Probing EEG Signals with Neural-Network Classifiers

Ren Li Yifei Song Hao Zhao

Abstract

We consider investigating the effect of gender in word production by classifying Electroencephalography (EEG) brain signals. The EEG signals are analyzed in temporal domain, spatial domain, and frequency domain using various neural-network (NN) models, e.g., Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN). The highest classification accuracy, 94.7%, is obtained in Beta band (14-30Hz). To understand the disparity of brain activities between female and male, we propose a gradient-based method to indicate the brain regions and time periods exhibiting dominant differences in the picture naming task, which helps us reveal dynamic changes of brain in a more accurate and easier fashion compared to the accuracy-based method and traditional analysis techniques.

1. Introduction and Related Works

We consider the problem of investigating the effect of gender in word production. The ability of speaking is one of unique developments of human in evolution. However, speech (or word) production is a complex mental process involving vision perception, memory retrieval, semantic embedding and decoding, etc. Usually, to understand this process, the picture naming task is used, where subjects are required to utter the name of the object shown on the picture. Meanwhile, the subjects' brain signals are recorded by non-invasive neuroimaging techniques, e.g., EEG, magnetoencephalography (MEG), or functional magnetic resonance imaging (fMRI), for further analysis.

It has been reported that the language abilities differ significantly across individuals [2, 5, 22]. Factors, such as gender, age, memory capacity and education, are believed to contribute to these differences. Previous works [3, 4, 11] have researched thoroughly on the age factor for word production, and found obvious differences in topographic patterns of EEG signals from subjects with various ages. However, the effect of gender is less investigated, though it is well known that female and male observe many disparities in brain structure and function and mental behaviours [8, 10, 16, 17]. To fill this gap, we focus on the study of the influence of gender in word production by analyzing EEG signals with the help of the neural-network classifier (as we have witnessed the advances of neural networks in the study of EEG [7, 12, 19, 23]).

Given the measurements of brain activity (i.e., EEG signals) and their corresponding labels (female and male), we ask 2 key questions- (a) *if we train a classifier on the measurements and obtain an accuracy (on test data) far above the chance, can we claim that there are distinctive patterns existing in EEG signals for female and male?*, and (b) *given the well trained classifier, can we use it as a probe to help us diagnose when and where these different patterns can happen?*

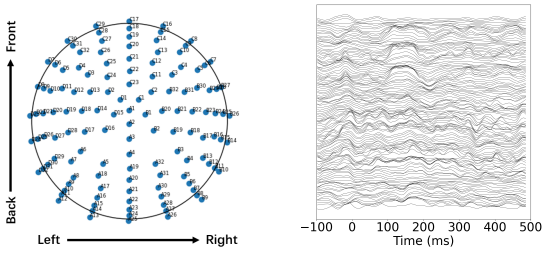
The answer to (a) will be yes under the assumption that the data is enough for training and the classifier is powerful enough to find the underlying patterns. Since the EEG signal is high-dimensional data with strong spatio-temporal correlation essentially, a powerful classifier is required to be able to find the features encoding these properties for EEG. Out of this reason, we consider a CNN model modeling the non-linear and spatio-temporal dynamics of EEG and show its superior performance in the classification accuracy when compared to other models. Next, with this well-trained CNN model, we propose an accuracy-based and a gradient-based methods to assist us in the detection of signal topographic differences between female and male. As will be discussed in Sec. 2.3, we argue that the accuracy and the gradient obtained from the classifier are the strong indicator of the degree of signal dissimilarity, which makes the realization of (b) feasible. Thanks to these methods, we can find significant differences in amplitude and latency existing between female and male, in certain regions (e.g., channel A29/D31) and after 100ms from picture onsite. When compared to the traditional analysis techniques, i.e., event-related potential (ERP) [13] and topographic analysis of variance (tANOVA) [14], our method reveals brain dynamic changes in a more accurate and easier fashion.

