Using articulated speech EEG signals for imagined speech decoding

Bras, Chris; Patel, Tanvina; Scharenborg, Odette

DOI
10.21437/Interspeech.2024-1289

Publication date
2024

Document Version
Final published version

Published in
Interspeech 2024

Interspeech 2024
1-5 September 2024, Kos, Greece

Using articulated speech EEG signals for imagined speech decoding

Chris Bras, Tanvina Patel, Odette Scharenborg

Multimedia Computing Group, Delft University of Technology, the Netherlands
chris.s.bras@gmail.com, tanvina.patel@gmail.com, o.e.scharenborg@tudelft.nl

Abstract

Brain-Computer Interfaces (BCIs) open avenues for communication among individuals unable to use voice or gestures. Silent speech interfaces are one such approach for BCIs that could offer a transformative means of connecting with the external world. Performance on imagined speech decoding however is rather low due to, amongst others, data scarcity and the lack of a clear starting and end point of the imagined speech in the brain signal. We investigate whether using electroencephalography (EEG) signals from articulated speech can be used to improve imagined speech decoding in two ways: we investigate whether articulated speech EEG signals can be used to predict the end point of the imagined speech and use the articulated speech EEG as extra training data for speaker-independent imagined vowel classification. Our results show that using EEG data from articulated speech did not improve classification of vowels in imagined speech, probably due to high variability in EEG signals amongst speakers.

Index Terms: Brain computer interfaces, covert (imagined) speech, electroencephalography (EEG).

1. Introduction

Neurodegenerative disorders such as Amyotrophic Lateral Sclerosis (ALS) or conditions like locked-in syndrome frequently result in profound muscular impairment, rendering patients incapable of voluntary muscle movement and consequently unable to articulate speech [1]. This profound physical debilitation presents significant obstacles for individuals affected by these conditions when attempting to engage in effective communication with their external environment.

Brain-computer interfaces (BCIs) have emerged as a potential avenue to address this issue [1]. By analysing brain activity, BCI systems could facilitate communication solely based on the patient's thoughts. A promising approach in this context involves using imagined speech (*covert speech*), wherein an individual imagines to produce speech without any muscle movement nor audible or articulated speech. By decoding and interpreting these neural signals, BCIs hold promise for enabling communication in patients e.g., affected by ALS and multiple sclerosis (MS), bypassing the physical limitations imposed by their conditions. Different methods are used to capture the electrical signals generated by neural activity, including electrocorticography (ECoG) [2], Magnetoencephalography (MEG) [3], and Electroencephalography (EEG) [4, 5], which can be analyzed to understand both imagined and articulated speech.

Research on imagined speech from EEG has focused on classifying small sets of stimuli, e.g., vowels (English [6], Dutch [7, 8], Japanese [9], Spanish [10]) and isolated words ("yes" and "no" [11], nine Russian words [12]). For the task of classifying EEG data of imagined speech, many different machine and deep learning techniques have been used, including, support vector machine [13], linear discriminant analysis [14], random forest [15], vanilla deep neural networks (DNNs) [16], and convolutional neural network (CNN) [17, 18]). However, DNNs require large amounts of data to properly generalize for a given problem without having issues with overfitting [19] which often is not available for this type of EEG data. Moreover, different discriminative features extracted from the EEG signals have been used (e.g., wavelet domain features [20, 21] and common spatial patterns (CSP) [22, 23]). Nevertheless, no combination of classifier and features has proven to consistently achieve high decoding performances [18]; although Residual Network (ResNet) algorithms [12, 16] have been found to outperform other CNN algorithms on speaker-dependent imagined speech classification tasks in both robustness and practicability.

