowledge.

Acknowledgements

I would like to acknowledge Professor Frank Rudzicz for supporting me through this thesis, and providing me with a challenging and very interesting project for me to explore research. In addition, I'd like to thank Demetres Kostas, the graduate student who provided incredibly valuable feedback and insight to my project and pushing me to achieve the best work I can.

0. Table of Contents

Abstract i

Acknowledgements ii

0. Table of Contents iii

1. Introduction 1

2. Literature Review and Background 2

2.1 MEG and existing methods 2

2.2 Speech and the Brain 2

2.3 Machine Learning and Neural Networks 3

2.3.1 Neural Networks and Sequence Data 3

2.3.2 Convolutions and Spatial Data 4

3. Methods 4

3.1 Dataset 4

3.2 Model Architecture 6

3.2.1 Spatial Summary Stage 6

3.2.2 Gated Recurrent Unit (GRU) 7

3.2.3 Output 7

3.3 Model Training Procedures 7

4. Discussion and Results 8

4.1 Accuracy of Classification 8

4.2 Confusion Matrix 9

4.3 High Confidence Examples 10

4.4 Model Activations 133

4.5 Further Discussion 144

5. Conclusion 155

5.1 Summary of Results 155

5.2 Next Steps 155

6. References 17

7. Appendix 19

A. All Maximized Activations 19

1. Introduction

The act of speaking is a fairly complex activity, with higher level language understanding and intent being related to the Wernicke's area, and speech articulation being believed to be related to parts of the brain such as Broca's area [1]. There is also evidence that shows that speech articulation separated from language engages areas such as the Rolandic cortex, integrating motor control and sensory information to articulate speech [2]. In this work, we consider a dataset that is a combination of 4 different speech tasks, a pair of non-word mono and multi-syllable utterances, a verb generation task, and a mouth-open task recorded using magnetoencephalography (MEG) to better understand the relationships of the various areas of the brain and speech.

Magnetoencephalography (MEG) is a neuroimaging technique to map brain activity that records the changing magnetic fields caused by the natural processes in the brain by utilizing magnetometers on the scalp. Other methods to record data for brain-to-speech related tasks involved intracortical sensors [1], [3]. In contrast, MEG does not require opening up the skull and is sufficiently accurate and has a high enough time resolution to be less invasive, and more practical for speech tasks.

Simple machine learning models have been used to work with MEG, particularly for brain-computer interfaces for silent speech tasks [4]–[6]. However, as with most data collected from the brain, the data can vary between subjects, and even day-to-day with the same subject. In addition, there is a low signal-to-noise ratio that must be overcome, through various pre-processing techniques such as cropping, normalization, band-pass filtering, principal component analysis (PCA) or independent component analysis (ICA) [7], [8]. These pre-processing techniques leverage expert knowledge about the brain and the collected data to distinguish important features and special characteristics of specific brain frequency bands, which, in and of itself can be a biasing factor that can reinforce underlying assumptions and established knowledge while potentially ignore unknown factors and features.

We propose that it is possible for a neural network model to distinguish between different non-word and word utterances from training on raw and mostly unprocessed MEG data. The models can achieve many of these pre-processing and feature engineering components intrinsically through training, and potentially uncover unknown features or relationships. We use raw MEG to train a neural network to classify the brain data from four different speech tasks.

2. Literature Review and Background

2.1 MEG and existing methods

MEG and electroencephalography (EEG) are neuroimaging techniques that record the changing magnetic fields or surface potentials caused by the natural processes of the brain. These techniques has the advantage of being mostly non-invasive, when compared to other techniques to analyse brain activity. While both techniques can be useful, MEG can be done much more quickly and precisely for the lack of electrodes attached to the scalp [9]. There has been many efforts for automated classification of EEG signals related to motor or language tasks [4], [6], while MEG has been used to find the spatial locations of brain signals[9]. There have been relatively fewer attempts and MEG data classification other than epileptic spike detection [10], [11].

MEG classification presents some challenges. First, the magnetic fields generated by the brain are extremely weak and are prone to be contaminated by external noise, leading to low signal-to-noise ratios. Secondly, MEG recordings generally have a high number of channels, in the order of 100 channels [9]. However, due to the cost, as well as fatigue of subjects, it is usually difficult to record many trials relative to the number of channels. We are fortunate to have access to a dataset with a significant number of subjects and trials.

Existing work using MEG related to verbal and language have used MEG to classify various words being spoken or mouthed but utilize techniques such as spatial principal components analysis (PCA), independent components analysis (ICA), second-order blind identification (SOBI), as well as filtering and feature engineering to try to separate the sources from the artifacts found in the MEG data and overcome the low signal-to-noise ratio [5], [10].

2.2 Speech and the Brain

Speech production and articulation is a very complicated process that integrates the auditory, somatosensory, and motor information in the temporal, parietal, and frontal lobes of the cerebral cortex [2]. There is the physical act of producing sounds and syllables in a desired way, which has been related to Broca's area, in addition to some evidence showing engagement in the Rolandic cortex, integrating motor control and sensory information to articulate speech. As well, in speech, there is the language processing component needed to make coherent sounds that relate to some meaning. This is considered to be related to Wernicke's area where higher language levels of

language understanding and intent is believed to be processed [1]. The hope is that in this work, we are able to see these areas of the brain activate during the speech tasks analysed.

2.3 Machine Learning and Neural Networks

2.3.1 Neural Networks and Sequence Data

When dealing with sequential data, it makes intuitive sense to process the data one step at a time and persist relevant information over time. While traditional neural networks cannot do this, recurrent neural networks (RNNs) can handle this. RNNs are networks with loops that can allow previous information to persist by passing a message to the successor [12]. These types of networks have been the building blocks for dealing with tasks dealing with data that is sequential such as language modelling, speech-to-text, or translation.

The problem with traditional RNNs is that during training, they suffer from the problem of vanishing or exploding gradients [13], which become particularly pronounced when the length of the subsequence to be considered becomes large. One might want the RNN to encode long-range dependencies throughout a data trial but increasing the length of the context to consider compounds the problem of vanishing/exploding gradients. When RNNs are trained, a technique called Backpropagation Through Time (BPTT) is used to compute the parameters and weights of the network, but as the length of the network is increased, the partial derivatives with respect to the error gradient either quickly drops to 0, or quickly overflows [12].

There are two typical methods that deal with the vanishing/exploding gradient problem of a traditional RNN. The first is the Long Short-Term Memory unit (LSTM), that was originally proposed by Hochreiter and Schmidhuber [13]. Unlike the units in an RNN, which computes a weighted sum of the inputs and applies a non-linearity, an LSTM unit maintains a memory, an output, and an output gate that controls how much of the previous output to send to the next timestep. The memory is also affected by another gate which determines how much of the existing memory should be forgotten, and how much new memory should be added. Intuitively, when the LSTM detects an important feature, it can store this information easily for long distances [13]. The second way to deal with the vanishing/exploding gradient problem is a gated recurrent unit (GRU). The GRU operates similarly to the LSTM with gates that control the flow of information inside the unit to the next but does not have a separate memory cell. Thus, the GRU has one less parameter to store at each timestep [14].

Both methods have an additive component, keeping the existing information, and adding new information on top of it, which creates a shortcut that allows the error to be back-propagated without

quickly vanishing (which is caused by passing through multiple bounded non-linearities) [15]. In this work, we utilize a GRU to deal with the sequence MEG data, due to its ability to better deal with long term dependencies while being slightly more computationally efficient due to having less parameters.

2.3.2 Convolutions and Spatial Data

Convolutions are a technique often used in neural networks to improve the performance on a variety of tasks. A convolution is a fairly simple operation where you start with a kernel (a small matrix of weights) that slides across the input data. At every point, the kernel performs an elementwise multiplication with the part of the data that the kernel is currently on and summing the data into a single output. This has the effect of reducing the size of the input data. An important distinction of the convolution is that the kernel of weights is held constant, which drastically reduces the number of parameters compared to a fully connected linear layer [16].

Convolutional neural networks (CNNs) are often used when dealing with image or spatial data and combine convolutions with non-linearities. This is potentially convolutions utilize a kernel that combines data from a smaller local area to form an output, and so is more invariant to small translations of data. Thus it is capable of finding features from local inputs, which is quite well suited for images as well as time-series data [16].

3. Methods

3.1 Dataset

The dataset is comprised of synchronized MEG and speech recordings, which were originally recorded to examine age- and sex-related developmental language differences in children by Yu et al [17]. and Doesburg et al [18]. The demographics are summarized in Table 1. Each of the subjects spoke English as their first language, and had no suspected history of language, speech, hearing, or development disorders, according to their guardians. The acquisition protocol was approved by the SickKids Research Ethics Board (REB #1000016645) which acts in accordance with the guidelines established by the Tri-Council Policy on ethical Conduct for Research Involving Humans. All participants or their guardians gave written informed consent, and children unable to read the consent form provided verbal assent.