2. Method

2.1. EEG Data Collection

The 128-channel EEG signals are recorded when subjects are undertaking the picture naming task. The subjects are 80 right-handed and French native speakers, 40 females

and 40 males. The EEG signals are sampled at 512Hz and band-pass filtered (low cut-off frequency 0.2Hz, high cut-off frequency 30Hz) with a second order acausal Butterworth filter. Epochs of 300 time-frames (TFs), 50 TFs before the picture onset and 250 TFs after, are kept (approximately 600ms). The trails with artifact are discarded, giving us 7256 validate data in total. All the data is z-scored prior to training and classification. More details about the data collection and preprocessing can be found in [3]. The data is split into 5 folds for cross validation (see Appendices A). We show the actual channel positions in Fig. 1(a), where the circle represents the head viewed from the top. In Fig. 1(b), we show an example of the raw 128-channel EEG signal, where time point 0 means the picture onsite.



(a) Channel locations (b) EEG waveform

Figure 1. The illustration of (a) the channel locations on the brain (top view) and (b) the waveform of an EEG sample.

2.2. Classifier

Since EEG signal is high-dimensional data with strong spatio-temporal correlation, we consider various models to explore different characteristics of EEG. We first start with a simple linear model (denoted by LIN), for which the input is a flattened vector of EEG with 128×300 (channel $\times$ TF) elements. To model the non-linearity of EEG, the second model we implement is a multilayer perceptron (MLP) having three fully connected (FC) layers with ReLU and dropout in between. To encode the temporal correlation, we use a model with a GRU layer followed by two FC layers. Inspired by [12], we also deploy a simple yet effective CNN model as shown in Fig. 2 to capture the spatio-temporal dynamics in EEG. The signal is first fed to a 1D CNN layer, which processes each of the 128 channels independently with eight 1D CNNs. This yields an output having 128×8 features per time point. The features at each time point then are processed by an FC layer to encode spatial dynamics into a 40-element code. After the average pooling along the time axis, we have the final feature that will be taken by the final FC layer for classification. ReLU and dropout are added after each layer, except the pooling layer and the last layer. (See Appendices A for more implementation details.) Note the output of our model is a 2-D one-hot vector $y = [y_0, y_1]^T$ instead of a single value for the binary gender classification (theoretically, they will give us the same classification results). The purpose of this design is to facilitate the gradient-based visualization introduced in Sec. 2.3.

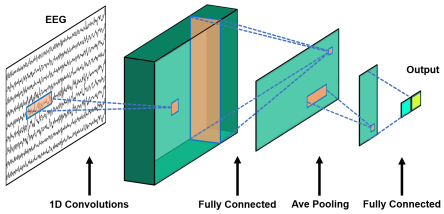


Figure 2. The architecture of our CNN model.

2.3. Topographical Significance Visualization

Typically, ERPs and tANOVA are used to analyze the underlying topographic differences between different groups (classes). ERPs are measured by means of EEG from the same class for each individual channel. People need to check different channels to locate where the components (positive/negative waveform peaks) have significant difference in latency or amplitude, which is tedious and laborious. tANOVA is a non-parametric randomization test measuring the global topographic dissimilarities between classes at each time point independently. However, it cannot reflect the channel-level difference and the correlated fluctuation over time [21]. Given these drawbacks, we propose an accuracy-based and a gradient-based methods to help us detect and visualize the underlying spatio-temporal topographic patterns in EEG.

Accuracy-based Significance. Intuitively, the classification accuracy of a specific classifier can be directly used as the indicator to measure the significance of dissimilarities-the higher the accuracy, the more difference the signals will exhibit in waveform patterns. Therefore, we propose a naive accuracy-based method. To measure the temporal dissimilarities, we use a sliding window with length l to extract the signal segment starting from i^{th} TF. We train our classifier on these segments, and the corresponding test accuracy is recorded as the significance value at $(i + l/2)^{th}$ TF. In this way, we can acquire the temporal change of signal dissimilarities. As for the spatial measurement, we use each single channel to train our classifier. Again, the test accuracy is recorded for each channel as the significance value, indicating the level of difference happening at that channel position. However, as will be discussed in Sec. 3.2, this method suffers from drawbacks like high time cost and inaccuracy. To address these issues, we propose the following gradient-based method that only needs to train the model once while being able to indicate the spatio-temporal significance more accurately and detailedly.

