DAIS: The Delft Database of EEG Recordings of Dutch Articulated and Imagined Speech

Dekker, Bo; Schouten, Alfred; Scharenborg, Odette

DOI
10.1109/ICASSP49357.2023.10096145

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096145

DAIS: THE DELFT DATABASE OF EEG RECORDINGS OF DUTCH ARTICULATED AND IMAGINED SPEECH

Bo Dekker[1], Alfred C. Schouten[1], Odette Scharenborg[2]

[1]Department of Biomechanical Engineering, Delft University of Technology, Delft, The Netherlands
[2]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Silent speech interfaces could enable people who lost the ability to use their voice or gestures to communicate with the external world, e.g., through decoding the person's brain signals when imagining speech. Only a few and small databases exist that allow for the development and training of brain computer interfaces (BCIs) that can decode imagined speech from recorded brain signals. Here, we present an open database consisting of electroencephalography (EEG) and speech data from 20 participants recorded during the covert (imagined) and actual articulation of 15 Dutch prompts.

A validation speaker-independent classification experiment using a ResNet-50 model with spatial-spectral-temporal features extracted from the EEG signals obtained an average accuracy of 70.6% for the classification of rest vs. covert vs. articulated speech trials. This and observed structural differences in the EEG signals between covert and articulated speech demonstrate that the EEG signals in the three classes contain discriminative information.

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

1. INTRODUCTION

People who lost the ability to speak and cannot use gestures due to severe neuromuscular diseases (e.g., severely paralysed people or patients of locked-in syndrome) are strongly impaired in communicating with the external world [1]. Brain-computer interfaces (BCIs) might enable communication with the external world [2]. These BCIs should convert the intended message from neural activity in the brain, e.g., through decoding the person's brain signals during covert speech [3]. Covert speech is imagining speaking without moving any of the articulators or making any sound, so without any actual motor activity.

The neural signals that give the best results in BCIs are obtained using electrocorticography (ECoG) [4][5][6]. ECoG is an invasive technique where electrode arrays are placed directly onto the patient's brain surface. Electroencepha-lography (EEG) is non-invasive: electrodes are placed on the head, typically with a cap, which is more user-friendly and cheaper, but at the cost of less good decoding performance.

EEG signals have shown some (limited) success in, e.g., decoding imagined articulation of vowels (English [7], Dutch [8], Japanese [9], Spanish [10]) and isolated words ("yes" and "no" [11], nine Russian words [12]). Typically machine learning algorithms are applied for training and decoding (e.g., support vector machine [13], linear discriminant analysis [14], random forest [15], vanilla deep neural networks (DNNs) [16], and convolutional neural network (CNN) [17][18]). Moreover, different discriminative features extracted from the EEG signals have been used (e.g., wavelet domain features [19][20] and common spatial patterns (CSP) [21][22]). 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 well performing CNN algorithms on covert speech classification tasks in both robustness and practicability.

Existing methods are hard to compare as they are typically trained and evaluated on different databases, which are often not available for other researchers. We are aware of only three open datasets: KARA ONE [23], Nguyen et al. [24] (both English), and Coretto et al. [15] (Spanish), which differ in several aspects, e.g., EEG signal quality, recording device, and set-up (see Table 1). Moreover, no internal quality check within the datasets was performed (e.g., check whether the participants truly perform covert speech), potentially leading to networks trained on poorly labelled data [12][13]. Moreover, most databases acquire EEG only from covert speech, so without acquiring EEG signals during articulated speech in the same trial, which makes it (even more) difficult to verify whether a subject truly performed the covert speech task. Of the open datasets, only the KARA ONE database acquires covert and articulated speech consecutively.

To build BCIs that can decode the intended message from neural activity, a thorough understanding of the relationship between neural signals, sounds, articulation, and imagined speech is crucial, and largely lacking. Here, we present a database of EEG recordings of participants during Dutch articulated and imagined speech (DAIS), recorded in the same trial, including their speech. We used prompts that consist of five vowels in isolation and as part of 10 C_1VC_2 (consonant-vowel-consonant) words, where also the reverse, C_2VC_1, are Dutch words. This allows for the investigation of the neural signatures of vowels in isolation vs. in context, and of consonants in two contexts. The collected EEG signals are validated visually by comparing event related potentials (ERPs) and by a speaker-independent classification task of pre-stimulus (rest) vs. covert vs. articulated speech.

