INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

SPEAK YOUR MIND!
Towards Imagined Speech Recognition With Hierarchical Deep Learning

Pramit Saha[1], *Muhammad Abdul-Mageed*[2], *Sidney Fels*[1]

[1]Human Communication Technologies Lab, University of British Columbia
[2]Natural Language Processing Lab, University of British Columbia

pramit@ece.ubc.ca, muhammad.mageed@ubc.ca, ssfels@ece.ubc.ca

Abstract

Speech-related Brain Computer Interface (BCI) technologies provide effective vocal communication strategies for controlling devices through speech commands interpreted from brain signals. In order to infer imagined speech from active thoughts, we propose a novel hierarchical deep learning BCI system for subject-independent classification of 11 speech tokens including phonemes and words. Our novel approach exploits predicted articulatory information of six phonological categories (e.g., nasal, bilabial) as an intermediate step for classifying the phonemes and words, thereby finding discriminative signal responsible for natural speech synthesis. The proposed network is composed of hierarchical combination of spatial and temporal CNN cascaded with a deep autoencoder. Our best models on the KARA database achieve an average accuracy of 83.42% across the six different binary phonological classification tasks, and 53.36% for the individual token identification task, significantly outperforming our baselines. Ultimately, our work suggests the possible existence of a brain imagery footprint for the underlying articulatory movement related to different sounds that can be used to aid imagined speech decoding.

Index Terms: Brain Computer Interface, hierarchical deep neural network, phonological categories, Imagined Speech recognition, spatio-temporal CNN, deep autoencoder.

1. Introduction

Speech-related Brain Computer Interface (BCI) technologies provide neuro-prosthetic help for people with speaking disabilities, neuro-muscular disorders and diseases. It can equip these users with a medium to communicate and express their thoughts, thereby improving the quality of rehabilitation and clinical neurology. Such devices also have applications for healthy individuals–in entertainment, preventive treatments, personal communication, games, etc.

Typical forms of daily human interaction involve verbal and non-verbal communication in the form of vocal speech (or sounds) and physical gestures. However, the majority of existing research focuses on motor imagery-based control of external devices [1,2] (e.g., wheelchair). For communicating expressions and thoughts, we need a control space equipped with more functionalities and higher degrees of freedom. For these reasons, the vocal space involving labial, lingual, naso-pharyngeal and jaw motion is arguably an alternative, multi-dimensional controlling paradigm.

The challenge is that speech production is a complex process, involving intricate muscular hydrostat structure movement (e.g., the tongue). Recently, deep neural networks have emerged as efficient tools for handling complex tasks. Yet, there is hardly any work investigating the applicability and performance of such deep learning techniques for speech imagery-based BCI.

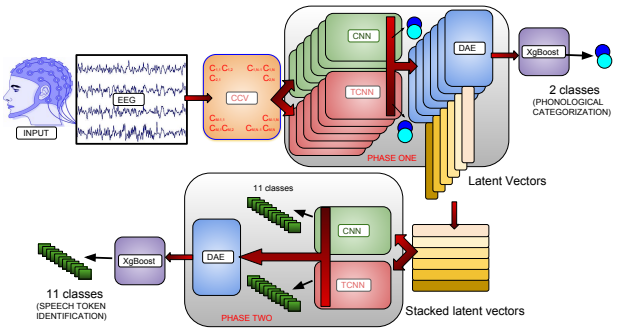

Figure 1: *Overall framework of the proposed approach*

Among the various brain activity-monitoring modalities in BCI, Electroencephalography (EEG) [3, 4] has been demonstrated as carrying promising signal for differentiating different brain activities (through measurement of related electric fields). However, these are high dimensional, and have poor Signal-to-Noise ratio, low spatial resolution, and plenty of artifacts. Besides, it is not entirely clear how to decode the desired information from the high-dimensional raw EEG signals.