Gradient-based Significance. It has been shown that the model gradient can be applied to interpret model's output [20, 18], which inspires us to use gradient as the significance value. Specifically, the classifier $f(x)$ is first trained

on the entire training set, where x is the input signal. Given a test EEG trail $\tilde{x}$ and its label $l \in \{0, 1\}$, we compute the gradient with respect to $\tilde{x}$ using Guided Backpropogation (GB) [20] [1] as $g_l = \frac{\partial y_l(\tilde{x})}{\partial \tilde{x}}$, where $y_l(\tilde{x})$ is the l^{th} element of the model output. GB backpropogates the gradients having positive influence in the prediction by masking out the negative values with the help of the forward-pass output of ReLU layers in the model. Since the input EEG signals can have both positive and negative values (note a positive value contributes positively via positive gradients while the negative one via negative gradients), the final significance value is obtained by

$$S(\tilde{x}) = max(0, g_l \odot sign(\tilde{x})), \quad (1)$$

where $\odot$ and $sign$ represent the element-wise multiplication and the sign function, respectively. $S(\tilde{x})$ depicts the significance of each input element in the contribution of the final correct prediction. In other words, it shows the degree of the signal dissimilarity at a specific space-time point- the higher the significance value, the higher the possibility of pattern dissimilarity existing. As long as the classifier is well-trained and able to capture the spatio-temporal dynamics of EEG, this gradient-based method can reveal the complex dynamic changes of brain across the spatial and temporal domains in a more accurate and fine-grained level.

3. Experiments

3.1. Gender Classification

We first present the gender classification results for EEG signals. In Table 1, we compare the performance of the models introduced in Sec. 2.2. The accuracy of the non-linear model MLP outperforms that of the linear model LIN by 4.1% when tested on the full-band signal (0.2-30Hz). The accuracy is improved further if we use the temporal model GRU. The best result 90.0% is obtained by the CNN model, which considers both the non-linearity and the spatio-temporal properties of EEG. Note that the parameter number of CNN is comparable to LIN's but much less than MLP's and GRU's. This demonstrates the efficacy of our CNN model. To investigate the model performance on different frequency bands, we filter the EEG data by second-order Butterworth filters to generate signals in Delta (0.2-4Hz), Theta (4-8Hz), Alpha (8-14Hz) and Beta (14-30Hz) bands. It is noteworthy that using the Beta band gives us the highest accuracy for CNN/GRU, while the model performance drops significantly on the Theta band (for CNN, 94.7% in Beta vs 76.5% in Theta). This indicates that the EEG signals of female and male are more distinguishable in the Beta band but less in the Theta band, which can be explained by the fact that beta waves are often associated with active thinking whereas theta waves tend to appear during inactive states [1], e.g., meditation and sleeping. However, LIN/MLP do not show the same tendency- they perform better in low frequencies. This can be explained by their incapability of modeling temporal or spatial information. Because of our CNN model's superior performance, we will use it as the visualization tool for the detection of signal topographic differences between female and male.

Classifier	LIN	MLP	GRU	CNN
Full (0.2-30Hz)	76.0%	80.1%	85.4%	90.0%
Delta (0.2-4Hz)	73.8%	80.8%	81.6%	82.2%
Theta (4-8Hz)	72.4%	73.4%	77.3%	76.5%
Alpha (8-14Hz)	70.9%	73.5%	82.0%	84.4%
Beta (14-30Hz)	67.1%	71.2%	92.9%	94.7%
#Param	38k	493k	263k	46k

Table 1. Classification accuracy for different frequency bands of EEG and the parameter numbers for different models.