Nevertheless, classification of imagined speech using EEG remains a challenging task, with classification results close to chance level [24]. Being a non-invasive solution, the recording made using an EEG is not optimal, inducing noise from, for example, blinks or muscle movement in the data [25]. Moreover, there is a lack of imagined speech EEG data. Although different databases have been released [6, 8, 15, 26], they differ in the number of speakers, language, the absence or presence of articulated speech, and differences in the recording set-up, e.g., the number of channels used to record the EEG signals (e.g., 6 for Coretto et al[15] and 62 for DAIS [8]). On top of that, EEG signals vary a lot, especially between different subjects [27]. This makes it difficult for models to generalize and validate their performance. A further difficulty is the lack of an accurate ground truth: it is not easily verifiable if the subject performed the imagined task correctly, as is the case with articulated speech. A roundabout way of testing whether participants complied with the task of imagining speech is to visually investigate whether structural differences exist between the event-related potentials (ERPs) of the EEG signals for rest and imagined speech or run a classification task predicting whether an EEG signal came from the rest state or imagined speech [8].

In this work, we focus on speaker-independent classification of EEG signals from imagined speech. Research suggests that imagined speech production can be seen as interrupted articulated speech production without the actual muscle movement required for producing sound [28]. Therefore, we investigate whether EEG data captured from articulated speech can be used to improve classification accuracy and generalization of EEG data captured from imagined speech in two ways: we investigate whether articulated speech EEG signals can be used to predict the end point of the imagined speech in the EEG signal and whether using the articulated speech EEG as additional training data improves speaker-independent imagined

10.21437/Interspeech.2024-1289

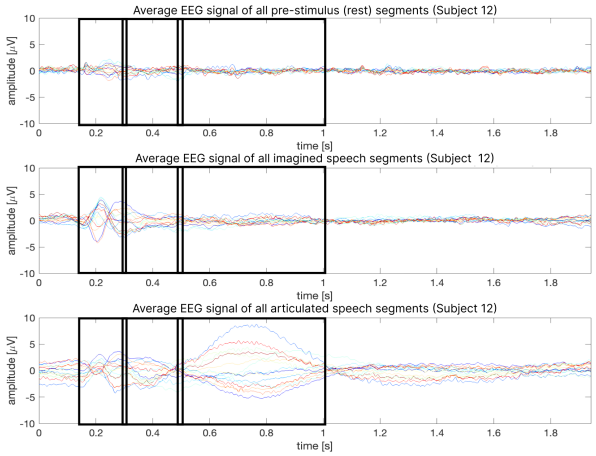

Figure 1: *Averaged event related potentials (ERP) for Participant 12 from the DAIS dataset for rest (top panel), imagined (middle panel) and articulated (bottom panel) speech[8].*

vowel classification. If successful, articulated speech EEG data from other speakers than the patient/user could be used to improve performance. In all experiments, we compare performance on different numbers of channels, to investigate whether there is an influence of the number and location of the electrodes from which the EEG signals are collected on imagined and articulated speech EEG classification. In the first experiment, we aim to predict the starting point of the articulated speech and use that as the end point of the imagined speech, and use this "pre-speech" part of the EEG signal to classify the EEG signals. In the second experiment, we add the articulated speech EEG as training data to the imagined speech EEG data for speaker-independent Dutch imagined vowel classification from EEG.

2. Methodology

2.1. Database

This paper uses the Delft Articulated and Imagined Speech (DAIS) dataset [8], which consists of EEG signals of imagined and articulated Dutch and speech from 20 native Dutch subjects, 6 male and 14 female [8]. The subjects were asked to imagine and articulate speech of 15 prompts: five vowels (a:, e:, i, o:, u where ":" indicates long vowels) and 10 words. The 5 vowels constitute the different corners of the Dutch vowel quadrant. The 10 words are 5 Dutch word-pairs that are also words when read backwards: taal, laat, leeg, geel, niet, tien, toon, noot, soep, poes (Eng: "language", "late", "empty", yellow", "not", "ten", "tone", "note", "soup", and "cat"). Each vowel is part of one word pair. The EEG was recorded over 62 channels, placed according to the standard 10-20 international system [29] using the TMSi SAGA 64+ and with a BrainWave EEG Cap at a sampling frequency of 1024 Hz and the TMSi SAGA interface for MATLAB. The SAGA docking station was located outside the sound-attenuating room. Impedances were kept below $50k\Omega$. The audio is sampled at 44.1 kHz.

