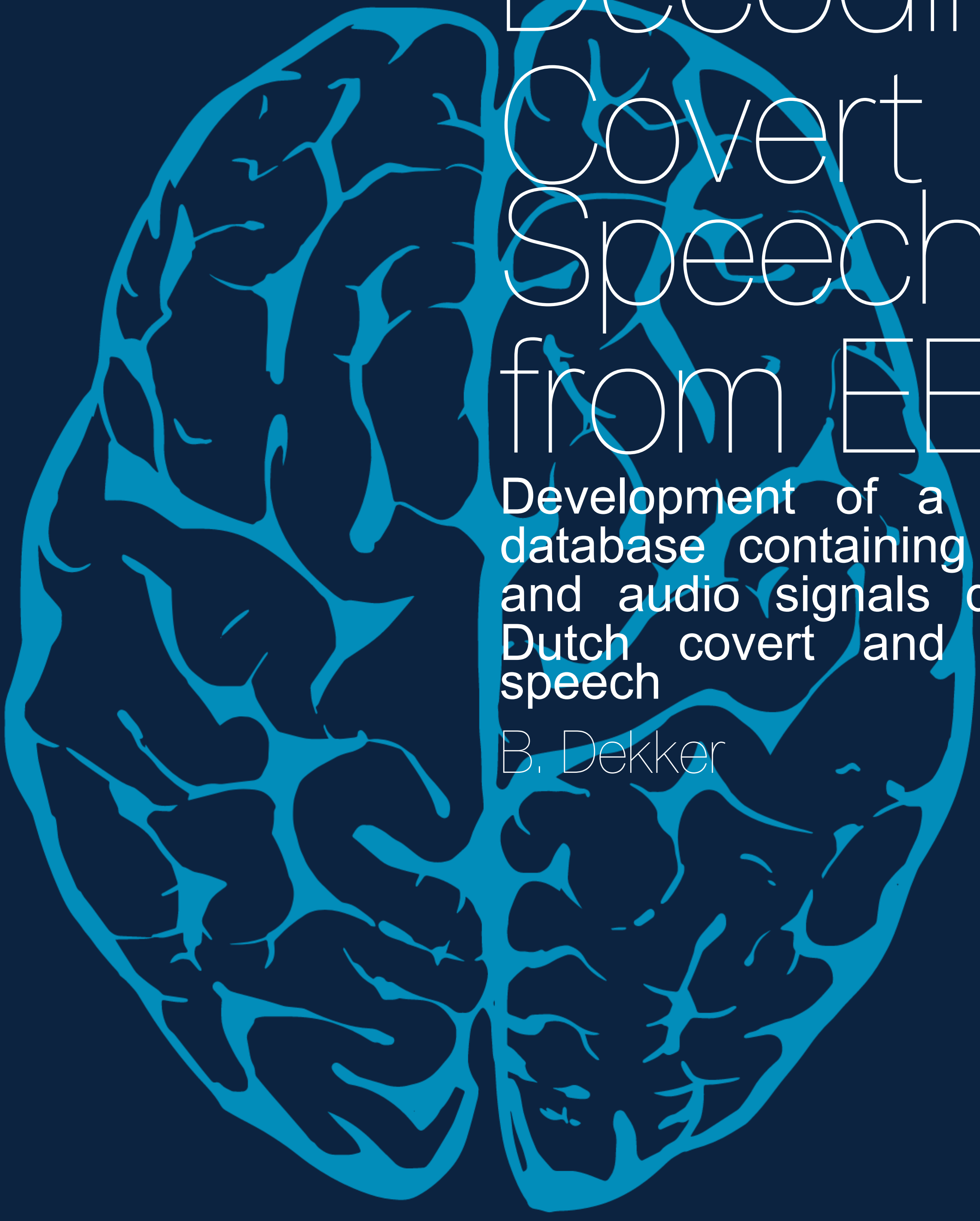

Decoding Covert Speech from EEG

Development of a novel database containing EEG and audio signals during Dutch covert and overt speech

B. Dekker

Delft University of Technology

Decoding Covert Speech from EEG

Development of a novel database containing EEG and audio signals during Dutch covert and overt speech

by

B. Dekker

to obtain the degree of Master of Science in Biomedical Engineering
at the Delft University of Technology,
to be defended publicly on Thursday October 25, 2022 at 2:00 PM.

Student number:	4560817	
Project duration:	March 10, 2022 – October 25, 2022	
Thesis committee:	Prof. dr. ir. A. C. Schouten,	TU Delft, supervisor
	Dr. O. E. Scharenborg,	TU Delft, supervisor
	Dr. H. Hung,	TU Delft
	Dr. M. Kok,	TU Delft

Decoding Covert Speech from EEG: development of a novel database containing EEG and audio signals during Dutch covert and overt speech

B. Dekker
Department of Biomedical Engineering
Faculty of Mechanical, Maritime, and Materials Engineering
Delft University of Technology

***Abstract*—To enable communication for patients who have lost the ability to speak due to severe neuromuscular diseases, covert speech based brain-computer interfaces (BCIs) might be used. These system use neural signals arising from covert speech and translate them into text or synthesised speech. Covert speech is imagining to speak without moving any of the articulators and therefore does not rely on actual motor activity. As recognizing covert speech from neural signals is extremely challenging, machine learning algorithms are deployed. To make use of the full potential of machine learning approaches in the field of decoding covert speech and to accommodate real-world deployment of a BCI, a large number of training samples is required to train the networks.**

In this study, a novel database is presented containing EEG and audio data from 20 subjects recorded during the covert and overt pronunciation of 15 Dutch prompts. To validate the recorded data, two speaker-independent classification tasks were performed using a ResNet-50 algorithm as classifier with spatial-spectral-temporal features extracted from the EEG signals. The speaker-independent three-class classification of pre-stimulus (rest) trials versus covert speech trials versus overt speech trials obtained an average accuracy of 70.6% and the speaker-independent five-class classification of five covert vowels ("aa", "ee", "oo", "ie", "oe") obtained an average accuracy of 19.6%. Even though the five-class classification task did not reach an above chance level accuracy, the high performance reached by the three-class classification task provides support of the existence of discriminative information in the covert speech segments to decode covert speech in the future.

Future research should focus on EMG artifact detection and on determining the performance per subject to improve the dataset. Furthermore, subject normalisation strategies should be investigated to address the challenges of subject-independent covert speech decoding.

***Index Terms*—brain-computer interface (BCI), convolutional neural network (CNN), Dutch covert speech, electroencephalography (EEG), ResNet-50.**

I. Introduction

People who lost the ability to speak and cannot use sign language 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]. In many of these patients, the cognitive abilities are preserved and the inability to communicate has a strong adverse effect on their quality of life. For these patients, brain-computer interfaces (BCIs) might enable communication with the external world [2].

BCIs can offer a way of communicating by using neural signals during covert speech. Covert speech is imagining speaking without moving any of the articulators or making any sound, so without relying on actual motor activity. The standard modality for measuring neural signals with BCIs is electroencephalography (EEG), mainly due to the non-invasive nature, low costs, user-friendliness, and good temporal resolution [3], [4]. For a covert speech based BCI, the EEG recordings need to be decoded through signal processing and classification algorithms to allow the user to communicate.

The relationship between covert (imagined) and overt (spoken) speech is still unclear. Generally, covert speech is considered as truncated overt speech (i.e., interrupted speech production). However, at what level this interruption exactly occurs is still subject of much debate [5], [6]. Both covert and overt speech tasks activate essential language areas (Broca's and Wernicke's areas, inferior parietal lobule) and several structures on the left and right hemisphere [5], [7], [8]. Some studies suggest that covert speech can be considered as overt speech minus articulatory motor execution [9], while other studies observed greater activity in several regions (e.g., middle temporal gyrus, left inferior frontal gyrus) during covert speech in comparison to overt speech [10], [11].

Accurate classification of covert speech from EEG is difficult [12], [13]. Previous studies have tried to overcome this difficulty by deploying many different traditional machine learning algorithms (e.g., support vector machine [14], linear discriminant analysis [15], and random forest [16]) and deep learning architectures (e.g., convolutional neural network (CNN) [17] and deep neural network (DNN) [18]). To classify covert speech from EEG signals, discriminative features must be extracted. Among the features used for covert speech decoding are statistical features (e.g., mean, variance, and standard deviation) [19], wavelet domain features [20], [21], and common spatial patterns (CSP) [22], [23]. Although previous studies have shown the potential of using machine learning algorithms for decoding covert speech from neural signals [12], [13], [24], [25], no combination of classifier and features has been proven to consistently achieve high decoding performances [26]. A systematic search in literature did show that CNNs provide the most promising results in decoding covert speech from EEG. More specifically, ResNet (Residual Network) algorithms [13], [27] outperformed other well performing CNN algorithms (e.g., DenseNet [28] and CNNeeg1-1 [12]) on covert speech classification tasks in both robustness and practicability. ResNet models are deep CNNs based on residual learning. The ResNet architecture uses residual blocks to solve the vanishing gradient problem [29]. Pre-trained ResNet models are pre-trained on more than a million images from the ImageNet database [30] to learn features from these images. These models can be used in many specific applications to reduce the need for sample size [31], [32]. The network has learned a rich set of features but can through fine-tuning still learn features specific to the new data.

To exploit the full potential of machine learning approaches in the field of decoding covert speech, a large amount of training data for a particular task is required to train the networks [33]. Multiple research teams have created these types of datasets and some of them are openly shared (e.g., KARA ONE database [34], Coretto et al. (2017) [16] database, and Nguyen et al. (2018) [35] database). The different datasets are poorly comparable because the BCI devices used have different number of channels, signal quality, and recording devices. Furthermore, there is no internal quality control of the data in the datasets (e.g., did the subjects truly perform covert speech), which leads to training networks on poorly labeled data [13], [14].

The main purpose of this research is to provide the scientific community with an multi-class EEG and audio database of covert and overt speech that could be used to better understand the related brain mechanisms and ultimately develop a BCI based on covert speech. As the research is conducted in the Netherlands and because native and non-native language processing differs [36], [37], the database will consists of Dutch prompts. While publicly available datasets for covert speech do for example exist for English [34], [35], [38] and Spanish prompts [16], [39], to the best of our knowledge there is no publicly available EEG dataset containing Dutch covert prompts.

In this study, an experimental design is set up and executed to provide a database containing EEG and audio recordings during overt and covert pronunciation of Dutch prompts. The Dutch prompts collected are a combination of vowels and words. Most studies on decoding covert speech acquire covert and overt speech data separately, which makes it difficult to verify whether a subject truly performed the covert speech task. In contrast to these studies, the covert and overt speech tasks in this study are collected consecutively in a single trial. By collected both covert and overt speech in one trial and by limiting the duration of the covert speech task, behavioral control can be applied to ensure the subject only imagines the articulation of the presented prompt [40].

The EEG signals collected in the database are analysed by performing two analyses aimed at demonstrating the potential use of the data: a speaker-independent three-class classification task of pre-stimulus (rest) versus covert speech versus overt speech and a speaker-independent five-class classification task of the covert vowels. For these classification tasks, spatial-spectral-temporal features are used to train a ResNet-50 algorithm.

II. Methods

A. Subjects

Twenty healthy volunteers, 14 women and 6 men (mean age: 24.6 ± 1.0 years, range 23-26), participated in the experiment, see Table VI (Appendix B). All subjects were adult native Dutch speakers without speech, language, or cognitive disorders, and with normal or corrected to normal vision. Two subjects reported to be left-handed. The handedness of the subjects is relevant due to the potential relationship between handedness and language dominant hemisphere [41], [42].

The experiments were conducted at the faculty of Mechanical, Maritime and Materials Engineering (3mE) at the Delft University of Technology. This study was approved by the Human Research Ethics Committee (HREC) of the Delft University of Technology (#2265). All subjects received a participation information letter and gave written informed consent prior to the start of the experiment. The participation information letter and the informed consent form can be found in Appendix C.

B. Experimental Set-Up

In the experiment, EEG and audio signals were collected during trials with both covert and overt speech. The subjects were seated in a comfortable chair in front of a microphone and a screen in a sound-attenuating room, see Figure 1. Visual cues were presented on the screen to inform the subjects about the specific task to perform. The visual cues were designed using the Psychtoolbox-3 [43] running in MATLAB (The MathWorks, Inc., USA). A webcam was used to provide a way for the subject to communicate through hand gestures and for the researcher to observe and intervene when articulatory movements were made during the covert speech segments. The choice to visually check for movements instead of using electromyography (EMG) (e.g., on the superior and inferior orbicularis oris) was made with the comfort of the subject in mind and to ensure that the overt speech was not negatively influenced by an overcomplicated experimental set-up.