2. METHODOLOGY

2.1. Participants

Twenty native Dutch speakers, 14 women and 6 men (mean age: 24.6 years, SD: 1.0, range 23-26) participated. No participants reported speech, language, or cognitive disorders, and all had normal or corrected to normal vision. Two persons reported to be left-handed. The study was approved by the Human Research Ethics Committee of the Delft University of Technology. All participants gave written informed consent prior to the experiment. The participants received no monetary reward.

2.2. Stimuli

The prompts consisted of five Dutch vowels (/a:, e:, o:, i, u/, where ":" indicates lengthening) and ten Dutch words. The five vowels

constitute the different corners of the Dutch vowel quadrant. The ten words are five 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 consonants are chosen to be diverse in their manner of articulation (i.e., nasals, plosives, and fricatives) while having a fairly similar pronunciation irrespective of position in the word (except /l/). This selection of prompts enables researchers to explore the effects of the phonetic environment on EEG signals.

2.3. Experimental set-up

Participants were recorded individually while seated in a comfortable chair in front of a microphone and a screen in a sound-attenuating room. Visual prompts (designed using the Psychtoolbox-3 [25] running in MATLAB) were presented on the screen to inform the participants which specific task to perform (rest, read, imagine speech, articulate speech).

2.3.1. EEG recordings

During the entire experiment, continuous EEG was collected from 64 electrodes using the TMSi SAGA 64+ (with a BrainWave EEG Cap using the 10-20 system; see Figure 2) 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 50 kΩ. During recording, all signals were amplified against the average of all connected channels (i.e., average reference amplifier).

2.3.2. Speech recordings

The articulated speech was recorded using an Audio Technica AT2020USB+ microphone (F_s= 44.1 kHz). To reduce popping sounds, a pop filter was placed between the microphone and the participant at 10 cm from the microphone. The mouth-to-mic distance was fixed at 30 cm.

2.4. Procedure

Each experiment consisted of 20 runs of 15 trials, one for every prompt (i.e., the 15 Dutch vowels and words). The prompts were randomised for each participant using a balanced Latin square to reduce order effects. A trial consisted of four successive segments, see Figure 1: pre-stimulus (rest; blank screen), reading of the prompt (either a vowel, denoted in orthographic script for ease of the participant, or a word), covert (imagined) speech (indicated with a thought balloon), and articulated speech (indicated with a speech balloon). Each run was followed by (another) 2s rest. Prompts and instructions were shown as black text on a dark-grey coloured background to minimise eye fatigue.

The task during the pre-stimulus (rest) segment was to relax, and participants were allowed to blink. During the reading segment, participants were to read the prompt only. During the covert speech segment, participants were to imagine the execution of the different articulatory gestures as if one were to articulate the prompt once without emitting sound or making any actual articulatory movement. During the articulated speech segment, participants were to articulate the prompt once. During all segments except pre-stimulus (rest), the participant was instructed to minimise moving, swallowing, and blinking to reduce the presence of artefacts.

The participants were instructed to perform the specific task once right after the visual cue appeared on the screen. By limiting the duration for the covert and articulated speech tasks to 2s and by collecting both tasks within a single trial, behavioural control can be applied to ensure that the participant only imagines the articulation of the presented prompt [26]. To retain attention and prevent fatigue, a three-minute break was scheduled after every five runs, additionally, participants could ask for additional breaks between runs. During the placement of the EEG electrode cap, the experimental protocol was explained and participants were familiarised with the different visual cues in test trials.

2.5. Pre-processing of the EEG data

During pre-processing, only channels that disconnected during the experiment were removed (see Section 3.1). Other potentially bad channels were kept. The EEG data was band-pass filtered (Hamming windowed sinc FIR filter) between 1 Hz and 70 Hz to remove low-frequency trends in the data and to remove artifacts related to EMG activity by excluding the high gamma band. A 49-51 Hz notch filter (Hamming windowed sinc FIR filter) was used to remove power line noise at 50 Hz and all data was re-referenced.

