MEG-driven Emotion Classification Using Convolutional and Recurrent Neural Networks

Jakub Orlinski
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
j.orlinski@student.utwente.nl

ABSTRACT

In this paper, a new multi-channel emotion classification method based on the novel magnetoencephalography (MEG) dataset CiNet is proposed. This paper falls into the field of Brain-Computer Interface (BCI) research, as it uses brain activity data for recognizing human emotions. It should prove a valuable contribution and a comparison, as most BCI research uses electroencephalography (EEG) data instead, primarily from the DEAP dataset. Using a combination of a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN), the system will analyze the high-fidelity data in an attempt to recognize the emotional state of the subject. The CNN encodes spatial information, while the RNN tracks changes over time. Each part is evaluated separately as well as in conjunction, so as to establish the contribution of each of the aspects of analysis. Those model variations are evaluated on both raw MEG signals and the Power Spectrum Density (PSD) extracted from the signal. The experimental results show that the best model is the CNN+RNN combination trained on raw signal data, and it achieves a mean accuracy of 56.5% on the valence/arousal classification task.

Keywords

Magnetoencephalography, Artificial Neural Network, Emotion Classification, Brain-Computer Interface, Valence-Arousal model

1. INTRODUCTION

Emotion recognition is a key aspect in developing Human-Computer interfaces as the users' needs and wants are heavily influenced by their emotional state. It could also provide an automated system for generating labels of subjects' emotional states during other experiments that require such an objective measure.

In recent years the field of emotion recognition based on EEG data has seen a lot of increased interest [20] and the leading dataset in this domain is the DEAP dataset [12][20]. In this research, however, we will use a new dataset that was acquired using MEG. While more expensive, MEG data is thought to be much better suited to the task of recognizing emotions as the magnetic fields propagate much further within the brain than electrical ones. For EEG this problem is compounded by the fact that the scalp itself reduces the signal strength by as much as two orders of magnitude. In contrast, MEG signals propagate and penetrate the brain and scalp much better, leading to much higher fidelity. This is a significant advantage especially in this task, as the limbic system, the part of the brain responsible for processing emotions, is located deeper within the brain. While EEG is capable of capturing signals at shallow cortical regions in the frontal lobe, which partly plays a role in the limbic system, MEG can record signals from much deeper regions where the center of emotional processing is located at, as shown in [18]. Another important advantage of the MEG readings is that the signal has a much better spatial resolution which, in combination with a much higher channel count than previous studies on EEG datasets, should yield an improvement due to superior source separation ability [15].

The other techniques in this area of research, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) aren't suitable for this task. PET is invasive and includes the use of radiation which disqualifies it completely. fMRI is much better than PET but still provides very low temporal resolution for this task - around 5 seconds [16]. This is above the approximate 1 second necessary for recording emotional response [23].

Having marked the difference between EEG and MEG, it is important to say that both signals stem from the same source - an assembly of neurons being activated in a coordinated way. Neuron activation is done through electric signals and the resulting electric and magnetic fields can be measured. Since the source is the same, it should be possible to leverage existing knowledge from the EEG-based research in developing algorithms for data preprocessing and for making informed assumptions about the design and architecture of the Neural Network.

This paper's contribution to the field of affective Brain-Computer Interface (aBCI) is two-fold. Firstly, it will provide a Neural Network model that uses MEG data from the novel CiNet dataset [22] for emotion classification. Secondly, it will investigate the paradigm known as Convolutional Recurrent Neural Networks (CRNN) for this task and explore the hyperparameter space to establish which aspects of the system and of the data are most important and provide best performance.

The rest of the paper is organized as follows. Firstly, to establish a baseline of knowledge for the reader, background of the topic and the dataset is given in Section 3. After that in Section 4, related work in this field is reviewed. Section 5 details the experimental steps, the results of which are presented in Section 6 and discussed in Section 7. Lastly, Section 8 gives conclusions from the research

33rd Twente Student Conference on IT July. 3rd, 2020, Enschede, The Netherlands.

and recommendations for further directions.

2. RESEARCH QUESTIONS

RQ1: Can an architecture be made that uses MEG data and classifies emotions better than a standard machine learning model?

RQ2: Which part of the proposed model are important and how does each impact the accuracy of the system?

