12-8-2020

Decoding Perception of Speech from Behavioral Responses using Spatio-Temporal CNNs

Kazi Ashraf Moinuddin

DECODING PERCEPTION OF SPEECH FROM BEHAVIORAL RESPONSES USING SPATIO-TEMPORAL CNNs

by

Kazi Ashraf Moinuddin

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

Major: Computer Engineering

The University of Memphis
December 2020

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Mohammed Yeasin for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Faruk Ahmed and Dr. Gavin M. Bidelman. Additionally, I express my gratitude to National Institutes of Health (NIH/NIDCD R01DC016267) for the support of this work.

ABSTRACT

Categorical perception (CP) of speech is a complex process reflecting individuals' ability to perceive sound and is measured using response time (RT). The cognitive processes involved in mapping neural activities to behavioral response are stochastic and further compounded by individuality and variations. This thesis presents a data-driven approach and develops parameter optimized models to understand the relationship between cognitive events and behavioral response (e.g., RT). We introduce convolutional neural networks (CNN) to learn the representation from EEG recordings. In addition, we develop parameter optimized and interpretable models in decoding CP using two representations: 1) spatial-spectral topomaps and 2) evoked response potentials (ERP). We adopt state-of-the-art class discriminative visualization (GradCAM) tools to gain insights (as oppose to the'black box' models) and building interpretable models. In addition, we develop a diverse set of models to account for the stochasticity and individual variations. We adopted weighted saliency scores of all models to quantify the learned representations' effectiveness and utility in decoding CP manifested through behavioral response. Empirical analysis reveals that the γ band and early ($\sim 0 - 200ms$) and late ($\sim 300 - 500ms$) right hemisphere IFG engagement is critical in determining individuals' RT. Our observations are consistent with prior findings, further validating the efficacy of our data-driven approach and optimized interpretable models.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Categorical perception (CP) of speech is a cogntive process of grouping sounds into small phonetic categories [Liberman et al. (1967)]. CP of speech is a complex process reflecting individuals' ability to perceive sound and can be measured using response time (RT). The cognitive processes involved in mapping neural activities to behavioral responses can be decoded through in-depth analysis of neurophysiological recordings such as EEG. Decoding categorical perception (CP) from EEG recordings involves analyzing spatial-spectral-temporal properties that define the underlying cognitive functions [Bashivan et al. (2014); Mahmud et al. (2020a); Bidelman et al. (2019)]. The spatial, spectral, and temporal aspects explain *'where'* in the brain, the type of operation (i.e., memory, attention) and *'when'* in time the neural activities occurs.

While hypothesis-driven analysis is being widely used in decoding CP, but the multivariate approach based on machine learning (ML) algorithms have been gaining momentum. For example, the ML-based approach reported in Bidelman et al. (2019); Mahmud et al. (2020a) show promising results in determining contributing factors in age-related hearing loss. In another work reported in Al-Fahad et al. (2020) used an ML-based apporach to decode functional connectivity patterns in CP. The mentioned studies uses classical ML, such as support vector machines (SVM) [Cortes and Vapnik (1995)] with stability selections [Meinshausen and Bühlmann (2010)] to model cognitive processes involved in CP. The feature selection process provides a limited interpretation of the causal relationship between neural activities and behavioral responses.

This thesis presents a data-driven approach and develops parameter optimized models to understand the relationship between cognitive events and behavioral responses (e.g., RT). We introduce convolutional neural networks (CNN) to learn the relevant features from EEG recordings using two representations: 1) spatial-spectral topomaps and 2) Event Related

Potentials (ERP) to model the spatial-spectral and temporal properties of CP. In addition, we develop a diverse set of deep CNN models to account for the stochasticity and individual variations. We have used bootstrap averaging of trials to generate ERPs in both spatial-spectral and temporal data generation. We utilize bootstrapping process as a data augmentation step to generate a larger number of samples to improve the generalization of CNN models. We use Bayesian hyperparameter optimization algorithm Tree-structured Parzen Estimator (TPE) [Bergstra et al. (2011)] to find best performing spatial-spectral and temporal CNN models, respectively. We have selected ten best performing spatial-spectral and temporal CNNs separately to analyze behavioral responses in relation to CP.

In deep learning (DL), model interpretation is still a challenge as these models contain millions of parameters and therefore are extremely difficult to interpret. Convolution Neural Networks (CNNs) are the only models in the DL arena, where insight into feature importance allocations is possible. The visual interpretation of models are achieved through class discriminative feature visualization techniques like Class Activation Maps [Zhou et al. (2016)], GradCAM [Selvaraju et al. (2017)], CNN-fixation [Mopuri et al. (2019)] and EigenCAM [Muhammad and Yeasin (2020)]. Studies like Jonas et al. (2019); Li et al. (2020); Wang et al. (2020a) shows that GradCAM does capture feature importance allocation by CNNs from data and therefore could be used to infer spatial-spectral-temporal properties underlying a cognitive event. Despite the successes in visual interpretation, it begs the question *"Are class discriminative feature visualizations alone enough to capture patterns dictating cognitive events from EEG data?"* To address this, we propose quantification of learned spatial-spectral-temporal representation from EEG data by CNN models.

We argue that consistent patterns over multiple models could be considered the neural correlates of CP. To this extent, we have proposed the computation of overall saliency score that allows us to find the prevalent spatial-spectral-temporal patterns consistent over multiple CNN models. We have defined two processes to compute overall saliency scores, 1) averaging of saliency scores across models 2) performance weighted averaging of saliency scores across models.

To understand the efficacy of CNN models, we performed mixed model ANOVA analysis on the saliency scores to determine the spatial-spectral-temporal differences in neurological actions that define the RT groups.

We empirically evaluate the CNN models using the CP data obtained from 50 participants. First, we cluster the RTs using Gaussian Mixture Model (GMM). We modeled spatial-spectral-temporal attributes of the neural activities defining three categories of RT (slow, medium, and fast) from EEG data. Employing the proposed process, we observe that early and late engagement in right-hemispheric frontal regions (presumably IFG) is crucial in determining listeners' decision speed. We also find that all three bands (α, β, γ) have active and passive roles while γ band is the most significant in driving listeners' RT. The significance of γ band suggests that auditory CP ability in individuals is the primary predictor of their decision speed. Our findings are coherent with recent and prior studies of brain-behavior function in auditory CP, a validation of our decoding process using CNNs.

The rest of the thesis is organized as follows: in chapter 2, we review existing decoding processes from EEG data using CNNs and the use of machine learning algorithms in decoding auditory CP. Chapter 3 provides a detailed description of our proposed modeling and decoding process, and in chapter 4, we present our modeling and decoding results. Finally, in chapter 5, we discuss our approach's novelty and the findings of the cognitive processing of behavioral responses in categorical speech perception.

CHAPTER 2. REVIEW OF LITERATURE

In this section, we look at some applications of DL models in modeling and decoding neural activities from EEG data. We also review some recent studies where ML algorithms are used to decode spatial-spectral-temporal properties of categorical speech perception.

2.1 Utilization of DL models in EEG studies

Deep learning models have been widely used to model neural processes from EEG data. According to Craik et al. (2019), there are four classification tasks where DL models have been predominantly applied,

- Motor imagery tasks.
  - In motor imagery (MI) tasks, participants are instructed to imagine certain muscle movement on limbs while their neurological activities are captured through EEG [Pfurtscheller and Neuper (2001)]. MI tasks are used in Brain-Computer Interface (BCI) systems to predict users' limb movements. DL models have been effective in modeling MI tasks from EEG data. For example, Kumar et al. (2016); Chiarelli et al. (2018) used deep neural network (DNN) while Amin et al. (2019); Tang et al. (2017); Olivas-Padilla and Chacon-Murguia (2019) used different CNN models to achieve significant results in modeling MI tasks from EEG data. Dai et al. (2019); Rezaeitabar and Halici (2017) combined CNNs with stacked and variational autoencoders to predict limb movements from EEG recordings. There also has been the use of recurrent neural networks such as Wang et al. (2018); Luo et al. (2018) utilized Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks in modeling MI tasks from EEG data.

- Mental workload tasks.
  - Mental workload (MW) or cognitive load tasks involve recording EEG of subjects while they are put through varying degrees of mental task complexity. [Craik et al. (2019)]. Bashivan et al. (2015) was one of the early application of DL in cognitive neuroscience where cognitive load was modeled using recurrent convolution neural network (RCNN) from spatial-spectral-temporal features extracted from EEG. Hajinoroozi et al. (2016) designed channel-wise convolution neural networks (CCNN) and CNN with Restricted Boltzmann Machine (CCNN-R) to model drivers' cognitive state from EEG data. Yin and Zhang (2017) proposed adaptive Stacked Denoising Auto Encoders (SDAE) to classify different levels of MW and showed higher performance in dealing with cross-session EEG features.
- Seizure detection tasks.
  - In studies of seizure detection, EEG recordings are carried out during seizure and seizure-free periods of epileptic patients [Craik et al. (2019)]. Modeling seizure detection tasks allows prediction of upcoming seizures in epileptic patients. Hosseini et al. (2017, 2016) are two such real-time seizure prediction systems using stacked autoencoder and CNN.
- Sleep stage scoring tasks.
  - In sleep stage scoring studies, EEG signals of patients are recorded overnight and into different sleep stages. The application of these studies is to automate the analysis of patients' sleep stages [Craik et al. (2019)]. DL models have been successful in the classification of sleep stages from neuropsychological data. For example, Chambon et al. (2018) designed a CNN that allows end-to-end learning of polysomnography (PSG) signals (EEG, EMG, and EOG) without handcrafted features. DeepSleepNet is another CNN model that allows automatic sleep stage scoring from raw EEG data [Supratak et al. (2017)].

Along with modeling neurological functions from EEG data, some studies have also evaluated learned representations by DL models. As mentioned in section 1, CNNs are the only models in DL domain, where insight into learned representation is possible through visual interpretations. Among the visual interpretation tools of CNN, GradCAM has been widely used for the validation of sensible feature learning by CNNs. Recent studies such as Jonas et al. (2019); Wang et al. (2020b); Chen et al. (2019); Aslan and Akin (2020) applied GradCAM to evaluate the learned features by CNN models from EEG data. Other than GradCAM, Ang et al. (2012) proposed network correlation maps for explaining learned spatial-spectral features by CNNs. Although visual interpretation of CNN models effectively depicts learned representation from EEG data, it is not enough to determine the neurological factors underlying a complex cognitive process such as speech perception. Due to the sensitive and stochastic nature of speech perception, proper quantification of the learned representation by CNN models is required to infer important neurological factors of CP.

2.2 Utilization of ML in Decoding CP

ML algorithms have been gaining popularity in decoding CP from neurological measures like EEG and fMRI. One of the most popular ML technique in decoding neurological functions is multivariate pattern analysis (MVPA) [Haxby et al. (2001)]. MVPA can capture varying brain states from a cortical region and encode different types of information from fMRI data [Haxby (2012)]. MVPA has been used in decoding CP from fMRI data; most notably, Lee et al. (2012) used MVPA to find the cortical regions responsible for speech processing. Arsenault and Buchsbaum (2015) also used MVPA to investigate distributed activation patterns in brain regions while processing phonological features. In another study, Zhang et al. (2015) describes the pattern of activations in subregions of the auditory cortex for sound categories.

Classical ML models also have been successful in decoding CP, especially from EEG data. Recently, Bidelman et al. (2019); Mahmud et al. (2020a) found the contributing factors related to hearing loss in older adults through SVM and [Cortes and Vapnik (1995)] with stability selections [Meinshausen and Bühlmann (2010)]. Al-Fahad et al. (2020) also used a similar framework for decoding individuals' behavioral response from functional connectivity measures. While there has been the application of ML models, there are no DL models applied to decode CP from either EEG or fMRI data.

CHAPTER 3. METHODS AND PROCEDURES

In this section, we will present our procedures of modeling and decoding speech categorization behavior. As stated earlier, we utilize CNNs to model the spatial, spectral, and temporal aspects of behavioral auditory CP. We model the spatial and spectral content together using a spatial-spectral topographic representation of the scalp surface. The temporal contents are modeled separately using ERPs. The CNNs responsible for modeling the spatial-spectral properties are called Spatial Spectral Models (SPSM), and the CNNs attributed to modeling the temporal aspect are called Temporal Models (TM). For decoding, we have used class discriminative feature visualization tools like GradCAM [Selvaraju et al. (2017)] that are used to represent feature importance learned by a trained CNN model. Our implementation of CNN and GradCAM uses Keras [Chollet et al. (2015)] and TensorFlow [Abadi et al. (2015)]. We also use the Bayesian Hyperparameter optimization library hyperopt [Bergstra et al. (2013)] to optimize the hyperparameters of CNNs. All of our implementations are provided in `https://github.com/kmnuddin/stable_hypothesis_selection_eeg_cnn`.

3.1 Data

3.1.1 Participants

The dataset consisted of 50 participants, which we used for modeling the behavioral aspect of CP. All of the participants were recruited from the University of Memphis student body and the Greater Memphis area. The experiment consisted of 15 males and 35 females aging between 18 and 60 years with a mean of ≈ 24 years. Participants were strongly right-handed (mean Edinburgh Hand Score ≈ 80.0), had acquired a collegiate level of education (mean ≈ 17 years), and had a median of 1 year of formal music training. All participants were paid for their time and gave informed consent in compliance with the Institutional Review Board at the University of Memphis. Figure 3.1 (A, B) shows the demographic of the participants.