Table 1. Overview of the different open datasets and DAIS: #P(articipants), #Ch(annel)s, #Prompts (V(owels), W(ords)), #Rep(etitions)/Pr(ompt), Notes, L(a)ng(uage).

Dataset	P	Ch	Prompts	Rep/Pr	Notes	Lng
Coretto [15]	15	5	5V + 6W	40 covert + 10 articulated	Not in same trial	SP
KARA ONE [23]	8	64	7V + 4W	12 covert + 12 articulated	In same trial	EN
Nguyen [24]	15	64	many	100 covert	4 consecutive repetitions of same prompt	EN
DAIS	20	64	5V + 10W	20 covert + 20 articulated	In same trial	NL

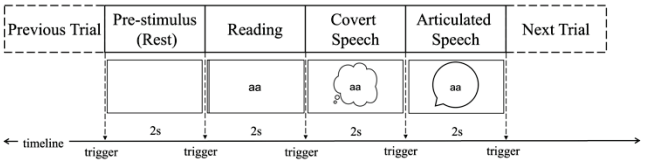

Figure 1. A trial consisted of four successive segments: pre-stimulus (rest), reading of the prompt, covert (imagined) speech, and articulated speech.

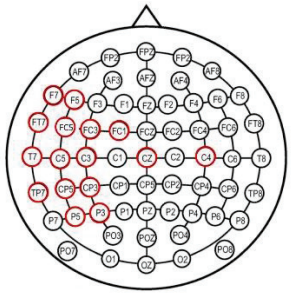

Figure 2. Visualisation of the electrode placement (10-20 system) with 62 channels. In red, the 16 EEG channels used in our validation study.

Each trial was segmented into its four 2s segments using triggers (see Figure 1) from the trigger channel. Segments containing eye blinks were marked using ERPLAB artefact detection (moving window peak-to-peak threshold). To preserve the data as much as possible for further research, eye blinks were only marked and not removed (e.g. by Independent Component Analysis).

3. THE DELFT ARTICULATED AND IMAGINED SPEECH (DAIS) DATABASE

EEG and the speech of in total 5993 trials of 20 participants were recorded. All participants completed the 20 runs of 15 trials (prompts), yielding 300 trials per participant, except for participant #2, who completed 19 runs, as run 7 was aborted halfway due to technical difficulties. DAIS is available to the community[1].

3.1. EEG and speech data

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. A total of 24370 segments were recorded for the 4 different tasks and the 15 different prompts. After pre-processing, 16510 segments (68%) were unmarked for eye blinks.

Table 2 gives an overview of the number of recorded segments for each task, the number and percentage of unmarked segments after pre-processing, and average number of unmarked segments per prompt per participant. For the pre-stimulus the max is 320 ((15 for each trial + 1 after each run) * 20 runs). For the other segments, the maximum is an average of 20 (1 for each run).

For the reading, and covert and articulated speech segments, a high percentage of segments was unmarked (all > 70%). For the pre-stimulus (rest) segments a considerably lower number of segments was unmarked (33%), which can be attributed to the fact that participants were allowed to blink during the pre-stimulus segments.

In total, 5993 speech files were recorded during the articulated speech segments: 300 for each participant, except for participant #2 293 speech files were recorded.

3.2. File name convention

The segmented EEG data and the speech files in the DAIS database are named using the EEG extension to the Brain Imaging Data Structure (BIDS) [27]. The EEG segments corresponding to a specific task and prompt are combined for each participant, and stored in .fdt and .set file format, and named following the format: sub-<participant ID>_task-<label>_eeg. The speech is stored in .wav format with the naming format: sub-<participant ID>_task-<label>_run-<number of run>_audio, where the *label* corresponds to the specific task and prompt, e.g., covert-aa.

Table 2. DAIS: Overview of the EEG data per task: the number of recorded segments, the number and % of unmarked segments after pre-processing, and the average number of unmarked segments per participant per prompt.

Task	Segments		Avg segments/ partic/prompt
	Recorded	*After preproc (%)*	
Pre-stimulus	6392	2108 (33%)	105
Reading	5993	4540 (76%)	15
Covert	5993	5550 (93%)	19
Articulated	5992	4312 (72%)	14

[1] https://doi.org/10.17026/dans-xc3-66ze.

