Decoding the Neural Correlates of Decision Speed in Speech Perception using CNNs

Anonymous Authors[1]

Abstract

Categorical perception (CP) is a complex cognitive process of grouping speech sounds into distinct categories. Decision speed in CP tasks reflects the perceptual difficulty of auditory stimuli and is captured by the listener's response time (RT). The cognitive processes involved in mapping neural activities to behavioral RT are subject to individual variations. This paper presents a data-driven approach and develops parameter-optimized models to understand the relationship between cognitive events and behavioral response (e.g., RT) of individuals. To investigate the neural correlates of decision speed in speech categorization, we have explored convolutional neural networks (CNN) to learn and decode CP behavior from EEG recordings. We applied GradCAM to gain insight into the learned representation of the CNN models and infer the neural orientation underlying CP behavior. We further show that activation values from Guided-GradCAM outputs can quantify learned representation of CNNs as analysis of them reveals neural patterns consistent with prior findings of CP behavior.

1. Introduction

Categorical perception (CP) of speech is a cognitive process of grouping sounds into small phonetic categories (Liberman et al., 1967). CP is critical in understanding the neural process of speech comprehension and learning. It is an effect that is found to be present in infants and evolves due to speech training (Eimas et al., 1971). Thus, investigating the neural organization of CP provides insight into the neural process of speech perception from an elementary level. The lexical processing paradigm of CP is well studied; however, the neural orientation of speech perception, which induces variation in individuals' decision speed, is obscure. In CP studies, the response time (RT) is captured by measuring the time lapse between the voice onset time (VOT) and the time of the actual response (need to cite). RTS effectively capture the decision speed of listeners (Pisoni & Tash, 1974) and analysis of RTs can reveal the cause of perceptual difference between subjects. Investigating the neural correlates of decision speed in speech categorization can also explain the variation in behavior across and within-subjects.

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 properties of CP, 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. Another work reported in (Al-Fahad et al., 2020) used an ML-based approach to decode functional connectivity patterns in CP. The mentioned studies uses classical ML, such as support vector machines (SVM) (Cortes & Vapnik, 1995) with stability selections (Meinshausen & Bühlmann, 2010) to model cognitive processes involved in CP. The feature selection process using stability selection provides a limited interpretation of the causal relationship between neural activities and behavioral responses.

To capture the relationship between cognitive events and behavioral responses (e.g., RT), we present a deep learning approach to first learn the underlying cognitive function and then infer the neural process of CP behavior from learned representations. Inferring regional or spatial orientation of cognitive processes from EEG recordings is difficult due to low spatial resolution and noise. To learn the spatial organization from EEG data, 1) it is required to have a spatial representation extracted from EEG signals, 2) and a robust

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

noise reduction technique. To this extent, we have utilized band power topomaps to represent the spatial contents from EEG data and a bootstrap averaging process to reduce noise in EEG signals. We randomly sample trials (with replacement) within a subject's RT category and average them to generate event-related potentials (ERP). ERPs contain less noise than the original EEG signals due to averaging of multiple trials. We then compute those ERPs' band powers and project them into a 2D topographic representation of the scalp surface known as topomaps. The bandpower topographical representation is introduced by (Bashivan et al., 2015) and captures the spatial-temporal-spectral contents from EEG signals. However, in our work, we have focused on the spatial-spectral representation since we want to decode the neural organization of CP rather than the temporal aspect.

We introduce convolutional neural networks (CNN) to learn the spatial-spectral representation from EEG recordings. CNNs have been successful in computer vision tasks and have proven capacity in learning spatial representations. CNNs have also been successful in modeling neural processes from EEG data. (Amin et al., 2019; Tang et al., 2017; Olivas-Padilla & Chacon-Murguia, 2019) used different CNN models to achieve significant results in modeling MI tasks from EEG data. (Dai et al., 2019; Rezaeitabar & Halici, 2017) combined CNNs with stacked and variational autoencoders to predict limb movements from EEG recordings. (Bashivan et al., 2015) was one of the early application of DL in cognitive neuroscience where the 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. (Dai et al., 2019; Rezaeitabar & Halici, 2017) combined CNNs with stacked and variational autoencoders to predict limb movements from EEG recordings. Inspired by these applications, we chose CNNs to learn the neural processes underlying speech categorization behavior.