Although the area of BCI based speech intent recognition has received increasing attention within the research community in the past few years, most research has focused on classification of individual speech categories in terms of discrete vowels, phonemes and words [5–13]. This includes categorization of imagined EEG signal into binary vowel categories like */a/*, */u/* and rest [5–7]; binary syllable classes like */ba/* and */ku/* [8–10, 14]; a handful of control words like *'up'*, *'down'*, *'left'*, *'right'* and *'select'* [13] or others like *'water'*, *'help'*, *'thanks'*, *'food'*, *'stop'* [11], Chinese characters [12], etc. Such works mostly involve traditional signal processing or manual feature handcrafting along with linear classifiers (e.g., SVMs). In our recent work [15], we introduced deep learning models for classification of vowels and words that achieve 23.45% improvement of accuracy over the baseline.

In this work, as an extension of our previous work [16] , our goal is to detect speech tokens from speech imagery (*active thoughts* or *imagined speech* [12]). Speech imagery is about representing speech in terms of sounds inside the human brain without overt vocalization nor articulatory movements. We hypothesize the existence of some sort of brain footprint for articulatory movements underlying related speech token imagery. Hence, we attempt to first predict phonological categories and then use these predictions to aid recognition of imagined speech at the token level (phonemes and words). We introduce our framework for solving this problem next.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-3041

2. Proposed Deep Learning Framework

2.1. Mathematical Formulation

We denote the multivariate time-series data as $X \in R^C * T$, with sets of labels $Y \in y_1, y_2, ..., y_{11}$ where X corresponds to the single trial EEG data, having a number of channels C, and for a number of time steps T. Y is a one-hot encoded vector of 11 labels corresponding to individual words and labels. In our case, C is 64 and time interval is represented in terms of 5,000 time steps. As discussed earlier, we essentially build our system in two consequent steps: The first step is binary classification of $X \in R^C * T$ into presence or absence of 6 phonological categories: $\{z_1, \bar{z_1}\}, \{z_2, \bar{z_2}\}, \{z_3, \bar{z_3}\}, \{z_4, \bar{z_4}\}, \{z_5, \bar{z_5}\}, \{z_6, \bar{z_6}\}$. The second step is tease apart the concatenated autoencoder latent vectors from these 6 classification models *viz.*, $W = \bigcup_{i=1}^{6} w_i$ into 11 classes: $\{y_1, y_2, ..., y_{11}\}$ where w_i corresponds to latent vector space corresponding to i^{th} phonological classification into $\{z_i, \bar{z}_i\}$.

2.2. Predicting Phonological Categories

We build on our hypothesis that the active thought process underlying covert speech does have some relevant features corresponding to the intended activity of nasopharynx, lips, tongue movements and positions etc. Hence, in the first phase, we target five binary classification tasks addressed in [17, 18], *i.e.* presence/absence of consonants, phonemic nasal, bilabial, high-front vowels and high-back vowels. Additionally, we add a voiced vs. voiceless classification task whose goal is to provide information about the intended involvement of vocal folds. In this way, rather than directly discriminating the individual phonemes and words, we first attempt to accurately classify imagined phonological categories on the basis of underlying intended articulatory movements.

Rather than using the raw multi-channel high-dimensional EEG data (which requires long training times and resources), we experimentally[1] found that it is a better strategy to first reduce the dimensionality of the EEG by capturing the joint variability of the electrodes. Crucially, our target was to model the directional relationship and dependency among the electrodes over the entire time interval. Hence, instead of the conventional approach of selecting a handful of channels as in [17, 18], we address this issue by computing the channel cross-covariance (CCV), resulting in positive, semi-definite matrices encoding the connectivity of the electrodes. We define CCV between any two electrodes c_1 and c_2 as: $Cov(X_t^{c_1}, X_{t+\tau}^{c_2}) = \mathbb{E}[X^{c_1}(t) - \mu_{X^{c_1}}(t)][X^{c_2}(t+\tau) - \mu_{X^{c_2}}(t+\tau)]$.

We use convolutional neural networks (CNNs) [19] to extract the spatial features from the covariance matrix. Each layer decodes non-linear spatial feature representations from the previous layer using convolutional filters and non-linear ReLU [20] activation functions applied to the resulting feature maps. We employ a four-layered 2D CNN stacking two convolutional and two fully connected hidden layers. This is the first level of hierarchy where the network is trained with the corresponding labels as target outputs, optimizing cross-entropy cost function. We describe architectural and hyper-parameter choices for our networks in Table 1.