The EEG data was collected using the TMSi SAGA 64+ at a sampling frequency of 1024 Hz and the TMSi SAGA interface for MATLAB. The docking station of the TMSi SAGA, located outside the sound-attenuating room, and the data recorder of the TMSi SAGA were connected using an optical fiber, see Figure 2. A 64-channel BrainWave EEG Cap infinity was used where the electrode placement follows the 10-20 system [44]. An appropriate capsize was chosen based on the head circumference of each subject and the gaps between

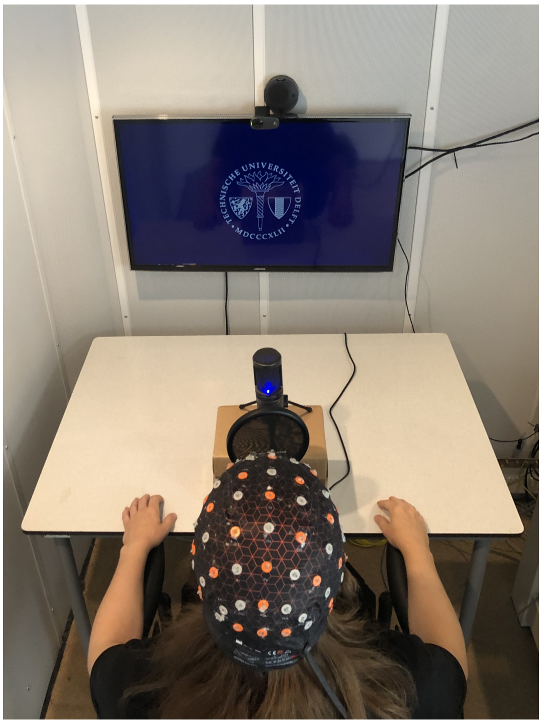

Fig. 1. Experimental set-up in the sound-attenuating room. The subject is wearing the EEG cap and sits in front of a microphone and a screen. The screen shows the visual cues during the experiment. In this illustrative figure, no visual cue is given.

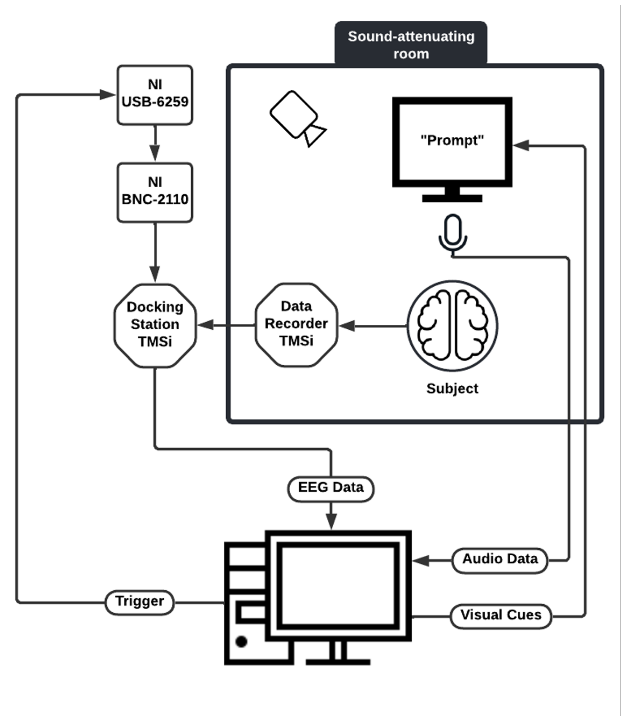

Fig. 2. Experimental set-up. The subject was seated in a sound-attenuating room in front of a microphone and a screen. A computer, located outside the sound-attenuating room, controlled the stimulation protocol, and received the sampled EEG and audio data from the acquisition systems. The EEG data was collected using the TMSi SAGA 64+. A National Instrument data acquisition set-up sent a trigger to the TMSi for each time a new visual cue is presented to the subject.

the scalp and the electrodes were filled with ABRALYT HiCl Abrasive Electrolyte-Gel. The electrode impedance was checked, and the experiment was only started if the impedance of all electrodes was less than 50 kOhm. During operation, the input was configured as an average reference amplifier, meaning all signals were amplified against the average of all connected channels.

The audio was recorded using an Audio Technica AT2020USB+ microphone at a sampling frequency of 44.1 kHz. To reduce popping sounds, a pop filter was placed between the microphone and the subject at 10 cm from the microphone, see Figure 1. The mouth-to-mic distance was approximately 30 cm and was kept relatively constant by fixating the position of the chair and the position of the microphone. Audio was solely recorded during the overt speech task.

A computer, located outside the sound-attenuating room, executes the stimulation protocol, and receives the sampled EEG and audio data from the acquisition systems. A National Instrument data acquisition set-up was used to send a trigger to the TMSi each time a new visual cue was presented to simplify the signal processing. Each prompt has a unique trigger value for the different tasks, see Table VII (Appendix B).

C. Stimuli

The subjects participated in one single session in which they were asked to perform covert and overt speech of 15 prompts. The prompts consist of five Dutch vowels and ten Dutch words. The vowels are “aa”, “ee”, “oo”, “ie”, “oe”. The specific vowels were chosen as these make up the different corners of the Dutch vowel quadrilateral [45]. The ten words are five Dutch word-pairs that turn in each other when read backwards. The ten Dutch words are: “taal”, “laat”, “leeg”, “geel”, “niet”, “tien”, “toon”, “noot”, “soep”, and “poes”, corresponding to the English words: “language”, “late”, “empty”, “yellow”, “not”, “ten”, “tone”, “note”, “soup”, and “cat”. The specific words contain the aforementioned vowels and contain different consonants (e.g., nasals, plosives, and fricatives) that are pronounced the same if they appear at the beginning or at the end of a word. This selection of prompts enables researchers to explore the effects of the phonetic environment in EEG signals and can be used to recognize the order of different phonemes.

D. Experimental Protocol

The full experimental protocol can be found in Figure 3. The experiment consisted of multiple trials in which one of the 15 different prompts was shown. As shown in the bottom half of Figure 3, each trial (i.e., showing of a single prompt) consisted of four successive segments: pre-stimulus (rest), reading, covert (imagined) speech, and overt (spoken) speech. The pre-stimulus (rest) segment was the period two seconds before the onset of the visual stimulus (blank screen), during which the subject was instructed to relax and was allowed to

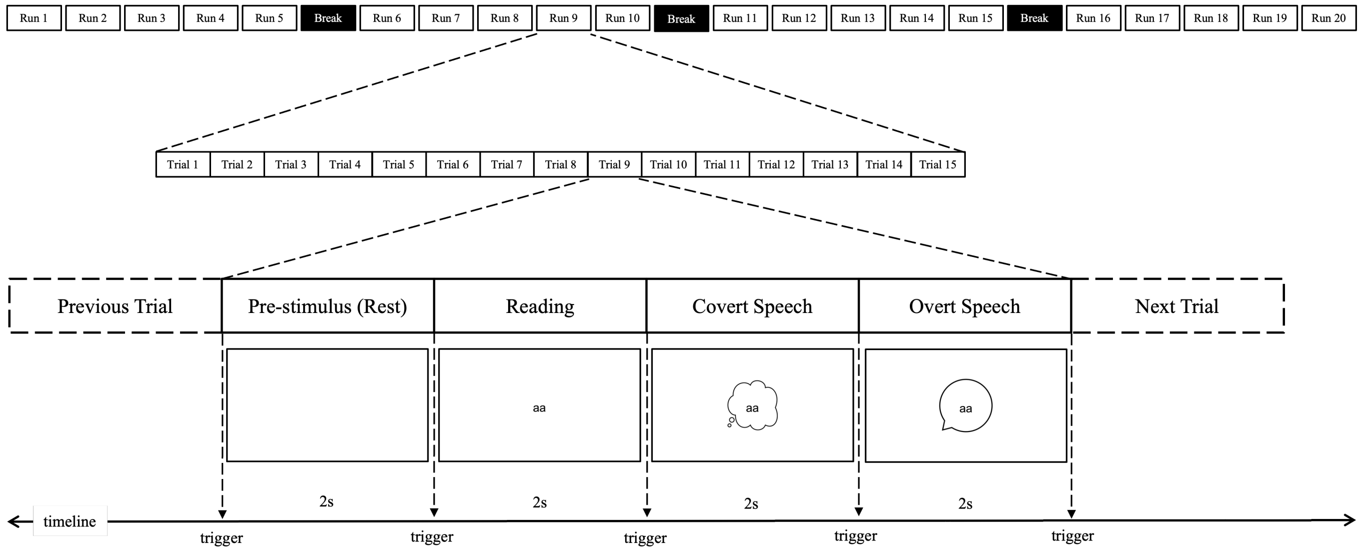

Fig. 3. Experimental protocol. The experiment consisted of 20 runs per subject. Each run consisted of 15 trials in which the different prompts were shown. The order of the prompts was randomized for each run. A trial consisted of four consecutive segments of two seconds: pre-stimulus (rest), reading, covert speech, and overt speech. At the start of each segments a different visual cue was shown on the monitor and at onset of this visual cue a trigger was sent to the TMSi SAGA. During the experiments, black text was used on a dark-grey coloured background for the visual cues.

blink. This segment was followed by a two second reading segment. For this segment, the prompt was shown on the screen and the subject was instructed to only read the prompt. During the following two second covert speech segment, the prompt was shown in a thought cloud and the subject was instructed to imagine the articulation of the prompt once without emitting sound or making any articulatory movement. The subject was told to focus on imagining the execution of the different articulatory gestures. Lastly, during the two second overt speech segment, the prompt was shown in a speech balloon and the subject was instructed to articulate the prompt once. During all segments except pre-stimulus (rest), the subject was instructed to avoid moving, swallowing, and blinking to reduce the presence of artifacts. To minimize eye fatigue, black text was used on a dark-grey coloured background for the visual cues. The subjects were instructed to perform the specific task once right after the visual cue appeared on the screen. By limiting the duration for covert speech task and by collecting covert and overt speech in a single trial, behavioral control can be applied to ensure that the subject only imagines the articulation of the presented prompt which they are expected to overtly pronounce in the same trial [40].

The top half of Figure 3 shows the schematic representation of the experimental timeline. Fifteen consecutive trials made up one single run. For each run, the prompts (i.e., the 15 Dutch vowels and words) were randomized using a balanced Latin square to reduce order effects and remove immediate carry-over effects [46], see Table X (Appendix B). A single session of the experiment consisted of 20 runs. Considering the mental effort required for executing the covert speech tasks, the experiment should not last more than approximately one hour per subject. Increasing the number of runs per sessions leads to fatigue and subsequent quality degradation of the recorded EEG data. After the 20 runs, each subject therefore performed each segments a total of 20 times per prompt. To prevent boredom and fatigue, a three-minute break was scheduled after every 5th run. Furthermore, the subject had the possibility to ask for a break between each run to relax and keep focus during the recording process. To ensure that the subjects were familiar with the experimental protocol, the experiment was explained and the different visual cues were shown during the placement of the EEG electrode cap using a test trial.

E. Data Pre-Processing

To clean, organise, and make the data ready for future use, the raw data was pre-processed. The EEG data was pre-processed and analysed with MATLAB (The MathWorks, Inc., USA) using custom scripts and functions from EEGLAB [47]. During pre-processing the following channels were deleted: the status (channel not used during data acquisition), counter (channel containing the sample numbers), M1, M2 (unused reference electrodes), and any channels that disconnected during the experiment. Other potentially bad channels (e.g., noisy channels) were not deleted. If no channels disconnected during the experiment, a total of 62 EEG channels remain per subject.

The 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 [25], [35]. A notch filter (Hamming windowed sinc FIR filter) between 49 and 51 Hz was applied to remove power line noise at 50 Hz. The filtering was done before data segmentation and artifact removal. The data was re-referenced after filtering using the average reference. Then, the data was segmented into trials (epochs) from 0.0 to +2.0 seconds after stimulus onset based on the trigger values from the trigger channel. Epochs containing eye blinks were marked using ERPLAB artifact detection (moving window peak-to-peak threshold) [48]. To preserve the original data as much as possible, as the relevant features of the covert speech paradigm are still unknown, and because there were enough epochs that did not contain any eye blinks for the covert and overt speech trials, it was decided not to use Independent Component Analysis (ICA) for the removal of eye artifacts. No EMG artifact detection was done in this study.