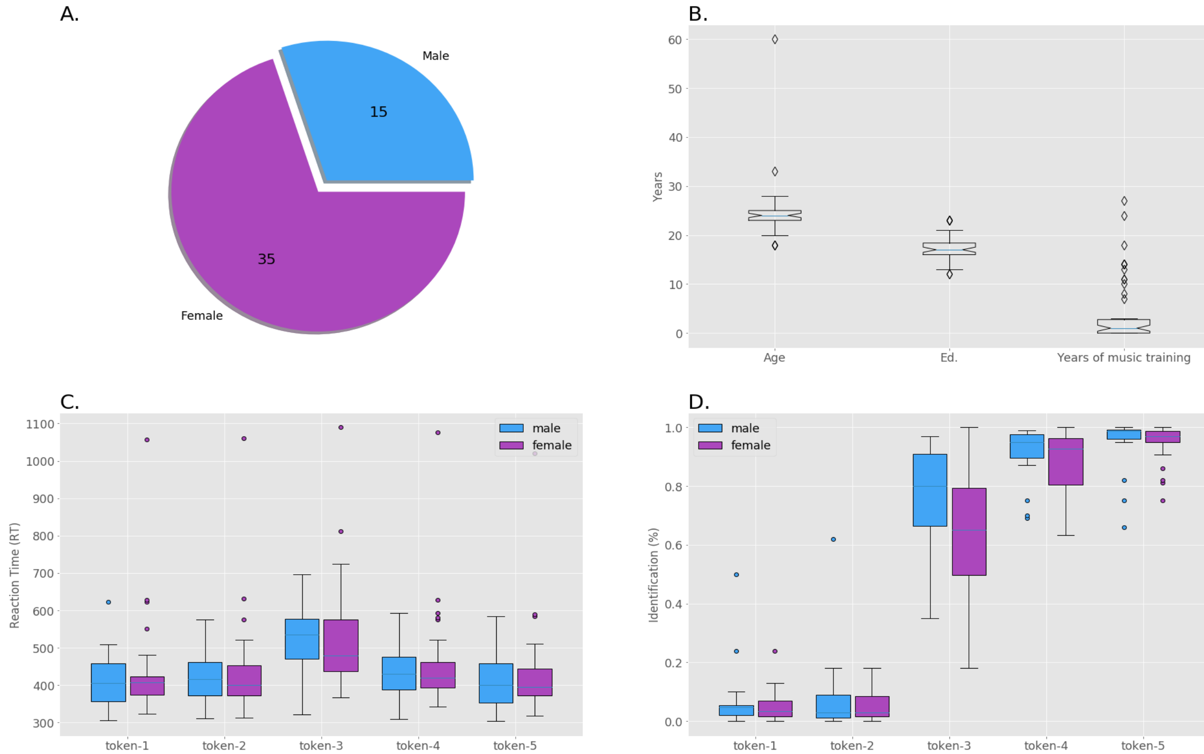

Figure 3.1: Demographics, token-wise identification and RT variation. A) Male-female ratio. B) Meta information of participants. C) RT variation on each tokens, token-3 shows overall large variability in RT. D) Identification rate of each tokens.

3.1.2 Experiment Design

During the experiment, the participants were instructed to listen from a five-step vowel continuum; each token of the continuum was separated by equidistant steps based on first formant frequency (F1) categorically perceived as /u/ to /a/. Tokens were 100 ms long, including 10 ms rise and fall time. The stimuli were delivered through shielded insert earphones; listeners heard 150-200 trials of individual tokens and were asked to label the sound as perceived through binary responses (‘u’ or ‘a’). Response times (RTs) were recorded as the difference between the stimulus onset and the behavioral response (labeling of tokens). Figure 3.1 (C, D) shows token-wise variability of RTs and identification of tokens. Simultaneous EEG recording was carried out using 64 sintered Ag/AgCI electrodes at standard 10-10 locations around the scalp during the trials. As subsequent preprocessing steps, ocular artifacts were corrected using principal component analysis (PCA), filtered (bandpass: 1-100 Hz; notch filter: 60 Hz), epoched (-200 to 800 ms) into single trials, and baseline corrected (-200 ms to 0 ms).

3.1.3 Behavioral Data Analysis

To classify behavioral CP, we opted to form categories within RTs from all the samples using the exact process in Al-Fahad et al. (2020). The idea is to use Gaussian Mixture Model (GMM) with expectation-maximization (EM) to identify the plausible number of clusters from the distribution of RTs. We found four clusters within the distribution of RT using the Bayesian Information Criterion (BIC) as a metric to select the optimal number of components (clusters, ranges from 1-14) and the type of covariance parameter (full, tied, diagonal, and spherical). The procedure concluded with an optimal of four clusters using covariance type 'spherical'. We inferred fast, medium, and slow RTs as the underlying categories based on the centroid and minimum, maximum range of each of these clusters. The fourth cluster was determined to be an outlier due to its low probability and was discarded from further analysis. Figure 3.2 illustrates the optimization of GMM, the RT distribution, the probability of each RT cluster, and the maximum, minimum range of each RT cluster.

3.1.4 Spectral-Spatial Data Generation

As explained earlier, we have opted to use bootstrapping to generate more examples appropriate for modeling using DL tools. We use the process of sampling trials with replacement in individual RT clusters and averaging them to generate ERPs. We sampled and averaged 50 trials at once in each RT cluster and repeated this process 500 times. This process produced 62525 ERPs, converting to power spectral densities (PSDs) and band powers. We compute PSDs focusing on three frequency bands: α (8-15 Hz), β (16-31 Hz), and γ (32–60 Hz). We used the

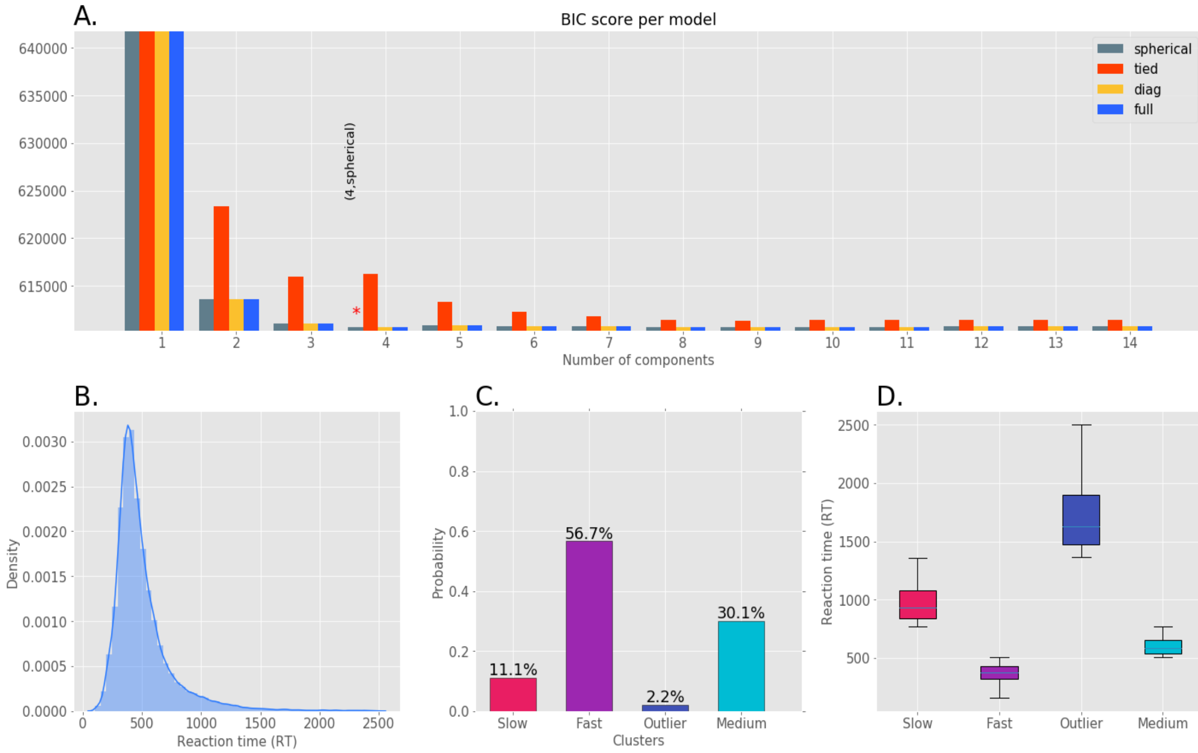

Figure 3.2: Clustering of RT data. A) Optimization of hyperparameters of GMM using Bayesian Information Criterion. B) Original RT distribution. C) The probability of each RT clusters using the optimum GMM. D) The range of each RT clusters

built-in *psd_welch function* provided in the open-source software package MNE-Python [Gramfort et al. (2013)] to compute the PSDs for the three distinct bands. Next, we average across each discrete frequencies within the bands to acquire average band power for each of the 64 channels. The first three steps of Figure 3.3 (A, B, C) depicts the band power calculation from the ERPs. We proceed to project these scaler band powers into a 2d topographical representation of the scalp known as topomap. The scaler band powers associated with each channel get mapped into the location of the channel in the topomap and extrapolated ('box') for crisp visual representation. We generate topomaps for the three-band powers (α, β, γ) individually, convert them to grayscale images, and stack them along the third dimension (RGB color channels) [Bashivan et al. (2015)]. In this way, each of the bands gets represented through different color channels (see Figure 3.3 (D)). We used the *plot_topomap* from MNE-Python to generate the topomaps from the average band powers.

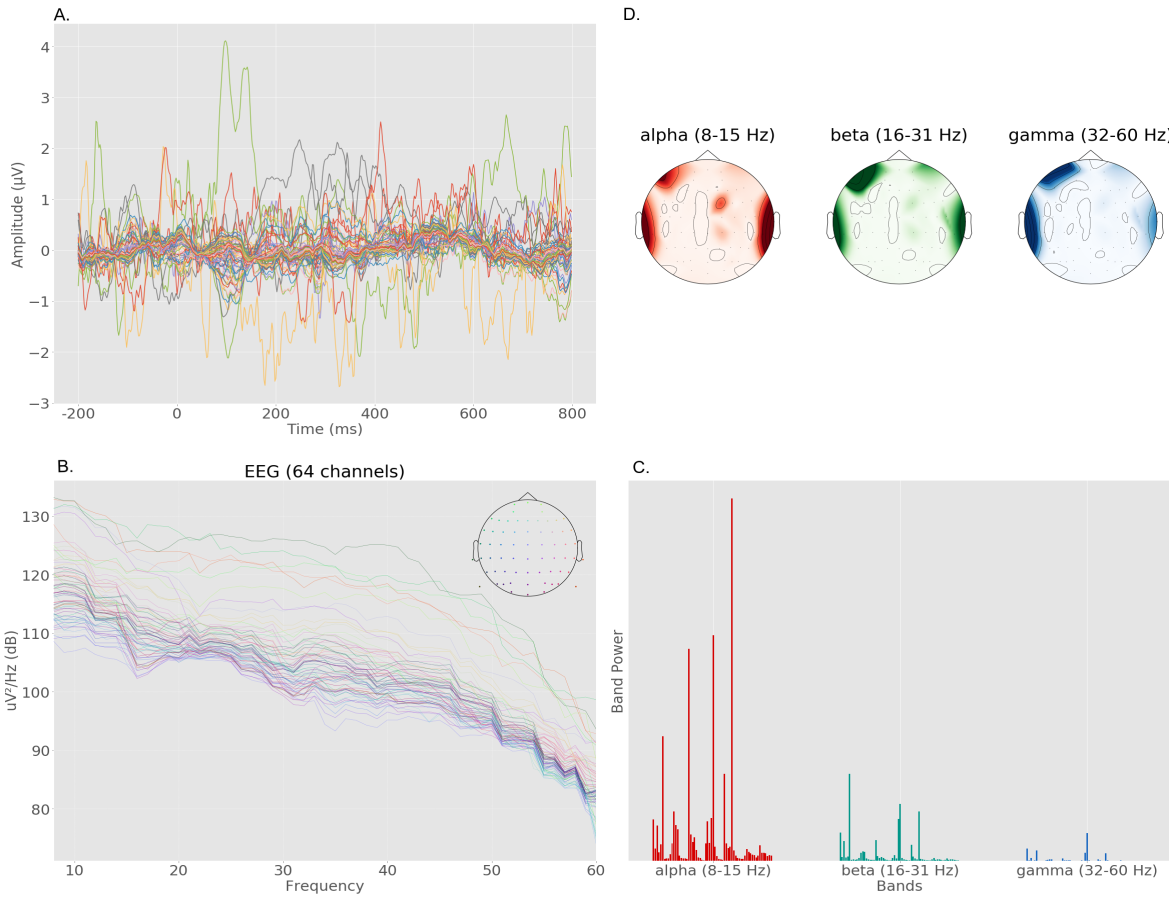

Figure 3.3: A) Sample ERP. B) PSDs of 64 spatial locations C) α, β, γ band powers D) Spatial topomaps of α, β, γ bands

3.1.5 Temporal Data Generation

We use ERPs to model and decode the temporal attributes of behavioral CP. We did not model with single trial EEG due to uneven class samples ($|slow| \approx 4000$, $|med| \approx 25000$, $|fast| \approx 30000$). ERPs have a very good temporal resolution (precision in milliseconds) [Friedman and Johnson Jr. (2000)] and therefore are perfect for answering the 'when in time' question. Usually, ERPs are computed by taking the average of all trials of an arbitrary event. However, since ERPs have a good temporal resolution, we opted to take a bootstrap and average approach. Another reason is that bootstrapping allows for more sample generation which is needed for effective learning through DL models, as explained in section 3.1.4. We did keep the rate of bootstrapping significantly lower than that of the topomap generation process due to the good resolution of ERPs. We take 10 trials with replacement and average to generate one ERP sample on each iteration; 500 iterations of bootstraping and averaging are carried out within each subjects' RT clusters. The bootstrap process produced a dataset of 74004 ERP samples with equal number of samples in each class.