3.2. Accuracy-based Significance Visualization

We next evaluate the accuracy-based method for significance visualization. The temporal and spatial significance is shown in Fig. 4. It is indicated that the signals around 200ms and in the red regions should observe obvious difference between female and male, which, however, is not completely consistent with the EEG amplitude distribution (see Appendices C). Besides, a smooth visualization requires retraining the model for each time point and each channel multiple runs, and then averaging the accuracy over these runs to eliminate the influence of noise caused by random seeds. Moreover, for the temporal significance, a higher temporal resolution can be achieved by a smaller step size but will increase the time cost further, and its result is sensitive to the choice of window length (Fig. 1 in Appendices); for the spatial significance, because it is computed independently for each channel, the inter-channel correlation is ignored, which also explains the low accuracy obtained for each channel. Given these issues, the accuracy-based method is not ideal for significance visualization.

3.3. Gradient-based Significance Visualization

We finally evaluate our gradient-based method and show how to reveal the intergroup differences in EEG with it. Fig. 5 shows the significance maps of female and male averaged over test EEG trails and time. It is noticeable that the signals from the center-back region contribute more to the prediction of female, while it is the region of the left-back for male. These regions cover the parietal lobe, the temporal lobe and the occipital lobe of brain, whose functions are highly related to language interpretation, vision perception and memory [6]. This is consistent with the fact that vision, language and memory of participants are highly involved in the picture naming task. So there can be significant signal differences in these regions.

[1] Our model architecture is not directly fitted to Grad-CAM [18], so we use Guided Backpropogation here.

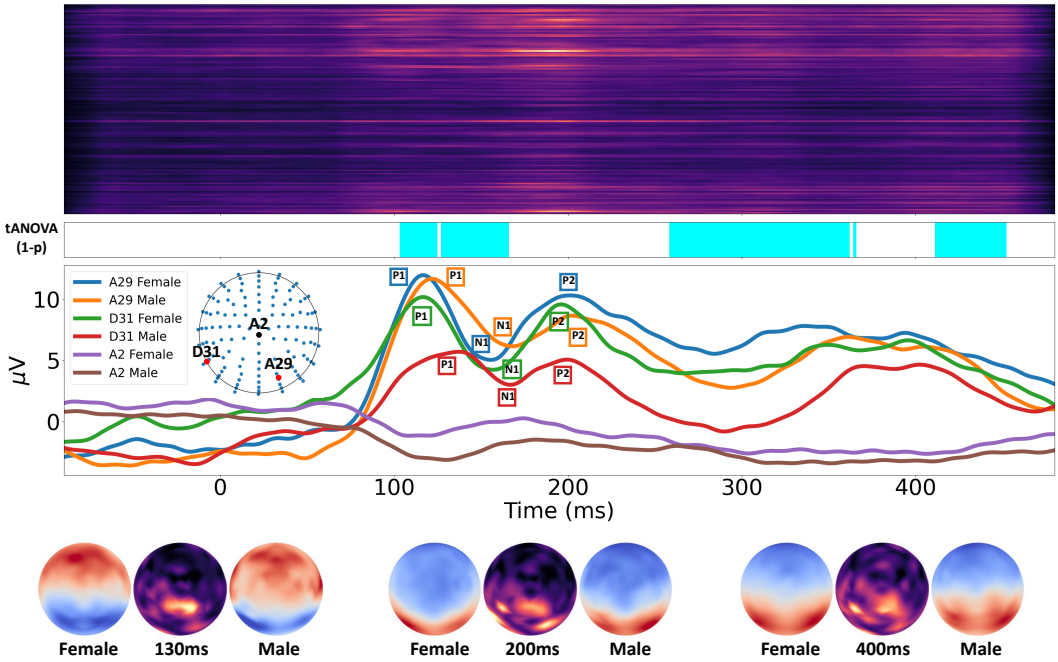
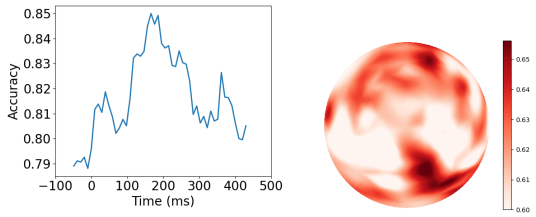
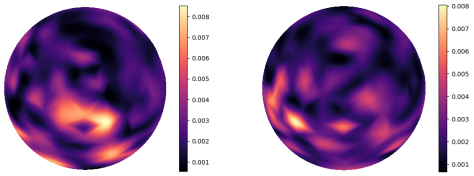