Task	Mean Age	Age Range	Subjects	Trials	M/F Split
/pah/	10.73	4.1 - 18.1	89	115	0.45/0.55
/pah tah kah/	11.01	4.1 - 18.1	83	115	0.45/0.55
VG	13.23	5.7 - 18.0	28	81	0.42/0.58
MO	10.73	4.1 - 18.1	89	115	0.45/0.55

Table 1. Dataset demographics by stimuli types

The children were given two standardized language tests: the Peabody Picture Vocabulary Test (PPVT) [19] and the Expressive Vocabulary Test (EVT)[20]. All subjects had scores that were at or above the average expected scores for their age.

Recordings were performed in a soundproof room using a CTF whole-head MEG system in a magnetically shielded room in the Neuromagnetic Lab of the Hospital for Sick Children in Toronto. To minimize movement of the head, padding was added to restrict space, and each subject was given very clear instructions to stay still, with any excess movement leading to restarting of trial acquisition.

There were 4 different speech tasks that the subjects were asked to perform. The first task is a monosyllabic non-word utterance */pah/*. The second task was the multisyllabic non-word utterance */pah tah kah/*. These tasks were simple enough for young children, and is part of the diadochokinetic rate (DDK) test, which can be used to evaluate motor speech disorders [18]. The experimenter demonstrated the production of each task without word-like or prosodic features. The third task is a verb generation (VG) task, where subjects were prompted with an image they are familiar with, and asked to produce a verb associated with the object. e.g. saying ‘throw’ when shown a ball. The final task was a task similar to the first two tasks, but instead the subject is instructed to hold their mouth open while saying the non-word utterance.

The system recorded 151 MEG channels with a sampling rate of 4 kHz. Some basic clean-up of the data was performed. The original data was low-pass filtered at 160 Hz and resampled to 80 Hz, which was determined to be sufficient to accommodate the range for typical α, β, γ, δ, θ activity. As well, we took 1s of data before the stimulus as a baseline that was subtracted from the data. Finally, the data was cropped to 1 second before the stimulus, to 2s after the stimulus to reduce the size of the data while still capturing most of the raw data. This reduced the size of each trial of each task per subject to 151x481, with a fixed length of time for each trial.

3.2 Model Architecture

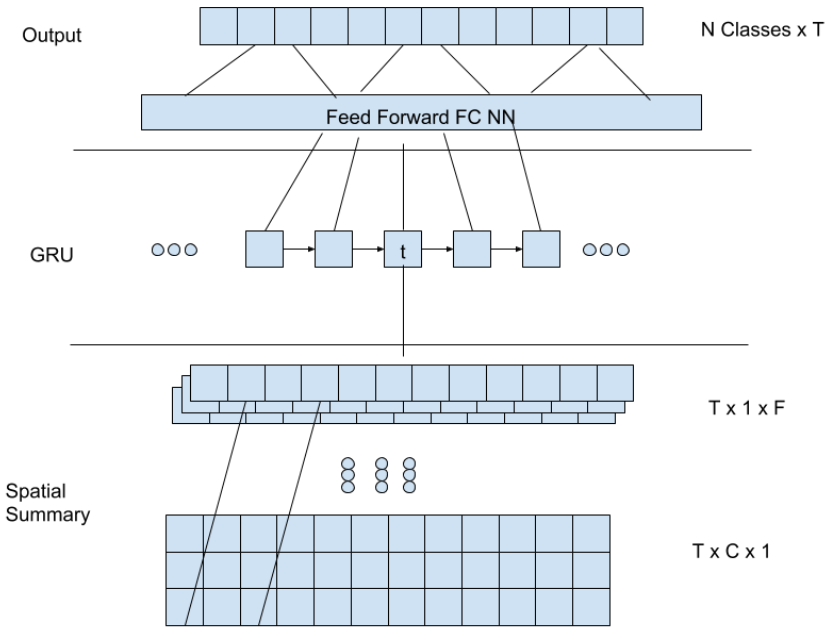


Figure 1. The model follows the above architecture.

The model can be summarized into three fairly distinct sections as shown in figure 1. The data is processed from the bottom to the top. Each trial has C channels and T samples (where C = 151, and T = 481 for the dataset), which is eventually processed into an N x T output, where the output of the last timestep is used as the output classes, which is determined by applying a SoftMax on the output. In our dataset, N = 4.

3.2.1 Spatial Summary Stage

The first layer of processing is the spatial summary stage of the model. This stage performs convolutions across the channels at each timestep (and thus not impacting the number of timesteps). In the model discussed in this report, there are 3 levels of convolutions. Each convolutional layer uses a kernel size of 51 and increasing the number of filters from 1 to 25, 50, and a final value of 75 filters in the output. In between each convolution, we apply a batch normalization, put the output through a ReLU unit, and finally apply dropout. Thus, this area transforms the initial input of size T x C x 1 into an output of T x 1 x F, where F was 75 in this case.

The idea behind this part of the model is that at every time step, while there are 151 channels of MEG, those channels likely correspond to some undetermined number of source signals that are

being recorded by the MEG. The goal of the spatial summary stage is the reduce the 151 channels of MEG data into F features (which are the filters from the CNN). It was originally considered to use a fully connected feed-forward network as the spatial summary stage. However, it was decided that convolutions might be more useful in finding more local features that are invariant to small translations, when compared to using a multi-layer perceptron. Thus, convolutions are used to model this part of the model, which acts as a part of the model that "learns" to do feature engineering on the original data without expert knowledge.

3.2.2 Gated Recurrent Unit (GRU)

The output of the spatial summary stage is used as an input to a GRU network. A GRU was selected to deal with the time-sequence nature of the input data. While we do not have to worry about varying sequence lengths since the data has the same fixed length, we do require the more robust long-term memory functionality that a GRU or Long-Short Term Memory (LSTM) provides when compared to a traditional RNN since the length of the sequence is relatively large (481), and thus are at risk of dealing with vanishing/exploding gradients. The GRU was selected over the LSTM due to it being potentially more computationally efficient due to each GRU cell having one less gate when compared to the LSTM. We utilized the GRU that comes packaged with PyTorch, using 3 hidden layers, with 50 hidden units at every layer.

3.2.3 Output

The last timestep of the output of the GRU is fed into a fully connected feed-forward network which has a SoftMax applied to it for final classification. The output of this layer is N x 1 where N is the number of expected classes.

3.3 Model Training Procedures

Nine subjects of the 95 total subjects are held out as a testing group, of which only 1 did not perform the verb generation task. The test subjects are selected to have a similar distribution of the number of trials in each age range. The remaining subjects' trials are distributed across 5 folds for cross-validation. The best performing model of the 5 generated models was selected to be evaluated against the testing set. In addition, there is an imbalance in the number of training examples for each trial. A loss penalty was applied in proportion to the number of trials for each task, and the number of total trials there were in the training set.

We utilize PyTorch to build all of the models described in this report and utilize a stochastic gradient descent optimizer with momentum. The rectified linear unit (ReLU) was used as the activation functions. All layers are batch-normalized after activation.

4. Discussion and Results

4.1 Accuracy of Classification

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev.
Test Acc.	31.07%	32.17%	32.65%	32.65%	33.24%	32.37%	0.727%

Table 2. Accuracy on the test set, of models trained with different folds held out as validation.

As can be seen in table 2, the accuracy of the model classification is not fantastic, where random guessing with four classes would be a 25% accuracy. In addition, the dominant-class threshold is an accuracy of 27.67%, (since there is a disparity in the number of available trials in each type of task, it would be a reasonable baseline to predict the majority class). The accuracy of the classification is fairly low.

Some reasons why the accuracy might be fairly low is that firstly, there is a fairly large discrepancy in the number of VG trials and the other task trials. The low number of examples makes this task the hardest to classify. As well, the differences between each class might be some slight differences in a few features, over only a few time steps. This might explain the fairly low accuracy and precision rates as seen below in tables 3 and 4.

However, the goal of this project was not to focus on getting the highest classification accuracy, but to gain some insight into the relationship between brain activity and the brain and diving deeper into the results provide some interesting results.