3. BACKGROUND

3.1 Emotion classification

Emotion classification as a field is split between two approaches. One uses proxy indicators, such as speech tone and facial expressions, to predict the current affection [13]. This field has been studied extensively but is limited due to similarity between signals of different emotions [26]. The other - physiological signals - offer a much more objective measure of emotional state, chief amongst which is the neural activity in the brain. The recorded magnetic fields used in this research serve as a much more objective manifest variable that provides a basis for inference of the emotional state.

In the field of BCI the standard for quantizing emotion has long been the Valence/Arousal paradigm [4]. There, the emotions are mapped onto a 2D plane with the x-axis being valence - how positive or negative the emotion is - and the y-axis being arousal - how strongly the person feels the emotion. This is further illustrated in Figure 1

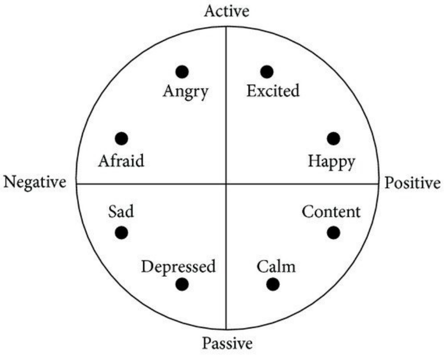


Figure 1. Valence/Arousal model as presented in [10]

3.2 CiNet dataset

CiNet is a new dataset aimed to further research into aBCI by providing higher than ever fidelity of data [22]. Its novel approach of capturing MEG rather than EEG information has already used in other domains [9] but also to supplement the limitations of facial emotion recognition techniques [26].

The data was gathered from 36 subjects over 6 trials each. A trial consists of 4 listening periods that are broken into a 5 second white noise primer and a subsequent 45 second song playback. After that, the subject has 20 seconds to rate his/her emotional response on a scale of 1 to 9 in the valence and arousal domains.

The MEG data consists of 102 magnetometers whose readings have been corrected to account for the position of the head using SSS algorithm [21] and other factors such as interference. This is a direct increase over previous datasets, the most used of which - DEAP – offers only 32 channels of measurement.

4. RELATED WORK

The field of aBCI has been extensively studied in recent years as the availability of advanced machinery for both data collection and analysis have increased [5][2].

This body of research can be broken down into multiple parts according to learning techniques used, data preprocessing steps and focus on domains of analysis reflected by system design.

Early research used shallow learning models to estimate the emotional state of the subjects. SVM's and kernel classifiers were used to limited effect of around 70% accuracy [3]. The other side of this division consists of Artificial Neural Network-based approaches.

As ANN's are sensitive to the data fed to them, preprocessing steps tend to have a substantial influence on their accuracy and ability to generalize. Researchers here seem divided mostly on two factors - whether to use raw signals as input [1][25] or whether to decompose the signal into its constituent frequencies [24][17][19][14]. The other factor determines how the data is to be structured.

The structural division yields many subcategories. The main one encompasses a large body of research that takes into account the multi-dimensionality of the input, a natural result of which is the choice of Convolutional Neural Networks. The dimensions in questions tend to reflect the spatial separation of the measuring nodes and the multiple frequency bands taken into account.

To account for the spatial separation of measuring nodes and brain region connectivity, some research groups use an adjacency matrix. The matrix is usually a 2D approximation of the spacing of electrodes or can take into account functional connectivity. The latter reflects the fact that different brain regions, while anatomically separate, can take part in the performance of the same brain function [6]. This can be done by computing a connectivity index measuring the coupling of signals from two different brain regions as in [17], but also can be the learning target of the network [19]. While this is outside the scope of this research, the target architecture will follow computational neuroscience principles and account for such effects.