Each participant completed 20 runs of 15 trials, one for every prompt (i.e., the 15 Dutch vowels and words), where a trial consisted of 4 successive segments: rest, reading of the prompt, imagining to produce the prompt, and articulating the prompt. Each run was followed by (another) 2s rest. Each EEG recording is divided into 2s segments, for each prompt [8].

2.2. Data Pre-processing: Filtering

First, the EEG data was band-pass filtered (Second order Butterworth filter) between 1 and 40 Hz to limit any electrical noise, such as power line noise, present in the signal. Artifacts such as blinks and muscle movement are removed where possible by using the low pass filter. If it is not possible to remove the artifact from the sample, the sample is discarded. Lastly, when dropping channels from the EEG data for different experiments, the channels must be re-referenced to each other by subtracting the average of all remaining channels combined from each channel.

2.3. Model Architecture

Based on pilot experiments in which we compared support vector machine (SVM), K-nearest neighbour (KNN) and random forest (RF) machine learning models as well as Long short-term memory (LSTM) and Convolutional Neural Networks (CNN) deep learning models on articulated speech vowel classification, we chose the best-performing model for the experiments reported here. The model used is a CNN model with an input size of 2048 timesteps with a variable number of features. The feature count per timestep depends on the number of EEG channels used. The model consists of three repeated convolutional layers followed by global average pooling and finally a fully connected prediction layer, similar to commonly used convolutional models for time series classification [30]. Each convolutional layer has batch normalization and a dropout of 25% to prevent overfitting with rectified linear activation at each layer.

Two versions of the CNN model were created: one for the classification task (Experiment 2 below) and one for the start-of-articulated-speech/end-of-imagined speech detection task (Experiment 1 below). The speech detection model, being a regression model, has one single output neuron that outputs the estimated start of speech time, while the classification model has one neuron for each vowel class to predict.

2.4. Predicting the End Point of Imagined Speech

Figure 1 shows the averaged ERP for participant #12 from the DAIS dataset for rest (top panel), imagined (middle panel) and articulated (bottom panel) speech [8]. Comparing the three ERP signals shows clear differences between the EEG data of the rest segment, imagined speech, and articulated speech. For rest (top panel) only background EEG activity was found. Important to our experiments, for both imagined and articulated speech activity was observed around 0.25 - 0.3 seconds, which was followed by a broad peak/trough (depending on the channel) starting at 500 ms for articulated speech. This corresponds to start of the articulated speech and is therefore associated with the movement of the articulators.

The aim is to predict the point where imagined speech ends. While we do not have a ground-truth for the endpoints in imagined speech, for articulated speech, we know when speech starts, and the time stamps of the speech are aligned with the EEG signal of the articulated speech. We assume that the starting point of articulated speech is the end of the preparations to speak, and we take that as the endpoint of the imagined speech. To determine at which timestamp speech starts for each data sample, SileroVAD, a voice activity detection (VAD) algorithm [31], is applied to the DAIS speech files. From this, a timestamp is extracted at which speech starts in the articulated speech. The period in the EEG signal prior to this point we refer to as "pre-speech". A CNN model was trained for 150 epochs with articulated speech EEG signal segments as input and the

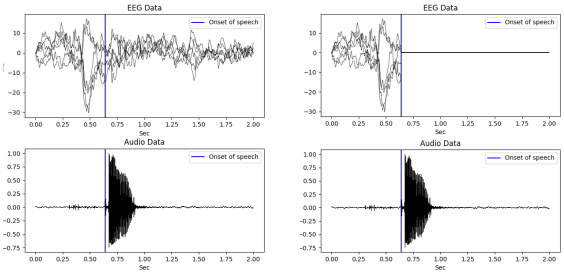

Figure 2: *Left panels: EEG segment of articulated speech (top) and the acoustic signal (bottom); Right panels: EEG signal of the "pre-speech", after cutting the segment at the start of speech (vertical line).*

timestamps obtained from the VAD as target. All code used in these experiments can be found at `https://github.com/ChrisSBras/imagined_vs_articulated_speech`.