4.2 Confusion Matrix

		Predicted				
		/PA/	**/PDK/**	**VG**	**MO**	**TOTAL**
Actual	**/PA/**	477	280	41	237	1035
	/PDK/	420	357	52	206	1035
	VG	234	190	102	114	640
	MO	391	281	51	307	1030
	TOTAL	1522	1108	246	864	

Table 3. Confusion matrix of best model on test set

In the above table we can see the confusion matrix of the best performing model. Before running the experiment, it was hypothesized that the model would have a fairly difficult time differentiating between /*pah*/ and /*pah tah kah*/ but we expected there to be more differentiation between the syllabic non-word tasks, and VG/MO. This is because PA and PDK's only major difference is in the number of syllables, with the first syllable being the same. Both tasks are of non-words with a fixation cross visual prompt rather than the image of an object, so we'd expect the tasks to be very similar, and distinct from the VG task.

The first thing that stands out from the results of table 3, is that while there is the expected cluster around /PA/ and /PDK/, the model had about as equally hard of a time differentiating between the PA/PDK as well as MO. This result is not as surprising as initially believed however. Since the task in MO was to try to say PA/PDK while keeping their mouth open, the fact that the model had a hard time differentiating between MO and PA/PDK might suggest that the model has learned some features that capture speech intent, and not just the motor function of the speech.

The second thing that stands out is that VG performs incredibly poorly in classification. While the precision of the VG predictions is relatively high, it has terrible recall as seen in table 4,

classifying more true VG trials incorrectly as PA than correctly as VG. Overall, looking at the distribution of the predicted classes we can see that VG is almost never predicted, even beyond the discrepancy in the test set class distribution. We expected that relative weighting of the loss functions would deal with the issue of the VG task being underrepresented in the training set, but this still appears to be a problem. This might be solved by using other techniques such as oversampling or generating synthetic data such as SMOTE, although it is potentially dubious to generate synthetic data given that what we are trying to understand is the relationship between the changes in the brain signals in the different classes, and brain signals being fairly noisy and not well understood.

	Precision	Recall	F-1 Score
/PA/	0.31	0.46	0.37
/PDK/	0.32	0.34	0.33
VG	0.41	0.16	0.23
MO	0.36	0.3	0.32

Table 4. Precision and Recall values for best model on test set, as well as F-1 Score, which is the harmonic mean between precision and recall

4.3 High Confidence Examples

To have a better understanding of what the model is learning to classify, a high confidence and correct example of /*pah*/ was selected. The resulting activations were visualized relative to the physical locations of the sensors in the brain using MNE's plot_topomap function which takes the input data, and an array defining each channel's physical position, and interpolates the recording over a scalp image. The idea is that the high confidence examples are a "prototypical" example of the various classes have very prominent and expected features that helps the model differentiate and classify the example, and that we might be able to pick out some of these features that we hope the model has learned to look for.

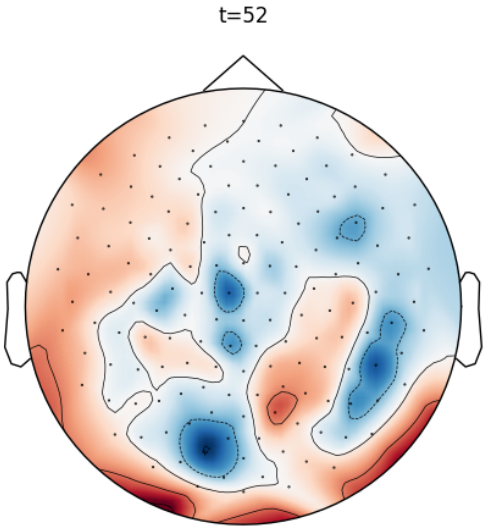


Figure 2. High confidence example, with high activations around the back of the head. Each trial has 481 time steps, and this was taken at the 52nd timestep.

In figure 2, we see dark red spots near the back of the head. This suggests that there is some visual stimulus that the subject is responding to, which can be explained by the visual cue that the subject receives before they perform the verbal task.

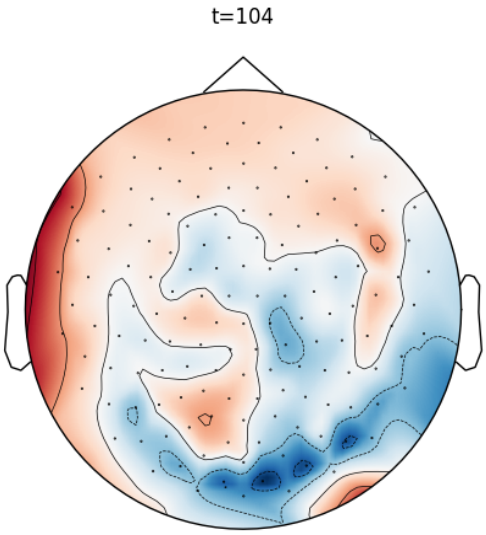


Figure 3. High confidence example with high activations around the left auditory cortex

In figure 3, we can see the highest activations around the left side of the brain near the ear. This is plausibly due to the left auditory cortex. This comes a few moments after the visual cue, and is likely the subject processing the auditory prompt.

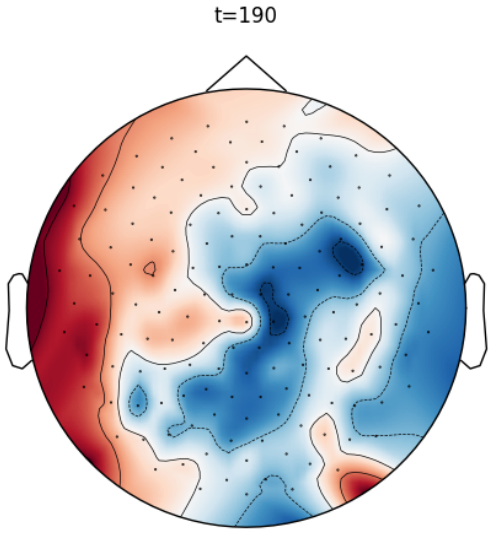
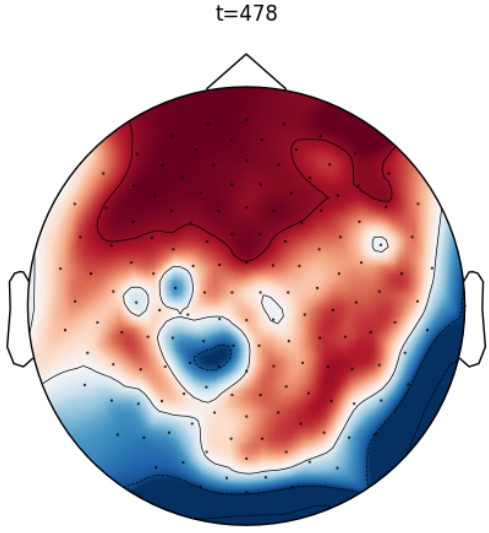


Figure 4. High confidence example with high activations around motor cortex and Broca's area

In figure 4, we see the higher activations move away from the edge of the left side of the head, and have higher activations starting roughly around broca's area and covering premotor and motor areas.. These are areas we expect to have high activations for a non-word task. As well, the area near the auditory cortex is very active as well, likely because the subject is hearing themselves speak as well.

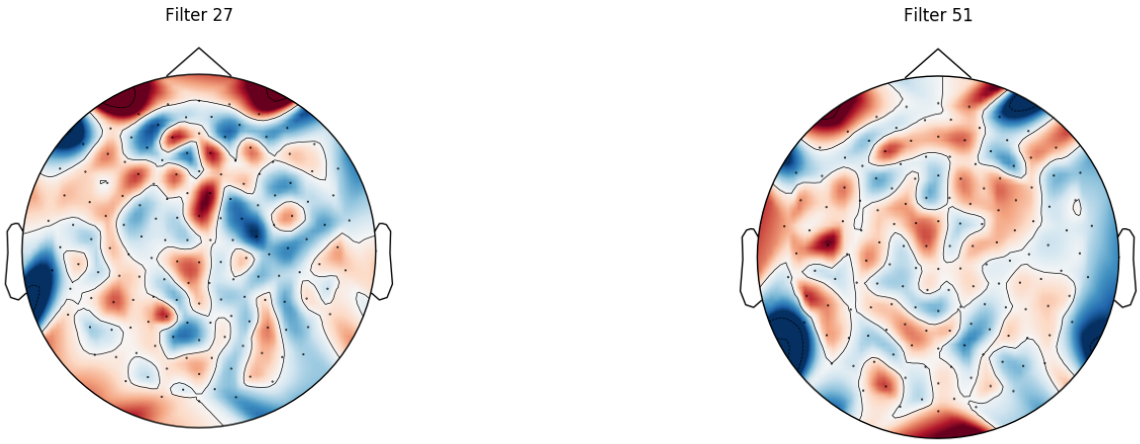
Figure 5. High confidence example, that demonstrates some muscle disturbance, possibly blinking



In figure 5, we see bright activations around all the front of the brain, which is likely due to some physical disturbance in the sample such as blinking. The hope is that because this is a high confidence example, the model has learned to ignore features/signals from disturbances such as this, and only focus on the more differentiating features.