4. VALIDATION

To investigate whether the participants complied with the task of imagining speech, we visually investigated whether structural differences exist between the pre-stimulus (i.e., rest), covert speech and articulated speech segments by inspected the event-related potentials (ERPs) of the EEG signals. Additionally, we ran a speaker-independent, three-class classification task with the task to predict whether the EEG signal came from the pre-stimulus (rest) state, covert speech, or articulated speech. If a covert speech EEG signal is distinct (enough) from rest EEG signals and articulated speech signals, it is more likely that participants indeed fulfilled the task of imagining speech.

4.1. Data

For the validation experiment, 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 was excluded 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.

The pre-stimulus, covert, and articulated speech EEG data of the remaining 15 participants was used in a leave-three-out cross-validation scheme for which five folds were created. For each fold, the data from 10, 2, and 3 participants were assigned to the training set, the validation set, and to the test set, respectively. The average number of EEG recordings used for training the model was 6088 (± 96; range: 5970-6185). For more details, see [28].

4.2. Methods

4.2.1. Event related potentials

The EEG signals obtained during the pre-stimulus (rest), covert speech, and articulated speech segments are visualised through event-related potentials (ERPs) in the time series. The ERPs are calculated by averaging the 16 pre-selected channels of each segment for each subject. Due to the high inter-subject variability in the timing of the speech, no grand average between subjects was calculated. An ERP curve is characterized by positive or negative peaks which occur after the stimulus [31]. Clear distinct ERPs between tasks are a first visual check of the data.

4.2.2. Classification experiment

Based on the involvement of specific areas of the cortex in language processing [16][24], 16 EEG channels were included in the validation study, i.e., the red electrodes in Figure 2: FC1 = Premotor cortex; FC3 = Premotor cortex; Cz = Motor cortex; C4 = Motor cortex; C3 = Motor cortex; FC5 = Broca's area; FT7 = Broca's area, inferior temporal gyrus; F5 = Broca's area; F7 = Broca's area; C5 = Wernicke's area, primary auditory cortex; T7 = Middle temporal gyrus, secondary auditory cortex; CP3 = Wernicke's area; CP5 = Wernicke's area; TP7 = Wernicke's area; P5 = Wernicke's area; P3 = Superior parietal lobule. For more details, see [28].

To use the frequency information of the cortex, the pre-processed EEG data from the 16 channels were converted to wavelet scalograms using the continuous wavelet transform (CWT) with Morlet wavelets [26] and combined in one 4-by-4 image of the scalograms. A pre-trained ResNet-50 model [29], trained on more than 1M images from the ImageNet database [30], was used for the

classification task. Since the scalograms are quite different from the images in the ImageNet database, we retrained all weights of all layers to tune the pre-trained models for the three-class classification task. Moreover, the last two layers of the ResNet were deleted (i.e., the fc1000 and classificationLayer-Predictions) and replaced with a new fully connected layer with the number of outputs equal to the number of classes (i.e., three) and a new classification layer.

4.3. Results

4.3.1. Event related potentials

Visual inspection of the ERPs of the participants revealed clear differences between the EEG data of pre-stimulus (rest), covert and articulated speech as illustrated in Figure 3, which shows the average EEG over time for Participant #12. The top panel shows that no ERPs were found for the rest task, only background EEG activity, which was expected as no stimulus is given. For both overt (middle panel) and articulated speech (bottom panel) clear ERP components (the peaks and troughs) are found following the onset of the visual cue, followed by approximately 100-200 ms of enhanced activity.