Figure 2 shows an example of a segment of articulated speech (bottom left) and the accompanying EEG signal (top left). The blue vertical line marks the onset of speech as provided by the VAD. The top right panel shows the EEG signal after removing the EEG signals after the speech onset, i.e., it only shows the "pre-speech" EEG signal.

3. Experimental Setup

For the experiments we used the vowel data /a:, e:, i, o:, u/. Data of five participants were excluded: Participants 9 and 13 were excluded because they are left-handed, participants 7 and 17 were excluded because their signals contained multiple noisy channels, and participant 2 because a large part of the articulated speech trials were rejected as they contained eye blinks, causing an imbalance in the number of covert speech trials vs. the articulated speech trials. For participant 1, channel FC2 disconnected during the experiment and was deleted. For the other 19 participants, data from all 62 EEG-channels is available.

After dropping faulty segments and subjects, a total of 1291 articulated speech segments and 1412 imagined speech segments remain. This gives per vowel an average of 258 segments for articulated speech and 282 segments for imagined speech and per subject an average of 86 segments for articulated speech and 94 segments for imagined speech.

The experiments were run in a speaker-independent scenario. Articulated and imagined data are both split in an 80% training split and 20% test split. This results in 1039 articulated and 1130 imagined speech training samples and 252 articulated speech and 282 imagined speech test samples. These sets are added together for the combined data experiments, resulting in a total of 2169 training samples and 534 test samples for those experiments. The same training and test sets were used for both the detection and classification experiment. Each experiment is repeated 5 after which an average accuracy is computed.

3.1. EEG Channel Selection

In both the speech detection and vowel classification experiments, we compare performance on the vowel EEG signals from four sets of electrodes:

- ***Channel Set 6***: 6 channels {F3, F4, C3, C4, P3, P4}. Following the Coretto database [15], we use 6 channels, which is the lowest number used in any database of imagined speech.
- ***Channel Set 8***: 8 channels {Fz, C3, Cz, C4, Pz, PO7, Oz, PO8} as used in [5]. These channels are chosen as they are close to Broca's and Wernicke's regions of the brain, which is assumed to produce good quality imagined speech-based EEG signals [5].
- ***Channel Set 16***: 16 channels {F7, F5, FT7, FC5, FC3, FC1, T7, C5, C3, Cz, C4, TP7, CP5, CP3, P5, P3} that are located on specific areas of the cortex that are known to be involved in language processing. These 16 channels were also used in the validation study reported in [8].
- ***Channel Set 62***: All available channels are used.

Each channel subset is run 5 times for each combination of train and test data (articulated, imagined and combined data) as discussed in Section 3.3.

3.2. Predicting end point of imagined speech

In the first experiment, we aim to predict the starting point of the articulated speech. Four models were trained to predict the start of the speech signal in the EEG of the articulated speech, one model for each channel set. The same training and test data are used for the different channel sets. We evaluate the models' performance on the articulated speech test set, in terms of the Mean Squared Error (MSE) of the difference in timestamp between the target timestamp and predicted timestamp in milliseconds, calculated over the five runs of each model.

3.3. Using articulated speech EEG for improving imagined speech EEG classification

We ran several experiments: 1) we trained and tested on the articulated speech EEG to set an upper-bound for the task of imagined Dutch vowel classification from EEG; 2) we trained and tested on the imagined speech EEG to set a baseline; 3) we carried out two cross-experiments where the model trained on articulated speech EEG is tested on the imagined speech EEG and vice versa. 4) To investigate whether using EEG from articulated speech during training improves imagined vowel classification, we combined the articulated speech EEG training data and the imagined speech EEG training data to train a combined model for each channel set. This model was tested on both the imagined and articulated speech test sets.

Under the assumption that the pre-speech part for imagined and articulated speech is similar, we predicted the end point of the imagined speech from the imagined speech EEG using the prediction model, and used the pre-speech part of the imagined speech and articulated speech EEG to train the four models. We then ran the same experiments as done with the full EEG signal, but only using the pre-speech of the imagined and the pre-speech of the articulated speech EEG. The models are evaluated using accuracy on the 5-vowels classification task.