4.4 Model Activations

The expectation is that the explainable factors from the prototypical example above are the features that the spatial summary stage is capturing, and that the GRU is learning to observe and using to classify the various speech tasks. By maximizing the output of the model's spatial summary stage with respect to the input of the network, we can gain some insight into the preferred characteristics of the input data, and understand the various features that the spatial summary stage has learned to capture. This method has been used in the past to show the characteristics of individual layers, and the preferred input for image classification networks [21]. We implemented artificial input maximization for the spatial summary stage for the model with the highest single fold test accuracy and visualized the inputs that maximized the outputs of each individual filter of the spatial summary stage. We could then visualize these maximized inputs in the same way as the prototypical example.



(a) (b)

Figure 6. Two examples of input channel relative intensities, using the channel locations. Each example maximized the output of specific components/filters of the spatial summary stage.

Figure 6 b), shows spatial mixing with strong relative intensity around channels near potentially the inferior frontal region. In contrast, figure 6 a) shows relatively high intensities near the front of the head, particularly focused around the eyes. While the exact locations of these plots are not completely accurate because the plots are formed using bilinear interpolation so the intensities are affected by edge effects. The hope is that the GRU is learning to ignore or remove the high activations around the eyes such as in filter/component 27, and accept some combination of other components

more similar to filter 51, although it is hard to fully predict or explain what the GRU is using to make its predictions.

There were no “perfect” or prototypical examples of desired features, but this is likely because the output of the spatial layer are weighted and combined in specific linear combinations to fully represent the desired features, and so these artificial components are possibly linear combinations of the features that we desire to observe.

4.5 Further Discussion

This work proposes a neural network architecture that can do some classification between various speech tasks with mixed success using raw MEG. In addition there is some evidence to show that the model is able to extract some features that reflect some non-trivial and relevant brain activity related to these speech tasks. There were many interesting results from observing the high confidence prototypical example, and the inputs that maximize the various components of the spatial summary, that suggests some degree of implicit feature engineering. However, it becomes more difficult to explain what the GRU is doing through time, other than making educated guesses, and so it is worth looking into other more explainable network architectures to deal with the time series data other than a GRU.

It would be worth revisiting these visualizations, as well as the confusion matrix again after making some improvements to the classification performance for VG, and in particular dealing with the underrepresented VG trials in the training data. Then we might be able to see potentially more insightful visualizations in both the prototypical examples, as well as the maximized inputs.

There likely needs to be more work done to show that the classification decisions are not made as a result of non-event/task related activity in the brain, such as the length of time that each region is active (e.g /*pah*/ is a longer speech task than /*pah tah kah*/) but overall, the main advantage of this method is a consolidated training process that did not require much domain knowledge. This approach is entirely trained through gradient-descent based optimizers without intermediate processing steps, and some of these trained parameters may be able to have interpretable and meaningful connections to the speech tasks, and brain activity.

5. Conclusion

5.1 Summary of Results

In the main classification task of this work, there were no major classification accuracy improvements, as it performed barely better than the majority class threshold. However, many of the mistakes made by the model were fairly understandable, having difficulties differentiating between very similar speech tasks /*pah*/ and /*pah tah kah*/. As well, the prototypical examples have some features which seem to be plausibly explained with existing knowledge about the different functions of the brain. This might suggest that there is some evidence that the model has learned to observe at least some of these features. In fact, an exploration of the components that maximize the activations of the spatial summary layer, show some differentiation of components that capture different spatial areas of the brain, suggesting some level of learning of features.

It requires more work to be done to show that the classification decisions being made by the network are actually using these qualitatively observed features and components, and not other non-event related activities in the brain, but overall this work shows some interesting and promising steps towards using explainable AI to better understand the complicated relationships within the brain.

5.2 Next Steps

Future work that improves the model and training to improve accuracy might result it more informative or interesting analysis of the learned features of the model. For example, by applying techniques such as oversampling or SMOTE to improve the recall of VG trials and have the predictions more representative of the original data, might result in more informative visualization of model maximal activations. As well, it would potentially be more informative to compare the prototypical examples of VG compared to the other three tasks to see what the differentiating features might be between the two tasks.

Another way to improve the model would be to include attention mechanisms to the spatial summary stage. Firstly, an attention mechanism can be added in the spatial dimension, to learn better how to weigh different channels/learned features as an input at each timestep. This might be more effective than simple convolutions because convolutions look for local patterns in smaller windows of data, but the order of the channels in the MEG data is fairly arbitrary, and thus an attention mechanism might be more effective in grouping related channels together, even if they are for example, at index 1 and 100, which would not be captured by the convolution, This could lead to more distinguishable and distinct maximal activations. Secondly, an attention mechanism can be

added as an enhancement to the GRU, by having an attention weighted input layer, which can now combine the inputs at different timesteps to produce the output at any given outputs. This could be a significant improvement to the accuracy of the model because the temporal attention weighting can now reflect relationships between various features/behaviour of different parts of the brain which have some time-delayed relationship. For example, high activity in one part of the brain can lead to higher activity in a different part of the brain, just a few timesteps later.

Finally, while the spatial summary stage of the model can be explained with relative ease by doing the maximal activations and visualizing these inputs and filters of the convolutions. However, the GRU is harder to explain what features are being used and compared to perform the classifications, as well as what data is being saved in the memory of the GRU. This leaves plenty of room to apply techniques such as Local Interpretable Model-Agnostic Explanations (LIME), or Anchor LIME (ALIME) [22] to better understand the classifications, instead of looking at the GRU as a black box. This method attempts to create local explanations for the predictions made by a classifier and can lead to more a more detailed explanation of what the model is using to classify the various speech tasks.

6. References

[1] A. E. Hillis, M. Work, P. B. Barker, M. A. Jacobs, E. L. Breese, and K. Maurer, "Re-examining the brain regions crucial for orchestrating speech articulation," *Brain*, vol. 127, no. 7, pp. 1479–1487, Jul. 2004.

[2] F. H. Guenther, "Neural control of speech movements," *Phonetics and phonology in language comprehension and production: Differences and similarities*, pp. 209–239, 2003.

[3] D. S. Kadis, E. W. Pang, T. Mills, M. J. Taylor, M. P. McAndrews, and M. L. Smith, "Characterizing the normal developmental trajectory of expressive language lateralization using magnetoencephalography," *J. Int. Neuropsychol. Soc.*, vol. 17, no. 5, pp. 896–904, Sep. 2011.

[4] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG Classification of Covert Speech for Brain–Computer Interfacing," *Int. J. Neural Syst.*, vol. 27, no. 08, p. 1750033, Dec. 2017.

[5] M. P. Guimaraes, D. K. Wong, E. T. Uy, L. Grosenick, and P. Suppes, "Single-trial classification of MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 3, pp. 436–443, Mar. 2007.

[6] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 992–996.

[7] Y. R. Tabar and U. Halici, "Brain Computer Interfaces for Silent Speech," *European Review*, vol. 25, no. 02. pp. 208–230, 2017.

[8] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, no. 5, pp. 787–798, May 1999.

[9] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Rev. Mod. Phys.*, vol. 65, no. 2, p. 413, 1993.

[10] R. Assadollahi and F. Pulvermüller, "Neural Network Classification of Word Evoked Neuromagnetic Brain Activity," in *Emergent Neural Computational Architectures Based on Neuroscience: Towards Neuroscience-Inspired Computing*, S. Wermter, J. Austin, and D. Willshaw, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 311–319.

[11] T. N. Lal *et al.*, "A Brain Computer Interface with Online Feedback Based on Magnetoencephalography," in *Proceedings of the 22Nd International Conference on Machine Learning*, Bonn, Germany, 2005, pp. 465–472.

[12] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *arXiv [cs.LG]*, 09-Aug-2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated Feedback Recurrent Neural Networks," *arXiv [cs.NE]*, 09-Feb-2015.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv [cs.NE]*, 11-Dec-2014.

[16] Y. LeCun, Y. Bengio, and Others, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[17] V. Y. Yu, M. J. MacDonald, A. Oh, G. N. Hua, L. F. De Nil, and E. W. Pang, "Age-related sex differences in language lateralization: A magnetoencephalography study in children," *Developmental Psychology*, vol. 50, no. 9. pp. 2276–2284, 2014.