In parallel with CNN, we apply a temporal CNN (TCNN) [21, 22] on the channel covariance matrices to explore the hidden temporal features of the electrodes. Namely, we flatten the lower triangular matrix of the CCV and feed the data of length

[1] We do not report these experiments here, for space limitation.

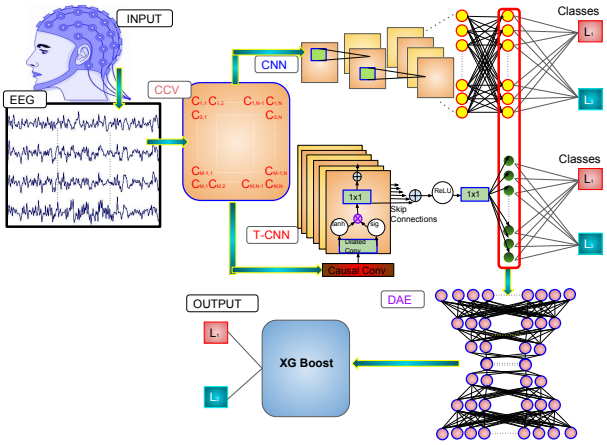

Figure 2: *Overview of phonological prediction of our novel architecture*

1,891 to the TCNN. In order to capture the long term dependencies and temporal correlations of the signal, we exploit a 6 layer stacked TCNN and train in a similar manner as CNN, using Adam [23] to optimize cross-entropy function. We use stacked dilation filters with a dilation factor of 2, resulting in exponential growth of receptive field with depth and increase in model capacity. This essentially enhances the non-linear discriminative power of the network, which is vital for our problem space. We concatenate the last fully-connected layer from the CNN with its counterpart in the TCNN to compose a single feature vector based on these two penultimate layers thereby forming a joint spatio-temporal encoding of the cross-covariance matrix. In order to further reduce the dimensionality of the spatio-temporal encodings and cancel background noise effects [24], we train an unsupervised deep autoenoder (DAE) [25] on the fused heterogeneous features produced by the combined CNN and TCNN information. The DAE forms our second level of hierarchy, with 3 encoding and 3 decoding layers, and mean squared error (MSE) as the cost function.

At the third level of hierarchy, the discrete latent vector representation of the deep autoencoder is fed into an Extreme Gradient Boost based classification layer [26, 27] motivated by [24]. The classifier receives its input from the latent vectors of the deep autoencoder and is trained in a supervised manner to output the final predicted phonological classes corresponding to speech imagery.

2.3. Predicting Speech Tokens

Next, our goal is to use the combined information available from all the six phonological categories to predict the 11 individual speech tokens present in our EEG dataset (introduced in Section 3.1). Such a hierarchical approach essentially differs from the direct speech classification approach as it imposes richer constraints on the information space by involving features from all the phonological categorization tasks. Our results show the utility of this approach as we report in Section 3.4. To this end, we first stack the bottleneck features of the autoencoders corresponding to the aforementioned six classification tasks, into a matrix of dimensions 6×256. In order to explicitly exploit phonological information in the imagined speech recognition task, we feed this stacked latent matrix as the input to our classification model similar to the first phase.

Table 1: *Selected parameter sets*

Parameters	CNN	TCNN	DAE
Epochs	50	50	200
Total layers	6	6	7
Hidden layers' details	Conv:32,64 masks:3x3 Dense: 64,128	mask: 5, Dilation : 2	E:1024,512,128 D:128,512,1024
Activations	ReLU, last-layer : softmax	sigm, tanh, ReLU, last-layer : softmax	ReLU, ReLU, sigm, sigm, ReLU, tanh
Dropout	.25, .50	.25, .50	.25, .25, .25
Optimizer	Adam	Adam	Adam
Loss	Categorical cross entropy	Categorical cross entropy	Mean Sq Error
l-rate	.001	.002	.001

3. Experiments

3.1. Dataset

We evaluate our models on a publicly available dataset, KARA ONE [17]. It is composed of multimodal data for stimulus-based, imagined and articulated speech state corresponding to 7 phonemic/syllabic (/iy/, /piy/, /tiy/, /diy/, /uw/, /m/, /n/) as well as 4 words (pat, pot, knew and gnaw). The study comprising the dataset consists of 14 participants, with each prompt presented 11 times to each individual. Since our intention is to classify the phonological categories from human thoughts, we discard the facial and audio information and only consider the EEG data corresponding to imagined speech. More details regarding the database can be found in [17].