Figure 3. The first row shows the spatio-temporal significance heat map generated by our gradient-based method (bright color means high significance). The second row shows periods of significant differences in the tANOVA analysis ($p < 0.01$, marked in cyan). The third row shows the ERPs of channel A29, D31 and A2 for female and male. The last row shows the comparison of EEG topographic maps (red - positive voltage, blue - negative voltage) and our significance heat maps at different time points.



(a) Temporal (b) Spatial

Figure 4. The illustration of (a) the temporal significance curve (window length 50 TFs, step size 5 TFs) and (b) the spatial significance map.



(a) Female (b) Male

Figure 5. The brain heat maps of female and male averaged over test EEG trails. Bright color means high significance.

The first row in Fig. 3 gives a more detailed illustration of how significance changes along with the time for each channel, where each line represents a single channel (the map is averaged over female and male). Given this heat map, we can easily localize the time periods and channels having significant differences by checking the color brightness. In general, the periods around 100ms and 200ms after the picture onsite show the highest significance. We plot the ERP waveform of two brightest channels (A29 and D31) and one darkest channel (A2) in the third row of Fig. 3 for a close inspection. We can find clear amplitude differences and latency shifts between female and male in A29 and D31- male has lower amplitude and higher latency for P1 and P2 components (1st/2nd positive peaks), and also large latency for the N1 component (the 1st negative peak). However, as implied by our heat map, the waveform of A2 for female and male shows almost no difference. The consistency between ERPs and our significance map indication demonstrates the accuracy and efficacy of our method. On the contrary, tANOVA (the second row of Fig. 3) is unable to have such high spatio-temporal resolution and even fail to highlight the period around 200ms. The last row in Fig. 3 displays EEG topographic maps and our significance maps at different time points. Our maps can directly tell us which regions of the brain we should pay attention to.

4. Conclusion

In this work, we consider to investigate the effect of gender in word production for EEG signals with NN classifiers. By means of experiments on different frequency bands, we prove the efficacy of our CNN model. With this model, we are able to design a gradient-based method to visualize the significance of signal difference across gender, which helps us find the significant dissimilarities in amplitude and latency between female and male from certain channels and time periods.

References

[1] S. M. Abdelfattah, K. E. Merrick, and H. A. Abbass. Theta-beta ratios are prominent eeg features for visual tracking tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1):21–25, 2016.

[2] J. Amunts, J. A. Camilleri, S. B. Eickhoff, S. Heim, and S. Weis. Executive functions predict verbal fluency scores in healthy participants. *Scientific reports*, 10(1):1–11, 2020.

[3] T. Atanasova, R. Fargier, P. Zesiger, and M. Laganaro. Dynamics of word production in the transition from adolescence to adulthood. *Neurobiology of Language*, 2(1):1–21, 2020.

[4] M.-J. Budd, S. Paulmann, C. Barry, and H. Clahsen. Brain potentials during language production in children and adults: An erp study of the english past tense. *Brain and Language*, 127(3):345–355, 2013.

[5] M. Daneman and I. Green. Individual differences in comprehending and producing words in context. *Journal of memory and language*, 25(1):1–18, 1986.

[6] J. D. E. Gabrieli, R. A. Poldrack, and J. E. Desmond. The role of left prefrontal cortex in language and memory. *Proceedings of the National Academy of Sciences*, 95(3):906–913, 1998.