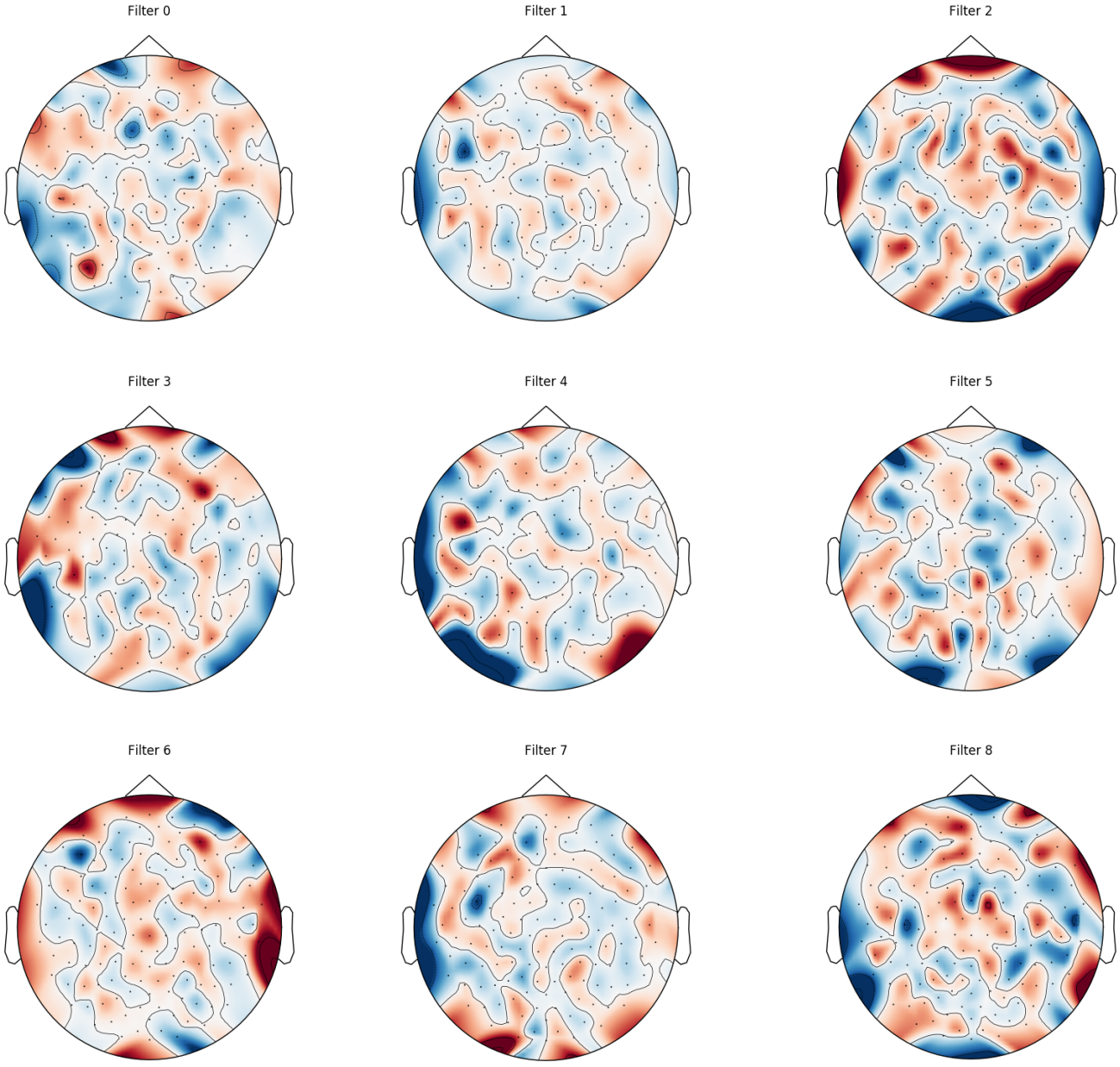
[18] S. M. Doesburg, K. Tingling, M. J. MacDonald, and E. W. Pang, "Development of Network Synchronization Predicts Language Abilities," *J. Cogn. Neurosci.*, vol. 28, no. 1, pp. 55–68, Jan. 2016.

[19] J. Campbell, "Book Review: Peabody Picture Vocabulary Test, Third Edition," *Journal of Psychoeducational Assessment*, vol. 16, no. 4. pp. 334–338, 1998.

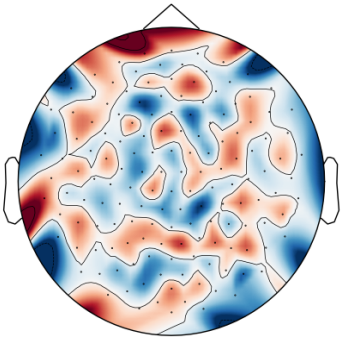
[20] K. T. Williams, *Expressive vocabulary test*. American Guidance Service, 1997.
[21] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding Neural Networks Through Deep Visualization," *arXiv [cs.CV]*, 22-Jun-2015.
[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv [cs.LG]*, 16-Feb-2016.