4. Experimental Results

4.1. Predicting the End point of Imagined Speech

Table 1 shows the MSE and standard deviation ($Stdev$) of the difference (in ms) between the ground-truth timestamp of start time in the articulated speech EEG signal and the predicted timestamp for the four models with different channel sets. First, with increasing number of channels, the MSE and Stdev reduce, although the smallest standard deviation was found for the model with only 6 channels. Importantly, all models show relatively good prediction results, with a maximum MSE of only 5.65 ms, which is a lot less than the duration of a single sound.

Table 1: *The MSE and $Stdev$ of the predicted start of the articulated speech (in ms).*

Channels	**Set 6**	**Set 8**	**Set 16**	**Set 62**
MSE (ms)	5.65	5.20	4.17	2.55
std dev (ms)	0.23	0.53	0.46	0.41

This indicates that the start of speech can be predicted from the EEG signal of articulated speech with a reasonable error.

4.2. Using articulated speech EEG for improving imagined speech EEG classification

Tables 2 and 3 show the classification results of the different models on the EEG of articulated and imagined speech in terms of accuracy together with the standard deviation for each channel subset when the full EEG signal is used (Table 2) and when only the pre-speech is used (Table 3) for the four different channel sets. The results are grouped by the type of training and test data used (articulated, imagined or a combination) and the number of channels.

The results for experiment 1 show a baseline of 53.8% accuracy for speaker-independent articulated Dutch vowel classification when all available channels and the full EEG segments are used. This can be viewed as an upperbound for the imagined vowel classification task. This performance drops to 43.9% when only the pre-speech part of the segments is used. The information related to articulation in the EEG signal is thus needed for improved classification of articulated vowels. The results for experiment 2 show a baseline of 24.8% accuracy for speaker-independent imagined Dutch vowel classification when all available channels and the full EEG segments are used. This increases to 27% accuracy when only using the pre-speech part of the EEG segment, although this still falls within 1 standard deviation from the baseline. Chance level in all cases is 20%.

The results from experiment 3 show a worse performance for models trained on the other type of data than with which they are evaluated. One thing to note for these cross-experiments however, is that the pre-speech only models perform better than the models trained on full EEG segments.

The final experiment investigated whether combining training data from articulated and imagined speech was beneficial for vowel classification. The results show that overall there is little benefit when training on both articulated and imagined speech EEG for classification of both imagined and articulated speech, with the exception of the imagined speech all-channels full-EEG model, which in fact gave the best performance on imagined vowel classification across all models.

Table 2: *Accuracy (in %) of the classification experiments using the full EEG frames, grouped by channel sets used.*

Test	**Training**	6 Chan.	8 Chan.	16 Chan.	All Chan.
art.	*art.*	28.4 ± 1.0	31.0 ± 0.9	45.5 ± 2.2	53.8 ± 4.9
art.	*img.*	25.5 ± 0.8	27.4 ± 1.4	26.5 ± 1.7	30.1 ± 4.1
art.	*combined*	28.8 ± 2.4	30.3 ± 1.8	42.0 ± 2.2	43.2 ± 2.9
img.	*art.*	25.3 ± 1.7	23.9 ± 0.5	23.6 ± 1.3	24.8 ± 2.1
img.	*img.*	24.1 ± 1.6	24.8 ± 1.2	27.1 ± 2.7	24.8 ± 1.8
img.	*combined*	25.8 ± 2.0	24.9 ± 1.6	25.1 ± 1.4	27.5 ± 1.7
combined	*combined*	26.1 ± 0.8	25.5 ± 1.5	30.5 ± 1.3	30.4 ± 1.0

5. Discussion and Conclusion

The aim of this paper is to use articulated speech EEG data to improve classification results on imagined speech EEG data. An onset of articulated speech detection model is created which is then used to mark the end of imagined speech in the imagined speech EEG segments. The experiment investigating the viability of such a model showed potential for a speech detection model based on EEG signals instead of audio signals when used on articulated speech EEG data in a speaker-independent way. We refer to the EEG signal before articulation onset as "pre-speech". One assumption made in this paper, is that this can then be used to mark the point in the EEG data where imagined speech ends. Due to the lack of ground truth data, we however cannot validate this assumption. Future research could investigate the performance of our approach in a speaker-dependent scenario, which will reduce the between-speaker variability in the EEG signals.

