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

Anonymous Authors[1]

Abstract

Categorical perception (CP) is a neural process of detecting phonetic categories in 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 theblack 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.

[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.

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 these properties of 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 & Vapnik, 1995)] with stability selections [(Meinshausen & Bhlmann, 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 & 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 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.

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 $\approx$ 24 years. 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. Figure **??** (A, B) shows the demographic of the participants.

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).

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 **??** illustrates the optimization of GMM, the RT distribution, the probability of each RT cluster, and the maximum, minimum range of each RT cluster.

Spatial-Spectral Representation: 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 (3260 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 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 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 **??** (D)). We used the *plot_topomap* from MNE-Python to generate the topomaps from the average band powers.

3. Modeling

Hyperparameter Optimization: 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 1 describes the hyperparameters optimized by TPE for SPSMs. 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 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 **??** illustrate the chosen hyperparameters during each trial with associated test accuracy.

Performance: The hyperparameter optimization for both temporal and spatial-spectral models are run for 35 trials. Figure **??** 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 $\approx$ 83% to 87%, while the range for the top 10 TMs is from $\approx$ 91% to 95%. Table **??** shows the performance of the top 10 SPSMs and TMs respectively.

4. Learned Representation

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 **??**). Figure **??**, **??**, **??** 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 **??** illustrates the overall and individual band saliency variation across samples as modeled by top 10 SPSMs.

Table 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.

Table 2. 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%

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.

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 & Walker, 2019; Bidelman & Lee, 2015b; 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 **??**). 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 **??**), 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 3). 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 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 & Bertrand, 1999)] and local network synchronization [(Giraud & 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 & Lee, 2015a)] 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 3. 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

5. Conclusion

If a paper is accepted, we strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, **do not** include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

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.

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. 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 dorsalventral 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.

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):273297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL https://doi.org/10.1023/A:1022627411411.

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

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 Hmlinen, 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.

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.

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 Bhlmann, 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.

Price, C. N., Alain, C., and Bidelman, G. M. 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, 2019. ISSN 0306-4522. doi: https://doi.org/10.1016/j.neuroscience.2019.10.044. URL http://www.sciencedirect.com/science/article/pii/S0306452219307432.

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.

Senkowski, D., Molholm, S., Gomez-Ramirez, M., and Foxe, J. J. 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, 12 2005. ISSN 1047-3211. doi: 10.1093/cercor/bhj091. URL https://doi.org/10.1093/cercor/bhj091.

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.

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.

Tzagarakis, C., West, S., and Pellizzer, G. Brain oscillatory activity during motor preparation: effect of directional uncertainty on beta, but not alpha, frequency band. *Frontiers in Neuroscience*, 9:246, 2015. ISSN 1662-453X. doi: 10.3389/fnins.2015.00246. URL https://www.frontiersin.org/article/10.3389/fnins.2015.00246.

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.