7. Appendix

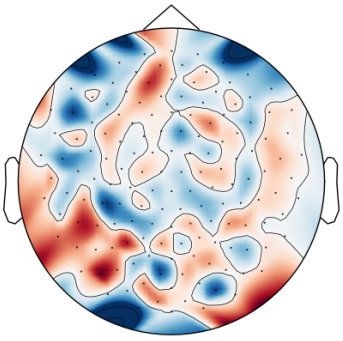
A. All Maximized Activations



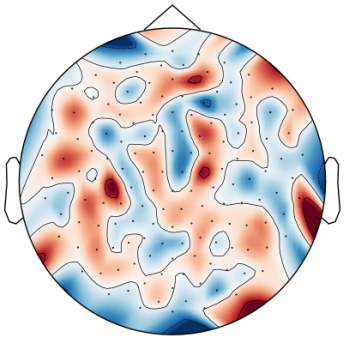
Filter 9



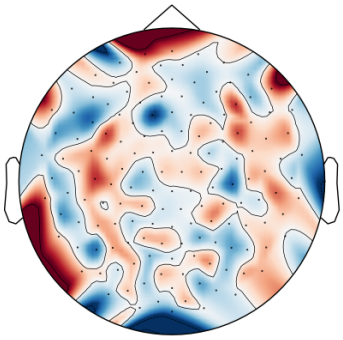
Filter 10



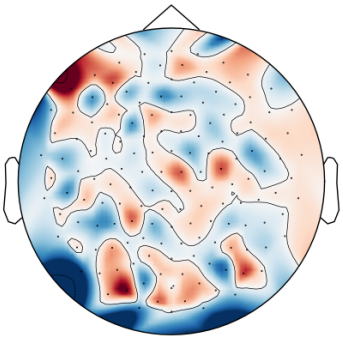
Filter 11



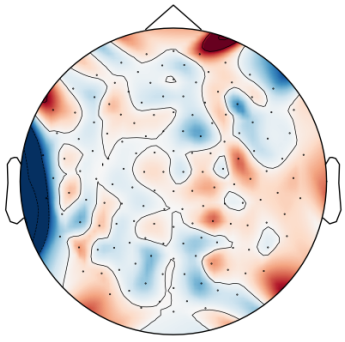
Filter 12



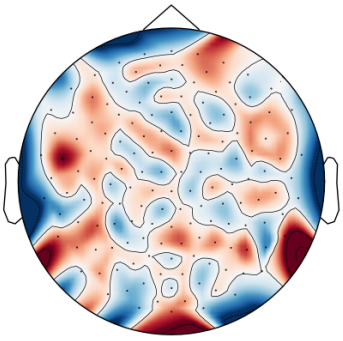
Filter 13



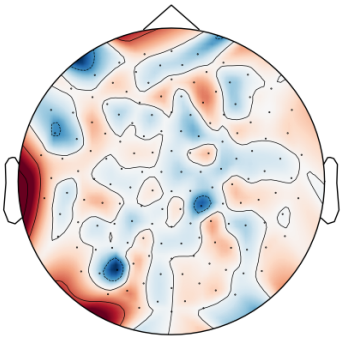
Filter 14



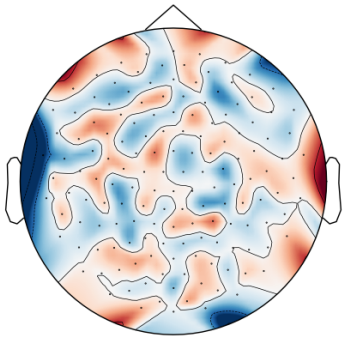
Filter 15



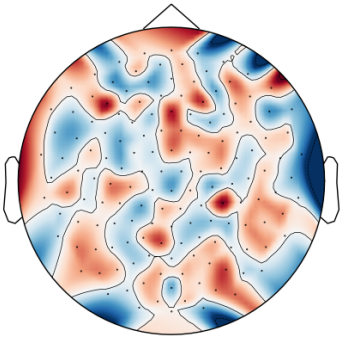
Filter 16



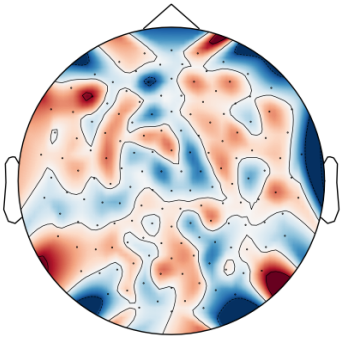
Filter 17



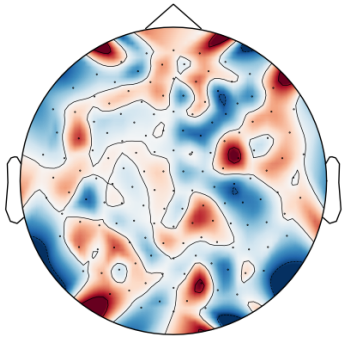
Filter 18



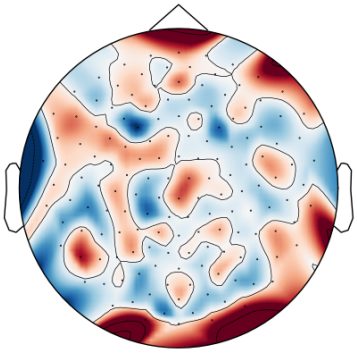
Filter 19



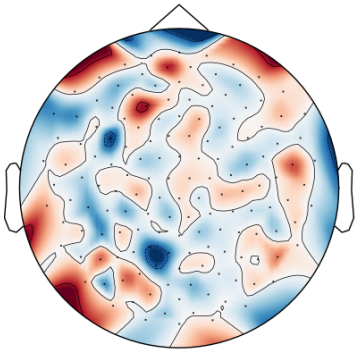
Filter 20



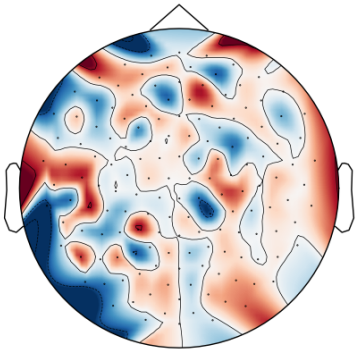
Filter 21



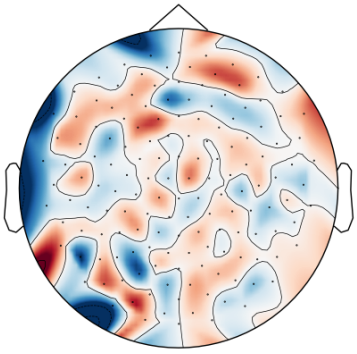
Filter 22



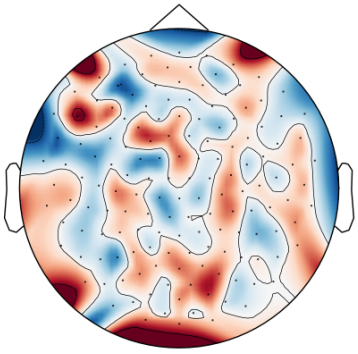
Filter 23



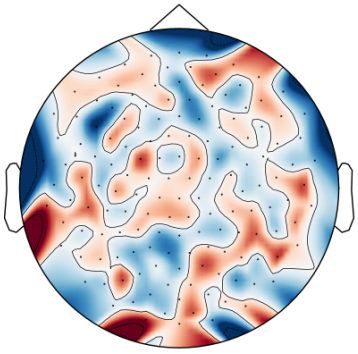
Filter 24



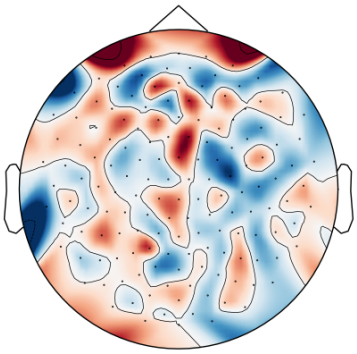
Filter 25



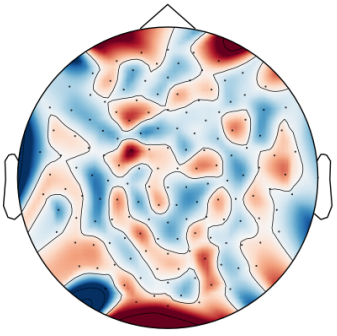
Filter 26



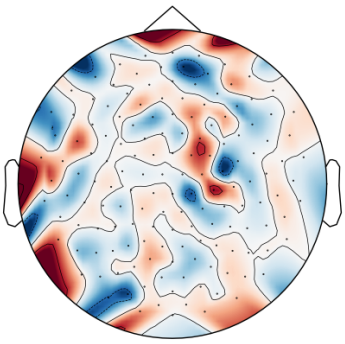
Filter 27



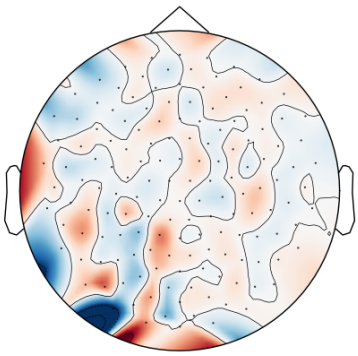
Filter 28



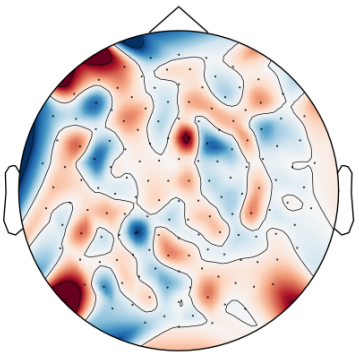