[7] Z. Gao, X. Sun, M. Liu, W. Dang, C. Ma, and G. Chen. Attention-based parallel multiscale convolutional neural network for visual evoked potentials eeg classification. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[8] A. N. Kaczkurkin, A. Raznahan, and T. D. Satterthwaite. Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology*, 44(1):71–85, 2019.

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

[10] M. E. Kret and B. De Gelder. A review on sex differences in processing emotional signals. *Neuropsychologia*, 50(7):1211–1221, 2012.

[11] M. Laganaro, H. Tzieropoulos, U. H. Frauenfelder, and P. Zesiger. Functional and time-course changes in single word production from childhood to adulthood. *NeuroImage*, 111:204–214, 2015.

[12] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind. The perils and pitfalls of block design for eeg classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2020.

[13] S. J. Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[14] M. M. Murray, D. Brunet, and C. M. Michel. Topographic erp analyses: a step-by-step tutorial review. *Brain topography*, 20(4):249–264, 2008.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[16] D. Reilly, D. L. Neumann, and G. Andrews. Gender differences in reading and writing achievement: Evidence from the national assessment of educational progress (naep). *American Psychologist*, 74(4):445, 2019.

[17] S. J. Ritchie, S. R. Cox, X. Shen, M. V. Lombardo, L. M. Reus, C. Alloza, M. A. Harris, H. L. Alderson, S. Hunter, E. Neilson, et al. Sex differences in the adult human brain: evidence from 5216 uk biobank participants. *Cerebral cortex*, 28(8):2959–2975, 2018.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[19] G. Sharma, A. Parashar, and A. M. Joshi. Dephnn: A novel hybrid neural network for electroencephalogram (eeg)-based screening of depression. *Biomedical Signal Processing and Control*, 66:102393, 2021.

[20] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[21] J. Yang, H. Zhu, and X. Tian. Group-level multivariate analysis in easyeeg toolbox: Examining the temporal dynamics using topographic responses. *Frontiers in neuroscience*, 12:468, 2018.

[22] M. J. Yap, D. A. Balota, D. E. Sibley, and R. Ratcliff. Individual differences in visual word recognition: insights from the english lexicon project. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1):53, 2012.

[23] C. Zhang, Y.-K. Kim, and A. Eskandarian. Eeg-inception: an accurate and robust end-to-end neural network for eeg-based motor imagery classification. *Journal of Neural Engineering*, 18(4):046014, 2021.

Appendices

A. Implementation Details

For MLP, the hidden size of its first two layers is 128. For GRU, the hidden size of the GRU layer is 128. The experiment results show a larger hidden size will lead to a less generalized classification performance because of the overfitting effect, and a smaller hidden size more likely towards underfitting. The outputs of the GRU layer at every 30th time point are concatenated as the temporal feature, and are forwarded to the next FC layer, whose input size is 128×10 consequently. This can give us a better classification result than only using the output from the last time point. Dropout is also added between each layer to strengthen the regularization ability, and ReLU is used as the activation function for FC layers. For CNN, the 1D CNN layer has 8 kernels of length 16 and stride 1. The kernel size and the stride for the average pooling layer is 8 and 4, respectively. The dropout probability is set to 0.5 for all models. Cross entropy loss and Adam optimizer [9] are used in the optimization of model parameters. We set the learning rate to 1e-3, the batch size to 128 and the maximum number for the epoch to 100 (we did not see any improvement with more epochs). The implementation is in PyTorch [15].

We randomly select 10% of the data as the test data. The rest data is split into 5 folds for cross validation. For each fold, the test accuracy is obtained using the model having the highest validation accuracy. Then, we average test accuracy over 5 folds to give final results reported in the paper.