To capture the general neural patterns of different CP behavior, we have chosen to form the problem irrespective of population. We have defined three RT classes (slow, medium, and fast), which is irrespective of population, i.e., the RT categories are formed using individual trials rather than subjects. To form RT categories, we apply Gaussian Mixture Model (GMM) clustering in the RTs of all available trials. The bootstrap and averaging process for generating ERPs are carried out within each RT class of a subject and converted to band power topomaps. We chose to represent the spatial orientation of α, β and γ frequency bands in the topomaps as they are highly correlated with CP (Bidelman, 2017; Bidelman & Lee, 2015a; Giraud & Poeppel, 2012).

We train CNNs using these band power topomaps to learn the representation of these different RT categories. We deploy the Bayesian hyperparameter optimization algorithm, the tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011) to find the best configuration for our CNN model. Finally, we use our best performing model to interpret the neural process of CP behavior from learned representation.

In deep learning (DL), model interpretation is still a challenge as these models contain millions of parameters and therefore are difficult to interpret. Visual interpretations are the only effective ways to get an insight into the learned features of a neural model. 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 & Yeasin, 2020). Studies like (Jonas et al., 2019; Li et al., 2020; Wang et al., 2020) shows that GradCAM does capture feature importance allocation by CNNs from EEG data and therefore could be used to infer spatial-spectral-temporal properties underlying a cognitive event. Despite the successes in visual interpretations, we acknowledge that it is impossible to infer any neural process using these visualizations alone. It is required to quantify the learned representations of neural models to infer any cognitive process from them. We have used activation values from Guided-GradCAM outputs to quantify the learned representation of our CNN models to address this issue. We analyze these activation values to explain the neural correlates of decision speed in speech categorization.

We empirically evaluate the CNN models using the CP data obtained from 50 participants. Our analysis using the activation values shows that the right frontal regions are crucial in determining listeners' decision speed. We also find that all three bands (α, β, γ) have active and passive roles, while the γ band is the most significant in driving listeners' RT. The significance of the 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.

2. Data

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. 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 **??** (A, B) shows the demographic of the participants.

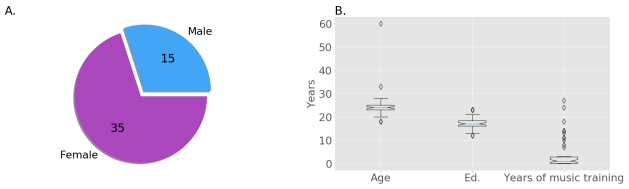

Figure 1. A) Gender distribution B) Demography of the participants (includes age, musical training, and education)

EEG Recording & Preprocessing: 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).

Clustering RTs: To learn 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, minimum, and 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 2 illustrates the optimization of GMM, the RT distribution, the probability of each RT cluster, and the maximum, minimum range of each RT cluster.

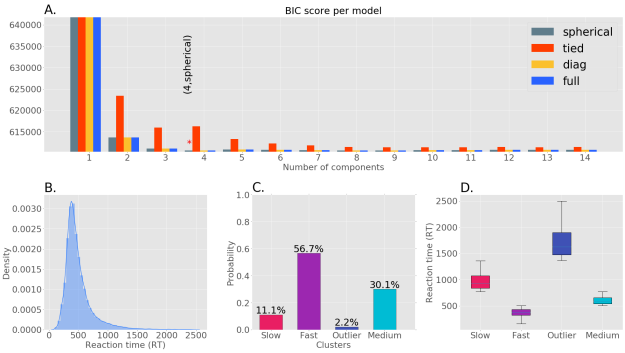

Figure 2. Clustering of RT data. A) BIC scores of models with different number of components and covariance type, the '*' denotes the model with the lowest BIC score. B) Original RT distribution. C) The probability of each RT clusters using the GMM with lowest BIC score. D) The RT range of each clusters (slow : 772 - 1360 ms, fast : 100 - 504 ms, outlier: 1364 - 2500 ms, medium: 506 - 770 ms

.