All files are organized and named using the EEG extension to the Brain Imaging Data Structure (BIDS) [49], [50]. The database structure can be found in Appendix A.

F. Data Analysis

1) Pre-Processing and Channel Selection: Subjects were excluded for the data analysis if they were left-handed, if their data contained multiple noisy channels, and if more than 40% of the covert or overt trials were marked (i.e., contain eye artifacts) during pre-processing. Only the pre-stimulus (rest), covert speech, and overt speech segments were used for the data analysis. The reading segments were not used because multiple subjects indicated that they found it difficult to differentiate between the reading task and the covert speech task. These subjects were already focused on the different articulatory gestures during the reading task in preparation to the covert speech task.

Due to the onset of the microphone occurring between the onset of the visual cue for the overt speech task and sending the trigger for the overt speech segment, the overt trigger was not sent to the TMSi SAGA directly after the visual cue of the overt speech was shown, but approximately 0.06 seconds later. As the trials were segmented based on these triggers, the overt speech epochs were segmented from 0.06 to +2.06 seconds after stimulus instead of 0.0 to +2.0 seconds after stimulus. To compensate for this delay and synchronise the timeline of the three different segments (pre-stimulus (rest), covert speech, and overt speech), the first 0.06 seconds of the pre-stimuli (rest) and covert speech segments and the last 0.06 seconds of the overt speech segments were neglected for the data analysis. The segment synchronisation approach is visualised in Figure 10 (Appendix B). The segments are not synchronised in the data saved in the database.

Based on the involvement of specific areas of the cortex in language processing [18], [35], [51]–[53], the following 16 EEG channels were chosen to be used for the analysis in this study (Figure 4):

1) FC1: Premotor cortex
2) FC3: Premotor cortex
3) Cz: Motor cortex
4) C4: Motor cortex
5) C3: Motor cortex
6) FC5: Broca's area
7) FT7: Broca's area, inferior temporal gyrus
8) F5: Broca's area
9) F7: Broca's area
10) C5: Wernicke's area, primary auditory cortex
11) T7: Middle temporal gyrus, secondary auditory cortex
12) CP3: Wernicke's area
13) CP5: Wernicke's area
14) TP7: Wernicke's area
15) P5: Wernicke's area
16) P3: Superior parietal lobule

The significance of the channels covering Broca's and Wernicke's areas for classifying covert speech has been shown by multiple studies using common spatial patterns (CSP) and event-related spectral perturbation (ERSP) [34], [35], [54]. Moreover, by discarding the EEG channels over the occipital lobe, the interference of the visual cues on the EEG recordings is greatly reduced as visual cues mainly elicit responses in the occipital lobe [25], [35].

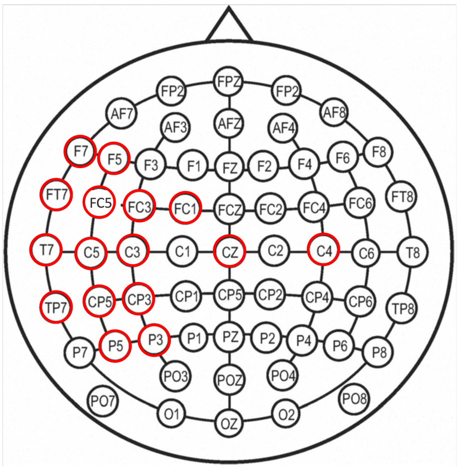

Fig. 4. Visualisation of the electrode placement according to the international 10-20 system with 62 channels. The sixteen EEG channels used in this study are colored red.

2) Event Related Potentials (ERP): As a method to validate the EEG signals collected in the database, the structural differences between the segments are visualised through averaging. The average of all epochs for the 16 pre-selected channels of each segment (i.e., pre-stimulus (rest), covert speech, and overt speech) for each subject were calculated to visualise event-related potentials (ERPs) in time series. Due to the high inter-subject variability in timing of speech, no grand average between subjects was calculated. A peak or a trough of the curve of the ERP waveshape is identified as an ERP component, which is thought to reflect the maximum activation of a brain process associated with a specific task in information processing. The ERP components can be divided into three main categories: exogenous, endogenous, and motor components. Exogenous, or early components (e.g., P1, N1, P2, and N2) are pre-attentive responses and are dependent on the physical characteristics of the stimulus. These exogenous components do not reflect cognitive processing and mostly occur withing 250 ms after the stimulus. Endogenous, or late components (e.g., P3) do reflect perceptual and cognitive aspects of information processing. Typically, the latency of P3 (or P300) response is between 250 and 400 ms. Endogenous components are completely task dependent. Motor components accompany both the preparation and the execution of a motor response. However, the boundaries between these categories are not always clear [55]–[57].

Visual cues mainly elicit responses in the occipital lobe [35]. As the EEG channels over the occipital lobe are not included in the pre-selected channels, the exogenous ERPs are less distinguishable in comparison to a plot containing all channels.

3) Classification Tasks: To validate the EEG signals collected in the database and to demonstrate the potential use of the data, a speaker-independent three-class classification task of pre-stimulus (rest) versus covert speech versus overt speech and a speaker-independent five-class classification task of the covert vowels were performed. Speaker-independent classification means that the training and testing data were from different subjects. For the covert and overt speech groups, the data from all covert prompts and all overt prompts (i.e., the five vowels and ten words) were combined. The trials were classified using a pre-trained ResNet-50 model.

4) ResNet: The pre-trained ResNet-50 model is pre-trained on more than a million images from the ImageNet database [30] to learn features from these images for a 1,000 class image classification task. The architecture of the ResNet-50 model is show in Figure 11 (Appendix B). The model consists of 48 convolution layers, 1 MaxPool, and 1 Average Pool layer. By re-training the pre-trained network on new data, the network can be fine-tuned to learn features specific to the new dataset. Fine-tuning the network for a relatively small dataset is faster than training an untrained network. The choice can be made to freeze (the weight of) layers in the network to speed up the learning even more and prevent overfitting when using a small dataset, but this does also prevent the network from learning in those frozen layers. As the new data is quite different from the images in the ImageNet database, the choice was made to not freeze the initial layers and thereby allow the weights of all layers to be updated during fine-tuning. Although no layers were frozen, fine-tuning a pre-trained network is still preferred over using an untrained network as it allows to build upon the generic features extracted by the initial layers and thereby speed up the process.

To tune these pre-trained models for the three-class (i.e., pre-stimulus vs. covert speech vs. overt speech) and five-class classification problem (i.e., five covert vowels) instead of the 1,000 class image classification problem the network was pre-trained for, the last two layers 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 five) and a new classification layer. These last two layers combine the features extracted by the model into class probabilities, a loss value, and the predicted label. As the classes for the new data differ from the data the model was pre-trained on, these layers need to be replaced.

5) Spatial-Spectral-Temporal Features: Discriminative features must be extracted from the EEG signals to classify the covert speech trials. Previous studies have demonstrated the discriminative value of spectral features of neural signals for decoding covert speech [40], [58], [59]. To use the frequency information of the cortex, wavelet scalograms of the pre-processed EEG signals were computed by performing continuous wavelet transform (CWT) with Morlet wavelets [40], [60], [61]. CWT can be described as a kind of template matching where the cross-covariance between the signal and a predefined wavelet is obtained by scaling and translating the mother wavelet (in this case the Morlet wavelet) across different scales. As the Morlet wavelet provides good resolution in both time and frequency domains [62] and because they are highly effective in capturing oscillatory neural activity [63], this wavelet was used to compute the scalograms. A scalogram is the time-frequency representation of the absolute value of the CWT coefficients of a signal, thereby providing spectral-temporal features of the recorded signal [40], [64]. Morlet scalogram images were generated from the 16 channels of single trial EEG signals, see Figure 5. Only the trials that did not contain eye blink artifacts were used. The frequency was plotted on a logarithmic scale and the maximum wavelet band-pass frequency was set on 70 Hz. The maximum magnitude was set on 10. To use the spatial information, a 4 x 4 matrix of the scalogram images from the 16 channels of single trial EEG signals were created, see Figure 6. The matrices of scalogram images were resized to 224 x 224 x 3 as per input requirement for the ResNet-50 network. By combining information from multiple channels, information transfer between different brain regions can be captured [25].

6) Data Splitting: Leave-three-out cross-validation was performed to avoid any bias in the division of the database, to prevent overfitting, and to increase the generalizability of the results. The included data was split into a training set, a validation set, and a test set, using the data from three subjects as part of the test set (i.e., leaving three out). The data from two subjects were assigned to the validation set, and the data from the remaining subjects were assigned to the training set. Subsequently, the model was trained using the training and validation set, and tested using the test set. This process was repeated multiple times with different subjects in each set (creating multiple folds) until the data from each subjects had been in the test set once. The test data set was completely new for the model (i.e., the test data did not contain any augmented data of the train or validation set).

7) Hyperparameters: The maximum number of epochs was set to 100, with early stopping if there was a continuous

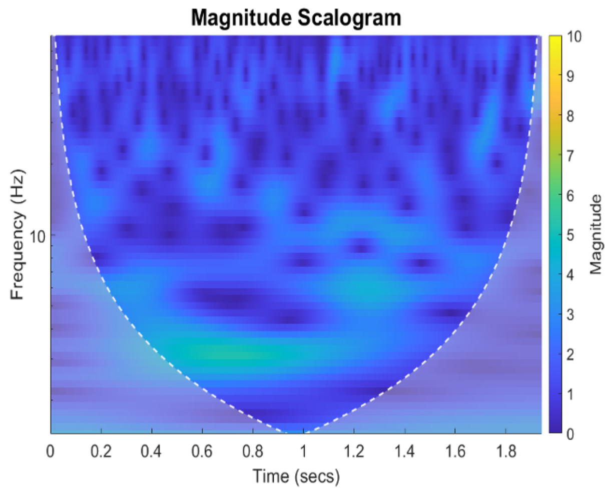

Fig. 5. Scalogram plot of the EEG signal of a single trial. This specific image represents the cone of influence (COI) plot of the scalogram representation of the neural signal obtained from channel FC5 during covert speech of the vowel ”aa” by subject 5 (trial 2). The color bar reflects the change of energy of the various CWT coefficients obtained with the Morlet wavelet.

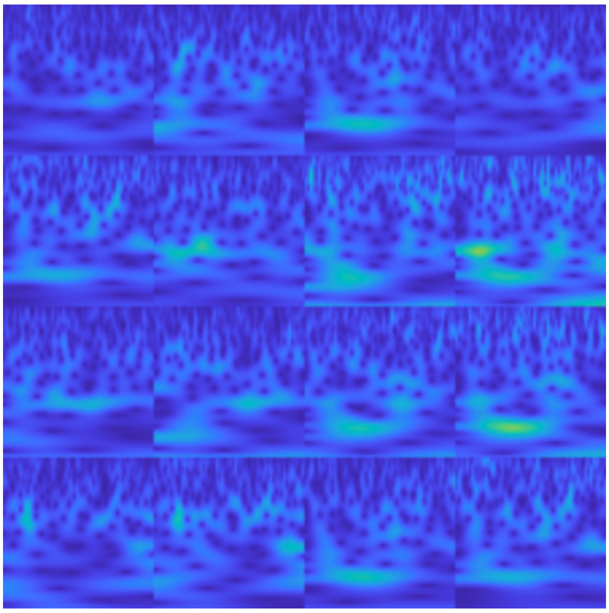

Fig. 6. Scalogram matrix. The image represents a 4 x 4 matrix of the scalograms from the 16 selected channels obtained from the sixteen selected channels during covert speech of the vowel ”aa” by subject 5 (trial 2).

increase in validation loss for more than 5 epochs (i.e., validation patience of 5). For the ResNet-50 model, an Adam optimizer, a minibatch size of 32, and an initial learning rate of 0.001 was used.