Table 3: *Accuracy (%) of the classification experiments using the pre-speech, grouped by channel sets used.*

Test	**Training**	6 Chan.	8 Chan.	16 Chan.	All Chan.
art.	*art.*	29.7 ± 1.7	32.5 ± 1.6	37.3 ± 3.3	43.9 ± 2.3
art.	*img.*	27.5 ± 2.7	25.8 ± 1.6	27.6 ± 2.7	29.4 ± 1.2
art.	*combined*	28.0 ± 2.1	29.8 ± 1.9	33.9 ± 2.5	39.4 ± 2.9
img.	*art.*	23.9 ± 0.8	24.4 ± 1.0	23.9 ± 1.7	26.1 ± 2.6
img.	*img.*	24.9 ± 1.3	25.6 ± 0.9	25.7 ± 1.4	27.0 ± 2.4
img.	*combined*	24.1 ± 0.8	25.9 ± 1.6	25.2 ± 0.8	25.9 ± 0.9
combined	*combined*	24.8 ± 0.7	26.2 ± 1.6	26.8 ± 1.6	27.8 ± 1.4

In a second experiment, we investigated whether combining articulated and imagined speech EEG improved imagined speech vowel classification, and whether there was a difference when using the full EEG signal or only the pre-speech EEG signal as a way to increase training data for imagined vowel classification. First, overall, our results on speaker-independent imagined vowel classification are in line with other research where the focus is on speaker-independent models [32].

Second, from the results of experiment 3 in Tables 2 and 3, it can be seen that there is usable information in the pre-speech part of articulated speech EEG data as all results are above chance level. Nevertheless, in general using pre-speech only did not improve vowel classification of imagined nor articulated speech. However, there are two interesting findings. For the all-channels model trained on imagined vowel EEG and tested for imagined vowel classification, using only the pre-speech gave an improvement over using the full EEG, suggesting there is information in the full signal that is unnecessary and reduces performance. This is in line with the second interesting finding: the 16-channel model trained on imagined speech outperforms the all-channels model on imagined speech vowel classification. Also here, using less data improved imagined vowel classification. Future research should focus on investigating what information is necessary for improved imagined vowel classification and what information is better removed from the EEG signal.

The lack of an improvement when using pre-speech EEG is not in line with our assumptions based on the literature: the difference between imagined speech EEG and the phase before speech in articulated speech EEG is too different to be used together to train a classifier. Although, of course, we cannot exclude that our speech onset detection in EEG algorithm is not working well enough for imagined speech. However, due to a lack of ground truth, this cannot be verified.

To conclude, the data scarcity problem in imagined speech EEG classification cannot easily be solved by adding more data from articulated speech EEG. Instead, research focus should lie on investigating what information in the EEG signal to use for imagined speech classification, and better ways to generalize the EEG signals across speakers.

6. References

[1] E. W. Sellers, D. B. Ryan, and C. K. Hauser, "Noninvasive brain-computer interface enables communication after brainstem stroke," *Science Translational Medicine*, vol. 6, no. 257, 2014.

[2] S. Komeiji, T. Mitsuhashi, Y. Iimura, H. Suzuki, H. Sugano, K. Shinoda, and T. Tanaka, "Feasibility of decoding covert speech in ecog with a transformer trained on overt speech," *Scientific Reports*, vol. 14, p. 11491, 2024.

[3] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nat Mach Intell*, vol. 5, p. 1097–1107, 2023.

[4] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, 2010.

[5] M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Imagined speech classification using EEG and deep learning," *Bioengineering*, vol. 10, no. 6, 2023.

[6] 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.

[7] L. Hausfeld, F. De Martino, M. Bonte, and E. Formisano, "Pattern analysis of EEG responses to speech and voice: Influence of feature grouping," *NeuroImage*, vol. 59, no. 4, pp. 3641–3651, 2012.