Spatial-Spectral Representation: As mentioned in section 1, we have opted to use bootstrapping to reduce noise in the EEG samples. 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. The bootstrapping process of generating ERPs also works as a data augmentation step since we can generate more diverse samples by averaging different trials in each iteration. This process produced 62525 ERPs; we then compute the power spectral density (PSD) and band powers of these samples. We compute the bandpowers of 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 (Gramfort et al., 2013) to compute the PSDs. Next, we average the PSDs across α, β and γ bands to acquire associated band power of the 64 spatial locations.

Figure 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 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 (see figure 4). We used the *plot_topomap* from MNE-Python to generate the topomap images (size: $128 \times 128 \times 3$) from the average band powers.

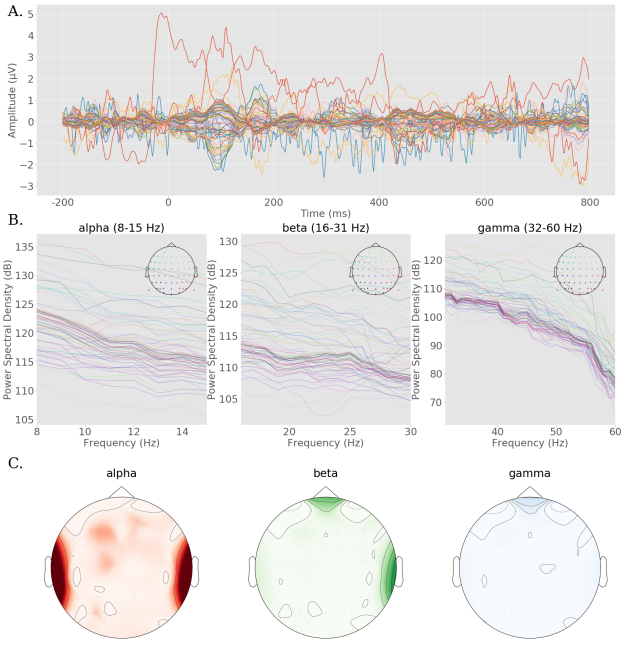

Figure 3. A) ERP. B) PSDs of 64 spatial locations of α, β, γ frequency bands. C) Corresponding spatial topomaps of α, β, γ bands. D) Combined band power topomap through stacking of the α, β, γ topomaps.

3. Modeling

Hyperparameter Optimization: Using the bootstrap and average process, we generated 62525 ERP samples, which we converted to bandpower topomaps as described above. We used 46893 (75%) samples for training and 15632 (25%) samples for testing each model produced by the Bayesian hyperparameter optimization algorithm. We optimize the architecture and the general hyperparameters (e.g., batch size, epochs, learning rate) of CNN using TPE; table 1 describes the hyperparameters optimized by TPE. We utilized Adam (Kingma & 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 Models. In the convolution layers, each layer contains twice the number of filters as 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 to find the best hyperparameters of our CNN model. Figure 3 shows the test accuracy of each model generated

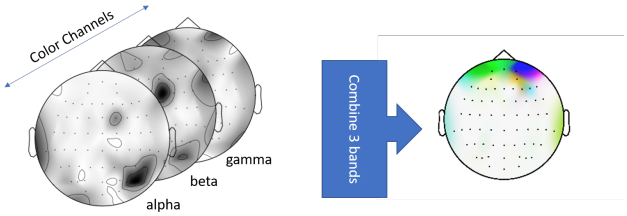

Figure 4. Combined band power topomap through stacking of the α, β, γ topomaps.

by the TPE algorithm. We implemented the TPE algorithm for optimizing CNN using the open-source library hyperopt (Bergstra et al., 2013), and Keras (Chollet et al., 2015).