8) Classification Evaluation: The final classification accuracy per classification task was calculated by averaging over the multiple folds. For both tasks, the confusion matrices were obtained by combining the results from the five folds of the three-class classification task and the five-class classification task respectively. The confusion matrix provides a summary of the prediction results on the classification task and can be used to evaluate the performance of a network. Each number in the confusion matrix indicated the number of observations of a class (true label) identified as any class (predicted label). A perfect network would create a diagonal confusion matrix. From the confusion matrix, the sensitivity and the specificity were calculated per class. Sensitivity, or the true positive rate (TPR), is the proportion of trials from a specific class that got predicted as being that specific class. Specificity, or the true negative rate (TNR), is the proportion of trials not from a specific class that did correctly get predicted as not being that specific class. The sensitivity and specificity were calculated using equation 1 and 2.

$$Sensitivity = \frac{TruePositive}{TruePositive+FalseNegative} \tag{1}$$

$$Specificity = \frac{TrueNegative}{TrueNegative+FalsePositive} \tag{2}$$

III. Results

A. Database Results

In total, the EEG and audio signals of 20 subjects were acquired using the proposed experimental protocol and set-up. Almost all subjects completed 20 runs of 15 prompts and thereby fulfilled the 20 repetitions per prompt. Subject 2 only completed 19 runs, as run 7 was disturbed due to technical difficulties. One channel (subject 1, channel FC2) disconnected during the experiment and this channel was therefore deleted during pre-processing from the data.

A total of 24.370 trials were recorded for the 4 different segments and the 15 different prompts. After pre-processing, 16.510 trials (68%) were retained from the recorded data. The number of trials per segments recorded, after data pre-processing, and average per prompt per subject can be found in Table I. For the reading, covert speech, and overt speech segments, a high percentage of trials was retained (all above 70%). For the pre-stimulus (rest) trials a considerable lower number of trials was retained (33%), which can be attributed to the fact that subjects were allowed to blink during the pre-stimulus (rest) segments. The low number of retained pre-stimulus (rest) equate to an average of 105 pre-stimulus (rest) trials per subjects which remains higher than the number of trials per prompt per subject for the other three segments. The number of covert speech and overt speech trials per prompt per subject after data pre-processing can be found in Table VIII and Table IX (Appendix B).

TABLE I
Number of trials recorded, after data pre-processing, and average per prompt per subject after pre-processing for the different segments (pre-stimulus (rest), reading, covert speech, and overt speech).

Task	Trials recorded after	Trials retained after pre-processing (%)	Average trials per subject per prompt
Pre-stimulus (rest)	6392	2108 (33%)	105
Reading	5993	4540 (76%)	15
Covert Speech	5993	5550 (93%)	19
Overt Speech	5992	4312 (72%)	14

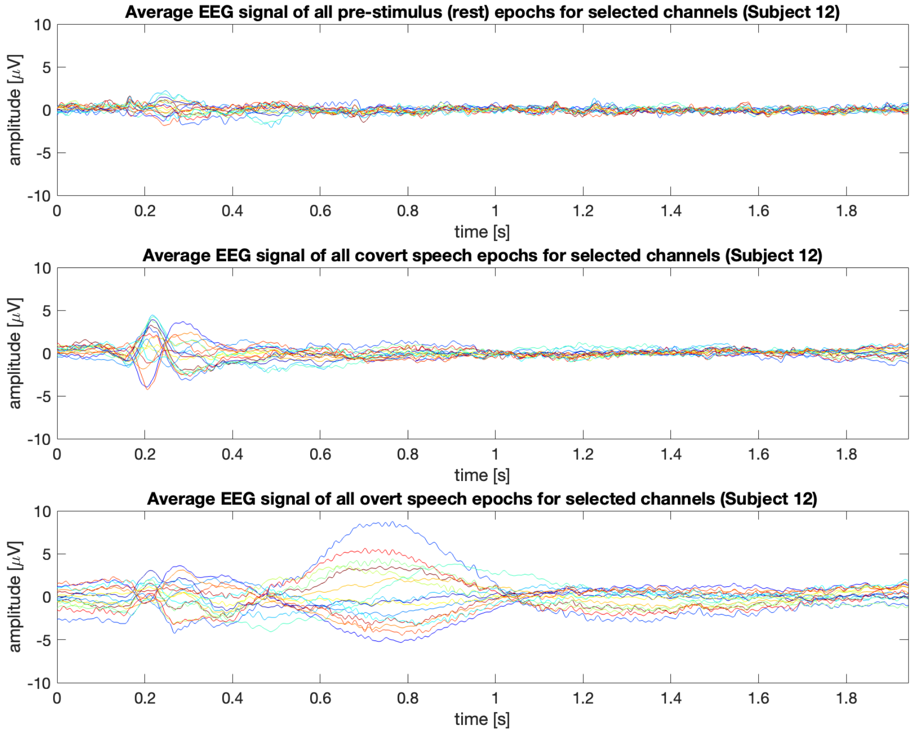

Fig. 7. Average EEG of all epochs for the synchronised segments (i.e., pre-stimulus, covert speech, and overt speech) for subject 12 to visualise event-related potentials (ERPs) in time series. The different lines correspond to the 16 selected channels.

B. Data Analysis Results

1) Excluded Subjects: Five of the 20 subjects were excluded from the data used for the data analysis. Subject 9 and subject 13 were excluded because the subjects were left-handed, subject 7 and subject 17 were excluded because their signals contained multiple noisy channels, and subject 2 was excluded because a large part of the overt speech trials were rejected as they contained eye blinks, causing an imbalance in the number of covert speech trials versus the overt speech trials.

2) Event Related Potentials (ERP): Visual inspection of the ERPs of the subjects for the different segments revealed clear differences between EEG data of pre-stimulus (rest), covert speech, and overt speech. The average EEG signals of subject 12 (chosen because of clarity) for the pre-stimulus (rest), covert speech, and overt speech segment are shown in Figure 7. For both the covert and the overt speech task exogenous and endogenous ERP components are found following the onset of the visual cue, recognised in the covert and overt segments by the peaks around 0.2 and 0.3 seconds, see Figure 7.

For the covert speech and the overt speech task, these peaks are followed by approximately 100-200 ms of enhanced activity. In the averaged EEG signals from subject 12, this enhance activity can be observed between approximately 0.3 and 0.5 seconds. For the overt speech task, this enhanced activity is followed by a broad peak/trough (depending on the channel) coinciding with the acoustic onset and therefore associated with voluntary movement of the articulators. For subject 12, this averaged peak/trough starts around 0.5 seconds and ends around 1.0 seconds, see Figure 7. The latency of this broader peak/trough differs greatly between subjects as it is associated with timing of speech. The broad peak/trough is not seen in the covert speech epochs. No apparent ERPs were found for the rest task, only background EEG activity.

3) Data Splitting: The leave-three-out cross-validation approach outlined in section II-F6 was used to split the data from the 15 remaining subjects. A total of five folds was created. For each fold the data from 10 subjects were assigned to the training set, the data from 2 subjects were assigned to the validation set, and the data from 3 subjects were assigned to the test set. An overview of how the data was split for each fold can be found in Figure 8. The same data splitting approach was used for both classification tasks.

The number of trials used for training and evaluating the deep learning algorithm per set and per fold can be found in Table II. The average number of trials used for training the three-class classification task and the five-class classification task was 6088 (± 96) and 943 (± 20) respectively. This equates to an average of 2029 training trials per class for the

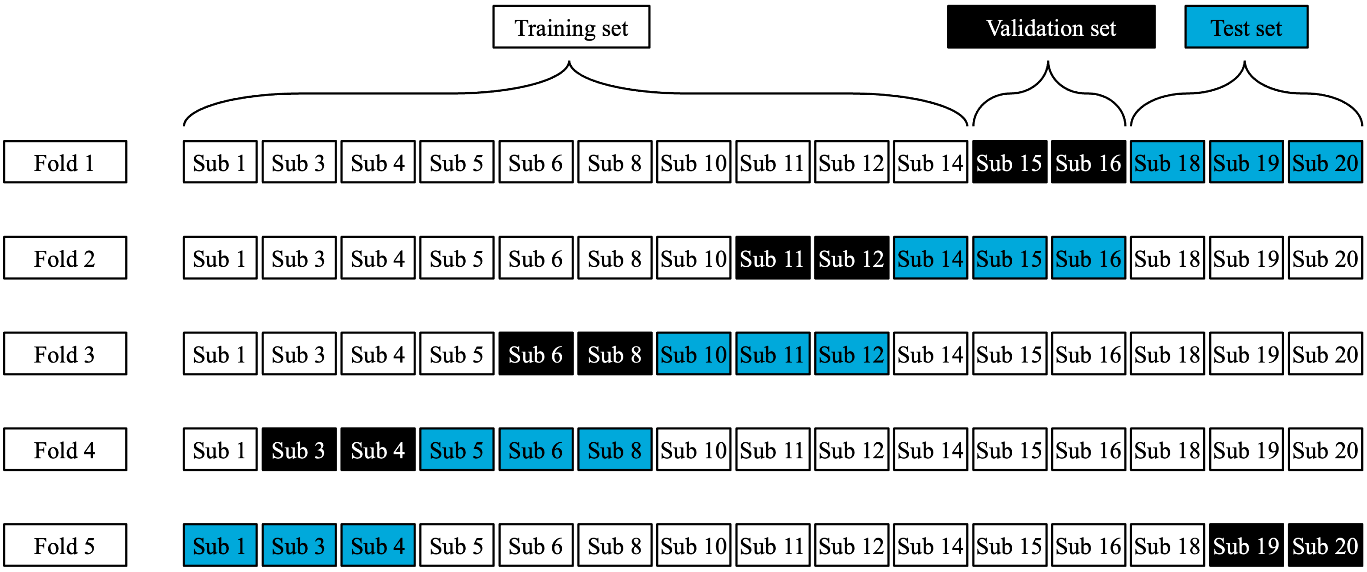
Training set
Validation set
Test set
Fold 1 Sub 1 Sub 3 Sub 4 Sub 5 Sub 6 Sub 8 Sub 10 Sub 11 Sub 12 Sub 14 Sub 15 Sub 16 Sub 18 Sub 19 Sub 20
Fold 2 Sub 1 Sub 3 Sub 4 Sub 5 Sub 6 Sub 8 Sub 10 Sub 11 Sub 12 Sub 14 Sub 15 Sub 16 Sub 18 Sub 19 Sub 20
Fold 3 Sub 1 Sub 3 Sub 4 Sub 5 Sub 6 Sub 8 Sub 10 Sub 11 Sub 12 Sub 14 Sub 15 Sub 16 Sub 18 Sub 19 Sub 20
Fold 4 Sub 1 Sub 3 Sub 4 Sub 5 Sub 6 Sub 8 Sub 10 Sub 11 Sub 12 Sub 14 Sub 15 Sub 16 Sub 18 Sub 19 Sub 20
Fold 5 Sub 1 Sub 3 Sub 4 Sub 5 Sub 6 Sub 8 Sub 10 Sub 11 Sub 12 Sub 14 Sub 15 Sub 16 Sub 18 Sub 19 Sub 20

Fig. 8. Leave-three-out cross-validation diagram (15 subjects included). For each fold, the data from 10 subjects are assigned to the training set, the data from 2 subjects are assigned to the validation set, and the data from 3 subjects are assigned to the test set. The test set of each fold consists of the data from different subjects.

TABLE II
NUMBER OF TRIALS PER FOLD USED FOR TRAINING, VALIDATING, AND TESTING THE NETWORK.

	Pre-stimulus vs. Overt vs. Covert			Covert vowel classification (five-class)		
	Training set	**Validation set**	**Test set**	**Training set**	**Validation set**	**Test set**
Fold 1	6141 (67%)	1122 (12%)	1878 (21%)	965 (68%)	186 (13%)	261 (18%)
Fold 2	6185 (68%)	1397 (15%)	1559 (17%)	937 (66%)	196 (14%)	279 (20%)
Fold 3	6000 (66%)	1103 (12%)	2038 (22%)	923 (65%)	193 (14%)	296 (21%)
Fold 4	6145 (67%)	1324 (14%)	1672 (18%)	927 (66%)	193 (14%)	292 (21%)
Fold 5	5970 (65%)	1177 (13%)	1994 (22%)	963 (68%)	165 (12%)	284 (20%)

three-class classification task and 189 training trials per class for the five-class classification task.

