Learning the Neural Organization of Speech Perception from Behavioral Responses: A Deep Learning Approach

1st Kazi Ashraf Moinuddin
Dept. of EECE
The University of Memphis
Memphis, TN, USA
kmnuddin@memphis.edu

2nd Mohammed Adel Bany Muhammad
Dept. of EECE
The University of Memphis
Memphis, TN, USA
mbnymhmm@memphis.edu

3rd Rakib Al Fahad
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Gavin M. Bidelman
School of Communication Sciences & Disorders
The University of Memphis
Memphis, TN, USA
gmbdlman@memphis.edu

5th Mohammed Yeasin
Dept. of EECE
The University of Memphis
Memphis, TN, USA
myeasin@memphis.edu

***Abstract*—**

***Index Terms*—**

I. Introduction

1) What is CP of speech?
2) What does individuals' behavior in speech perception tells us about CP?
3) How response time (RT) is a measure of individuals' behavior in speech perception?
4) **Our goal:** To learn the neural organization that drives individuals' RT in CP from EEG recordings using CNN.
5) **Motivation:**
   - Success of DL models in BCI tasks (Motor Imagery, Sleep Scoring, Seizure Detection, Mental Workload).
   - ML based decoding of CP.
   - Lack of application of DL models in decoding CP.
6) **Objective:**
   - Acquire spatial representation of features from EEG data.
   - Model the spatial properties of CP from EEG recordings using CNN.
   - Explore diverse set of CNN models of CP.
   - Use visual interpretation tools (GradCAM) to get insight into the learned representation of CP by CNNs.
   - Infer the neural organization of CP through model interpretation.
7) **Short description of our process:**
   - Spatial representation of features through bandpower topomaps.
   - Data Augmentation through bootstraping.
   - Learning optimal hyperparameters of CNN from the data through Bayesian Hyperparameter Optimization (TPE).
   - Select 10 best performing CNN models from the Hyperparameter Optimization step for further analysis.
   - Select most diverse subset of models from the set of best performing models to design a deep ensemble model of CP.
   - GradCAM to get insight into the learned representation by CNN models.
   - Using the activation values from GradCAM to explain the neural organization of CP.
8) **Short description of results:**
   - Performance of the 10 best CNN models.
   - Performance analysis of the ensemble model.
   - Neural organization of CP:
     - Significance analysis of the frequency bands.
     - Regions of Interest driving speech categorization behavior.

II. Method & Procedures

A. Data

1) 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). 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).

2) 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 $\approx$ 24. Participants were strongly right-handed (mean Edinburgh Hand Score $\approx$ 80.0), had acquired a collegiate level of education (mean $\approx$ 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.

3) Behavioral Data: To classify behavioral CP, we opted to form categories within RTs from all the samples. 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. GMMs are unsupervised models that assume that the input data is generated from a mixture of a finite number of Gaussian distributions (components). GMM allows the selection of the covariance structure of the components alongside the number of components. However, selecting the number of Gaussian components and associated covariance structure is challenging. Bayesian Information Criterion (BIC) can be used as a criterion for selecting these parameters of GMM; BIC is an index used in Bayesian statistics for choosing between multiple models. The model with the lowest BIC score is considered the best among a set of models. We tried different numbers of components (clusters, ranges from 1-14) and covariance types (full, tied, diagonal, and spherical) of GMM using a brute-force approach. By comparing the BIC score of each model, we find that the model with 4 components with 'spherical' covariance type achieves the lowest BIC score. We infer the slow, medium, and fast clusters of RTs based on the range of RT within each cluster through this model. The cluster with the lowest probability is deemed an outlier and discarded from the rest of the analysis.

4) Bandpower Topomap Generation: 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, we compute the power spectral densities (PSDs) of each of these samples. We compute the bandpowers by averaging across three frequency bands: α (8-15 Hz), β (16-31 Hz), and γ (32–60 Hz). We used the built-in *psd_welch function* provided in the open-source software package MNE-Python [[**?**]] 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 **??** (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 channels' location 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) [[**?**]]. In this way, each of the bands gets represented through different color channels (Figure **??** (C, D)). We used the *plot_topomap* from MNE-Python to generate the topomaps from the average bandpowers.

B. Modeling

1) Base CNN Classifiers:

- TPE Optimization
  - Description of hyperparameters.
  - Chosen hyperparameters at each step of the optimization.
  - Selection of best performing models (based on test accuracy).

2) Ensemble Classifier:

- Selecting most diverse subset of models among the best performing models as base classifiers for ensemble model.
  - Based on non-pairwise diversity measures.
- Use three ensembling techniques:
  - Unweighted Average.
  - Majority Voting.
  - Super learners.

3) Model Interpretation:

- GradCAM and Guided-GradCAM Visualization.
- Extraction of Activation Values from Guided-GradCAM outputs.
- Significance analysis on the activation values.
  - ANOVA for significance of differences in neural organization between the RT classes.
  - Tukey HSD test for significance of frequency bands in distinguishing the RT classes.

III. Results

A. Model Performance

- Classification reports for the best performing models.
- Loss and accuracy curves.
- Non-pairwise diversity measures among the best performing models.
- Classification reports for three types of Ensemble models.
- One vs. all ROC curves (testing the skill of the ensemble models).

B. Learned Representation

- Overall rank of spatial features by activation values.
- Rank of spatial features by RT classes.
- Rank of spatial features by frequency bands.
- Results from statistical analysis.

IV. Discussion

A. The Modeling Perspective

- Novelty and contribution.
- Why bandpower topomaps?
- Sanity of the bootstraping process.
  - Divesity.
  - Data Augmentation.
  - Uneven class samples.
- Why CNN?
  - Powerful.
  - Feature extraction / selection through convolution.
  - Visual Interpretation tools.
- Comparison of the base and ensemble classifiers.

B. Neural Organization of Behavioral CP

- ROIs in dictating individuals' behavior in speech perception.
- Functional roles of the ROIs (defined by the frequency bands).
- Validation through contemporary studies.
- Summary of findings.

V. Conclusion

A. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
Table Type Styles

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References