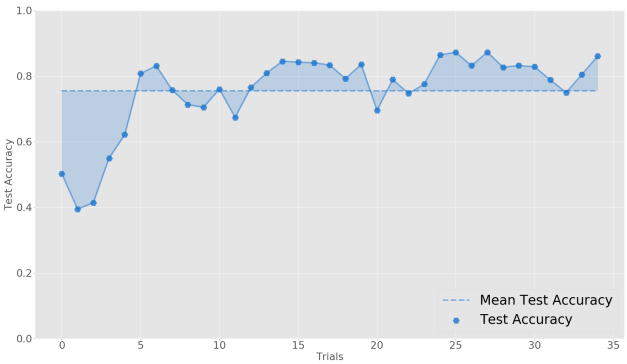

Figure 5. Test accuracy of each models during the steps of hyperparameter optimization

Performance: In this section, we evaluate our best CNN model (Model-10) of CP behavior. Among the 35 CNN models generated by the TPE algorithm, the mean test accuracy was 75.52%, with the best performing model achieving an accuracy of 87.28%. Figure 3 shows our best model's loss and accuracy curve during the training session. We achieved an average (macro) of 87.70% precision, 87.95% recall, and 87.79% f1-score, suggesting effective learning of neural patterns of different CP behavior. The slow, fast, and medium RTs achieved an average class probability score of 95%, 93%, and 92%, respectively. The slow RT class was classified more accurately (precision: 89.63%, recall: 91.4% and f1-score: 90.52%) compared to the other RT classes. Figure 3 shows the one vs. all precision-recall curve (PR curve), normalized and non-normalized confusion matrices. The area under the PR curve for the three RT categories (slow: 0.97, fast: 0.95, med: 0.94) suggests that the model can determine each category and distinguish between them skillfully. Table 2 shows the performance comparison of the ten best models generated in the hyperparameter optimization process.

Table 1. The hyperparameter optimized for CNNs 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.

Table 2. Performance comparison of 10 best CNN models generated by the TPE optimization algorithm.

MODEL	PRECISION	RECALL	F1 SCORE	ACCURACY
MODEL-1	82.96%	84.80%	83.59%	83.22%
MODEL-2	84.25%	84.22%	84.05%	83.35%
MODEL-3	84.23%	84.25%	84.16%	83.58%
MODEL-4	84.97%	84.46%	84.69%	84.05%
MODEL-5	84.43%	85.29%	84.83%	84.25%
MODEL-6	84.90%	85.24%	85.01%	84.53%
MODEL-7	86.21%	87.10%	86.60%	86.09%
MODEL-8	87.12%	87.02%	87.07%	86.46%
MODEL-9	87.54%	88.03%	87.75%	87.24%
MODEL-10 (BEST)	87.70%	87.95%	87.79%	87.28%

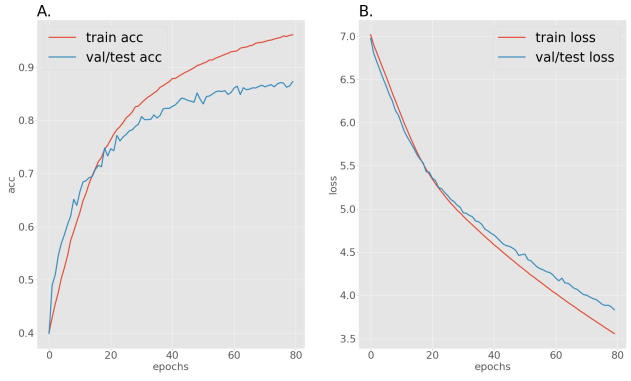

Figure 6. Accuracy (A) and loss curve (B) of best performing CNN during training.

4. Learned Representation

GradCAM and Guided-GradCAM: In this section, we present the learned representation of Model-10 (see table 2), which is our best CP behavior model. We use GradCAM and Guided-GradCAM visualization to get an insight into the learned spatial-spectral features of our best CNN. 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 class flowing into the final convolution layer to produce such visualizations. Guided-GradCAM is another class discriminative activation map that combines Guided-Backpropagation (Springenberg et al., 2015) with GradCAM to produce channel-wise activation maps. Figure 4 illustrates some GradCAM and Guided-GradCAM visualization of the learned features by CNN. As mentioned in section 1, visualization alone is not enough to infer a models' learned representation. Therefore, we have taken the activation values from the Guided-GradCAM output to measure the learned representation of the CNN. Since we have 64 spatial locations in our input topomaps, we need 64 activation values to quantify the learned features of CNN. We extract the activation values surrounding the neighborhood of each of these spatial locations by applying a median filter (20×20). Figure 4 show the channel (band) wise extraction of activation values through median filters.