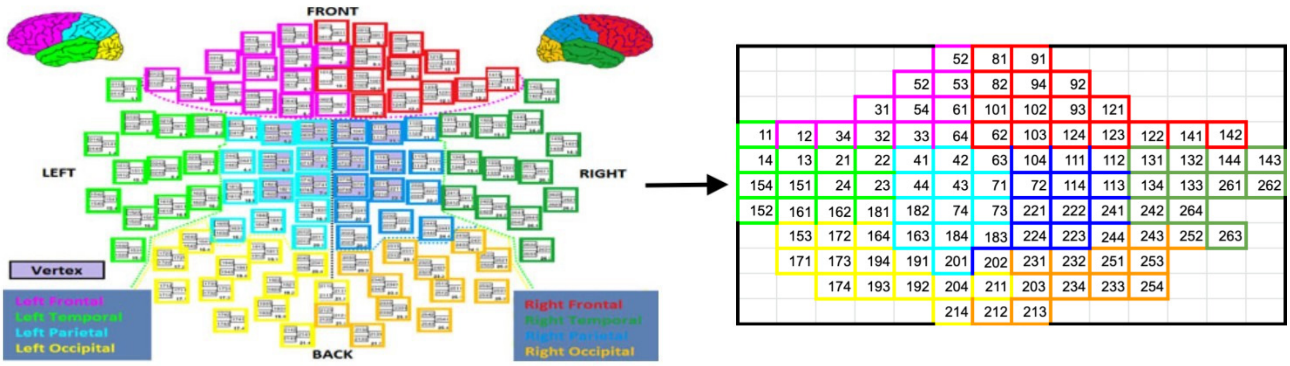
Last but not least, to reflect a focus on the time domain of the experiments, some research investigates the viability of the Recurrent Neural Network architecture pattern. The viability of this approach is demonstrated in [1], but can also serve as an auxiliary part of a larger system as shown in [25], resulting in an architecture that is responsive to both spatial and temporal context of the information. This is especially important as a single snapshot of brain activity can't represent the whole temporal spectrum of relevant activity which is crucial in analyzing emotional response, as indicated in [7].

There are other ways of extracting the temporal information from the signals such as fractal-dimension or zero-crossing value. However, it has been proven that Neural Networks can approximate any function [8]. This means that with the right architecture and enough time the RNN part of the network will approximate the most optimal function for extracting that temporal information.

5. METHODS

5.1 Preprocessing

The MEG data from the CiNet dataset was first imported into Python arrays and reshaped according to the mapping shown in Figure 2. This yielded the corresponding 11 x 14 matrix for each timestep of the experiment. The



					52	81	91						
				52	53	82	94	92					
			31	54	61	101	102	93	121				
11	12	34	32	33	64	62	103	124	123	122	141	142	
14	13	21	22	41	42	63	104	111	112	131	132	144	143
154	151	24	23	44	43	71	72	114	113	134	133	261	262
152	161	162	181	182	74	73	221	222	241	242	264		
	153	172	164	163	184	183	224	223	244	243	252	263	
	171	173	194	191	201	202	231	232	251	253			
		174	193	192	204	211	203	234	233	254			
					214	212	213						

Figure 2. MEG electrode spatial mapping onto a 2D 11 x 14 matrix. Position of electrodes was taken from [11]

data was then downsampled 8-fold from 1kHz, as the relevant frequencies are between 1 and 45Hz. With 45Hz being the Nyquist frequency, the sample rate of 125Hz is sufficient. The resulting array of timesteps per song per subject constituted the training examples for the system.

One of the variations on the proposed training scheme was to compute the Power Spectrum Density (PSD) of the data with the help of the Fourier Transform. This was done using the sliding window technique - Welch's method. The size of the window from which the Power Spectrum Density was derived was determined to be 1 second, with the step between applications of the Fourier Transform was defined to be 10 timesteps (80 milliseconds). The bands taken into account are Delta (1 to 4Hz), Theta (4 to 8Hz), Alpha (8 to 12Hz), Beta (12 to 30Hz) and Gamma (30 to 45Hz).

In regards to the training labels, the 1 to 9 scores on the valence and arousal scales were translated into high/low valence and arousal scores, with scores 1-4 being low and 5-9 taken to be high. This is done commonly in research in this field, due to the resulting dramatic decrease in complexity. These 4 scores form the multi-class output labels for the training set - [1010] encodes low valence low arousal and [0110] encodes high valence low arousal.

5.2 Architecture

The architecture explored in this paper was inspired by the work in [25]. It was one of the few papers to take into account both of the crucial dimensions of the data - the spatial and the temporal. This research adapts the model proposed there and experiments with the introduction of the novel ConvLSTM2D layers available in Keras. These layers perform similar operations to a regular LSTM cell but can operate on 2D data.