3.2 Modeling using CNN

To model spatial-spectral and temporal properties of speech categorization behavior, we utilized 2D-CNN and 1D-CNN, respectively. In order to acquire the best performing model for each of these tasks, we applied the Bayesian hyperparameter optimization algorithm Tree-structure Parzen Estimator (TPE) [Bergstra et al. (2011)]. TPE is a smart hyperparameter optimization technique that gradually improves the performance of any algorithm iteratively. As mentioned earlier, we have used multiple models for consistent feature selection, so the TPE algorithm is utilized here as a hyperparameter optimizer and also to come up with multiple performing models. In this section, we present the hyperparameters for modeling spatial-spectral and temporal data and the general configuration of modeling.

3.2.1 Spatial-Spectral Modeling

We use topomaps to model the spatial-spectral attributes of the behavioral CP, as mentioned in section 3.1.4. Among the 62525 topomaps generated, we used 46893 (75%) samples for training and 15632 (25%) for testing on each model optimized by the TPE algorithm. We optimize the architecture and the general hyperparameters (e.g., batch size, epochs, learning rate); table 3.1 describes the hyperparameters optimized by TPE for SPSMs. We utilized Adam [Kingma and Ba (2014)], Nadam [Dozat (2016)] and RMSprop as the optimizers (learning algorithm) and ReLU [Hahnloser et al. (2000), Jarrett et al. (2009)] or ELU [Clevert et al. (2016)] as activation functions during TPE optimization of SPSMs. In the convolution layers, each layer contains twice the number of filters than the previous layer. If there are more than four layers, then the number of filters on each layer is iteratively increased with a constant value (the initial number of filters chosen by TPE). The kernel size of filters in convolution layers and residual layers are fixed (3×3) with single strides (1, 1). The pooling size in max-pooling layers after convolution layers is also fixed (2×2) with single strides (1, 1). We ran 35 trials of the TPE optimization of spatial-spectral modeling and chose the top 10 SPSMs (based on test accuracy) among 35 for analysis (see section 3.4 for rationale). Figure 3.4 illustrate the chosen hyperparameters during each trial with associated test accuracy.

Table 3.1: The hyperparameter optimized for SPSMs with TPE

Hyperparameter	**Description**
batch_size	The batch size during training.
epochs	The number of epochs during training.
first_conv	The number of stacked convolution layers in the bottom of the network.
nb_conv_pool_layers	The number of consecutive convolution and max-pool layers.
conv_hiddn_units_mult	The number of filters in the 1st convolution layer ($40 \times mult$).
conv_dropout_drop_proba	The dropout probability of convolution filters.
residual	The number of residual layers, inspired by ResNet [He et al. (2016)].
conv_pool_res_start_idx	The layer to start the residual connections.
fc_units_1_mult	The number of neuron in the 1st fully connected (fc) layer ($750 \times mult$).
fc_dropout_drop_proba	The dropout probability of neurons in the fully connected layers.
one_more_fc	The number of neurons in the 2nd layer of the fc layers ($750 \times mult$).
l2_weight_reg_mult	The $l2$ regularization parameter ($\lambda = 0.0007 \times mult$).
lr_rate_mult	The learning rate parameter ($lr = 10^{-5} \times mult$).
use_BN	The use of batch normalization in convolution layers.
activation	The activation function in the convolution and fc layers.
optimizer	The optimization algorithm.

3.2.2 Temporal Modeling

We model ERPs with 1D-CNN or temporal CNN for analyzing the temporal aspect of auditory CP. Just like our spatial-spectral modeling, we use TPE for hyperparameter optimization. We took 75% of the ERP samples for training and 25% for testing. Most of the hyperparameters regarding the architecture design of 1D-CNN are the same as SPSM, except we limit the search to two types of architectures only. We chose between a vgg [Simonyan and Zisserman (2015)] like architecture as depicted in Jonas et al. (2019) and a normal CNN i.e. convolution in each layer followed by max-pooling (an architecture like that of LeNet [Lecun

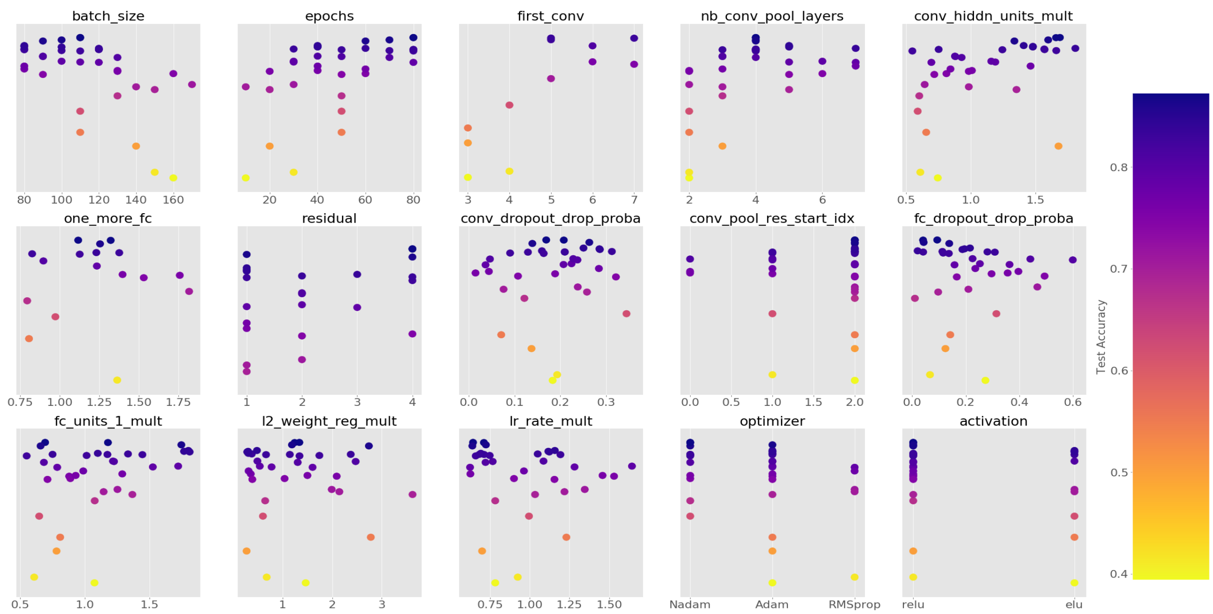

Figure 3.4: Hyperparameters chosen during TPE optimization of SPSMs.

et al. (1998)]). Also, our hyperspace for TPE algorithm is configured to build Recurrent Convolution Neural Network (RCNN) and CNNs with fully connected layers only. We use bidirectional LSTM [Schuster and Paliwal (1997)] in our RCNNs as they have superior ability to capture long term dependency than uni-directional LSTM [Hochreiter and Schmidhuber (1997)]. Similar to our spatial-spectral modeling approach, we kept the kernel size in the convolution layers fixed (kernel size of 3) but chose between a pool size of 2 or 4. We chose the top 10 models based on test accuracy for temporal analysis. Table 3.2 provides the description of the hyperspace for TMs and figure 3.5 illustrates the selected hyperparameters during TPE optimization.

3.3 Decoding using GradCAM

3.3.1 GradCAM and Guided-GradCAM

GradCAM is a visual interpretation tool that depicts a coarse localization map of an image detected by CNN w.r.t a class or label [Selvaraju et al. (2017)]. GradCAM uses gradients of a

Table 3.2: The hyperparameter optimized for TMs with TPE

Hyperparameter	Description
batch_size	The batch size during training.
epochs	The number of epochs during training.
nb_conv_pool_layers	The number of consecutive convolution and max-pool layers.
conv_hiddn_units_mult	The number of filters in the 1st convolution layer ($20 \times mult$).
conv_dropout_drop_proba	The dropout probability of convolution filters.
no_stack_vgg	The number of stacked convolutions in a layer of a vgg type network.
pool_size	Pool size in max-pool layers.
lstm_layer	The number of units in the lstm layer ($200 \times mult$).
lstm_dropout_drop_proba	The dropout probability of lstm units.
fc_layer	The number of neuron in the 1st fc layer ($200 \times mult$).
fc_dropout_drop_proba	The dropout probability of neurons in the fully connected layers.
one_more_fc$_l$ayer	The number of neurons in the 2nd layer of the fc layers ($100 \times mult$).
l2_weight_reg_mult	The $l2$ regularization parameter ($\lambda = 0.0007 \times mult$).
lr_rate_mult	The learning rate parameter ($lr = 10^{-3} \times mult$).
use_BN	The use of batch normalization in convolution layers.
arch_type	The type of architecture to utilize.
optimizer	The optimization algorithm.

class flowing into the final convolution layer to produce such visualizations. Guided-GradCAM is another class discriminative activation map proposed in the same study that combines Guided-Backpropagation [Springenberg et al. (2015)] with GradCAM to produce channel-wise class activation maps.

We now present the mathematical formulation of GradCAM, Guided-Backpropagation, and Guided-GradCAM. For computing GradCAM, $L^c_{GradCAM} \in \mathbf{R}^{u \times v}$ of width u and height v for any class c, the first step is to compute the gradient for class c, y^c, w.r.t activations A^k of a convolutional layer. The gradients are then global-average-pooled over the width i and height j dimensions to acquire the neuron importance weights α^c_k.

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\delta y^c}{\delta A^k_{ij}} \tag{3.1}$$

The next step is to perform a weighted combination of forward activation maps and following it up with a ReLU to acquire $L^c_{GradCAM}$ which is of the same size as the convolutional feature maps.

$$L^c_{GradCAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{3.2}$$

Guided-Backpropagation is an approach derived from the deconvolutional net (deconvnet) introduced by Zeiler and Fergus (2014). The deconvnet performs a forward pass of the network

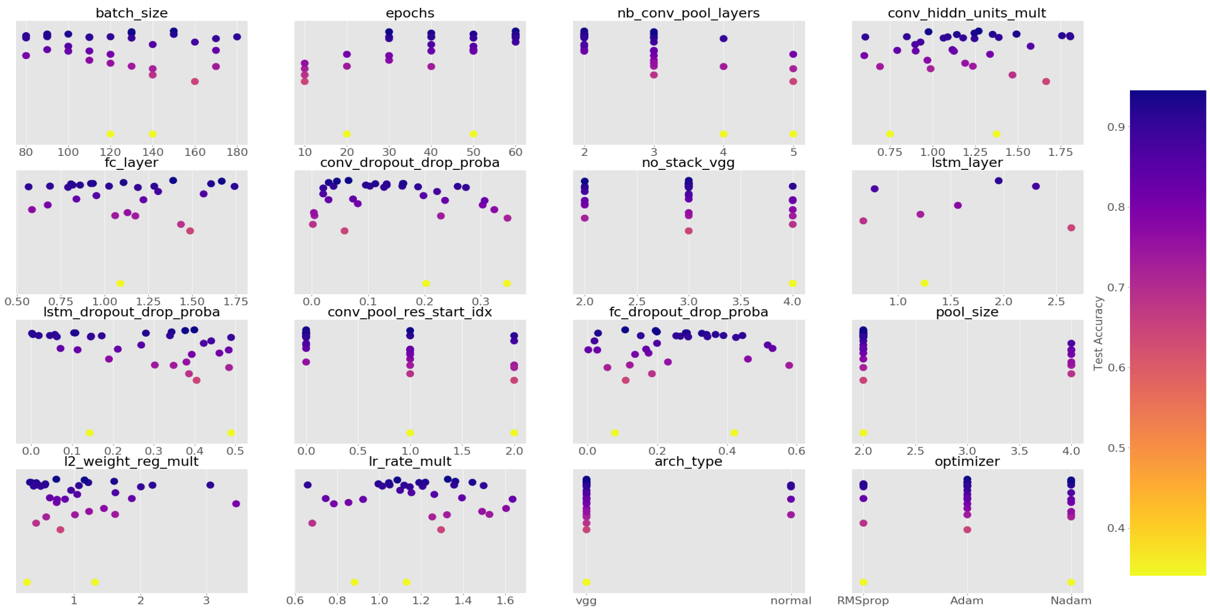

Figure 3.5: Hyperparameters chosen during TPE optimization of TMs.

to compute 'switches,' which are maxima positions within each pooling region. These 'switches' are then used to obtain a distinctive reconstruction. The reconstruction is based on a backward pass after the forward pass, and the backward pass is about going down from a neuron activation in a top layer to the input of the network. The backward pass is put through ReLU in each step to signify that only non-negative gradients flow from top to bottom layers. However, this approach shows noise in the reconstruction mostly because higher layers learn more representative features and cannot be pinpointed to a single input that activates them. Hence in Guided-Backpropagation, a slight adjustment is made by adding a guidance signal from higher layers to lower layers, the guidance signal could be interpreted as the signal that paves a path from top layers to the bottom input, and that path consists of only neurons that are maximally positively influenced by the input.