Statistical Analysis: We apply ANOVA analysis on the activation values to observe the significance of learned spatial-

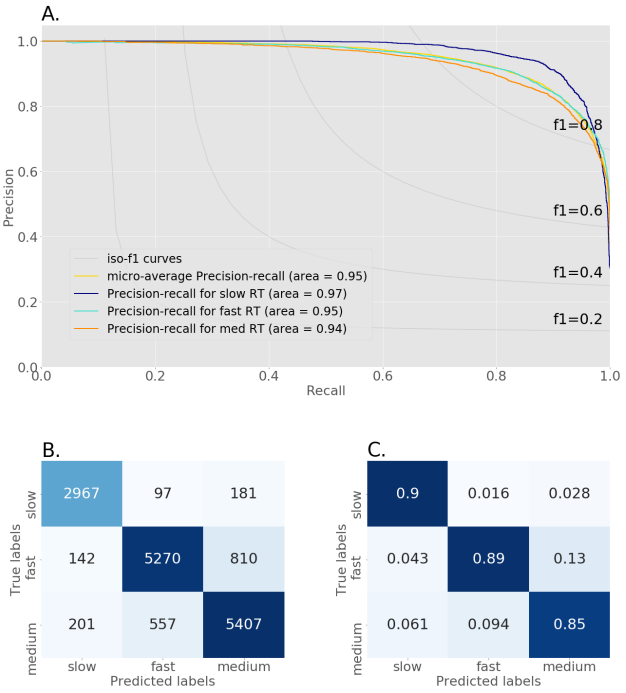

Figure 7. A) Precision-recall curve B) Confusion matrix C) Normalized confusion matrix

spectral representations between the RT categories. The analysis shows a significant difference in spatial activation between the bands ($p = 2.2e^{-16}$), validating that our spatial-spectral representation was accurate. It also confirms that our model was able to capture the different regional operations of individual bands. We also find different band activation between the RT categories ($p = 0.024$), implying that our model effectively learned these frequency bands' role in CP tasks. Another observation from the analysis is that there is significant ($p = 0.0056$) variation within the RT categories, further confirming that the neural process underlying speech categorization speed is subject to high variation. Overall, the ANOVA analysis showed that our best model was able to effectively capture the variable neural factors contributing to speech categorization speed. To infer from the learned representation, we have conducted Tukey HSD tests on the activation values. We focus on finding the spatial and spectral factors that contribute to different decision speed in CP.

Neural Correlates of CP Behavior: Auditory categorization in the human brain is revealed to use a distributed frontal-temporal-parietal network by contemporary EEG studies (Bidelman & Walker, 2019; Bidelman & Lee, 2015b; Al-Fahad et al., 2020). The canonical language processing is left hemisphere (LH) predominantly. However, through analysis of the activations from Guided-GradCAM, we find

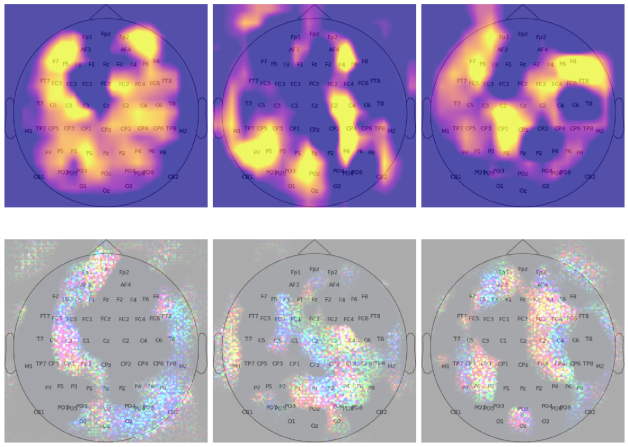

Figure 8. Activation maps from GradCAM (1st) and Guided-GradCAM (2nd)