The data flow is depicted in Figure 3 and proceeds as follows. First, timesteps from t to t+S (S is the number of considered timesteps, e.g. 1 sec * 125Hz) are fed through a series of parallel Conv2D layers. To preserve the dimensionality of the initial input, ”same” padding was used and the feature space is increased by doubling the filter size from one layer to the next. The kernels at each step are of the size 11 x 14 to make sure that functional connections are not lost - it is possible that regions that are spatially separate work together to provide some output. The outputs of the CNN layers are concatenated and then the dimensionality reduced by additional Conv2D layers.

At the same time, in the RNN part of the model, ConvLSTM2D layers are fed the MEG data. Here, only one series of layers is needed, as RNNs are specifically made for time-series data. After processing the data, this part of the network outputs the hidden state at the last timestep.

The resulting outputs from both parts of the system are 11 x 14 x 18 matrices (the first two axes are the original dimensions of the input, while the third is a value used in [25] and taken here as a reference), that are then concatenated and flattened. At the end, 3 Dense layers of size 1024 follow each other to provide enough complexity to encode the data. The last one outputs the vector of length 4, with the first two indices of which are valence scores and the second the arousal ones.

To answer the second research question, the hyperparameters tested were the impact of extracting the PSD from the data and the importance of the constituent parts of the above-described model. Therefore, the architecture was adjusted to provide 3 models - a CNN-only one, an RNN-only one and a hybrid one. Then, for each of them, the data was adjusted to be either the raw MEG signals or the PSD of the signals, yielding 6 total experiments.

6. RESULTS

6.1 Training & Validation

For each of the subjects, a model was constructed and then trained on 18 out of the 24 trials, leaving the last 6 as validation data. Then, the validation accuracy was averaged across subjects to get the model accuracy presented in Table 1. This is, therefore, a leave-one-out validation scheme. While it does limit the possible convergence of the models due to the reduction in the amount of training data, it provides as close a measure of accuracy as possible. K-fold cross-validation was at first considered but, due to the similarity of data after extracting the PSD, it was deemed to be unsatisfactory, because that similarity would yield incorrectly high accuracy. Another consequence of extracting the PSD from the signal is the reduction of the number of samples available for training and validation which could skew the resulting accuracy.

To provide a baseline of performance a Logistic Regression (LR) model was trained by extracting the PSD from 1 second of data into the 5 channels specified. This yielded an input matrix of (samples, channels, bands) that was squashed to conform to the shape accepted by the Scikit Learn implementation of the LR - (samples, channels * bands). The solver selected was SAGA.