4) Three-Class Classification Task: The three-class classification task of pre-stimulus (rest) trials versus covert speech trials versus overt speech trials reached an average classification accuracy of 70.6% (± 4.4%), which is significantly higher than chance level accuracy (33.3%) (1-tail t-test, $p < .05$). The classification accuracies per fold are presented in Table III. Table IV shows the overall confusion matrix obtained by combing the results from the five folds of the three-class classification task. The sensitivity and specificity deducted from the confusion matrix for each class indicate a high classification performance for both the overt speech class (sensitivity = 74.8%, specificity = 91.7%) and the covert speech class (sensitivity = 78.6%, specificity = 71.1%). These values indicate that the network was able to correctly identify the covert and overt speech trials, and that the network was able to reduce to number of false positives for the covert and overt speech segments. Overt speech trials were most often misclassified as covert speech trials (19%) and covert speech trials were most often misclassified as pre-stimulus (rest) trials (14%). In contrast, the classification performance for the pre-stimulus (rest) class was considerably lower (sensitivity = 33.4%, specificity = 89.2%). These values indicate that the network failed to correctly identify the pre-stimulus trials, but was capable of reducing the number of false positives for the pre-stimulus trials. Pre-stimulus trials were most often misclassified as covert speech trials (54%).

5) Five-Class Classification Task: The five-class classification task of the five covert vowels (“aa”, “ee”, “oo”, “ie”, “oe”) reached an average classification accuracy of 19.6% (± 2.1%), which is not significantly different than chance level accuracy (20%) (2-tail t-test, $p < .05$). The classification accuracies per fold are presented in Table III. Table V shows the overall confusion matrix obtained by combing the results from the five folds of the five-class classification task. The sensitivity and specificity of all classes are listed in the caption of Table V. The sensitivity values are consistently low, ranging from 12.7% for class “ie” to 28.9% for class “aa”, while the specificity values are consistently high, ranging from 65.9% for class “oe” to 81.7% for class “oo”. For class “aa” and “oe” the sensitivity is higher is comparison to the other classes (“ee”, “ie”, “oo”) indicating that the network slightly favoured these classes. However, based on the around chance level accuracy and the consistent low sensitivity and high specificity in all classes, the model performed similar to a random/naive model.

TABLE III

Classification results for the subject-independent three-class classification task (pre-stimulus vs. covert speech vs. overt speech) and the five-class classification task of covert vowels. The final classification accuracy is computed by averaging over the five folds.

	Pre-stimulus vs. Covert vs. Overt		**Five class covert vowel classification**	
	Validation accuracy	**Test accuracy**	**Validation accuracy**	**Test accuracy**
Fold 1	81.1%	68.4%	19.4%	18.4%
Fold 2	67.7%	78.0%	19.4%	18.6%
Fold 3	69.8%	70.8%	18.1%	23.0%
Fold 4	69.0%	68.9%	18.1%	20.2%
Fold 5	74.5%	66.7%	20.0%	18.0%
Average	72.4 ± 5.5%	70.6 ± 4.4%	19.0 ± 0.8%	19.6 ± 2.1%

TABLE IV

Confusion matrix obtained by combining the results from the five folds of the three-class classification task (pre-stimulus vs. covert speech vs. overt speech). For overt speech. sensitivity = 74.8%, specificity = 91.7%; for covert speech, sensitivity = 78.6%, specificity = 71.1%; for pre-stimulus (rest), sensitivity = 33.4%, specificity = 89.2%.

			Predicted Class	
		Overt	**Covert**	**Pre-stimulus**
True Class	**Overt**	2598 (75%)	654 (19%)	222 (6%)
	Covert	299 (7%)	3364 (79%)	615 (14%)
	Pre-stimulus	174 (13%)	751 (54%)	464 (33%)

TABLE V

Confusion matrix obtained by combining the results from the five folds of the five-class classification task (covert vowels). For prompt “aa”, sensitivity = 28.9%, specificity = 72.1%; for prompt “ee”, sensitivity = 15.6%, specificity = 77.1%; for prompt “ie”, sensitivity = 12.7%, specificity = 74.9%; for prompt “oe” , sensitivity = 27.4%, specificity = 65.9%; for prompt “oo”, sensitivity = 13.3%, specificity = 81.7%.

				Predicted Class		
		aa	**ee**	**ie**	**oe**	**oo**
True Class	**aa**	83 (29%)	44 (15%)	48 (17%)	60 (21%)	52 (18%)
	ee	78 (28%)	44 (16%)	39 (14%)	71 (25%)	50 (18%)
	ie	83 (29%)	44 (15%)	36 (13%)	69 (24%)	52 (18%)
	oe	75 (26%)	39 (14%)	40 (14%)	79 (27%)	55 (19%)
	oo	78 (29%)	41 (15%)	39 (14%)	77 (28%)	36 (13%)

IV. Discussion

A. Database

The aim of this study was to provide a novel database consisting of EEG and audio recordings during the covert and overt pronunciation of 15 Dutch prompts. In total, 5993 covert speech trials and 5992 overt speech trials were recorded from 20 subjects using 64 EEG channels. After data pre-processing, an average of 19 covert speech and 14 overt speech trials per prompt per subject were retained from the 20 recorded trials per prompt per subject. To the best of our knowledge this is the first database containing Dutch covert prompts.

There are five databases that are deployed by multiple different articles in the current literature on decoding speech [16], [34], [35], [38], [39]. When comparing our database to these often-employed databases, a few things stand out. Firstly, only one database [39] contains data from more subjects (27 subjects) than our database. Secondly, the electrode density is equal to the highest electrode density found in the often-employed databases (64 channels). The higher electrode density provides an increase in spatial resolution, which translates to improved potential localization and more information captured. Thirdly, most databases recorded more trials per prompt per subject (ranging from 33 to 100 trials) than our database (20 trials). Only one database recorded less trials per prompt per subject (12 trials). The lower number of trials per prompt per subject in comparison to the other databases is a result of the experimental protocol used. The other databases do not contain overt speech trials, used repeated covert speech in a single trial, and/or had an extreme long experiment duration (3.5 hours). Moreover, the number of prompts in our database (15 prompts) is higher than the number of prompts in the often-employed databases (ranging from 5 to 12 prompts), which also decreases the number of trials per prompt. Increasing the number of prompts leads to a higher number of degrees of freedom for the dataset. This in turn increases the usability for different possible analyses.

To sum things up, our database contains data of a high number of subjects acquired with a high electrode density.

The number of trials per prompts per subjects is lower, but the database does contain a higher number of different prompts.

B. Event Related Potentials (ERP)

The distinct differences between the averaged EEG data of all epochs for the 3 segments (pre-stimulus (rest), covert speech, and overt speech) show the discriminative value of the EEG signals from the 16 channels in classifying trials of the three segments. The difference between the averaged EEG data for the pre-stimulus (rest) trials and the covert speech trials indicates that the subjects did actively engage in cognitive processing during the covert speech segments and did not simply relax (i.e., resting state). It furthermore demonstrates the existence of information in the covert speech trials that can be used to decode covert speech. The similarities between the covert speech and the overt speech trials indicate that the two tasks both activate specific areas involved in language processing. The major noticeable difference between the covert and overt speech trials is the broad peak/trough found in the overt speech segments but not in the covert speech segments. As this peak is associated with voluntary movement of the articulators, this strongly suggests that no structural articulatory movements were made during the covert speech task. Although it is extremely difficult to verify whether the subjects performed true covert speech during the covert speech segments, the difference between the averaged EEG data does indicate that the subjects performed a mental activity distinctly different than overt speech and rest.

C. Classification

The EEG signals collected in the database were analysed by performing two analyses aimed at demonstrating the potential use of the data: a speaker-independent three-class classification task of pre-stimulus (rest) versus covert speech versus overt speech and a speaker-independent five-class classification task of the covert vowels (“aa”, “ee”, “oo”, “ie”, “oe”).

The ResNet-50 algorithm with spatial-spectral-temporal features reached a classification accuracy of 70.6% (± 4.4%) for the speaker-independent three-class classification task of pre-stimulus (rest) versus covert speech versus overt speech. This result indicates that the EEG signals from the 16 channels contains discriminative value in classifying trials of the three segments. The performance for both the covert speech (sensitivity = 78.6%, specificity = 71.1%) and the overt speech class (sensitivity = 74.8%, specificity = 91.7%) was high. This indicates that the classifier is able to distinguish covert and overt speech trials from each other and from pre-stimulus (rest) trials. The high classification performance for the covert speech segments provides further evidence that the subjects structurally performed covert speech during the covert speech trials. In contrast, the classification performance for the pre-stimulus (rest) trials was considerably lower (sensitivity = 33.4%, specificity = 89.2%). Something that could have contributed to this low classification accuracy is the experimental paradigm for the pre-stimulus (rest) segment. During the pre-stimulus (rest) segment, the subjects were allowed to relax and think without constraints. The subjects were instructed to begin this rest state directly after the blank screen appeared. However, due to the short duration of the pre-stimulus (rest) segment (2 seconds), the presumably resting state might be influenced by the overt speech segments of the trial prior to the pre-stimulus (rest) segment (carryover effect). This makes it more difficult to distinguish between trials of the different segments due to overlapping tasks. To achieve true resting state, the duration of the segment should be in magnitude of minutes instead of seconds [65], [66]. Previous studies that performed binary classification of covert speech versus rest achieved a similar classification performance. Lee et al. [38] and Sereshkeh et al. [67] reached an average subject-dependent classification accuracy of 79.65% and 75.94% respectively for the binary classification of covert speech versus rest. The high classification results of the three-class classification task show that the ResNet-50 model using spatial-spectral-temporal features is suitable for classifying covert and overt speech trials.

The ResNet-50 algorithm with spatial-spectral-temporal features reached a classification accuracy of 19.6% (± 2.1%) for the speaker-independent five-class covert vowel classification task, which is not significant different from chance level (20%). The classifier performed badly for all classes.

There are two studies that similarly use a ResNet algorithm to decode covert speech in the literature [13], [27]. Panachakel and Ganesan [27] used Resnet-50 and performed data augmentation to increase the sample size (100 trials per prompt) with a factor 17. Their approach reached an above 80% subject-dependent decoding accuracy for long words (“independent”, “cooperate”), short words (“in”, “out”, “up”), vowels (“a”, “i”, “u”), and short-long words (“in”, “cooperate”). However, when the classifier was trained for the short-long words classification task without data augmentation, the classification accuracy dropped to chance level for all subjects. This result suggests that 100 trials per prompt per subject is not enough to properly train the ResNet algorithm for these tasks. Vorontsova et al. [13] trained a ResNet-18 model on the data from 1, 2, 32, and 256 subjects. The nine-class (nine Russian words) classification task reached an above 80% accuracy when trained on the data from 1 and 2 subjects. The classification accuracy dropped below 20% when trained on 32 and 256 subjects. When tested on out-of sample data (i.e., training and testing data were from different subjects), the classification accuracy dropped to chance level for all training sizes. The big difference between the subject-dependent and the subject-independent classification performance shows that features learned on a limited dataset might not be transferable to the general population.

Through comparison of our work to the two studies employing ResNet as classifier, it can be deduced that the low classification performance of the five-class classification task can be mainly attributed to two factors: the subject-independent training and the number of trials per prompt per subject. Other studies that deploy a type of CNN for covert vowel classification [12], [17], [26], [68] all focus on subject-dependent classification, where training and testing data were from the same subject. Even though these studies report above chance level classification accuracies per subject, the features learned by these networks might be subject-specific and therefore not generalizable across the population. As shown by Vorontsova et al. [13] and stated by Panachakel and Ganesan [25], these well performing subject-dependent classifiers are likely to perform poorly for data from unseen subjects. This is consistent with our results for the subject-independent five-class covert vowel classification task. Performing subject-independent classification is very challenging given the cognitive variance between subjects [69]. The cognitive variance between subjects is so strong that training a BCI on a single subject is more effective for that specific subject than training the BCI on a larger dataset collected on a group of subjects despite the much smaller amount of training data [13].

The average number of trials per class used for training the three-class classification task and the five-class classification task was 2029 and 189 respectively. In comparison, the ResNet-50 model was pre-trained on more than a million images for 1,000 classes in the ImageNet database (i.e., an average of 1,000 images per class) [30]. As the new data is quite different from the images in the ImageNet database, it is reasonable to assume that the network requires data of a similar magnitude to fine-tune the model. The training set for the three-class classification task is therefore sufficient for fine-tuning the model. However, the training set for the five-class classification task can be considered as a relatively small sample dataset for proper fine-tuning. Especially when taking into account the cognitive variance between subjects, a higher number of data than the average of 19 trials per prompt per subjects is required to properly fine-tune the ResNet-50 algorithm [27], [40]. Multiple studies that have a similar number of trials to their disposal employ data augmentation techniques to address the lack of enough data for training the deep networks [12], [27], [35], [40].