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 responses. (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. We have similar findings as to 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 4). The Tukey HSD test on the spatial activations revealed that CP1 and P2 locations significantly discriminate the slow RTs from the other two classes. CP1 and P2 are part of a distributed activation that we observe across the parietal lobe (in both hemispheres for slow and med RTs, predominantly in RH for fast RTs). The parietal lobe has the function of managing short-term memory and retrieving encoded verbal material during speech perception (Jonides et al., 1998). Therefore, it is evident from this finding that delay in retrieval of encoded materials is also a factor contributing to slow RTs.

In terms of perceptual encoding of speech, we also find our spatial results to be coherent as (Bidelman & 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 4), which explains the functional lateralization of RH. In slower RT, we find more distributed region activations, especially lateralization of the LH, and lesser activation in frontal regions of RH. The lateralization of LH suggests increased use of language processing units, whereas low activation in the distributed

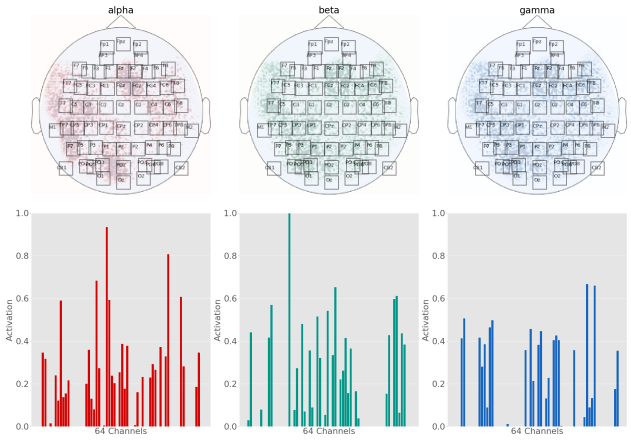

Figure 9. Band-wise activation value extraction from Guided-GradCAM outputs, the 1st row show the median filters on the 64 spatial location on each frequency bands, the 2nd row shows the extracted median activation values from the surroundings of those locations.

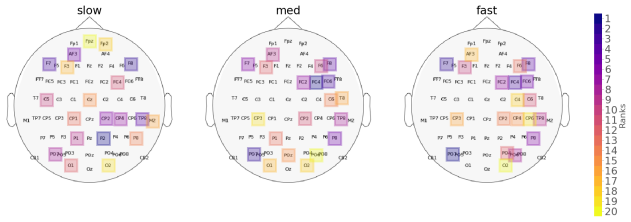

Figure 10. Top 20 ranked spatial features by mean activation values, the features are ranked by taking the mean of actiavtion across all test samples of a RT class.

frontal regions of RH implies a lack of attentional control.

We assess through pairwise Tukey HSD test that γ band significantly distinguishes between the fast-med ($p = 0.0069$) and med-slow ($p = 0.0098$) group. γ band also achieved the highest mean activation (0.0093) and is thus more predictive of participants' decision time. The significance of the γ band is coherent with the recent study of (Mahmud et al., 2020b) where γ band modulations are found to be more correlated with listeners' behavioral CP. γ is found to be responsible for auditory object construction (Tallon-Baudry & Bertrand, 1999) and local network synchronization (Giraud & Poeppel, 2012; Haenschel et al., 2000; Si et al., 2017). Since we have the effect of musical training in our participants, the saliency of the γ band might also reflect experience-dependent enhancement in CP. γ band is found to be a reflection of enhancement in individuals' CP dictated by their musical experience.