6.2 Scores

The results for the different architectures can be found in

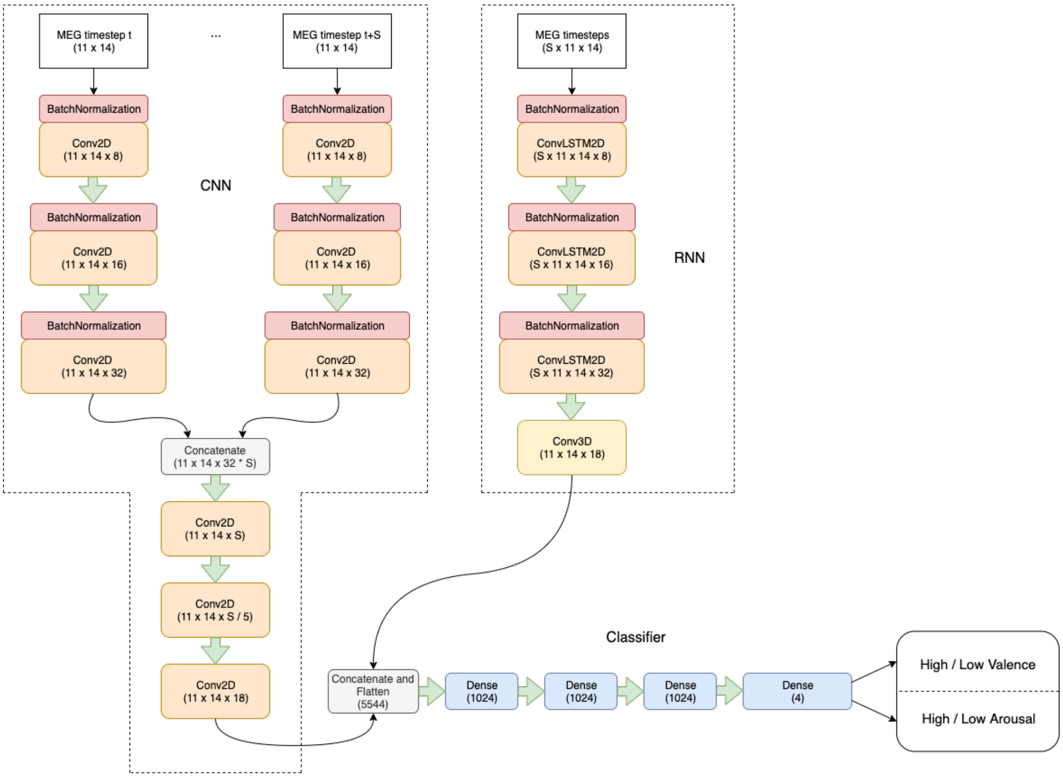


Figure 3. Final CNN + RNN architecture for emotion classification. The dashed line shows the separation between the constituent CNN and RNN networks. S is the number of considered timesteps, e.g. 1 sec * 125Hz

Model	PSD	Accuracy	St. Deviation
RNN	-	49.5%	25.3%
CNN	-	50.4%	27.6%
CNN + RNN	-	56.5%	27.9%
RNN	+	53.2%	26.3%
CNN	+	54.6%	23.8%
CNN + RNN	+	48.6%	29.8%
LogisticRegression	+	36.3%	5.3%
Random	-	25%	-

Table 1. Results

Table 1. Provided as a baseline is the random expected accuracy at 25% and the LR, which achieved an accuracy of 36.3% at a surprisingly low standard deviation, at only 5.3%. This is of course better than random, but below any expectation of a model in this task. However, this does help put the gains in performance achieved by our architecture into context.

In regards to the models, while the differences between accuracies are not striking, they do provide interesting insight. Firstly, the constituent parts of the model seem to perform quite similarly to each other, wit only a 1 percentage point difference. It is also quite clear that the CNN and RNN parts do perform better on the PSD extracted from the signal.

Secondly, while the CRNN system works better together on raw data, its performance on the PSD extracted data is rather sub par - a 5.3 percentage point drop from the mean of the two sub-networks. And as the standard deviation is not wildly different from the other models', it is rather striking.

7. DISCUSSION

The results achieved, while not on par with the state-of-the-art in the field, do still provide valuable insight into both the emotion classification task and some of the Neural Network performance.

The first research question is therefore answered - an architecture was found that can classify emotions based on brain activity data better than random and better than a basic Logistic Regression model.

The similarity in accuracy of the CNN and RNN models is surprisingly similar. This could be interpreted to mean that a series of concatenated convolutional outputs, each analyzing one time-point, provide similar functionality to the new ConvLSTM2D layers. This is a little surprising, even though the ConvLSTM2D layers were indeed made for such a purpose. The takeaway here might be that using the usual Conv2D layers is after all better, as, even though the number of parameters grows quickly, the convolutional operations can be done in parallel rather than sequentially. This provides a boost in training and prediction time - the CNN-only model took 382ms and 45 ms per training step, whereas the RNN-only model took 1sec and 212ms per training step, on raw and PSD data respectively.

As for the performance of the combined model, it is no surprise that it does better than the constituent parts as it provides more complexity to encode the data. The surprise comes in the form of the PSD-trained CRNN model which does perform much worse. This could be attributed to the limitation on the number of data points due to PSD extraction, but it should be investigated further.

8. CONCLUSIONS AND FURTHER RESEARCH

In this paper, an easily extendable network for classifying emotions was designed. Then, the raw MEG signals were reshaped and prepared for training. Lastly, multiple variants of the hybrid network were evaluated to select the one best suited for the task at hand. Experimental results show that the best model achieved an accuracy of 56.5% on the validation data selected in a leave-one-out approach.

My recommendation for further research would be to take into account recent developments for time-series analysis by introducing Transformers as the RNN part of the model and more experimentation with the model - e.g. varying the amount of time taken into account, trying different techniques for combining sub-network outputs - and data preprocessing - e.g. employing different methods for frequency extraction, removing the baseline of the signal as done in [25]. More work should also be done in developing techniques for the interpretability of AI. With high accuracy given by models and the ability to track how the model is able to provide its results, it might be possible to settle the debate over the patterns of brain activity that determine emotions and their structural and temporal aspects.