[8] B. Dekker, A. C. Schouten, and O. Scharenborg, "DAIS: The delft database of EEG recordings of dutch articulated and imagined speech," in *Proceedings of ICASSP*, 2023.

[9] N. Yoshimura, A. Nishimoto, A. N. Belkacem, D. Shin, H. Kambara, T. Hanakawa, and Y. Koike, "Decoding of covert vowel articulation using electroencephalography cortical currents," *Frontiers in Neuroscience*, vol. 10, 2016.

[10] L. C. Sarmiento, S. Villamizar, O. López, A. C. Collazos, J. Sarmiento, and J. B. Rodríguez, "Recognition of EEG signals from imagined vowels using deep learning methods," *Sensors*, vol. 21, no. 19, p. 6503, 2021.

[11] M. A. Lopez-Gordo, E. Fernandez, S. Romero, F. Pelayo, and A. Prieto, "An auditory brain–computer interface evoked by natural speech," *Journal of Neural Engineering*, vol. 9, no. 3, p. 036013, 2012.

[12] D. Vorontsova, I. Menshikov, A. Zubov, K. Orlov, P. Rikunov, E. Zvereva, L. Flitman, A. Lanikin, A. Sokolova, S. Markov, and A. Bernadotte, "Silent EEG-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification," *Sensors*, vol. 21, no. 20, p. 6744, 2021.

[13] C. Cooney, R. Folli, and D. Coyle, "Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG," in *2018 29th Irish Signals and Systems Conference (ISSC)*, 2018, pp. 1–7.

[14] A. Jahangiri, D. Achanccaray, and F. Sepulveda, "A novel EEG-based four-class linguistic BCI," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 3050–3053.

[15] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in *12th International Symposium on Medical Information Processing and Analysis*, vol. 10160. SPIE, 2017, p. 1016002.

[16] J. T. Panachakel, A. Ramakrishnan, and T. Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in *2019 IEEE 16th India Council International Conference (INDICON)*, 2019, pp. 1–4.

[17] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG," *Sensors*, vol. 20, no. 16, p. 4629, 2020.

[18] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves cnn generalization and transfer learning for imagined speech decoding from EEG," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 1311–1316.

[19] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[20] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. Torres-García, *Tensor Decomposition for Imagined Speech Discrimination in EEG*. Springer International Publishing, 2018, p. 239–249.

[21] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," *Biomedical Engineering Letters*, vol. 10, no. 2, p. 217–226, 2020.

[22] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2647–2659, 2020.

[23] Y. Zhao, Y. Liu, and Y. Gao, "Analysis and classification of speech imagery EEG based on chinese initials," *J. Beijing Inst. Tech*, vol. 30, pp. 44–51, 2021.

[24] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods*, vol. 250, pp. 126–136, 2015.

[25] R. Croft and R. Barry, "Removal of ocular artifact from the EEG: a review," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 30, no. 1, pp. 5–19, 2000.

[26] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using riemannian manifold features," *Journal of Neural Engineering*, vol. 15, no. 1, p. 016002, 2017.

[27] E. Gibson, N. J. Lobaugh, S. Joordens, and A. R. McIntosh, "EEG variability: Task-driven or subject-driven signal of interest?" *NeuroImage*, vol. 252, p. 119034, 2022.

[28] E. D. Palmer, H. J. Rosen, J. G. Ojemann, R. L. Buckner, W. M. Kelley, and P. S. E., "An event-related fMRI study of overt and covert word stem completion." *NeuroImage*, vol. 14, no. 1, p. 182–193, 2001.

[29] E. Niedermeyer and F. H. Lopes Da Silva, Eds., *Electroencephalography*, 5th ed. Philadelphia, PA: Lippincott Williams and Wilkins, 2004.

[30] A. Taherkhani, G. Cosma, and T. M. McGinnity, "A deep convolutional neural network for time series classification with intermediate targets," *SN Computer Science*, vol. 4, no. 6, 2023.

[31] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[32] D. Dash, A. Wisler, P. Ferrari, and J. Wang, "Towards a speaker independent speech-BCI using speaker adaptation," in *Interspeech 2019*. ISCA, 2019.