A forward pass is defined as $f_i^{l+1} = relu(f_i^l) = (f_i^l > 0)$, of any arbitrary layer l in the network travelling from the bottom input to consecutive deeper layers. The backward pass is progression from the top or the deepest layer to the input and is defined as $(R_i^{l+1} > 0).R_i^{l+1}$, for any arbitrary layer l. Guided-Backpropagation uses the forward pass f_i^{l+1} as the guidance signal and $(R_i^{l+1} > 0).R_i^{l+1}$ is the same backward pass of the deconvnet which ensures negative gradients does not flow through the bottom layers.

$$R_i^l = (f_i^l > 0).(R_i^{l+1} > 0).R_i^{l+1} \tag{3.3}$$

It should be noted that Guided-Backpropagation does not produce class discriminative visualization even though it is conditioned on the input. Guided-GradCAM is another visual interpretation of CNN which is suggested in Selvaraju et al. (2017). Guided-GradCAM is a component-wise multiplication of the saliency map from GradCAM and feature map from Guided-Backpropagation, thus it provides insight into saliency among all extracted features. As we mentioned earlier, the saliency map from GradCAM, $L^c_{GradCAM}$ is of the same size as the convolutional feature map, whereas the feature map from Guided-Backpropagation is of the size of the input image. For computing Guided-GradCAM, we need to upsample (by any image interpolation method) the saliency map to match the Guided-Backpropagation feature map. Let $L^{c'}_{GradCAM}$ be the upsampled saliency map of size $m \times n$ which is the same size of the input image, so the class discriminative map from Guided-GradCAM L^{*c} is defined as,

$$L^{c^*} = ReLU(L^{c'}_{GradCAM} \times R_i^c) \tag{3.4}$$

We have opted to clamp any negative values from the GradCAM output since these negative values correspond to features belonging to other classes [Selvaraju et al. (2017)]. Figure 3.6 illustrates some examples of class discriminative maps from GradCAM and Guided-GradCAM.

3.3.2 Spatial-Spectral Decoding

We stated earlier that class discriminative feature visualization tools like GradCAM and Guided-GradCAM do not allow us to reach conclusive inference due to lack of quantification. Therefore, we introduce a kernel-based extraction method that extracts feature importance values from class discriminative maps. Note that in our case, the class discriminative map L^{c^*} from equation 3.4 is a colored image which is represented as a matrix of size $m \times n \times c$, where

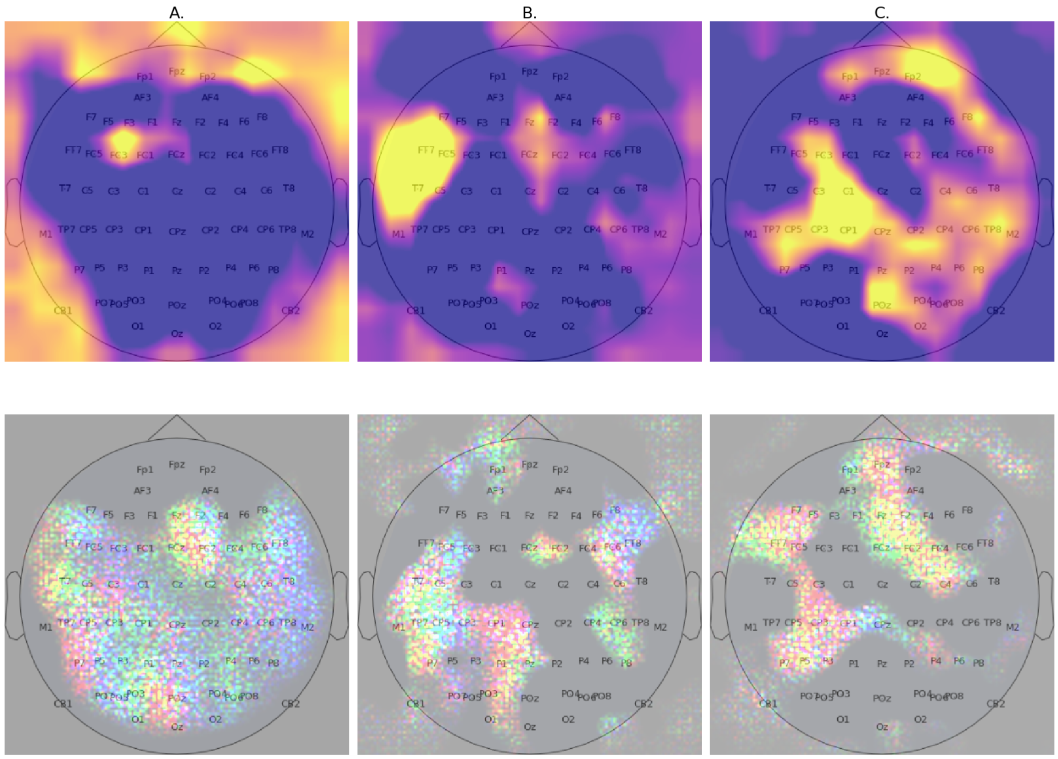

Figure 3.6: Sample GradCAMs and Guided GradCAMs. A) Noisy saliency map from GradCAM but with Guided-GradCAM shows a more noise free feature visualization. B) Comparitively less noisy GradCAM and Guided-GradCAM output. C) A perfect saliency map detecting features in the perpheri of the scalp.

m, n, c are height, width and RGB color channels respectively. The RGB channels correspond to the three distinct frequency bands of α, β, γ, and we have opted to apply the kernel-based extraction method on the band channels separately. Figure 3.7 shows an example of a band-wise class discriminative map. The band wise feature importance let us compute the spectral saliency (see equation 3.6) from CNN models.

Let the positions of the electrodes in the input topomaps of the model be $pos = \{p_1, p_2, p_3, ..., p_n\}$ where $n = |electrodes|$ and $p = (x, y)$, then for each $p \in pos$ we apply a median extractor kernel of size $k \times k$. Figure 3.8 shows an example for median kernel extraction from Guided-GradCAM. In our experiment, we have used a kernel size of $k = 20\,(20 \times 20)$.

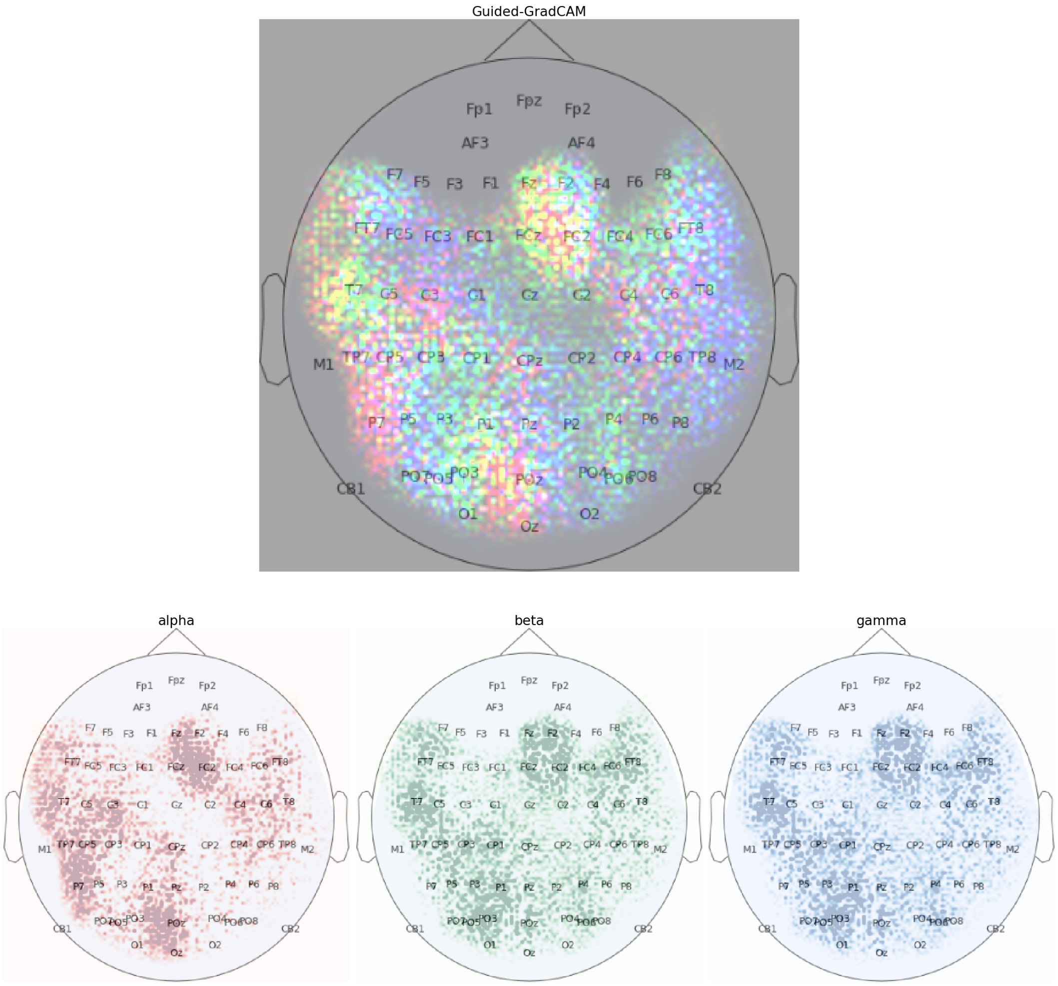

Figure 3.7: Band-wise Guided-GradCAM visualization.

$$S = median_{p \in pos}(L^{c^*}_{p_{k \times k}}) \tag{3.5}$$

S is a vector of size $n \times |bands|$ (in our case 64×3) and is called the Spatial-Spectral Saliency Score that contains the band wise spatial feature importance. The spatial saliency score S_e is the maximum n values of S across bands. To acquire spectral saliency score $S_{\mathbf{f}}$, we average S across each frequency bands.

$$S_{\mathbf{f}} = \frac{\sum_{e \in electrodes} S_{e,\mathbf{f}}}{n}, \quad \mathbf{f} \in \{\alpha, \beta, \gamma\} \tag{3.6}$$

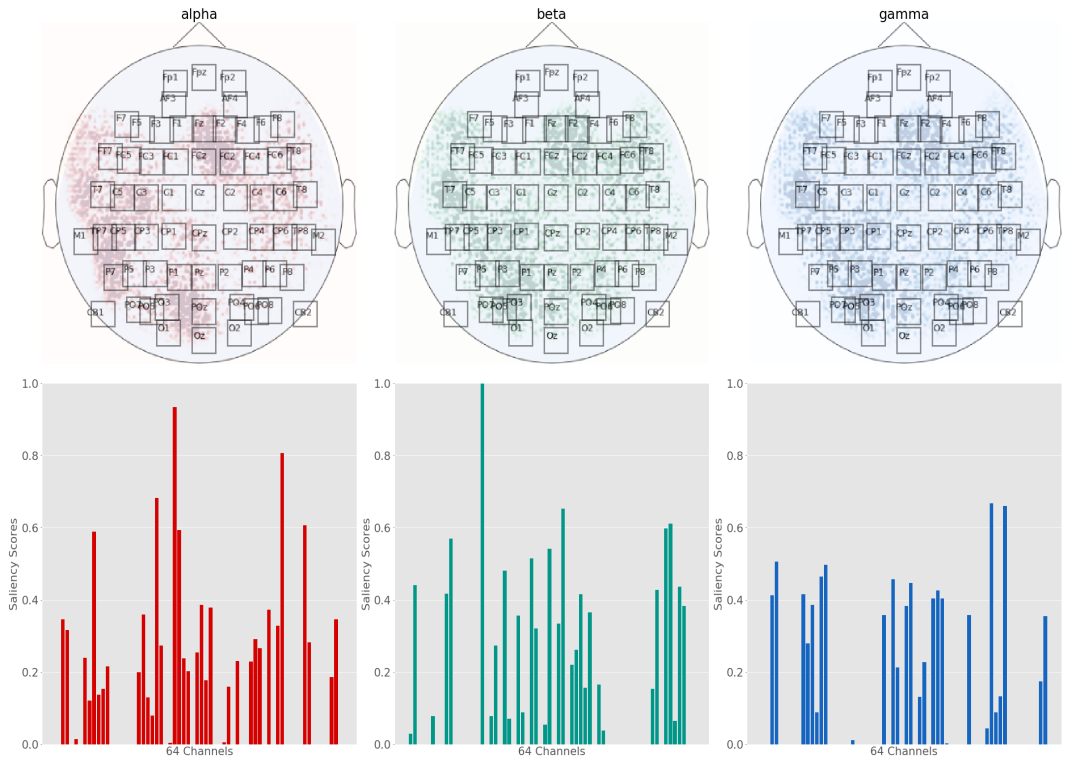

Figure 3.8: Extraction of band-wise selected spatial features from Guided-GradCAM.