9. REFERENCES

[1] S. Alhagry, A. Aly, and R. El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *International Journal of Advanced Computer Science and Applications*, 8, 10 2017.

[2] A. Alnafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil. Review and classification of emotion recognition based on eeg brain-computer interface system research: A systematic review. *Applied Sciences*, 7:1239, 11 2017.

[3] J. Atkinson and D. Campos. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers. *Expert Systems with Applications*, 47, 11 2015.

[4] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49 – 59, 1994.

[5] R. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *T. Affective Computing*, 1:18–37, 01 2010.

[6] K. Friston. Functional and effective connectivity: A review. *Brain connectivity*, 1:13–36, 01 2011.

[7] B. Giordano, W. Whiting, N. Kriegeskorte, S. Kotz, P. Belin, and J. Gross. From categories to dimensions: spatio-temporal dynamics of the cerebral representations of emotion in voice, 2018.

[8] R. Gonzalez-Díaz, M. Gutiérrez-Naranjo, and E. Paluzo-Hidalgo. Two-hidden-layer feedforward networks are universal approximators: A constructive approach. 07 2019.

[9] C. Hasegawa, T. Ikeda, Y. Yoshimura, H. Hiraishi, T. Takahashi, N. Furutani, N. Hayashi, Y. Minabe, M. Hirata, M. Asada, and M. Kikuchi. Mu rhythm suppression reflects mother-child face-to-face interactions: A pilot study with simultaneous meg recording. *Scientific Reports*, 6:34977, 10 2016.

[10] S. Jirayucharoensak, S. Pan-ngum, and P. Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *TheScientificWorldJournal*, 2014:627892, 09 2014.

[11] M. Khomami Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses". *IEEE Transactions on Affective Computing*, PP:1, 01 2015.

[12] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3:18–31, 12 2011.

[13] C. Latha and M. Priya. A review on deep learning algorithms for speech and facial emotion recognition. *APTIKOM Journal on Computer Science and Information Technologies*, 1:92–108, 11 2016.

[14] W. Liu, W.-L. Zheng, and B.-L. Lu. Emotion recognition using multimodal deep learning. volume 9948, 10 2016.

[15] F. Lopes da Silva. Eeg and meg: relevance to neuroscience. *Neuron*, 80:1112–28, 12 2013.

[16] R. Lystad and H. Pollard. Functional neuroimaging: A brief overview and feasibility for use in chiropractic research. *Journal of the Canadian Chiropractic Association*, 53:59–72, 04 2009.

[17] S.-E. Moon, S. Jang, and J.-S. Lee. Convolutional neural network approach for eeg-based emotion recognition using brain connectivity and its spatial information. pages 2556–2560, 04 2018.

[18] F. Pizzo, N. Roehri, S. Medina Villalon, A. Trébuchon, S. Chen, S. Lagarde, R. Carron, M. Gavaret, B. Giusiano, A. McGonigal, F. Bartolomei, J.-M. Badier, and C. Bénar. Deep brain activities can be detected with magnetoencephalography. *Nature Communications*, 10:971, 02 2019.

[19] T. Song, W. Zheng, P. Song, and Z. Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, PP:1–1, 03 2018.

[20] M. Soroush, K. Maghooli, K. Setarehdan, and A. Motie Nasrabadi. A review on eeg signals based emotion recognition. *International Clinical Neuroscience Journal*, 4:118–129, 10 2017.

[21] S. Taulu and J. Simola. Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. *Physics in medicine and biology*, 51:1759–68, 05 2006.

[22] N. Thammasan, A. Uesaka, T. Kimura, K.-i. Fukui, and M. Numao. Feasibility study of magnetoencephalographic inter-subject synchrony during music listening. pages 1–3, 5 2020.

[23] X. Wang, D. Nie, and B.-L. Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 04 2014.

[24] Y. Yang, Q. Wu, Y. Fu, and X. Chen. Continuous convolutional neural network with 3d input for eeg-based emotion recognition. 10 2018.

[25] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen. Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network. pages 1–7, 07 2018.

[26] K. Zhao, M. Liu, J. Gu, F. Mo, X. Fu, and C. H. Liu. The preponderant role of fusiform face area for the facial expression confusion effect: An meg study. *Neuroscience*, 433, 03 2020.