3.2. Procedure and Model Training

We randomly shuffle and divide the data (1913 signals from 14 individuals) into train (80%), development (10%) and test sets (10%). The architectural parameters and hyperparameters listed in Table 1 were selected through an exhaustive grid-search based on the development set. We conduct a series of empirical studies starting from single hidden-layered networks for each of the blocks and, based on the validation accuracy, we increase the depth of each given network and select the optimal parametric set from all possible combinations of parameters. For the gradient boosting classification, we fix the maximum depth at 10, number of estimators at 5,000, learning rate at 0.1, regularization coefficient at 0.3, subsample ratio at 0.8, and column-sample/iteration at 0.4. We did not find any notable change of accuracy while varying other hyperparameters while training gradient boost classifier. For the phonological categorization task, input data for CNN and TCNN (covariance matrix) is of length 61×61 and 1,891 respectively, while for the speech recognition task, the input data (phonological features) is of length 6×256 and 1,536 respectively. The input data for deep autoencoders pertaining the two tasks is of length 2,915 (1,891 TCNN + 1,024 CNN features).

3.3. Baselines

We use two baselines, one based on an individual LSTM and another based on an individual CNN. In each case, we pass the data from the cross-variance matrix and classify directly based on output from each of these networks. In addition, we compare to previous works on the same dataset [17, 18]. For meaningful comparisons, since these previous works follow a cross-validation set up (14-fold where the model is trained on 13 subjects' data and tested on the 14^{th}), we mimic the same data splits and report accuracy. To establish a benchmark for computationally costly deep learning work, we choose our 80%, 10%, 10% data splits after shuffling the data.

Table 2: *Results in accuracy on 10% test data for phonological prediction.* ***C-L-D:*** *CNN+LSTM+DAE*

Method	± Bilab	± Nasal	C/V	± /uw/	± /iy/	Avg
LSTM	46.07	45.31	45.83	48.44	46.88	46.51
CNN	59.16	57.20	67.88	69.56	68.60	64.48
CNN+LSTM	62.03	60.89	70.04	72.76	63.75	65.89
C-L-D	78.65	74.57	87.96	83.25	77.30	80.35
Our model	**81.67**	**78.33**	**89.16**	**85.00**	**87.20**	**84.27**

Table 3: *Classification Performance metrics on 10% test data in phonological prediction task*

Metrics	Precision	Recall	Specificity	f1 score	Kappa
± Bilab	72.09	75.61	84.81	73.81	63.34
± Nasal	67.44	70.73	82.28	69.05	56.66
C/V	86.36	65.52	96.7	74.51	78.32
± /uw/	77.27	56.67	94.44	65.39	70.00
± /iy/	86.04	78.72	91.78	82.22	74.40
± Voiced	78.95	86.96	68.63	82.76	58.32

3.4. Results of phonological category prediction

To demonstrate the significance of the hierarchical CNN-TCNN-DAE method, we also conduct separate experiments with the individual networks and summarize the results in Table 2. From the average accuracy scores, we observe that our proposed network performs much better than individual networks. A detailed analysis on repeated runs further shows that in most of the cases, LSTM alone does not perform better than chance. CNN, on the other hand, is heavily biased towards the class label from which it sees more training data. Although the situation improves with combined CNN-LSTM, our analysis clearly shows the necessity of a better encoding scheme to utilize the combined features rather than mere concatenation of the penultimate features of both networks. CNN-LSTM-DAE improves classification accuracy by a significant margin, thus demonstrating the utility of the autoencoder contribution towards filtering out the unrelated and noisy features from the concatenated penultimate feature set. Replacing the LSTM block with TCNN block endows the network with more temporal discriminative power, resulting in an increase of 3.93% mean accuracy as shown in Table 2. In addition to accuracy, we provide the precision, recall, specificity, f1 score and Kappa coefficients of our method for all the six classification tasks in Table 3. Kappa coefficients offer a metric for evaluating the utility of classifier decisions beyond mere chance [28]. Here, a higher mean kappa