B. Classification for Data in the Wild

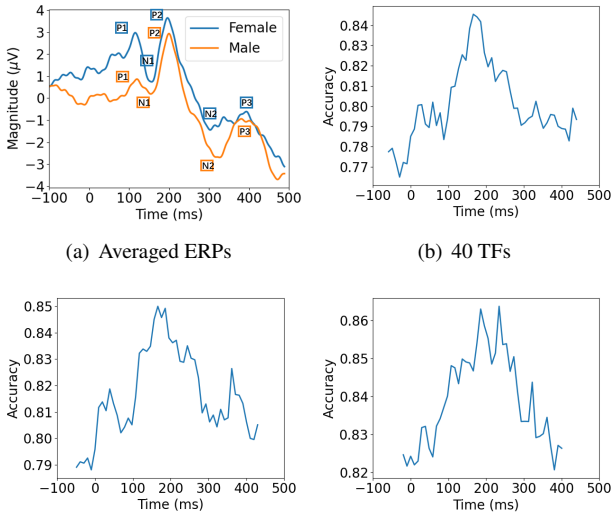
In this section, we want to see how the trained models will perform on EEG data from unseen subjects. That is, we are testing model's generalization capability. Since it is impossible for us to record EEG data from new subjects, we make modifications on the data splits to simulate this cross-subject setting. For the testing, we randomly select 4 subjects (2 females and 2 males) and use all of their EEG data (385 EEG trails) to form the test set (previously the test data is randomly selected from all subjects). The rest subjects will be randomly selected to form the validation set and the training set. We still use 5 folds for cross validation. In this way, it is guaranteed that the test data is 'in the wild' and from unseen subjects. Table 1 reports the classification results with this new test setting. We can notice the huge drop on the classification accuracy when compared to the results of Table 1 in the main paper (for CNN on full band data, 90.0% against 66.3%). However, the accuracy is still far above the chance (50%) - 30% improvement. Unlike nowadays' large-scale training which uses millions of data, we only have 65 subjects for training. Therefore, it is fair for us to argue that our CNN model is able to capture some general underlying patterns of EEG that are different between female and male. We believe the accuracy will improve when training with EEG data from more subjects,

Classifier	LIN	MLP	GRU	CNN
Full (0.2-30Hz)	55.5%	53.5%	62.4%	66.3%
Delta (0.2-4Hz)	50.8%	53.2%	57.3%	55.2%
Theta (4-8Hz)	50.1%	54.7%	55.4%	53.4%
Alpha (8-14Hz)	52.9%	61.2%	63.9%	61.9%
Beta (14-30Hz)	51.7%	56.8%	63.9%	63.1%

Table 1. Classification accuracy for unseen subjects.

C. Issues of the Accuracy-based Method

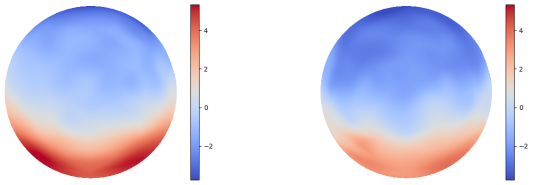
In Fig. 1, we compare the temporal significance curves of various window length with the averaged ERPs. We can notice the peaks around 200ms (which is corresponding to P2 in ERPs) from all curves. According to Fig. 1(a), there are dissimilarities between female and male at P1, N2 and N3, but these positions are not clearly pointed by the temporal significance curves. Besides, the curves generated by different window length are not completely consistent. As for the spatial significance map shown in Fig. 4(b) of the main paper, when compared to the EEG topographic maps as Fig. 2, we can find regions, such as the left-back, are not highlighted as expected. Therefore, the naive accuracy-based method is inaccurate and inappropriate for the significance visualization.



(a) Averaged ERPs (b) 40 TFs

(c) 50 TFs (d) 80 TFs

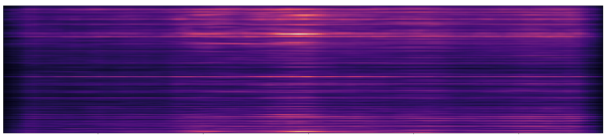
Figure 1. The comparison between (a) ERPs of female and male (averaged over channels) and the classification accuracy curves with different window sizes - (b) 40 TFs, (c) 50 TFs and (d) 80 TFs.