D. Limitations

The limitations of this study can be subdivided into limitations of the developed database and limitations of the performed classification tasks.

The first limitation of the database is that there is no guarantee that the subject did in fact execute the correct mental activity during the covert speech segments, despite the efforts to ensure that the subjects only imagines the pronunciation of the presented prompt. All subjects were naive BCI users and even though they all received the same instruction to imagine the articulation of the different prompts without emitting sound or making any articulatory movements, the interpretation of the instructions and the mental activity executed may differ between subjects [14], [40], [70].

Another limitation is the number of trials recorded per prompt per subject. By collecting covert and overt speech consecutively in a single trial, the duration of the trials is considerably extended. In combination with the relatively high number of prompts, the total number of trials recorded per prompt per subject is reduced.

In regards to the limitation of the classification task, only 16 EEG channels were used for the data analysis based on the involvement of the specific areas of the cortex in the production of speech. Although it has been shown that that both covert and overt speech tasks activate the essential language areas covered by these channels, other important centres for speech and language are spread widely throughout the brain [71], [72]. By only using the 16 channels, less information is captured, which could have affected the classification accuracy [25].

The second limitation of the classification task is the inclusion of the cone of influence (COI) in the scalogram images used for creating the 4 x 4 scalogram matrices. The COI shows the area of the scalogram where edge effects might have distorted the scalogram. These effects arise due to finite-length time series and affect the areas where the scaled wavelet extends beyond the edge of the finite signal [73], [74]. By including the COI in the matrix, edge effects might have influenced the classification results.

E. Future Research

Future research into improving the database should focus on detecting and removing EMG artifacts. Besides eye blink detection, no further EMG artifact detection was done in this study. Some of the movement artifacts might have been detected by the moving window peak-to-peak threshold deployed for the eye blink detection. However, especially the overt speech trials are probably contaminated with movement artifacts [70]. Furthermore, as no subject-dependent classification was performed in this study, the performance per subject was not determined. The data from a single subject who misunderstood the covert speech task could have introduced significant distortions into the dataset and subsequently have negatively influenced the results of the subject-independent classification tasks. Future research should look into the performance per subject and determine whether the subject exclusion criteria were valid.

Future research into improving the classification performance should investigate subject normalization strategies to address the challenges of subject-independent covert speech decoding. Techniques to normalize the EEG acquired from different subjects can help decrease the variability between the EEG signals and subsequently improve the classification accuracy. Other strategies to help improve the performance of classifiers should also be explored (especially for small-data size problems), such as transfer learning [26], [68], [75] and data augmentation [27], [40]. Transfer learning incorporates knowledge learnt from one domain to improve the classification performance of a new domain. Data augmentation techniques (e.g., overlapping window) can be used to generate more data from the already existing data. Although data augmentation has been demonstrated to be an effective method to increase the data size, it also induces the variability (i.e., introduces data bias) [40]. Also, to further increase the classification accuracy, research should focus on finding the best combination of features and classifier for covert speech classification tasks.

Finally, to better understand the true performance of systems, further research should be done using patients with neuromuscular diseases instead of only using healthy subjects. This will subsequently improve the practical applicability of the BCI.

V. Conclusion

In this study, a novel database containing EEG and audio data of Dutch covert and overt speech is presented. The database is structured according to the BIDS and contains data from 20 subjects acquired with a high electrode density of 64 channels. Our database provides a starting point for future research and facilitates the development of classification algorithms.

The usability of the EEG signals was demonstrated by a speaker-independent three-class classification task of pre-stimulus (rest) versus covert speech versus overt speech and a speaker-independent five-class classification task of covert vowels. Although the five-class classification task did not provide an above chance level accuracy, the high accuracy obtained during the three-class classification task is encouraging and demonstrates the existence of discriminative information in the covert speech trials to decode covert speech in the future.

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, 2021, pp. 389–392.

[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] N. Birbaumer, "Brain-computer-interface research: coming of age." 2006.

[4] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: a focused review," *Frontiers in neuroscience*, vol. 10, p. 429, 2016.

[5] M. Perrone-Bertolotti, L. Rapin, J.-P. Lachaux, M. Baciu, and H. Loevenbruck, "What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring," *Behavioural brain research*, vol. 261, pp. 220–239, 2014.

[6] S. Patel, "From speech to voice: on the content of inner speech," *Synthese*, vol. 199, no. 3, pp. 10 929–10 952, 2021.

[7] V. N. Kiroy, O. Bakhtin, E. Krivko, D. M. Lazurenko, E. Aslanyan, D. Shaposhnikov, and I. V. Shcherban, "Spoken and inner speech-related eeg connectivity in different spatial direction," *Biomedical Signal Processing and Control*, vol. 71, p. 103224, 2022.

[8] C. J. Price, "A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading," *Neuroimage*, vol. 62, no. 2, pp. 816–847, 2012.

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

[10] L. I. Shuster and S. K. Lemieux, "An fmri investigation of covertly and overtly produced mono-and multisyllabic words," *Brain and language*, vol. 93, no. 1, pp. 20–31, 2005.

[11] S. Basho, E. D. Palmer, M. A. Rubio, B. Wulfeck, and R.-A. Müller, "Effects of generation mode in fmri adaptations of semantic fluency: paced production and overt speech," *Neuropsychologia*, vol. 45, no. 8, pp. 1697–1706, 2007.

[12] 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.

[13] D. Vorontsova, I. Menshikov, A. Zubov, K. Orlov, P. Rikunov, E. Zvereva, L. Flitman, A. Lanikin, A. Sokolova, S. Markov *et al.*, "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.

[14] 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)*. IEEE, 2018, pp. 1–7.

[15] 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)*. IEEE, 2019, pp. 3050–3053.

[16] 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.

[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] 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)*. IEEE, 2019, pp. 1–4.

[19] M. Chen, Y. Liao, J. Liu, W. Fang, N. Hong, X. Ye, J. Li, Q. Tang, W. Pan, and W. Liao, "Comparison of sexual knowledge, attitude, and behavior between female chinese college students from urban areas and rural areas: a hidden challenge for hiv/aids control in china," *BioMed research international*, vol. 2016, 2016.

[20] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. Torres-García, "Tensor decomposition for imagined speech discrimination in eeg," in *Mexican International Conference on Artificial Intelligence*. Springer, 2018, pp. 239–249.

[21] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," *Biomedical Engineering Letters*, vol. 10, no. 2, pp. 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," *JOURNAL OF BEIJING INSTITUTE OF TECHNOLOGY*, vol. 30, no. zk, pp. 44–51, 2021.

[24] 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.

[25] J. T. Panachakel and R. A. Ganesan, "Decoding covert speech from eeg-a comprehensive review," *Frontiers in Neuroscience*, p. 392, 2021.

[26] 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)*. IEEE, 2019, pp. 1311–1316.

[27] J. T. Panachakel and R. A. Ganesan, "Decoding imagined speech from eeg using transfer learning," *IEEE Access*, vol. 9, pp. 135 371–135 383, 2021.

[28] M. M. Islam and M. M. H. Shuvo, "Densenet based speech imagery eeg signal classification using gramian angular field," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*. IEEE, 2019, pp. 149–154.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[31] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[32] J. Wang, J. Yang, Y. Wang, Y. Bai, T. Zhang, and D. Yao, “Ensemble decision approach with dislocated time–frequency representation and pre-trained cnn for fault diagnosis of railway vehicle gearboxes under variable conditions,” *International Journal of Rail Transportation*, pp. 1–19, 2021.

[33] J. T. Panachakel, A. Ramakrishnan, and T. Ananthapadmanabha, “A novel deep learning architecture for decoding imagined speech from eeg,” *arXiv preprint arXiv:2003.09374*, 2020.

[34] S. Zhao and F. Rudzicz, “Classifying phonological categories in imagined and articulated speech,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 992–996.

[35] 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.

[36] H. Clahsen and C. Felser, “How native-like is non-native language processing?” *Trends in cognitive sciences*, vol. 10, no. 12, pp. 564–570, 2006.

[37] L. Dekydtspotter, B. D. Schwartz, and R. A. Sprouse, “The comparative fallacy in l2 processing research,” in *Proceedings of the 8th generative approaches to second language acquisition conference (GASLA 2006)*, vol. 3340. Cascadilla Proceedings Project Somerville, MA, 2006.

[38] S.-H. Lee, M. Lee, J.-H. Jeong, and S.-W. Lee, “Towards an eeg-based intuitive bci communication system using imagined speech and visual imagery,” in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 4409–4414.

[39] A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, and G. García-Aguilar, “Implementing a fuzzy inference system in a multi-objective eeg channel selection model for imagined speech classification,” *Expert Systems with Applications*, vol. 59, pp. 1–12, 2016.

[40] D. Dash, P. Ferrari, and J. Wang, “Decoding imagined and spoken phrases from non-invasive neural (meg) signals,” *Frontiers in neuroscience*, vol. 14, p. 290, 2020.

[41] S. Knecht, B. Dräger, M. Deppe, L. Bobe, H. Lohmann, A. Flöel, E.-B. Ringelstein, and H. Henningsen, “Handedness and hemispheric language dominance in healthy humans,” *Brain*, vol. 123, no. 12, pp. 2512–2518, 2000.

[42] B. Stemmer and H. A. Whitaker, *Handbook of the Neuroscience of Language*. Academic Press, 2008.

[43] M. Kleiner, D. Brainard, and D. Pelli, “What’s new in psychtoolbox-3?” 2007.

[44] H. H. Jasper, “The ten-twenty electrode system of the international federation,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 370–375, 1958.

[45] C. Beverley, M. Inger M *et al.*, “The phonetics of english and dutch,” 2016.

[46] A. D. Keedwell and J. Dénes, *Latin squares and their applications*. Elsevier, 2015.

[47] A. Delorme and S. Makeig, “Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis,” *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[48] J. Lopez-Calderon and S. J. Luck, “Erplab: an open-source toolbox for the analysis of event-related potentials,” *Frontiers in human neuroscience*, vol. 8, p. 213, 2014.

[49] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko *et al.*, “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[50] C. R. Pernet, S. Appelhoff, K. J. Gorgolewski, G. Flandin, C. Phillips, A. Delorme, and R. Oostenveld, “Eeg-bids, an extension to the brain imaging data structure for electroencephalography,” *Scientific data*, vol. 6, no. 1, pp. 1–5, 2019.

[51] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, “Eeg signal classification of imagined speech based on riemannian distance of correntropy spectral density,” *Biomedical Signal Processing and Control*, vol. 59, p. 101899, 2020.

[52] B. Alderson-Day, S. Weis, S. McCarthy-Jones, P. Moseley, D. Smailes, and C. Fernyhough, “The brain’s conversation with itself: neural substrates of dialogic inner speech,” *Social cognitive and affective neuroscience*, vol. 11, no. 1, pp. 110–120, 2016.

[53] W. D. Marslen-Wilson and L. K. Tyler, “Morphology, language and the brain: the decompositional substrate for language comprehension,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1481, pp. 823–836, 2007.

[54] 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.

[55] D. Brandeis and D. Lehmann, “Event-related potentials of the brain and cognitive processes: approaches and applications,” *Neuropsychologia*, vol. 24, no. 1, pp. 151–168, 1986.

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

[57] 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.

[58] S. Datta and N. V. Boulgouris, “Recognition of grammatical class of imagined words from eeg signals using convolutional neural network,” *Neurocomputing*, vol. 465, pp. 301–309, 2021.

[59] H.-L. Halme and L. Parkkonen, “Comparing features for classification of meg responses to motor imagery,” *PloS one*, vol. 11, no. 12, p. e0168766, 2016.

[60] B. Chinta and M. Moorthi, “Brain computer interface-eeg based imagined word prediction using convolutional neural network visual stimuli for speech disability,” 2022.

[61] J. Moon, S. Orlandi, and T. Chau, “A comparison and classification of oscillatory characteristics in speech perception and covert speech,” *Brain Research*, vol. 1781, p. 147778, 2022.