Filter 30



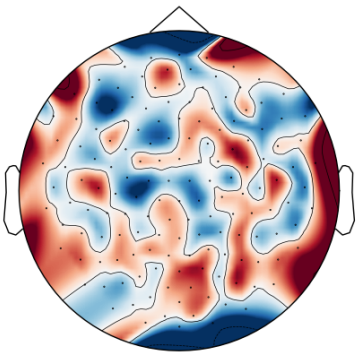
Filter 29



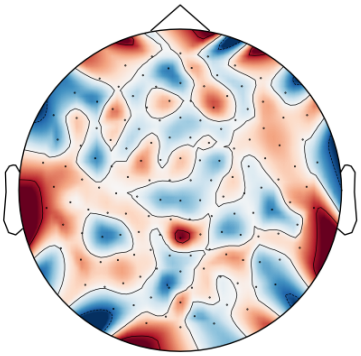
Filter 31



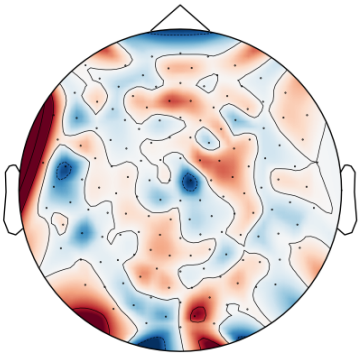
Filter 32



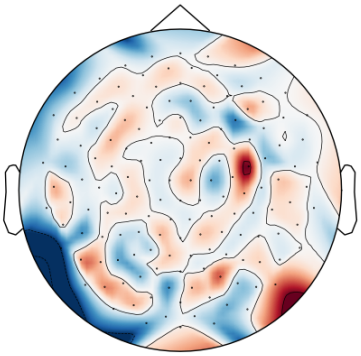
Filter 33



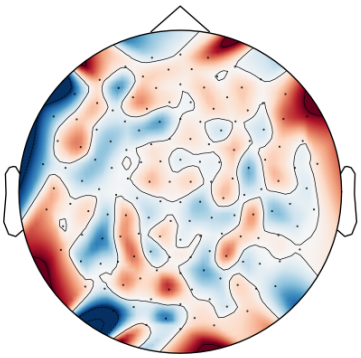
Filter 34



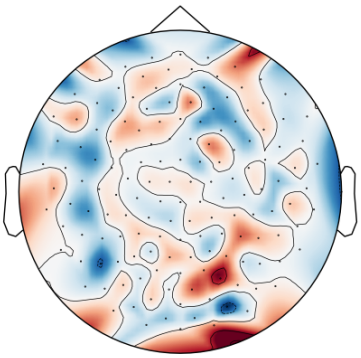
Filter 35



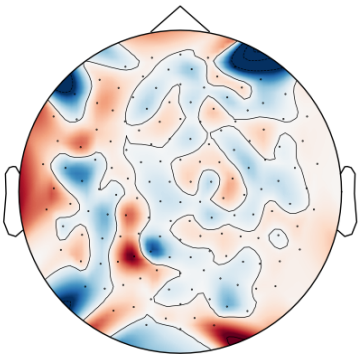
Filter 36



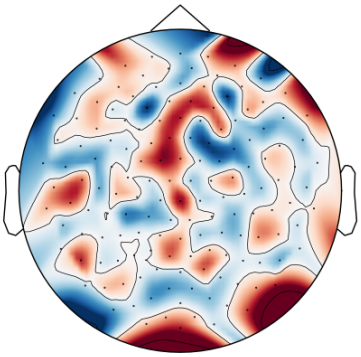
Filter 37



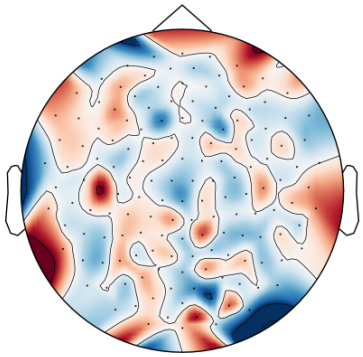
Filter 38



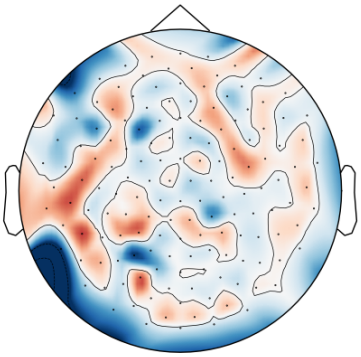
Filter 39



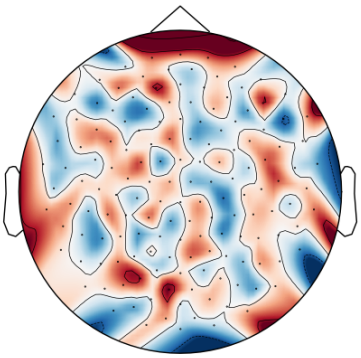
Filter 40



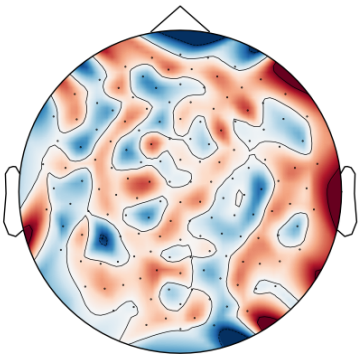
Filter 41



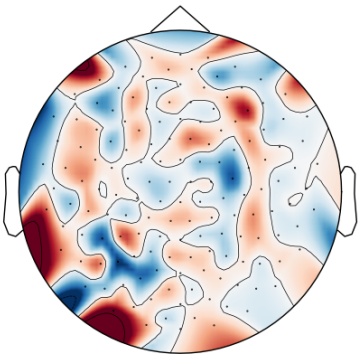
Filter 42



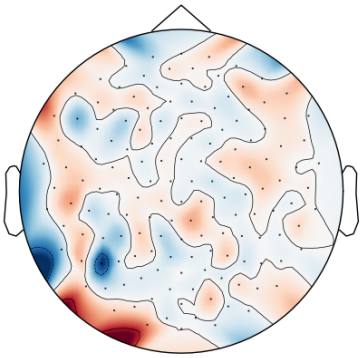
Filter 43



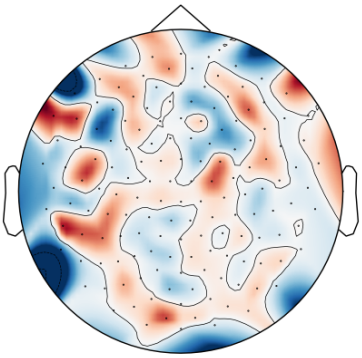
Filter 44



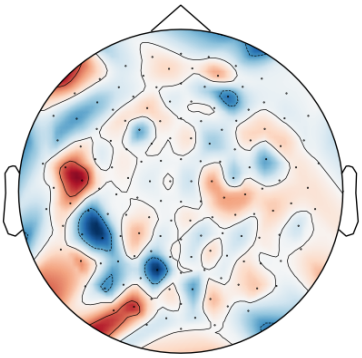
Filter 45



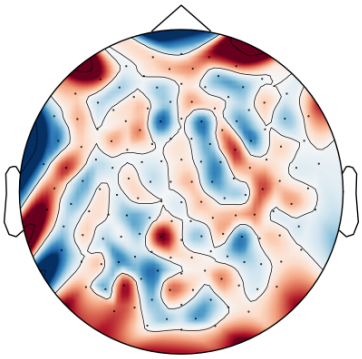
Filter 46



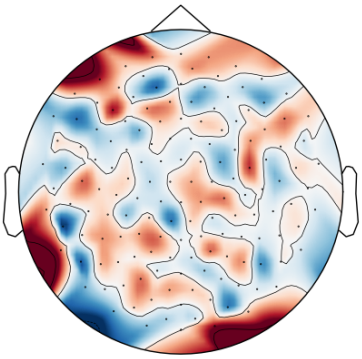
Filter 47



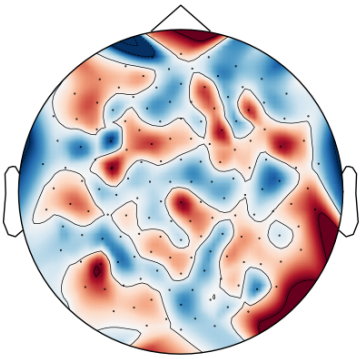
Filter 48



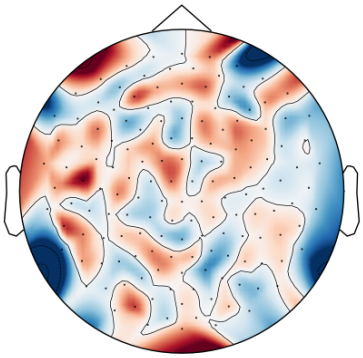
Filter 49



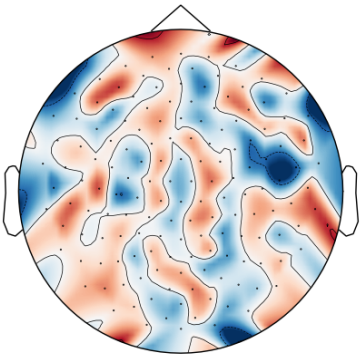
Filter 50