3.3.3 Temporal Decoding

Although we used 1D-CNN for temporal modeling, there is not much difference in the application of GradCAM except we work with convolutional feature maps, which are one dimensional. Naturally, the convolutional feature maps are not of the same size as the input; thus, we upsample the class discriminative maps from GradCAM to the original input size of $n \times |timesteps|$, where $n = |electrodes|$. Note we increase the height of the saliency maps to match $|electrodes|$ for visual inspection only. Since we are using 1D-CNN, the variability in the saliency maps is across the temporal dimension only [Jonas et al. (2019)], thus during the upsampling, the values are copied across the *electrodes*. Figure 3.9 shows the average saliency maps from the top ten temporal models and the overall saliency map of these ten models.

We use a similar kernel base extraction procedure for the temporal saliency maps to acquire saliency values for quantative analysis. Let L_t^c be the upsampled saliency map from GradCAM

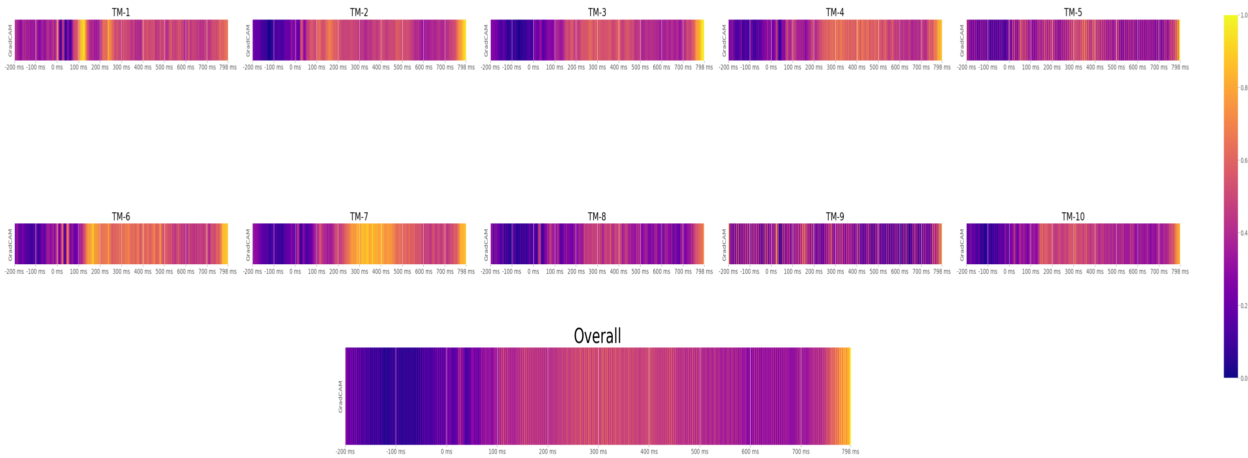

Figure 3.9: Individual and overall saliency maps of 10 best performing TMs.

belonging to class c, we then use a sliding median kernel extractor of size τ across the temporal dimension to obtain the Temporal Saliency Score $S_{\mathbf{t}}$,

$$S_{\mathbf{t}} = median_{\tau}(W(t-\tau)(L_t^c)),\ where\ t = \{t_0, t_1, t_2, t_3, ..., t_N\}\ and\ t_N = t_0 + N\Delta t \qquad (3.7)$$

Δt represents the interval in which the EEG signals are epoched. EEG signals in our data are epoched in 2 ms ($\Delta t = 2\,ms$) interval, so the kernel size τ is equivalent to two times the original timesteps. For example, we use $\tau = 50$ in our extraction step, which means taking the median values of 100 ms durations in the epoched EEG signal.

3.4 Overall Saliency Score

So far, we have illustrated the quantification process of learned representation from spatial-spectral and temporal modeling through saliency scores. To get insight of the most prevalent spatial-spectral-temporal features selected by these models we have computed the unweighted and weighted mean of the saliency scores of the 10 best SPSMs and TMs. If $S_M = \{S_{m_1}, S_{m_2}, S_{m_3}, ..., S_{m_{\mathbf{N}}}\}$ is a set of saliency scores extraced from $\mathbf{N}$ corresponding models, $M = \{m_1, m_2, m_3, ..., m_{\mathbf{N}}\}$, then the overall saliency score S^* is a weighted average of all the candidates in S_M,

$$S^* = \frac{P_m \sum_{m \in M} \mathbf{U}(S_m)}{\mathbf{N}} \tag{3.8}$$

P_m is a scaler metric that characterizes the performance of model m and $\mathbf{U}$ is a min-max normalization function. We use min-max normalization to ensure that saliency scores from all the models are within the same range (0 to 1). The performance metric P_m ensures that more weight is given to the models with better performance (see section 5.2 for rationale). We set $P_m = 1$ when computing overall saliency scores through unweighted averaging.

CHAPTER 4. RESULTS

The current section elaborates on the results of all the experiments conducted in this study. We have used ten best performing SPSMs and TMs for analysis. First, we present the modeling performance in classifying the three RT groups. Then we present a learned representation of individual models and the consistent features selected among them.

4.1 Modeling

The hyperparameter optimization for both temporal and spatial-spectral models are run for 35 trials. Figure 4.1 illustrates the test accuracy of SPSMs and TMs during the trials. The TPE algorithm iteratively chooses hyperparameters that gradually improves the modeling of some arbitrary function. Among the 35 SPSMs and TMs, the mean test accuracy was 75.52 and 82.66, respectively. The top 10 SPSMs has a range of test accuracy from ≈ 83% to 87%, while the range for the top 10 TMs is from ≈ 91% to 95%. Table 4.1 shows the performance of the top 10 SPSMs and TMs respectively.

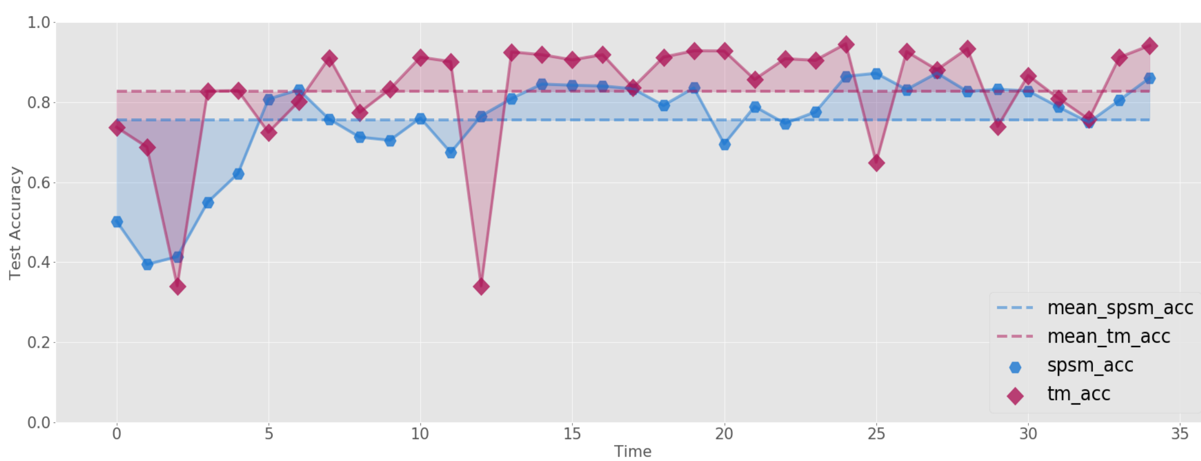

Figure 4.1: Test accuracy during TPE optimization of SPSMs and TMs

Figure A.3 and A.6 shows the overall and individual confusion matrices from the top 10 SPSMs and TMs respectively. The overall confusion matrices suggest effective learning of spatial-spectral and temporal patterns from EEG data in identifying the RT groups. The overall better performance in TMs than SPSMs is due to the excellent temporal resolution of ERPs.

Table 4.1: Performance metrics for top 10 SPSMs and TMs respectively.

(a) Performance of SPSMs

Model	Precision	Recall	F1 Score	AUC	Accuracy
SPSM-1	82.96%	84.80%	83.59%	95.86%	83.22%
SPSM-2	84.25%	84.22%	84.05%	95.86%	83.35%
SPSM-3	84.23%	84.25%	84.16%	95.57%	83.58%
SPSM-4	84.97%	84.46%	84.69%	95.60%	84.05%
SPSM-5	84.43%	85.29%	84.83%	95.87%	84.25%
SPSM-6	84.90%	85.24%	85.01%	95.92%	84.53%
SPSM-7	86.21%	87.10%	86.60%	96.76%	86.09%
SPSM-8	87.12%	87.02%	87.07%	96.79%	86.46%
SPSM-9	87.54%	88.03%	87.75%	97.16%	87.24%
SPSM-10	87.70%	87.95%	87.79%	97.07%	87.28%

(b) Performance of TMs

Model	Precision	Recall	F1 Score	AUC	Accuracy
TM-1	91.32%	91.33%	91.31%	98.30%	91.27%
TM-2	91.89%	91.95%	91.91%	98.45%	91.88%
TM-3	91.90%	91.93%	91.91%	98.54%	91.87%
TM-4	91.89%	91.33%	91.31%	98.30%	92.58%
TM-5	92.68%	92.71%	92.70%	98.75%	92.65%
TM-6	92.84%	92.88%	92.86%	98.82%	92.81%
TM-7	92.87%	92.91%	92.89%	98.76%	92.85%
TM-8	93.41%	93.45%	93.42%	98.98%	93.39%
TM-9	94.18%	94.21%	94.20%	99.22%	94.16%
TM-10	94.62%	94.65%	94.63%	99.23%	94.59%

4.2 Decoding

In this section, we present individual and overall learned representations across SPSMs and TMs through saliency score. The spatial, spectral, and temporal saliency score (denoted by $S_e, S_{\mathbf{f}}, S_{\mathbf{t}}$ respectively) quantifies the features selected by the models on each of these aspects. To observe the consistent learned representation across models, we have computed the overall saliency score through weighted-averaging of saliency scores of all the models (see equation 3.8). Figure 4.2, 4.3, 4.4 illustrates the spatial, spectral and temporal feature importance given by each of the respective models as well as consistent feature detected across them. The spectral and temporal difference between RT groups is inferred through pairwise Tukey HSD test and mixed-model ANOVA analysis on the respective overall saliency scores. By comparing RT groups within each band and timesteps using these tests, we were able to observe *'how'* and *'when'* the neural activities varies in dictating individuals RT.

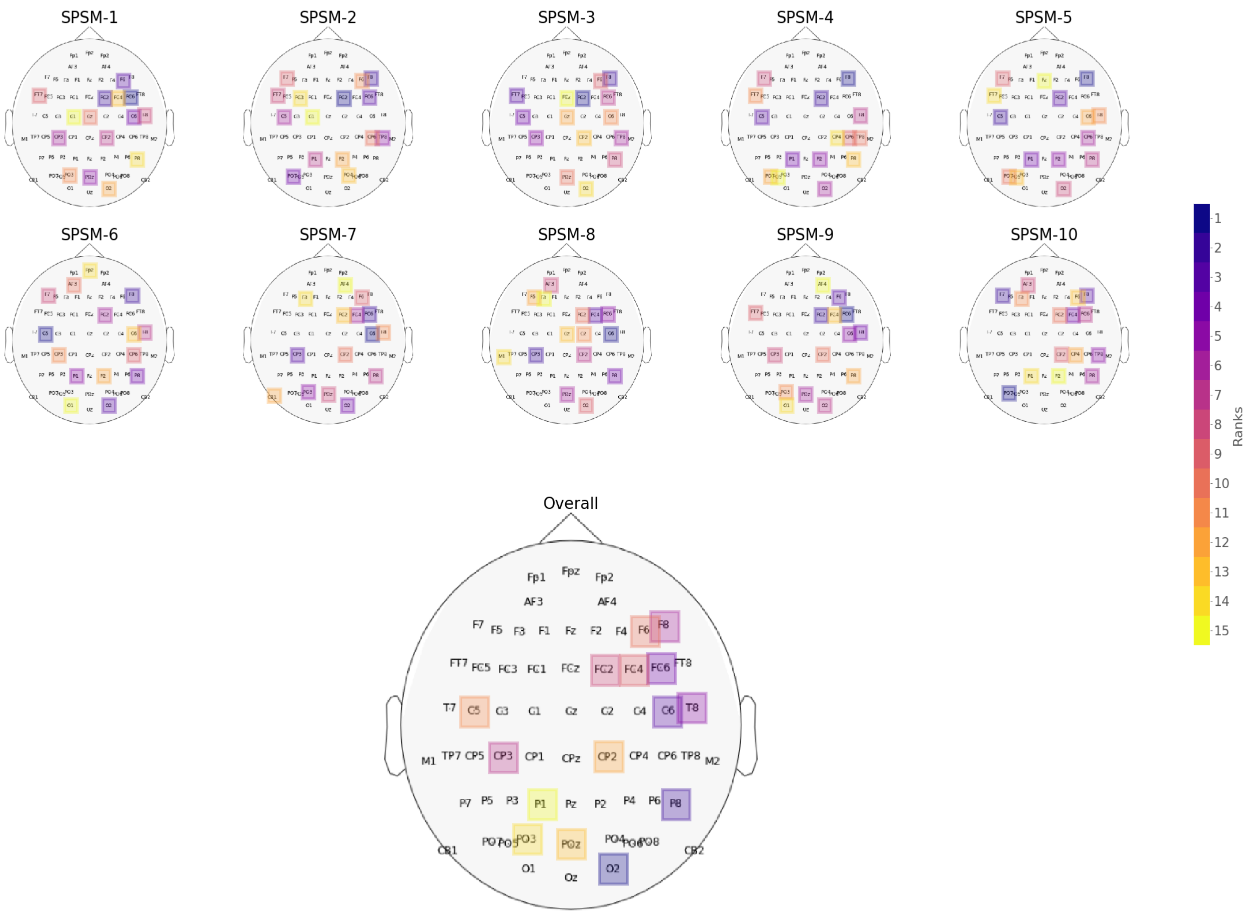

Figure 4.2: Individual and overall top 15 salient spatial features identified by 10 best performing SPSMs.