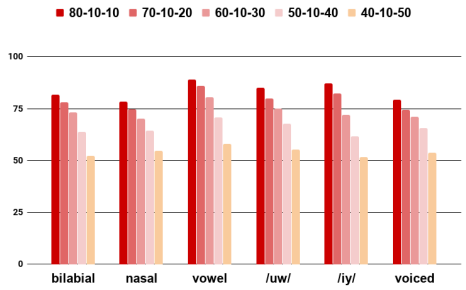

Figure 3: *Variation of performance accuracy of phonological prediction with varying training-validation-test data ratio*

Figure 4: *Inter-subject confusion matrix for speech token prediction with covariance data (left) and with phonological feature data (right)*

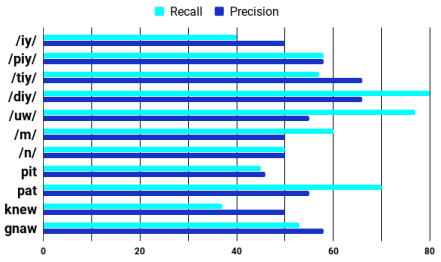

Figure 5: *Precision and recall metrics corresponding to each speech token on 10% train data*

value corresponding to a task implies that the network is able to find better discriminative information from the EEG data beyond random decisions. The maximum above-chance accuracy (78.32%) is recorded for presence/absence of the vowel task and the minimum (56.66%) is recorded for the $\pm nasal$.

Further, to evaluate the robustness of our model against availability of data, we run a set of experiments varying the train-test ratio of the data (results shown in Figure 3). As Figure 3 shows, even with less training data (40%) and more, and potentially more diverse test data (50%), our model performs above chance, which indicates its reliability even under these extreme data distribution condition.

We next compare our phonological prediction to [17] and [18]. As shown in Table 4, since the model encounters the unseen data of a new subject for testing, and given the high inter-subject variability of the EEG data, a reduction in the accuracy is expected. However, our network still manages to achieve an improvement of **18.91**, **9.95**, **67.15**, **2.83** and **13.70 %** over [17]. Besides, our best model shows more reliability compared to previous works: The standard deviation of our model's classification accuracy across all the tasks is reduced from 22.59% [17] and 17.52% [18] to a mere 5.41%.

3.5. Results of speech token prediction

We provide performance of the baseline methods on direct covariance data and phonological feature data in Table 5. For a closer look at the results, we report sample confusion matrix of our model on a leave-one-subject-out classification strategy in Figure 4. In this step, we essentially train the network on the data of 13 subjects and test on the 14^{th} subject, to check the inter-subject variability of our model. As it is evident from the figure, with direct covariance data, the predicted classes corresponding to each true label are widely distributed throughout the matrix and hardly gives any significant information about the actual speech token. However, involvement of the phonological categorization as an intermediate step increases the prediction accuracy. Interestingly, the false negatives corresponding to each of the tokens also inform us about the respective structure of the word or phoneme. For example, the misclassification of `/n/` as `/m/`, `'knew'` and `'gnaw'` in a few cases, show that while the network gets strong discriminative features from the other five networks, features pertaining to the nasal category require more discriminative ability to more accurately categorize the phoneme `/n/`. Such an observation indeed proves that the phonological features play a significant role for achieving an accurate classification of the speech tokens.Furthermore, Figure 5 records the precision and recall scores of all the speech tokens on 80-10-10 train-dev-test split. In Figure 6, we again vary the train-test ratio of data and present the performance accuracy for speech token prediction corresponding to the top 4 models as indicated in Table 5.