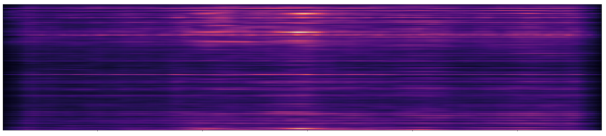
(a) Female (b) Male

Figure 2. The EEG topographic maps (amplitude distribution) of female and male averaged over time.

D. Gradient-based Significance Map



(a) Eq. 1



(b) Eq. 2

Figure 3. The comparison of significance maps generated by different strategies.

The significance value of our gradient-based method is calculated as

$$S(\tilde{x}) = max(0, g_l \odot sign(\tilde{x})), \quad (1)$$

where $\tilde{x}$ and g_l are the input EEG signal and the corresponding gradient respectively. Instead of using the sign of $\tilde{x}$, another way to compute the significance value can be

$$S(\tilde{x}) = max(0, g_l \odot \tilde{x}), \quad (2)$$

which directly uses the value of $\tilde{x}$. But as illustrated in Fig. 3, we observe no obvious difference in the maps generated by Eq. 1 and 2.

E. T-Test

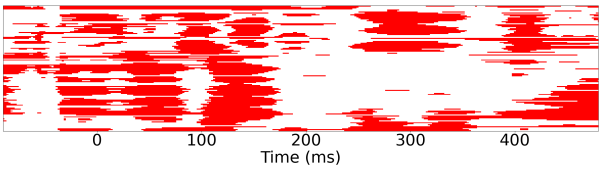


Figure 4. The t-test analysis of EEG at each channel and time point.

T-test or analysis of variance (ANOVAs) is also used for EEG Data analysis. Here we run t-test of each channel and for each time-point on amplitudes of EEG signals. The result is shown in Fig. 4 ($p < 0.001$ marked in red). This analysis only measures amplitude difference at a specific point, so it is not able to reflect the underlying spatio-temporal patterns. As can be seen from Fig. 4, the period before picture onsite (0ms) is even highlighted. Signals in this period usually have different amplitude but similar waveform for female and male (see the third row of Fig. 3 in the main paper), which are just some arbitrary brain states and irrelevant to word production. Consequently, t-test in itself is not accurate enough to show task-related disparity.

F. Behavioral Data Results

Mean (STD)	Female	Male
Accuracy	0.839(0.079)	0.837(0.067)
Reaction Time	943.8(128.6)	962.7(127.0)

Table 2. The mean and standard variance of female and male for production accuracy of the picture-naming task and the reaction time (ms) of cognitive assessments.

Here we report behavioral data results for female and male. In [3], an reaction time test is conducted for the cognitive assessment, where subjects hit different keys according to the content shown on the screen and the corresponding reaction time is recorded. They also record each subject's accuracy in the picture naming task. From Table 2 we can find that female and male have close values of mean and standard deviation (STD) for both accuracy and reaction time. Based on Kruskal-Wallis test, the p-value for accuracy and reaction time is 0.819 and 0.503 respectively, which are far larger than 0.05. This also confirms that the cognitive states of subjects are similar, and consequently eliminates its possibility to be one of the factors causing the differences in EEG.

G. Gender Distribution across Age

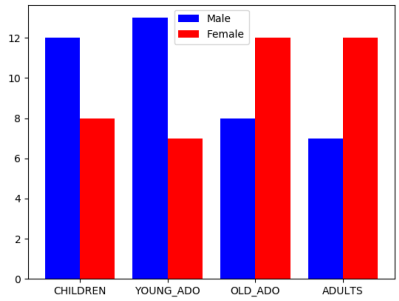


Figure 5. The gender distribution over each age group.

As stated in the introduction from the main paper, subjects' age is an influencing variable for word production. In

Fig. 5, we report the gender distribution for each age group as divided in [3]: children (10-12 years), young adolescents (14-16 years), older adolescents (17-18 years), and adults (20-30 years). We can notice the numbers of female and male in each group are not equal, but the gaps are relatively small. So the age is not likely to be the significant factor that leads to the results we obtained in the main paper.