Figure 4.2 show the individual and overall top 15 ranked spatial features based on the spatial saliency score of the top 10 SPSMs. The overall spatial saliency suggests (denoted by 'Overall', in figure 4.2) that frontal regions in the right hemisphere (RH) are consistent in differentiating between the RT groups. However, activation in the left hemisphere (LH) shows variability across SPSMs.

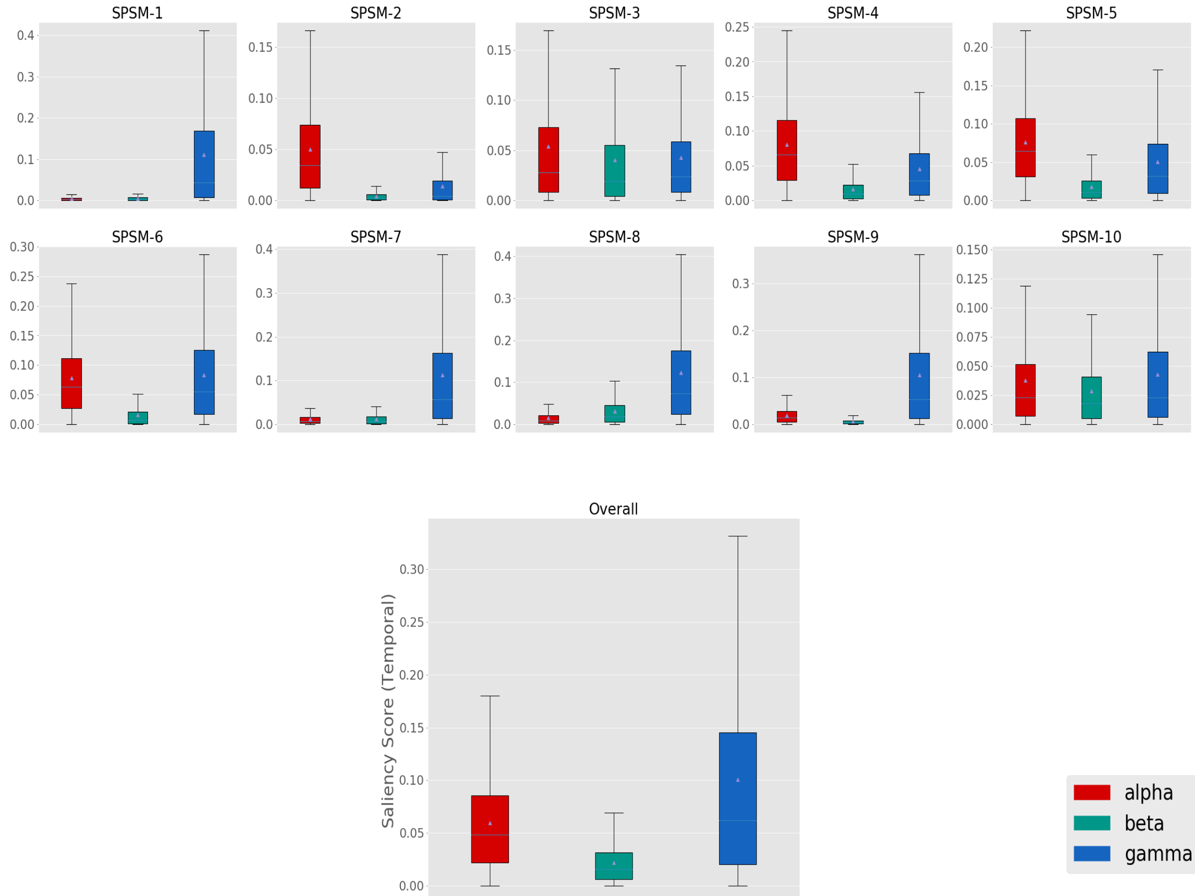

Figure 4.3: Comparison of individual and overall spectral saliency score across samples in 10 best SPSMs.

Figure 4.3 illustrates the overall and individual band saliency variation across samples as modeled by top 10 SPSMs. Primary observation suggests that the γ band is the most prominent in determining speech categorization behavior, although some SPSMs suggest that the α band is the most salient. But through overall spectral saliency score we see that γ band is associated with the highest score ($S_\alpha = 0.015, S_\beta = 0.006, S_\gamma = 0.026$). It is also clear from the analysis of spectral saliency scores that different models learn different spectral patterns.

The temporal saliency of the top 10 TMs is shown in figure 4.4 as the mean of temporal saliency scores ($S_{\mathbf{t}}$) across samples on each RT groups. The pairwise Tukey HSD test on each

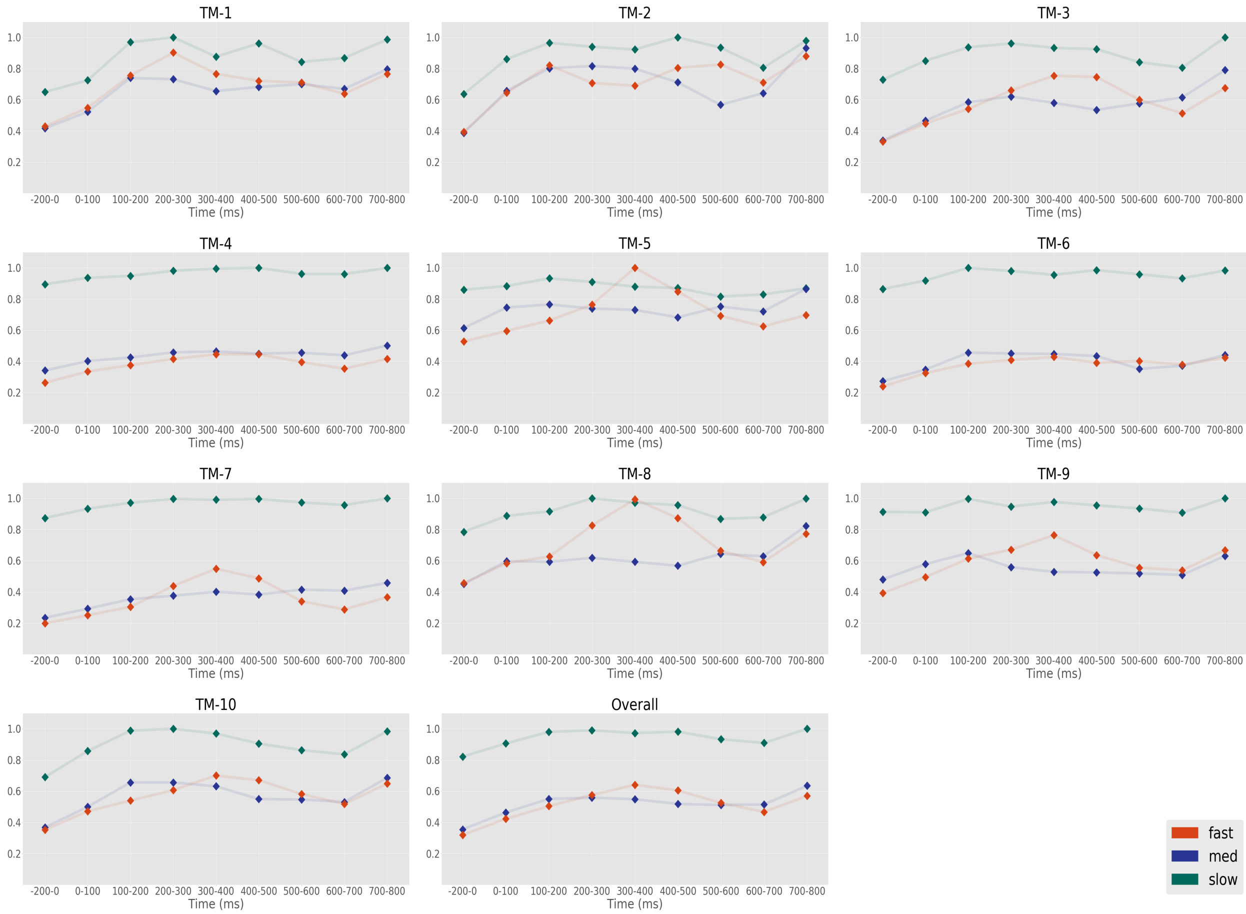

Figure 4.4: Individual and overall comparison between mean temporal saliency score of RT groups across 10 best TMs.

timestep within the RT groups using overall saliency score reveals that 0-200 ms, 300-500 ms, and 600-800 ms are the duration where the RT groups are most distinguishable.

CHAPTER 5. DISCUSSION AND CONCLUSION

We present further arguments for our decoding process using CNN, GradCAM, and saliency scores in the current section. We also look at our finding of speech categorization behavior with contemporary studies as a final validation step of our process.

5.1 Novelty of Our Decoding Process

In our study, we have demonstrated a novel approach to decode neural functionalities from EEG studies. Our proposed approach is an entirely data-driven procedure without the effects of hand-engineered features or prior assumptions. To our best of knowledge, this is the first EEG decoding framework using CNNs that allows insight into spatial, spectral, and temporal properties of a neurological process. This study has shown that the scoring of learned features by CNN models from EEG data allows us to reveal spatial-spectral-temporal patterns of complex cognitive processes like categorical speech perception. Our decoding using different CNN models shows that variable patterns are learned by different models even when there is little to no significant changes in their performance. As stated earlier, this aspect is a direct attribution to the stochastic nature of speech perception behavior and requires consideration of multiple conjectures. Through the overall saliency score, we have acquired the most consistent features learned across the model and can be considered as a unification of multiple conjectures.

5.2 Speech Categorization Behavior

Decoding response time (RT) in speech categorization reveals perceptual differences that drive speech identification ability among individuals [Al-Fahad et al. (2020)]. Auditory categorization in the human brain is revealed to use a distributed frontal-temporal-parietal network by contemporary EEG studies [Bidelman and Walker (2019); Bidelman and Lee (2015a);

Al-Fahad et al. (2020)]. The canonical language processing is left hemisphere (LH) predominantly. However, through the consensus of the best performing SPSMs that right hemisphere (RH) engagement is responsible for decoding RT of categorical speech processing. Especially, frontal regions in RH (F8, F6, FC2, FC4, FC6) are significant in mapping speech to the behavioral response. Hampshire et al. (2009, 2010) found through fMRI experiments that right inferior frontal gyrus (IFG) activation is responsible for attentional control and detection of task-relevant cues. Our results through overall saliency scores also suggest similar findings as the fast and medium RT groups show more importance in the F6, F8, FC6, FC8 spatial locations (presumably IFG) implying more attentional power in speech categorization decision (see figure 5.1). In terms of perceptual encoding of speech, we also find our spatial results to be coherent as Bidelman and Howell (2016) found that audio stimuli of lower SNR cause increased engagement of primary auditory cortex (PAC) and IFG in RH. Participants in our experiment predominantly reacted faster when given clear tokens (TK 1, 2, 4, 5) than the ambiguous one (TK. 3) (see figure 3.1), which explains the functional lateralization of RH. In the case of slower RT, we find more distributed region activations. Specifically, specfically we see a lesser activation in the frontal region (presumably IFG) in RH, which suggests lack of attentional control is responsible for driving slower RT. Al-Fahad et al. (2020) found in decoding RT from functional connectivity measures that activities outside the CP hub are the leading cause for slower RTs. We also find a similar pattern in our inference through overall spatial saliency as fast and medium RTs show a clear frontal-temporal-parietal (F5, F7, M1, P1, PO3, PO7) activation in LH. In contrast, the slower RT groups show no significant activations in LH frontal and temporal regions.