[62] V. Grubov, E. Sitnikova, A. Pavlov, A. Koronovskii, and A. Hramov, “Recognizing of stereotypic patterns in epileptic eeg using empirical modes and wavelets,” *Physica A: Statistical Mechanics and its Applications*, vol. 486, pp. 206–217, 2017.

[63] F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy, “Brainstorm: a user-friendly application for meg/eeg analysis,” *Computational intelligence and neuroscience*, vol. 2011, 2011.

[64] V. Bostanov, “Bci competition 2003-data sets ib and iib: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram,” *IEEE Transactions on Biomedical engineering*, vol. 51, no. 6, pp. 1057–1061, 2004.

[65] F. Kamal, K. Campbell, and V. Taler, “Effects of the duration of a resting-state eeg recording in healthy aging and mild cognitive impairment,” *Clinical EEG and Neuroscience*, vol. 53, no. 5, pp. 443–451, 2022.

[66] T. Mussigmann, B. Bardel, and J.-P. Lefaucheur, “Resting-state electroencephalography (eeg) biomarkers of chronic neuropathic pain. a systematic review,” *NeuroImage*, p. 119351, 2022.

[67] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, “Online eeg classification of covert speech for brain–computer interfacing,” *International journal of neural systems*, vol. 27, no. 08, p. 1750033, 2017.

[68] M.-O. Tamm, Y. Muhammad, and N. Muhammad, “Classification of vowels from imagined speech with convolutional neural networks,” *Computers*, vol. 9, no. 2, p. 46, 2020.

[69] D. Dash, P. Ferrari, and J. Wang, “Spatial and spectral fingerprint in the brain: Speaker identification from single trial meg signals.” in *INTERSPEECH*, 2019, pp. 1203–1207.

[70] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, “Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition,” *Scientific Data*, vol. 9, no. 1, pp. 1–17, 2022.

[71] B. Gick, I. Wilson, and D. Derrick, *Articulatory phonetics*. John Wiley & Sons, 2013.

[72] S. Guy, “A path worth exploring - neurolinguistics: An introduction to spoken language processing and its disorders, by john c. l. ingram. 2007. new york: Cambridge university press, 420 pp., $99.00 (hb); $48 (pb),” *Journal of the International Neuropsychological Society*, vol. 15, no. 1, p. 162–163, 2009.

[73] C. Torrence and G. P. Compo, “A practical guide to wavelet analysis,” *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998.

[74] A. Grinsted, J. C. Moore, and S. Jevrejeva, “Application of the cross wavelet transform and wavelet coherence to geophysical time series,” *Nonlinear processes in geophysics*, vol. 11, no. 5/6, pp. 561–566, 2004.

[75] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. A. Torres-García, “Transfer learning in imagined speech eeg-based bcis,” *Biomedical Signal Processing and Control*, vol. 50, pp. 151–157, 2019.

[76] Q. Ji, J. Huang, W. He, and Y. Sun, “Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images,” *Algorithms*, vol. 12, no. 3, p. 51, 2019.

Appendix

A. Brain Imaging Data Structure (BIDS)

The final dataset folder is called ”Decoding Speech Database” and is structured and named using the EEG extension to the Brain Imaging Data Structure (BIDS) [49], [50], see Figure 9. The dataset folder is composed of a subfolder containing the source data (data before file format conversion), 20 subfolders containing the raw data (each subfolder corresponding to the data from a different subject), and a subfolder containing the derived data (derivatives).

Note that the <index> ”value” of sub-<index> in real data file names correspond to the unique identifier of that subject (e.g., 01), the <label> of task-<label> correspond to the specific task and prompt (e.g., covert-aa or overt-geel), the <index> of run-<index> corresponds to the run number, the <index> of trial-<index> to the trial number, and the <label> of channel-<label> to the name of the specific channel (e.g., FC1).

1) sourcedata: The sourcedata subfolder contains the continuous EEG recording per run for all 64 channels for each subject in .Poly5 file format. The sourcedata subfolder also contains the audio recordings in .wav file format. As the audio was solely recorded during the overt speech task, each file contains the audio signal from a single trial.

2) raw data: The subfolders corresponding to the data from different subjects contain an EEG subfolder. This subfolder contains the continuous EEG recording per run for all 64 channels in .fdt and .set file format. The .fdt files contain the raw data ([channels x samples]) and the files with extension .set contain the metadata of the raw data. As the raw audio data is equal to the source audio data and no further processing has been done, no audio subfolder exists in the raw data folders.

3) derivatives: The derivatives folder contains the derivatives of the raw data. The folder of each subject in the derivatives folder contains two subfolders. The first subfolder (’eeg’) contains the epoched EEG data from that specific subject for each task and prompt after pre-processing as described in section II-E, stored in .fdt and .set files. The dimension of the data in the epoched EEG data files is [62 x 2048 x trials]. Each epoch contains 2048 samples of 62 channels which corresponds to 2.0 s of signal acquisition with a sampling rate of 1024 Hz. Epochs containing artifacts are marked for rejection and are shaded.

The second subfolder (”scalogram”) contains the scalogram of each channel and scalogram matrices for each trial per task and prompt computed as described in section II-F5. No scalograms have been computed for subject 9 and 13 as it was decided early on that these subjects were not to be included in the data analysis.

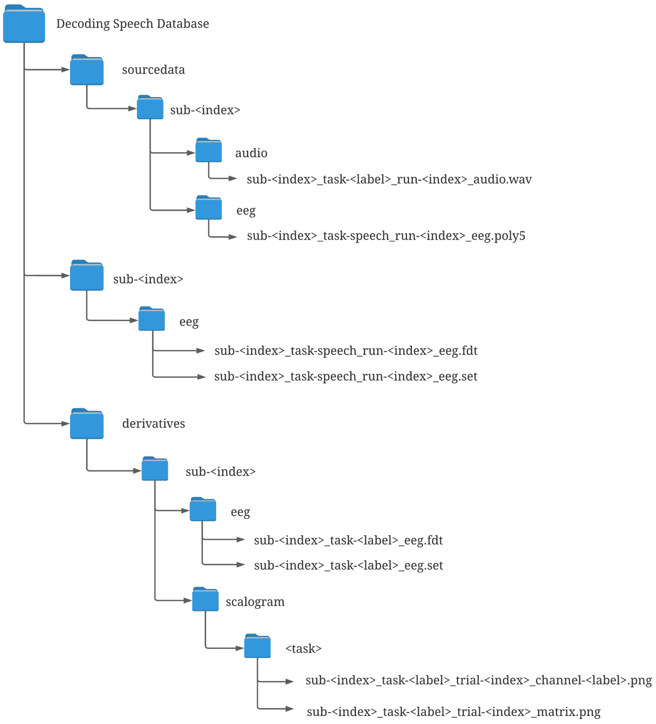
Decoding Speech Database
sourcedata
sub-<index>
audio
sub-<index>_task-<label>_run-<index>_audio.wav
eeg
sub-<index>_task-speech_run-<index>_eeg.poly5
sub-<index>
eeg
sub-<index>_task-speech_run-<index>_eeg.fdt
sub-<index>_task-speech_run-<index>_eeg.set
derivatives
sub-<index>
eeg
sub-<index>_task-<label>_eeg.fdt
sub-<index>_task-<label>_eeg.set
scalogram
<task>
sub-<index>_task-<label>_trial-<index>_channel-<label>.png
sub-<index>_task-<label>_trial-<index>_matrix.png

Fig. 9. Final dataset structure

B. Additional Figures

TABLE VI
PERSONAL INFORMATION OF THE SUBJECTS.

ID	Gender	Age	Handedness
sub-01	Female	24	Right
sub-02	Male	25	Right
sub-03	Female	23	Right
sub-04	Female	25	Right
sub-05	Female	25	Right
sub-06	Female	26	Right
sub-07	Female	23	Right
sub-08	Female	25	Right
sub-09	Female	25	Left
sub-10	Female	25	Right
sub-11	Male	26	Right
sub-12	Male	25	Right
sub-13	Male	25	Left
sub-14	Female	24	Right
sub-15	Male	24	Right
sub-16	Female	23	Right
sub-17	Female	26	Right
sub-18	Female	26	Right
sub-19	Female	25	Right
sub-20	Male	23	Right

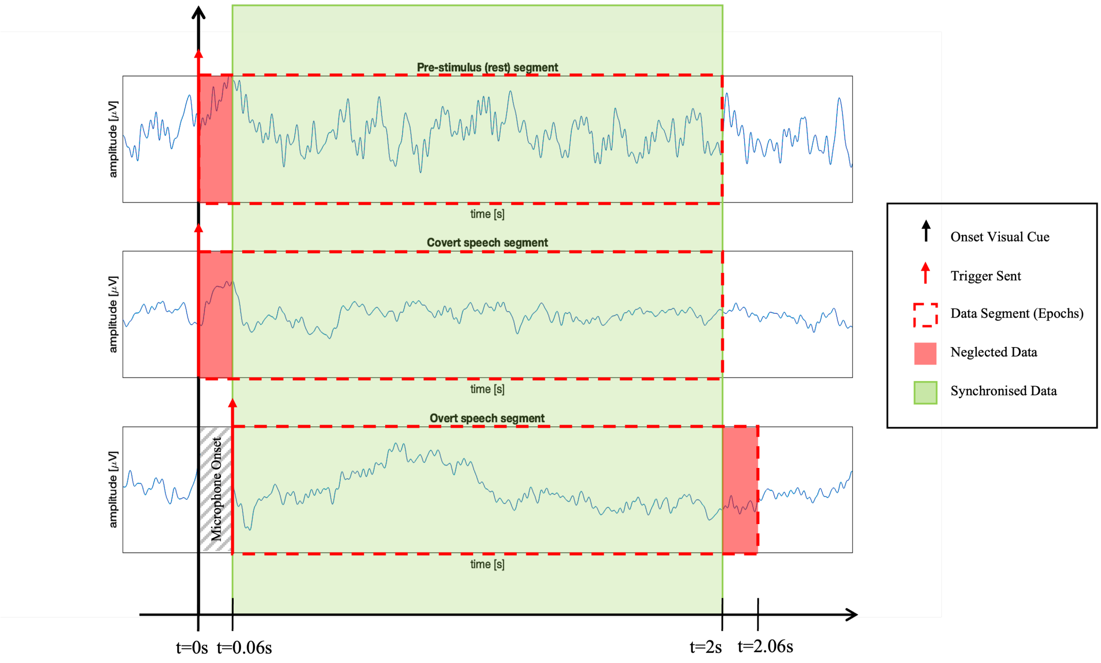

Fig. 10. Segment synchronisation approach for the data analysis to compensate for the delay (0.06 seconds) caused by the onset of the microphone occurring between the onset of the visual cue for the overt speech task and sending the trigger for the overt speech segment. The EEG signals in the figure are for illustrative purposes only and do not represent true segments.