Table 4: *Comparison in accuracy with Z_R: [17] and S_Q: [18]*

	± Bilabial	± Nasal	C/V	± /uw/	± /iy/
Z_R	56.64	63.5	18.08	79.16	59.6
S_Q	53	47	25	74	53
Ours	**75.55**	**73.45**	**85.23**	**81.99**	**73.30**

Table 5: *Comparison of accuracy on 10% test data for speech token prediction task*

Method	EEG data	Phonological features
LSTM	8.45	15.83
CNN	8.88	16.02
CNN+LSTM	12.44	22.10
CNN+LSTM+DAE	23.45	49.19
Our model	**28.08**	**53.36**

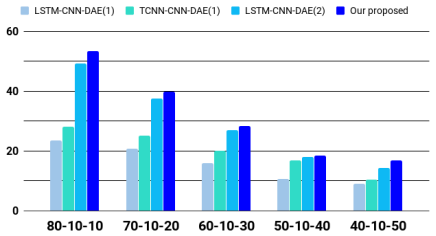

Figure 6: *Variation of performance accuracy of speech token prediction for top 4 algorithms with varying training-validation-test data ratio*

4. Conclusion and Contribution

We report a novel hierarchical deep neural network architecture composed of parallel spatio-temporal CNN and a deep autoencoder for phonological and speech token prediction from imagined speech EEG data. Overall, we made the following contributions: (1) we proposed a novel method for embedding the high dimensional EEG data into a cross-covariance matrix that captures the joint variability of the electrodes. Rather than attempting to directly decode speech thoughts into speech tokens, (2) we exploited the cross-covariance matrix to successfully classify the phonological attributes of these thoughts into 6 categories; and (3) we used these predicted phonological categories to identify speech tokens. Ultimately, (4) our work suggests the existence of a brain imagery footprint for underlying ariculatory movements representing speech tokens.

5. References

[1] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[2] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for eeg-based motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 4, pp. 317–326, 2008.

[3] S. Machado, F. Araújo, F. Paes, B. Velasques, M. Cunha, H. Budde, L. F. Basile, R. Anghinah, O. Arias-Carrión, M. Cagy *et al.*, "Eeg-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation," *Reviews in the Neurosciences*, vol. 21, no. 6, pp. 451–468, 2010.

[4] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.

[5] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[6] C. S. DaSalla, H. Kambara, Y. Koike, and M. Sato, "Spatial filtering and single-trial classification of eeg during vowel speech imagery," in *iCREATe '09*. ACM, 2009, p. 27.

[7] B. M. Idrees and O. Farooq, "Vowel classification using wavelet decomposition during speech imagery," in *SPIN, 2016*. IEEE, 2016, pp. 636–640.

[8] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "Eeg classification of imagined syllable rhythm using hilbert spectrum methods," *Journal of neural engineering*, vol. 7, no. 4, p. 046006, 2010.

[9] J. Kim, S.-K. Lee, and B. Lee, "Eeg classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition," *Journal of neural engineering*, vol. 11, no. 3, p. 036010, 2014.

[10] K. Brigham and B. V. Kumar, "Imagined speech classification with eeg signals for silent communication: a preliminary investigation into synthetic telepathy," in *iCBBE, 2010*. IEEE, 2010, pp. 1–4.

[11] K. Mohanchandra and S. Saha, "A communication paradigm using subvocalized speech: translating brain signals into speech," *Augmented Human Research*, vol. 1, no. 1, p. 3, 2016.

[12] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery eeg for bci," *Biomedical signal processing and control*, vol. 8, no. 6, pp. 901–908, 2013.

[13] E. F. González-Castañeda, A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, "Sonification and textification: Proposing methods for classifying unspoken words from eeg signals," *Biomedical Signal Processing and Control*, vol. 37, pp. 82–91, 2017.

[14] M. DZmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward eeg sensing of imagined speech," in *HCI*. Springer, 2009, pp. 40–48.

[15] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from eeg," *AAAI, 2019*. 2 pg abstract.

[16] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the eeg manifold for phonological categorization from active thoughts," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2762–2766.

[17] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *ICASSP, 2015*. IEEE, 2015, pp. 992–996.

[18] P. Sun and J. Qin, "Neural networks based eeg-speech models," *arXiv:1612.05369*, 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[21] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio." in *SSW*, 2016, p. 125.

[22] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

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

[24] X. Zhang, L. Yao, Q. Z. Sheng, S. S. Kanhere, T. Gu, and D. Zhang, "Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals," in *2018 PerCom*. IEEE, 2018, pp. 1–10.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[26] T. Chen, T. He, M. Benesty *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[28] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of eeg motor imagery signals," *Journal of neural engineering*, vol. 14, no. 1, p. 016003, 2016.