We assess through pairwise Tukey HSD test on the overall spectral saliency scores that α and γ band distinguishes between the fast-med ($p < .0001$) and fast-slow ($p < .0001$) group while β band is solely capable of characterizing the difference between med-slow ($p = 0.0461$) RT groups (see table 5.1). These findings corroborate different theories about neurological processes in association with auditory CP. Our study shows that γ band is more predictive of participants decision time as it acquire the the highest overall spectral saliency score. This is coherent with

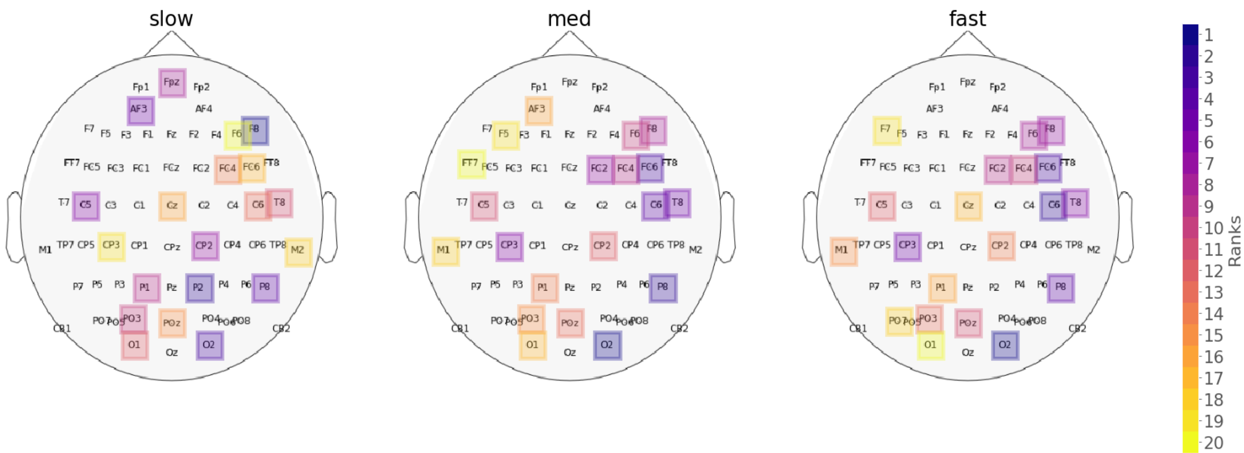

Figure 5.1: Top 20 ranked spatial features for each RT groups by overall spatial saliency score.

the recent study of Mahmud et al. (2020b) suggesting γ band modulations are more correlated with listeners' behavioral CP. So, we can hypothesize that auditory object construction [Tallon-Baudry and Bertrand (1999)] and local network synchronization [Giraud and Poeppel (2012); Haenschel et al. (2000); Si et al. (2017)] is crucial in determining listeners' RT as γ is found to be responsible for these tasks. Our result also suggests that β band is associated with large difference in RTs (fast-slow) of listeners. We conclude that listeners' speech identification capacity [Bidelman and Lee (2015b)] and representational memory [Bashivan et al. (2014)] also plays a pivotal role in dictating the extreme ends of behavioral responses. The effect of β band in the difference of medium and slow RTs is limited in our results. We assume the β band is only significant in late medium and early slow RT ranges ($\approx 700 - 1000ms$). Our assumption is based on the comparatively insignificant effect of β band ($p = 0.0461$) on the distinction between these RT groups. Nevertheless, we conclude that the effect of β band on this matter either could be related to motor-related activity and uncertainty in decision tasks [Senkowski et al. (2005); Tzagarakis et al. (2015)] or reflection of weak hearing capacity as Price et al. (2019) found top-down β connectivity increases for impoverished auditory inputs with minimal behavioral changes. The findings in Bidelman (2017) support the role of α band in discriminating fast-med and fast-slow RT groups where early evoked α oscillations were found to be fundamental in distinguishing behavioral responses between trained and untrained listeners (i.e., musicians vs. non-musicians). So, the effect of the α band in our data might reflect listeners' attentional control capacity dictated by their musical training experience.

Table 5.1: Significance of α, β, γ band in distinguishing RT groups.

RT Groups	**p-value**
α	
fast - med	$< .0001$
fast - slow	$< .0001$
med - slow	0.2638
β	
fast - med	0.1135
fast - slow	$< .0001$
med - slow	0.0461
γ	
fast - med	$< .0001$
fast - slow	$< .0001$
med - slow	0.2816

The general consensus of the best TMs through overall temporal saliency scores ($S_\mathbf{t}$) shows that 0-200 ms, 300-500 ms and 600-800 ms (figure 5.2, 4.4) are the duration where the RT groups are most distinguished. Recent study from Carter and Bidelman (2020) found that early ($\sim 250\,ms$) and late ($\sim 450\,ms$) engagement of right IFG during categorical processing. As we have found the role of the right IFG to be important, we presume 0-200 ms and 300-500 ms are the duration when the engagement of the right IFG is occurring. We can also conclude that since 0-200 ms is still during the "encoding" of the sound, the saliency of speech representation itself at an early perceptual level drives the later decision speed. The rearing end duration ($\sim 600 - 800\,ms$) found in our analysis might be related to the uncertainty in motor-related activity.

In summary, our results indicate early ($\sim 0 - 200\,ms$) and late ($\sim 300 - 500\,ms$) engagement in the right hemisphere (presumably PAC and IFG) are the primary indicator of individuals behavioral response. While all three bands have active and passive roles, γ band modulations are the main predictor of listeners' behavioral response. This indicates that auditory CP ability in listeners' dictates their RT.

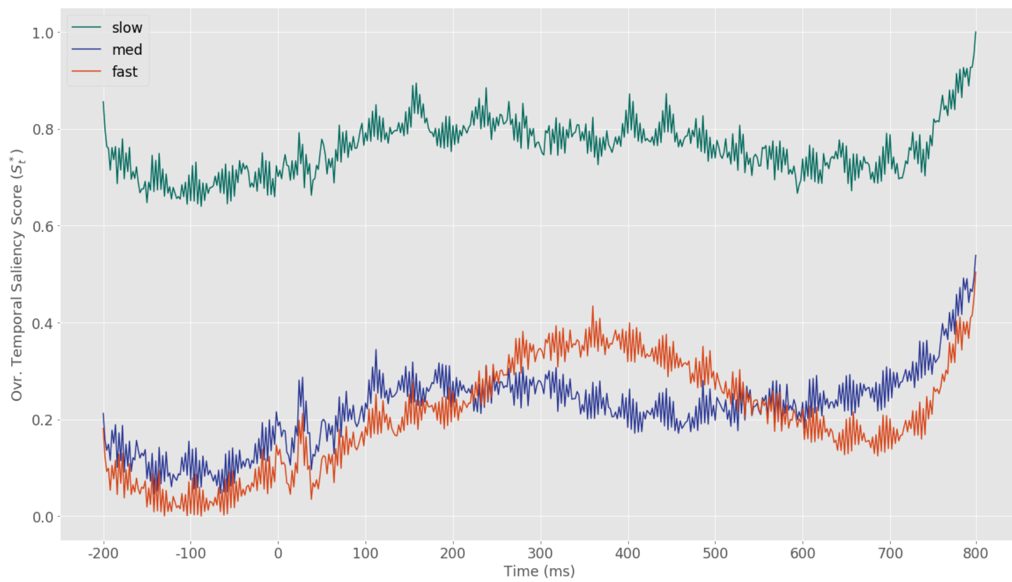

Figure 5.2: Overall temporal saliency scores of each RT groups (without the effect of median kernel).

5.3 Conclusion

In the prescribed study, we have demonstrated a novel way to decode neural activities dictating individuals' RT from EEG data using CNNs. Our data-driven approach is a bias-free decoding process since we have designed a framework for cultivating a consensus from multiple models, a reflection of only the most common and strong pattern underlying a cognitive task. We have found the efficacy of our approach by further confirming several supporting hypotheses of speech categorization behavior. Although the science of interpreting CNN models is still in its early steps, we show that existing tools like GradCAM and Guided-GradCAM can be used to explain the neurological properties of behavioral auditory CP. Our proposed process could be extended to decode other cognitive functions from EEG data.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Al-Fahad, R., Yeasin, M., and Bidelman, G. M. (2020). Decoding of single-trial EEG reveals unique states of functional brain connectivity that drive rapid speech categorization decisions. *Journal of Neural Engineering*, 17(1):016045.

Amin, S., Alsulaiman, M., Muhammad, G., Amine, M., and Hossain, M. S. (2019). Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. *Future Generation Computer Systems*, 101.

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., and Zhang, H. (2012). Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in Neuroscience*, 6:39.

Arsenault, J. and Buchsbaum, B. (2015). Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience*, 35:634.

Aslan, Z. and Akin, M. (2020). Automatic detection of schizophrenia by applying deep learning over spectrogram images of eeg signals. *Traitement du Signal*, 37:235–244.

Bashivan, P., Bidelman, G., and Yeasin, M. (2014). Spectrotemporal dynamics of the eeg during working memory encoding and maintenance predicts individual behavioral capacity. *European Journal of Neuroscience*, 40.

Bashivan, P., Rish, I., Yeasin, M., and Codella, C. F. N. (2015). Learning representations from eeg with deep recurrent-convolutional neural networks. *International Conference on Learning Representations*.

Bergstra, J., Yamins, D., and Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. ICML'13, page I–115–I–123. JMLR.org.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger,

K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc.

Bidelman, G. and Howell, M. (2016). Functional changes in inter- and intra-hemispheric cortical processing underlying degraded speech perception. *NeuroImage*, 124:581–590.

Bidelman, G., Mahmud, M. S., Yeasin, M., Shen, D., Arnott, S., and Alain, C. (2019). Age-related hearing loss increases full-brain connectivity while reversing directed signaling within the dorsal–ventral pathway for speech. *Brain Structure and Function*, 224.

Bidelman, G. M. (2017). Amplified induced neural oscillatory activity predicts musicians' benefits in categorical speech perception. *Neuroscience*, 348:107 – 113.

Bidelman, G. M. and Lee, C.-C. (2015a). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage*, 120:191 – 200.

Bidelman, G. M. and Lee, C.-C. (2015b). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage*, 120:191 – 200.

Bidelman, G. M. and Walker, B. (2019). Plasticity in auditory categorization is supported by differential engagement of the auditory-linguistic network. *NeuroImage*, 201:116022.

Carter, J. A. and Bidelman, G. M. (2020). Auditory cortex is susceptible to lexical influence as revealed by informational vs. energetic masking of speech categorization. *bioRxiv*.

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., and Gramfort, A. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769.

Chen, H., Song, Y., and Li, X. (2019). Use of deep learning to detect personalized spatial-frequency abnormalities in EEGs of children with ADHD. *Journal of Neural Engineering*, 16(6):066046.

Chiarelli, A., Croce, P., Merla, A., and Zappasodi, F. (2018). Deep learning for hybrid eeg-fnirs brain-computer interface: application to motor imagery classification. *Journal of Neural Engineering*, 15.

Chollet, F. et al. (2015). Keras.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Craik, A., He, Y., and Contreras-Vidal, J. (2019). Deep learning for electroencephalogram (eeg) classification tasks: A review. *Journal of Neural Engineering*, 16.

Dai, M., Zheng, D., Na, R., Wang, S., and Zhang, S. (2019). Eeg classification of motor imagery using a novel deep learning framework. *Sensors (Basel, Switzerland)*, 19.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Friedman, D. and Johnson Jr., R. (2000). Event-related potential (erp) studies of memory encoding and retrieval: A selective review. *Microscopy Research and Technique*, 51(1):6–28.

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature neuroscience*, 15:511–7.

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7:267.

Haenschel, C., Baldeweg, T., Croft, R. J., Whittington, M., and Gruzelier, J. (2000). Gamma and beta frequency oscillations in response to novel auditory stimuli: A comparison of human electroencephalogram (eeg) data with in vitro models. *Proceedings of the National Academy of Sciences*, 97(13):7645–7650.

Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.

Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T., and Huang, Y. (2016). Eeg-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, 47:549 – 555.

Hampshire, A., Chamberlain, S., Monti, M., Duncan, J., and Owen, A. (2010). The role of the right inferior frontal gyrus: inhibition and attentional control. *Neuroimage*, 50:1313–1319.

Hampshire, A., Thompson, R., Duncan, J., and Owen, A. (2009). Selective tuning of the right inferior frontal gyrus during target detection. *Cognitive, Affective, & Behavioral Neuroscience*, 9:103–112.

Haxby, J. (2012). Multivariate pattern analysis of fmri: The early beginnings. *NeuroImage*, 62:852–5.

Haxby, J., Gobbini, M. I., Furey, M., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425 – 2430.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Hosseini, M., Pompili, D., Elisevich, K., and Soltanian-Zadeh, H. (2017). Optimized deep learning for eeg big data and seizure prediction bci via internet of things. *IEEE Transactions on Big Data*, 3(4):392–404.

Hosseini, M., Soltanian-Zadeh, H., Elisevich, K., and Pompili, D. (2016). Cloud-based deep learning of big eeg data for epileptic seizure prediction. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1151–1155.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153.

Jonas, S., Rossetti, A. O., Oddo, M., Jenni, S., Favaro, P., and Zubler, F. (2019). Eeg-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Human Brain Mapping*, 40(16):4606–4617.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kumar, S., Sharma, A., Mamun, K., and Tsunoda, T. (2016). A deep learning approach for motor imagery eeg signal classification. In *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pages 34–39.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, Y., Turkeltaub, P., Granger, R., and Raizada, R. (2012). Categorical speech processing in broca's area: An fmri study using multivariate pattern-based analysis. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32:3942–8.

Li, Y., Yang, H., Li, J., Chen, D., and Du, M. (2020). Eeg-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-cam. *Neurocomputing*, 415:225 – 233.

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74:431–61.

Luo, T.-j., Zhou, C., and Chao, F. (2018). Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. *BMC Bioinformatics*, 19.

Mahmud, M. S., Ahmed, F., Al-Fahad, R., Moinuddin, K. A., Yeasin, M., Alain, C., and Bidelman, G. M. (2020a). Decoding hearing-related changes in older adults' spatiotemporal neural processing of speech using machine learning. *Frontiers in Neuroscience*, 14:748.

Mahmud, M. S., Yeasin, M., and Bidelman, G. M. (2020b). Speech categorization is better described by induced rather than evoked neural activity. *bioRxiv*.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Mopuri, K. R., Garg, U., and Venkatesh Babu, R. (2019). Cnn fixations: An unraveling approach to visualize the discriminative image regions. *IEEE Transactions on Image Processing*, 28(5):2116–2125.