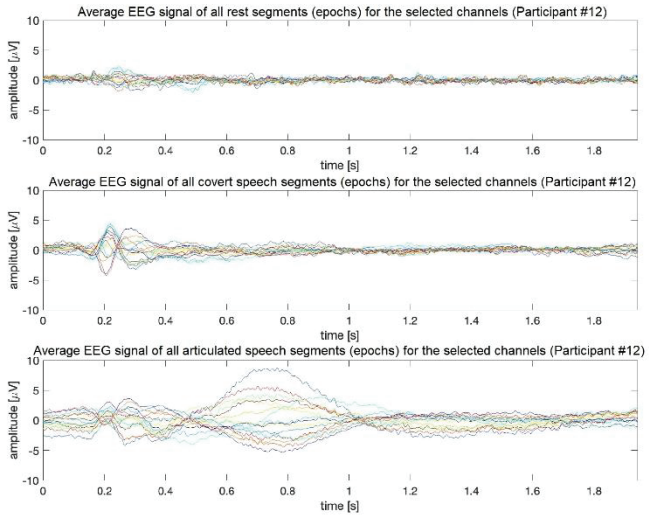

Figure 3. Average EEG for the synchronised segments for participant #12 to visualise the ERPs over time for rest, covert, and articulated speech. The different lines correspond to the 16 selected channels.

Table 3. Classification accuracies (%) per fold.

	Validation set	**Test set**
Fold 1	81.1	68.4
Fold 2	67.7	78.0
Fold 3	69.8	70.8
Fold 4	69.0	68.9
Fold 5	74.5	66.7
Average	72.4 ± 5.5	70.6 ± 4.4

Table 4. Confusion matrix (averaged over the five folds) of the number (and percentage) of predicted class labels per true class. Bold indicates highest percentage predicted labels per true class.

		Predicted class		
		Rest	*Covert*	*Articulated*
True class	*Rest*	464 (33)	**751 (54)**	174 (13)
	Covert	615 (14)	**3364 (79)**	299 (7)
	Articulated	222 (6)	654 (19)	**2598 (75)**

This similarity in activity around 250-300 ms for the covert and articulated speech tasks suggests that participants also carried out a language processing task during the covert speech segment [32][33].

For the articulated speech task, this enhanced activity is followed by a broad peak/trough (depending on the channel) starting at ~500 ms which coincides with the acoustic onset and is therefore associated with the voluntary movement of the articulators. The absence of a similar activity during the covert speech segment strongly suggests that no articulatory movements were made during the covert speech task.

4.3.2. Classification results

Table 3 shows the classification results for the validation and test sets averaged over pre-stimulus, covert, and articulated speech. The average accuracy over all folds of 70.6% is well above chance (33.3%). Table 4 shows the confusion matrix where each number in the confusion matrix indicates the number of segments in a class (true label) identified as any class (predicted label). Covert speech EEG is best classified, followed by articulated speech. Rest is most often classified as covert speech. Since covert speech is hardly classified as rest or articulated speech, we can conclude that the covert EEG signals contain information that is distinct from rest and articulated speech, strongly suggesting that the participants indeed performed the task of imagining speech.

5. DISCUSSION AND CONCLUSION

We presented the Delft Articulated and Imagined Speech (DAIS) database, consisting of the EEG recordings of 20 native Dutch speakers of the imagined (covert) and articulated speech of 15 Dutch prompts, and the speech recordings of the articulated speech. The database has recordings of more participants and repetitions per prompt than the only other database that recorded both covert and articulated speech in the same trial (KARA ONE).

Visualisation of the ERP showed distinct structural differences between the averaged EEG data for the pre-stimulus (rest), covert speech and articulated speech segments which strongly suggest that the participants carried out different tasks during those segments, as instructed. Moreover, it shows that the participants actively engaged in cognitive processing (and likely in a language processing task) during the covert speech segments and did not simply relax. To extract the frequency information, scalograms were made for all included individual segments, which were subsequently used by the classification algorithm.

The validation classification experiment showed that an off-the-shelf ResNet, fine-tuned on the EEG data, was able to correctly classify imagined and articulated speech EEG with an accuracy well above chance. The observed structural differences in the EEG signals of the three tasks can thus be used for classification of rest vs. covert speech vs. articulated speech.

Taking together the structural differences in the ERP and the limited number of confusions of the imagined speech segments with rest segments strongly suggest that the imagined speech EEG is different from the rest state EEG, suggesting that the participants performed the task of imagining speech.

Finally, the careful selection of prompts and the recording of EEG during both imagined and articulated speech yields a database that paves the way to a thorough understanding of the relationship between the neural signals, sounds, articulation, and imagined speech, which is crucial for developing brain-computer interfaces.