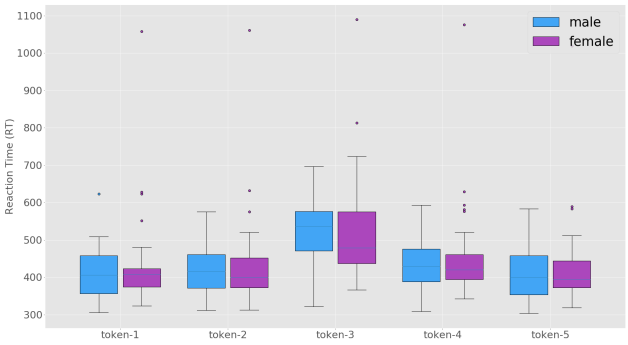

Figure 11. Variation of RTs across tokens

5. Conclusion

In the prescribed study, we have demonstrated a novel way to decode neural activities dictating individuals' RT from EEG data using CNN. 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.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Al-Fahad, R., Yeasin, M., and Bidelman, G. M. 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, feb 2020. doi: 10.1088/1741-2552/ab6040. URL https://doi.org/10.1088%2F1741-2552%2Fab6040.

Amin, S., Alsulaiman, M., Muhammad, G., Amine, M., and Hossain, M. S. Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. *Future Generation Computer Systems*, 101, 07 2019. doi:

10.1016/j.future.2019.06.027.

Bashivan, P., Bidelman, G., and Yeasin, M. Spectrotemporal dynamics of the eeg during working memory encoding and maintenance predicts individual behavioral capacity. *European Journal of Neuroscience*, 40, 10 2014. doi: 10.1111/ejn.12749.

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

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

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2546–2554. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf.

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

Bidelman, G., Mahmud, M. S., Yeasin, M., Shen, D., Arnott, S., and Alain, C. Age-related hearing loss increases full-brain connectivity while reversing directed signaling within the dorsal–ventral pathway for speech. *Brain Structure and Function*, 224, 07 2019. doi: 10.1007/s00429-019-01922-9.

Bidelman, G. M. Amplified induced neural oscillatory activity predicts musicians' benefits in categorical speech perception. *Neuroscience*, 348:107 – 113, 2017. ISSN 0306-4522. doi: https://doi.org/10.1016/j.neuroscience.2017.02.015. URL http://www.sciencedirect.com/science/article/pii/S0306452217300982.

Bidelman, G. M. and Lee, C.-C. Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage*, 120:191 – 200, 2015a. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.06.087. URL http://www.sciencedirect.com/science/article/pii/S1053811915005923.

Bidelman, G. M. and Lee, C.-C. Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage*, 120:191 – 200, 2015b. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.06.087. URL http://www.sciencedirect.com/science/article/pii/S1053811915005923.

Bidelman, G. M. and Walker, B. Plasticity in auditory categorization is supported by differential engagement of the auditory-linguistic network. *NeuroImage*, 201:116022, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116022. URL http://www.sciencedirect.com/science/article/pii/S1053811919306032.

Chollet, F. et al. Keras, 2015. URL https://github.com/fchollet/keras.

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

Cortes, C. and Vapnik, V. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL https://doi.org/10.1023/A:1022627411411.

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

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

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. Speech perception in infants. *Science*, 171(3968):303–306, 1971. ISSN 0036-8075. doi: 10.1126/science.171.3968.303. URL https://science.sciencemag.org/content/171/3968/303.

Giraud, A.-L. and Poeppel, D. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature neuroscience*, 15:511–7, 03 2012. doi: 10.1038/nn.3063.

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7:267, 2013. ISSN 1662-453X. doi: 10.3389/fnins.2013.00267. URL https://www.frontiersin.org/article/10.3389/fnins.2013.00267.

Haenschel, C., Baldeweg, T., Croft, R. J., Whittington, M., and Gruzelier, J. 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, 2000. doi: 10.1073/pnas.120162397. URL https://www.pnas.org/content/97/13/7645.

Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, June 2000. ISSN 1476-4687. doi: 10.1038/35016072. URL https://doi.org/10.1038/35016072.

Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T., and Huang, Y. Eeg-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, 47:549 – 555, 2016. ISSN 0923-5965. doi: https://doi.org/10.1016/j.image.2016.05.018. URL http://www.sciencedirect.com/science/article/pii/S0923596516300832.

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

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

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

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

Jonas, S., Rossetti, A. O., Oddo, M., Jenni, S., Favaro, P., and Zubler, F. Eeg-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Human Brain Mapping*, 40(16):4606–4617, 2019. doi: 10.1002/hbm.24724.