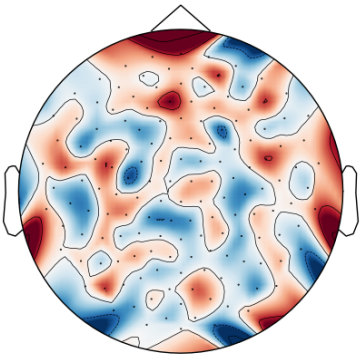
Filter 51



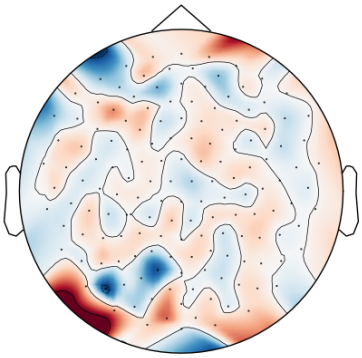
Filter 52



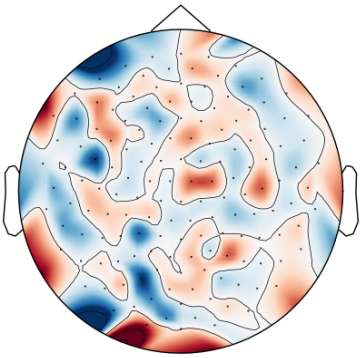
Filter 53



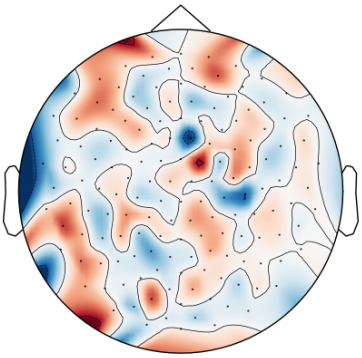
Filter 54



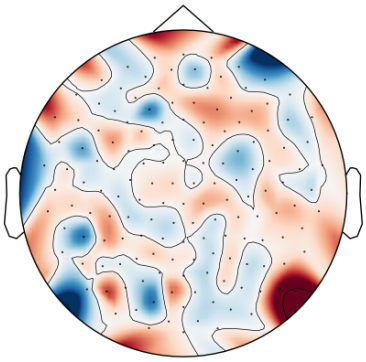
Filter 55



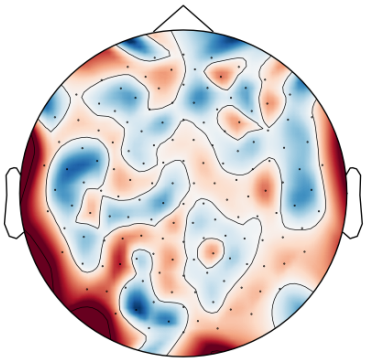
Filter 56



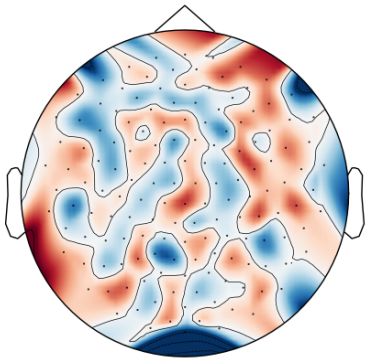
Filter 57



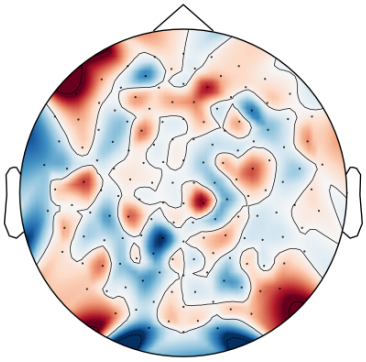
Filter 58



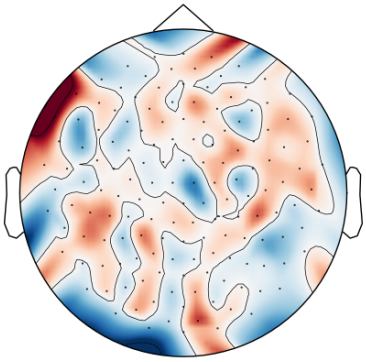
Filter 59



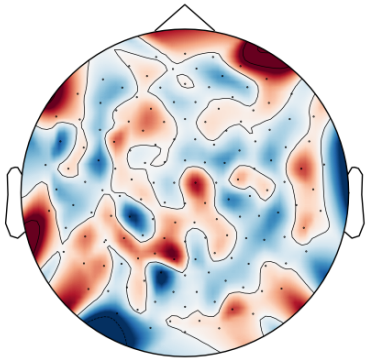
Filter 60



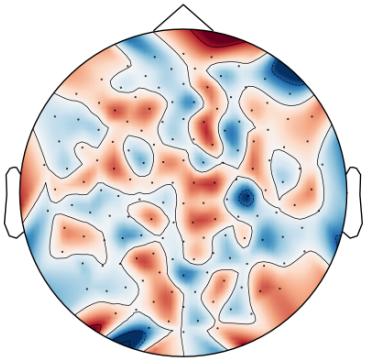
Filter 61



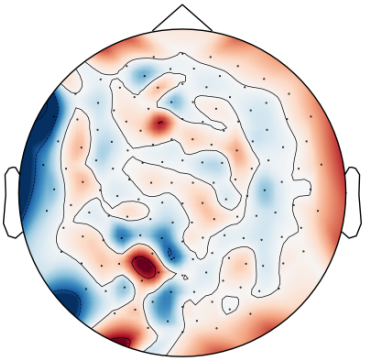
Filter 62



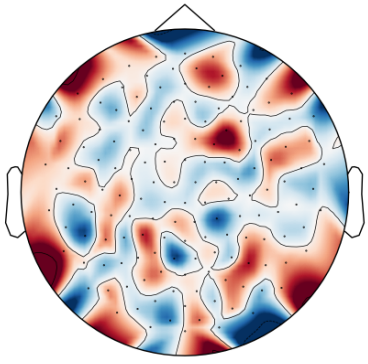
Filter 63



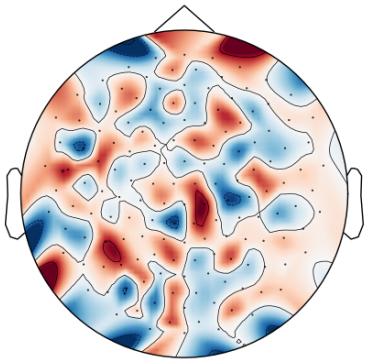
Filter 64



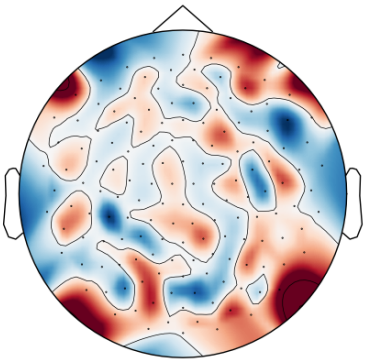
Filter 65



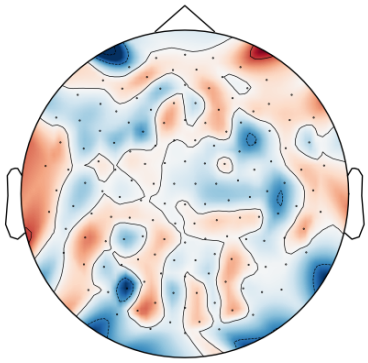
Filter 66

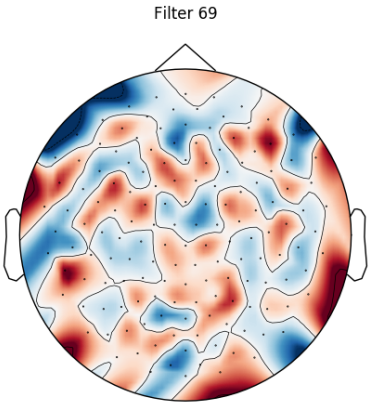


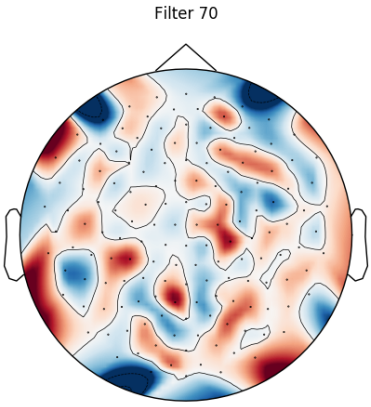
Filter 67

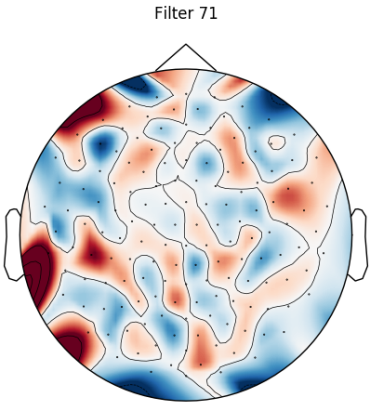


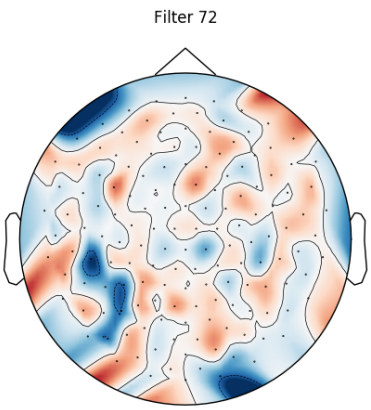
Filter 68

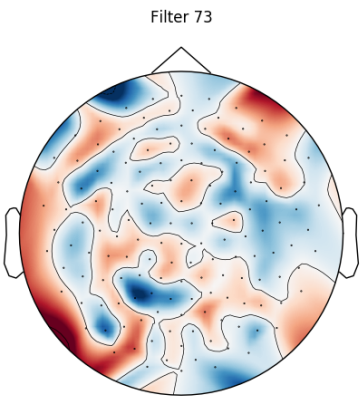


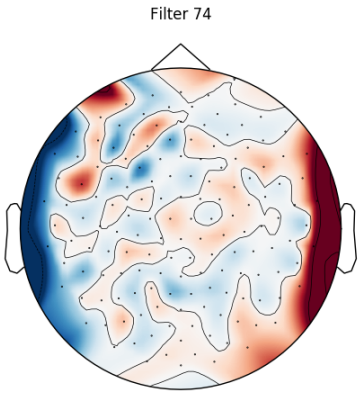

Filter 69


Filter 70


Filter 71


Filter 72


Filter 73


Filter 74