6. REFERENCES

[1] O. Scharenborg and M. Hasegawa-Johnson, "Position paper: Brain signal-based dialogue systems," in Increasing Naturalness and Flexibility in Spoken Dialogue Interaction, Springer, pp. 389–392, 2021.

[2] 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, pp. 257re7–257re7, 2014.

[3] O. Iljina, et al., "Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication." Brain-Computer Interfaces, 4, pp. 186–199, 2017.

[4] S. Martin, et al., "Word pair classification during imagined speech using direct brain recordings." Sci. Rep., 6, Art no. 25803, 2016.

[5] E. C. Leuthardt, et al., "Using the electrocorticographic speech network to control a brain–computer interface in humans." J. Neural Eng., 8, 2011.

[6] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans." J. Neural Eng., 8, p. 46028, 2011.

[7] C. S. Dasalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns." Neural Netw., 22, pp. 1334–1339, 2009.

[8] L. Hausfeld, F. De Martino, M. Bonte, and E. Formisano, "Pattern analysis of EEG responses to speech and voice: influence of feature grouping." Neuroimage, 59, pp. 3641–3651, 2012.

[9] Y. Natsue, et al., "Decoding of Covert Vowel Articulation Using Electroencephalography Cortical Currents." Frontiers in Neuroscience, 10, p. 175, 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, 21 (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." J. Neural Eng., 9, p. 036013, 2012.

[12] D. Vorontsova, et al., "Silent EEG-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification," Sensors, 21 (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 Syst. Conf. (ISSC). IEEE, pp. 1–7, 2018.

[14] A. Jahangiri, D. Achanccaray, and F. Sepulveda, "A novel EEG-based four-class linguistic BCI," in 2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. IEEE, pp. 3050–3053, 2019.

[15] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in 12th Int. Symp. on Med. Inf. Process. and Anal., SPIE, 10160, p. 1016002, 2017.

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

[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, 20 (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," IEEE Int. Conf. on Syst., Man and Cybernetics (SMC), pp. 1311–1316, 2019.

[19] J. S. Garcia-Salinas, L. Villase˜nor-Pineda, C. A. Reyes-Garcia, and A. Torres-Garcia, "Tensor decomposition for imagined speech discrimination in EEG," in Mexican Int. Conf. on Artif. Intell. Springer, pp. 239–249, 2018.

[20] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," Biomed. Eng. Letters, 10 (2), pp. 217–226, 2020.

[21] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," IEEE Trans. on Neural Syst. and Rehabil. Eng., 28 (12), pp. 2647–2659, 2020.

[22] Y. Zhao, Y. Liu, and Y. Gao, "Analysis and classification of speech imagery EEG based on Chinese initials," Journal of Beijing Institute of Technology, 30, no. zk, pp. 44–51, 2021.

[23] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 992–996, 2015.

[24] 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, 15 (1), p. 016002, 2017.

[25] M. Kleiner, D. Drainard, D. Pelli, A. Ingling, R. Murray, and C. Brousard, "What's new in psychtoolbox-3," Perception, 36, no. ECVP Abstract Suppl., 2007.

[26] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (MEG) signals," Frontiers in Neuroscience, 14, p. 290, 2020.

[27] C. R. Pernet, et al., "EEG-bids, an extension to the brain imaging data structure for electroencephalography," Scientific data, 6 (1), pp. 1-5, 2019.

[28] B. Dekker, "Decoding covert speech from EEG: development of a novel database containing EEG and audio signals during Dutch covert and overt speech," MSc thesis Delft University of Technology, Delft, The Netherlands.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE Conf. Comput. Vis. Pattern Recog., pp. 770-778, 2016.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," IEEE Conf. Comput. Vis. Pattern Recog., pp. 248-255, 2009.

[31] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014.

[32] D. Brandeis and D. Lehmann, "Event-related potentials of the brain and cognitive processes: approaches and applications," Neuropsychologia, 24 (1), pp. 151-168, 1986.

[33] D.-D. Tao, Y.-M. Zhang, H. Liu, W. Zhang, M. Xu, J. J. Galvin III, D. Zhang, and J.-S. Liu, "The p300 auditory event-related potential may predict segregation of competing speech by bimodal cochlear implant listeners," Frontiers in Neuroscience, p. 843, 2022.