Jonides, J., Schumacher, E. H., Smith, E. E., Koeppe, R. A., Awh, E., Reuter-Lorenz, P. A., Marshuetz, C., and Willis, C. R. The role of parietal cortex in verbal working memory. *Journal of Neuroscience*, 18(13):5026–5034, 1998. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.18-13-05026.1998. URL https://www.jneurosci.org/content/18/13/5026.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Li, Y., Yang, H., Li, J., Chen, D., and Du, M. Eeg-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-cam. *Neurocomputing*, 415:225 – 233, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.07.072. URL http://www.sciencedirect.com/science/article/pii/S0925231220311863.

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. Perception of the speech code. *Psychological review*, 74:431–61, 12 1967. doi: 10.1037/h0020279.

Mahmud, M. S., Ahmed, F., Al-Fahad, R., Moinuddin, K. A., Yeasin, M., Alain, C., and Bidelman, G. M. Decoding hearing-related changes in older adults' spatiotemporal neural processing of speech using machine learning. *Frontiers in Neuroscience*, 14:748, 2020a. ISSN 1662-453X. doi: 10.3389/fnins.2020.00748. URL https://www.frontiersin.org/article/10.3389/fnins.2020.00748.

Mahmud, M. S., Yeasin, M., and Bidelman, G. M. Speech categorization is better described by induced rather than evoked neural activity. *bioRxiv*, 2020b. doi: 10.1101/2020.10.20.347526. URL https://www.biorxiv.org/content/early/2020/10/21/2020.10.20.347526.

Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. doi: https://doi.org/10.1111/j.1467-9868.2010.00740.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00740.x.

Mopuri, K. R., Garg, U., and Venkatesh Babu, R. Cnn fixations: An unraveling approach to visualize the discriminative image regions. *IEEE Transactions on Image Processing*, 28(5):2116–2125, May 2019. ISSN 1941-0042. doi: 10.1109/TIP.2018.2881920.

Muhammad, M. B. and Yeasin, M. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, July 2020. doi: 10.1109/IJCNN48605.2020.9206626.

Olivas-Padilla, B. E. and Chacon-Murguia, M. I. Classification of multiple motor imagery using deep convolutional neural networks and spatial filters. *Applied Soft Computing*, 75:461 – 472, 2019. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2018.11.031. URL http://www.sciencedirect.com/science/article/pii/S156849461830663X.

Pisoni, D. and Tash, J. Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15:285–290, 1974.

Rezaeitabar, Y. and Halici, U. A novel deep learning approach for classification of eeg motor imagery signals. *Journal of Neural Engineering*, 14:016003, 02 2017. doi: 10.1088/1741-2560/14/1/016003.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, Oct 2017. doi: 10.1109/ICCV.2017.74.

Si, X., Zhou, W., and Hong, B. Cooperative cortical network for categorical processing of chinese lexical tone. *Proceedings of the National Academy of Sciences*, 114(46):12303–12308, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1710752114. URL https://www.pnas.org/content/114/46/12303.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. URL http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a.

Tallon-Baudry, C. and Bertrand, O. Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences*, 3(4):151 – 162, 1999. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(99)01299-1. URL http://www.sciencedirect.com/science/article/pii/S1364661399012991.

Tang, Z., Li, C., and Sun, S. Single-trial eeg classification of motor imagery using deep convolutional neural networks. *Optik*, 130:11 – 18, 2017. ISSN 0030-4026. doi: https://doi.org/10.1016/j.ijleo.2016.10.117. URL http://www.sciencedirect.com/science/article/pii/S0030402616312980.

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., and Coleman, S. Emotion recognition with convolutional neural network and eeg-based efdms. *Neuropsychologia*, 146:107506, 2020. ISSN 0028-3932. doi: https://doi.org/10.1016/j.neuropsychologia.2020.107506. URL http://www.sciencedirect.com/science/article/pii/S0028393220301780.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016. doi: 10.1109/CVPR.2016.319.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.