Muhammad, M. B. and Yeasin, M. (2020). Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Olivas-Padilla, B. E. and Chacon-Murguia, M. I. (2019). Classification of multiple motor imagery using deep convolutional neural networks and spatial filters. *Applied Soft Computing*, 75:461 – 472.

Pfurtscheller, G. and Neuper, C. (2001). Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134.

Price, C. N., Alain, C., and Bidelman, G. M. (2019). Auditory-frontal channeling in α and β bands is altered by age-related hearing loss and relates to speech perception in noise. *Neuroscience*, 423:18 – 28.

Rezaeitabar, Y. and Halici, U. (2017). A novel deep learning approach for classification of eeg motor imagery signals. *Journal of Neural Engineering*, 14:016003.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Senkowski, D., Molholm, S., Gomez-Ramirez, M., and Foxe, J. J. (2005). Oscillatory Beta Activity Predicts Response Speed during a Multisensory Audiovisual Reaction Time Task: A High-Density Electrical Mapping Study. *Cerebral Cortex*, 16(11):1556–1565.

Si, X., Zhou, W., and Hong, B. (2017). Cooperative cortical network for categorical processing of chinese lexical tone. *Proceedings of the National Academy of Sciences*, 114(46):12303–12308.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*.

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008.

Tallon-Baudry, C. and Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences*, 3(4):151 – 162.

Tang, Z., Li, C., and Sun, S. (2017). Single-trial eeg classification of motor imagery using deep convolutional neural networks. *Optik*, 130:11 – 18.

Tzagarakis, C., West, S., and Pellizzer, G. (2015). Brain oscillatory activity during motor preparation: effect of directional uncertainty on beta, but not alpha, frequency band. *Frontiers in Neuroscience*, 9:246.

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., and Coleman, S. (2020a). Emotion recognition with convolutional neural network and eeg-based efdms. *Neuropsychologia*, 146:107506.

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., and Coleman, S. (2020b). Emotion recognition with convolutional neural network and eeg-based efdms. *Neuropsychologia*, 146:107506.

Wang, P., Jiang, A., Liu, X., Shang, J., and Zhang, L. (2018). Lstm-based eeg classification in motor imagery tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2086–2095.

Yin, Z. and Zhang, J. (2017). Cross-session classification of mental workload levels using eeg and an adaptive deep learning model. *Biomedical Signal Processing and Control*, 33:30 – 47.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Zhang, F., Wang, J.-P., Kim, J., Parrish, T., and Wong, P. (2015). Decoding multiple sound categories in the human temporal cortex using high resolution fmri. *PloS one*, 10:e0117303.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.

APPENDIX A. Accuracy and Loss Curves

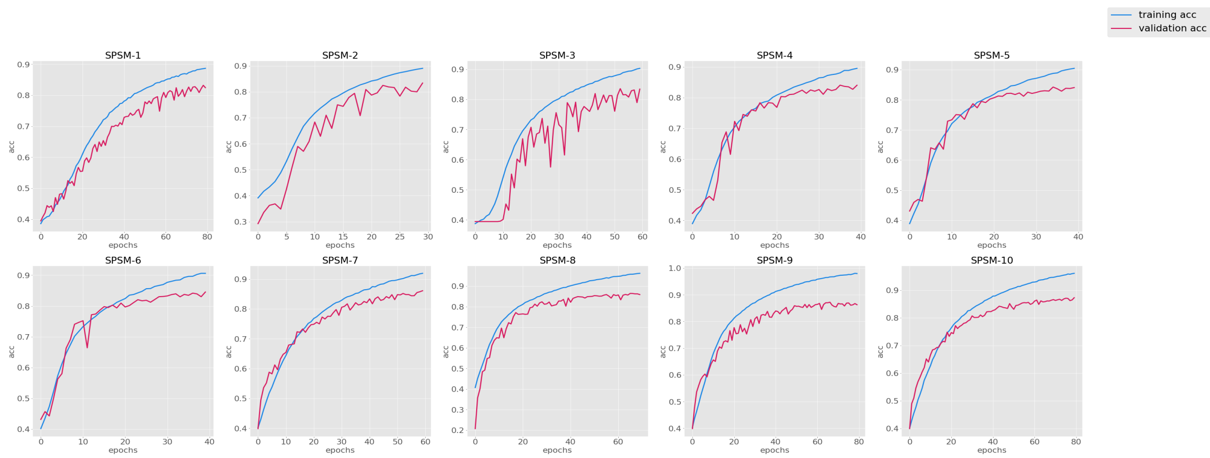

Figure A.1: Accuracy curves of 10 best SPSMs during training.

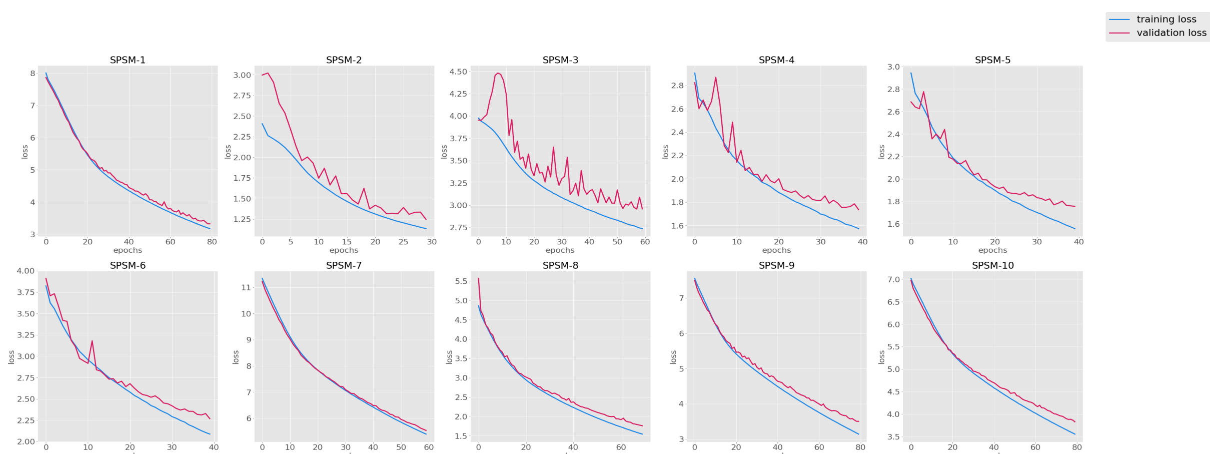

Figure A.2: Loss curves of 10 best SPSMs during training.

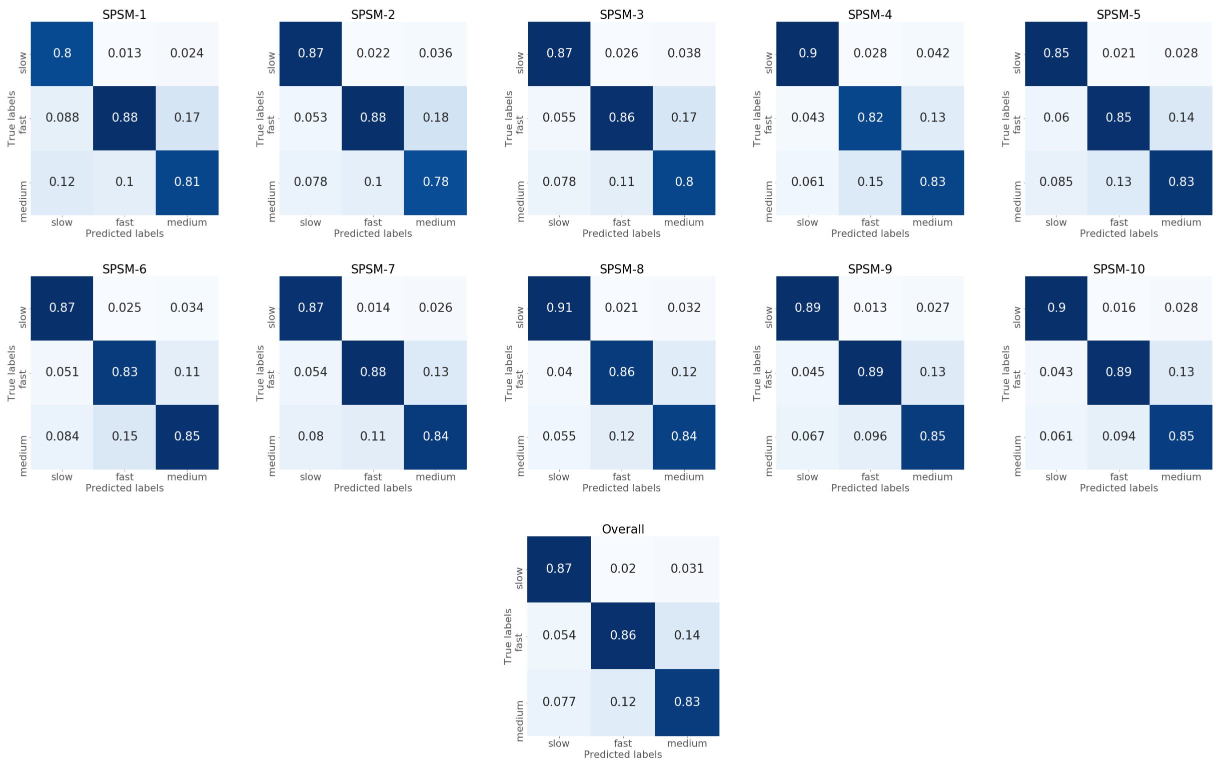

Figure A.3: Individual and overall confusion matrices of 10 best SPSMs.

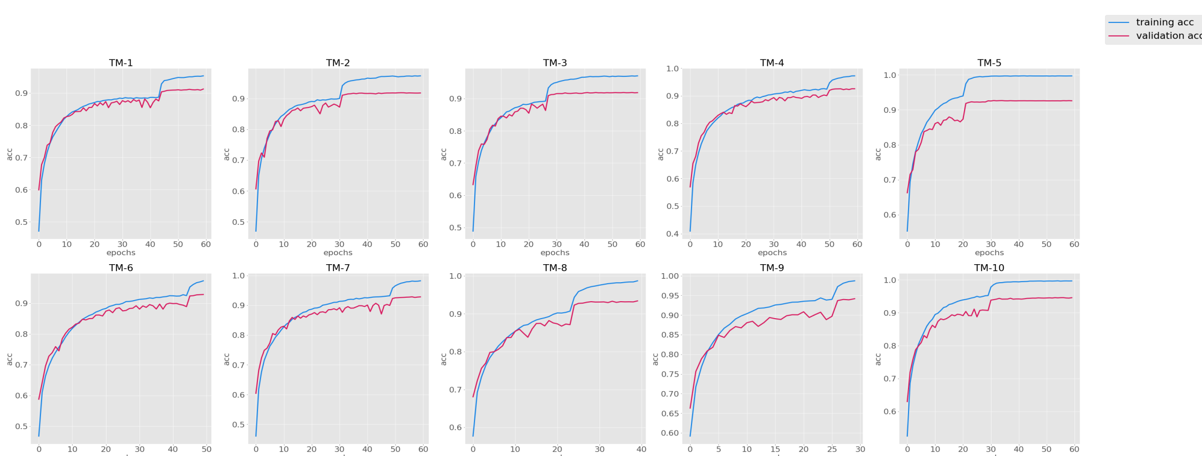

Figure A.4: Accuracy curves of 10 best TMs during training.

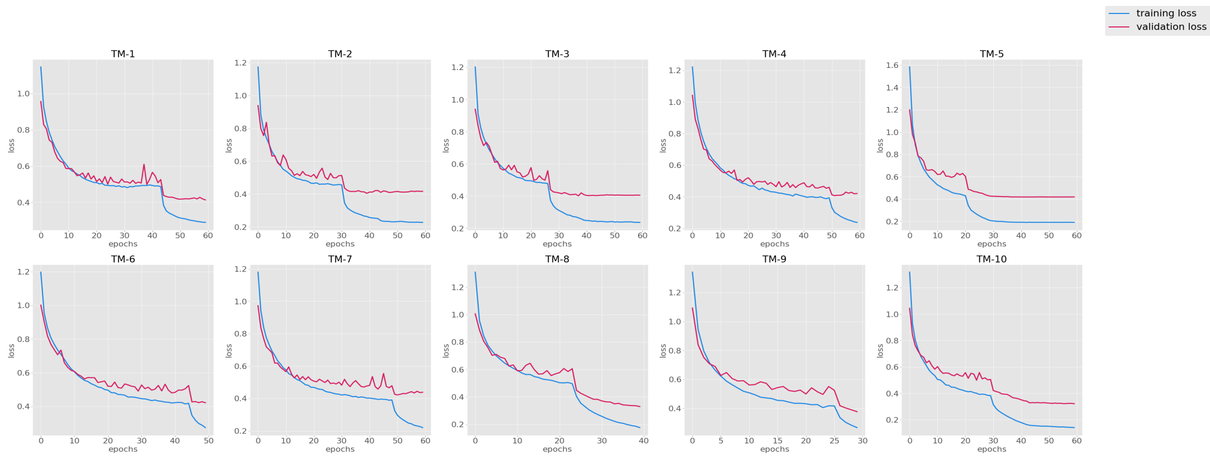

Figure A.5: Loss curves of 10 best TMs during training.

Figure A.6: Individual and overall confusion matrices of 10 best TMs.