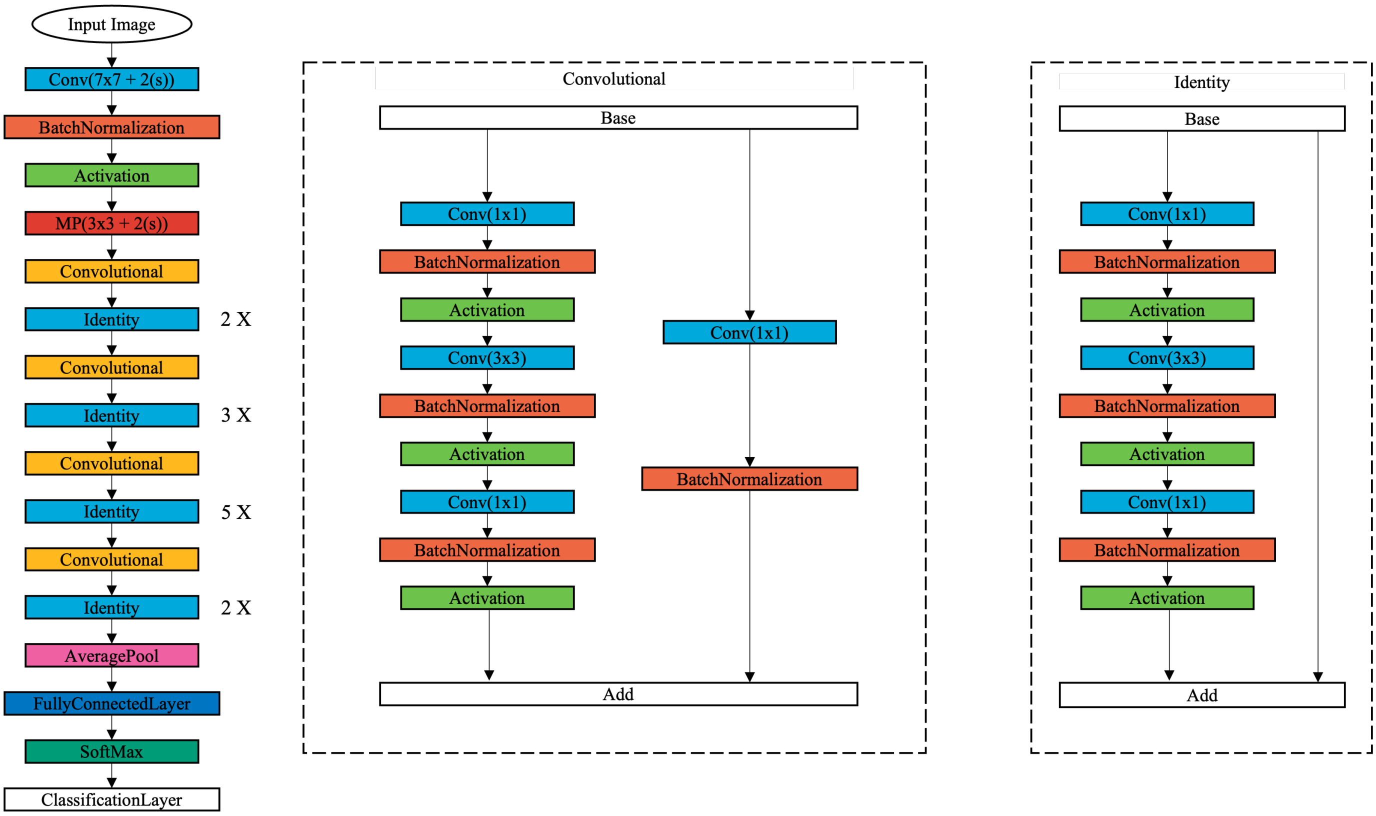

Fig. 11. ResNet-50 architecture (left), convolutional block (middle), and identity block (right). Adapted from [76].

TABLE VII
RAW DATA TRIGGERS AND DESCRIPTIONS.

Trigger ID	**Description**	
	Task	**Prompt**
1	Pre-stimulus	n.a.
16	Perception	aa
17	Perception	ee
18	Perception	ie
19	Perception	oo
20	Perception	oe
21	Perception	taal
22	Perception	laat
23	Perception	leeg
24	Perception	geel
25	Perception	niet
26	Perception	tien
27	Perception	toon
28	Perception	noot
29	Perception	soep
30	Perception	poes
32	Covert speech	aa
33	Covert speech	ee
34	Covert speech	ie
35	Covert speech	oo
36	Covert speech	oe
37	Covert speech	taal
38	Covert speech	laat
39	Covert speech	leeg
40	Covert speech	geel
41	Covert speech	niet
42	Covert speech	tien
43	Covert speech	toon
44	Covert speech	noot
45	Covert speech	soep
46	Covert speech	poes
48	Overt speech	aa
49	Overt speech	ee
50	Overt speech	ie
51	Overt speech	oo
52	Overt speech	oe
53	Overt speech	taal
54	Overt speech	laat
55	Overt speech	leeg
56	Overt speech	geel
57	Overt speech	niet
58	Overt speech	tien
59	Overt speech	toon
60	Overt speech	noot
61	Overt speech	soep
62	Overt speech	poes
63	Start/stop	n.a.

TABLE VIII

Number of EEG trials remaining per subject for covert and overt pronunciation of five vowels after removal of trials containing eye blinks. A total of 20 trials was recorded per subject for both covert and overt pronunciation of each vowel.

ID	**‘aa’**		**‘ee’**		**‘ie’**		**‘oo’**		**‘oe’**	
	C	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**
sub-01	18	6	18	10	19	10	17	15	19	13
sub-02	20	0	18	0	20	1	20	8	18	8
sub-03	19	16	20	14	19	16	20	14	18	19
sub-04	19	13	20	15	20	15	18	18	20	16
sub-05	20	6	20	15	19	14	20	16	20	13
sub-06	19	12	18	15	19	15	19	13	20	19
sub-07	16	4	17	12	15	8	13	6	16	8
sub-08	20	18	19	18	20	19	19	15	20	20
sub-09	20	19	20	20	19	17	20	19	19	18
sub-10	20	14	20	15	20	16	20	14	20	16
sub-11	20	14	19	12	20	15	20	19	19	19
sub-12	19	19	20	18	20	19	20	19	19	19
sub-13	13	7	11	5	12	8	11	9	10	6
sub-14	19	6	18	8	18	11	20	15	18	13
sub-15	20	17	18	16	17	15	19	14	20	16
sub-16	18	11	18	13	17	15	20	16	19	17
sub-17	16	12	14	11	17	13	17	13	20	14
sub-18	18	20	20	20	19	15	19	20	19	18
sub-19	20	17	20	18	20	20	20	17	20	17
sub-20	17	13	14	16	17	15	18	16	19	16
Total	**371**	**244**	**362**	**271**	**367**	**277**	**370**	**296**	**373**	**305**

TABLE IX

Number of EEG trials remaining per subject for covert and overt pronunciation of ten words after removal of trials containing eye blinks. A total of 20 trials was recorded per subject for both covert and overt pronunciation of each word.

ID	**‘taal’**		**‘laat’**		**‘leeg’**		**‘geel’**		**‘niet’**		**‘tien’**		**‘toon’**		**‘noot’**		**‘soep’**		**‘poes’**	
	C	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**	**C**	**O**
sub-01	19	11	18	12	19	13	18	13	18	13	16	8	18	11	18	14	18	14	18	14
sub-02	18	2	18	0	16	0	18	3	15	2	18	5	18	11	19	6	18	4	20	2
sub-03	18	18	20	19	18	18	16	17	20	17	19	20	19	19	18	18	20	15	20	15
sub-04	19	16	20	13	19	17	20	18	19	15	19	17	19	19	19	17	19	18	20	17
sub-05	19	14	20	15	19	15	19	15	19	16	19	12	18	15	19	20	20	17	19	15
sub-06	17	12	18	11	20	13	17	15	19	18	19	16	18	15	18	14	18	16	16	20
sub-07	16	7	17	7	19	8	18	6	18	7	18	9	17	9	17	9	18	13	14	9
sub-08	20	16	19	19	19	19	19	19	20	17	19	15	20	19	20	20	20	19	20	18
sub-09	18	20	20	18	18	18	19	19	20	17	20	19	20	18	20	19	20	19	19	20
sub-10	20	16	20	11	19	13	17	13	19	17	20	18	19	19	19	16	19	17	19	15
sub-11	20	16	19	12	19	17	20	17	20	18	20	18	20	14	19	17	20	16	19	19
sub-12	20	19	20	16	19	19	20	19	20	19	20	18	20	16	20	18	20	18	20	19
sub-13	8	6	11	7	13	6	10	7	9	9	14	9	11	8	15	10	13	6	11	2
sub-14	18	8	18	9	20	10	18	8	19	11	19	5	19	5	19	12	19	11	19	11
sub-15	20	19	18	6	19	15	20	17	20	19	18	18	19	18	20	16	20	16	19	15
sub-16	16	14	18	15	17	16	18	17	19	16	18	17	17	16	17	18	18	17	17	15
sub-17	16	11	14	7	17	9	15	13	15	10	18	11	19	10	16	9	17	14	18	12
sub-18	20	18	20	20	20	20	20	17	20	18	20	19	20	18	20	18	20	19	20	19
sub-19	20	19	20	17	20	18	19	18	20	18	20	17	20	18	20	19	20	19	19	18
sub-20	17	12	14	14	17	14	19	14	20	18	18	16	17	14	17	16	16	14	17	12
Total	**359**	**274**	**362**	**248**	**367**	**278**	**360**	**285**	**369**	**295**	**372**	**287**	**368**	**292**	**370**	**306**	**373**	**302**	**364**	**287**

TABLE X

Balanced Latin square used for the experiments. Each subject performed twenty runs and for each subject another part of the Balanced Latin square was used. All runs consist of 15 consecutive trials (T#).

Run	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
1	aa	ee	ie	oo	oe	taal	laat	leeg	geel	niet	tien	toon	noot	soep	poes
2	niet	geel	tien	leeg	toon	laat	noot	taal	soep	oe	poes	oo	aa	ie	ee
3	ie	oo	ee	oe	aa	taal	poes	laat	soep	leeg	noot	geel	toon	niet	tien
4	toon	tien	noot	niet	soep	geel	poes	leeg	aa	laat	ee	taal	ie	oe	oo
5	oe	taal	oo	laat	ie	leeg	ee	geel	aa	niet	poes	tien	soep	toon	noot
6	soep	noot	poes	toon	aa	tien	ee	niet	ie	geel	oo	leeg	oe	laat	taal
7	laat	leeg	taal	geel	oe	niet	oo	tien	ie	toon	ee	noot	aa	soep	poes
8	aa	poes	ee	soep	ie	noot	oo	toon	oe	tien	taal	niet	laat	geel	leeg
9	geel	niet	leeg	tien	laat	toon	taal	noot	oe	soep	oo	poes	ie	aa	ee
10	ie	ee	oo	aa	oe	poes	taal	soep	laat	noot	leeg	toon	geel	tien	niet
11	tien	toon	niet	noot	geel	soep	leeg	poes	laat	aa	taal	ee	oe	ie	oo
12	oe	oo	taal	ie	laat	ee	leeg	aa	geel	poes	niet	soep	tien	noot	toon
13	noot	soep	toon	poes	tien	aa	niet	ee	geel	ie	leeg	oo	laat	oe	taal
14	laat	taal	leeg	oe	geel	oo	niet	ie	tien	ee	toon	aa	noot	poes	soep
15	poes	aa	soep	ee	noot	ie	toon	oo	tien	oe	niet	taal	geel	laat	leeg
16	geel	leeg	niet	laat	tien	taal	toon	oe	noot	oo	soep	ie	poes	ee	aa
17	ee	ie	aa	oo	poes	oe	soep	taal	noot	laat	toon	leeg	tien	geel	niet
18	tien	niet	toon	geel	noot	leeg	soep	laat	poes	taal	aa	oe	ee	oo	ie
19	oo	oe	ie	taal	ee	laat	aa	leeg	poes	geel	soep	niet	noot	tien	toon
20	noot	toon	soep	tien	poes	niet	aa	geel	ee	leeg	ie	laat	oo	taal	oe
21	taal	laat	oe	leeg	oo	geel	ie	niet	ee	tien	aa	toon	poes	noot	soep
22	poes	soep	aa	noot	ee	toon	ie	tien	oo	niet	oe	geel	taal	leeg	laat
23	leeg	geel	laat	niet	taal	tien	oe	toon	oo	noot	ie	soep	ee	poes	aa
24	ee	aa	ie	poes	oo	soep	oe	noot	taal	toon	laat	tien	leeg	niet	geel
25	niet	tien	geel	toon	leeg	noot	laat	soep	taal	poes	oe	aa	oo	ee	ie
26	oo	ie	oe	ee	taal	aa	laat	poes	leeg	soep	geel	noot	niet	toon	tien
27	toon	noot	tien	soep	niet	poes	geel	aa	leeg	ee	laat	ie	taal	oo	oe
28	taal	oe	laat	oo	leeg	ie	geel	ee	niet	aa	tien	poes	toon	soep	noot
29	soep	poes	noot	aa	toon	ee	tien	ie	niet	oo	geel	oe	leeg	taal	laat
30	leeg	laat	geel	taal	niet	oe	tien	oo	toon	ie	noot	ee	soep	aa	poes

C. HREC Forms

This study was approved by the Human Research Ethics Committee (HREC) of the Delft University of Technology on May 31, 2022 (#2264).

Participation Information Letter

Concerning a study on decoding covert speech using electroencephalography (EEG).
Version date: 24/05/2022

Dear potential participant,

You have been asked to participate in a study in which EEG and audio is recorded during a speaking task. It is your decision whether you wish to participate. Before you decide, it is important to know more about the study. This information sheet provides detailed information about the study. Read this information letter thoroughly and discuss it with your partner, friends, or family. Please get in touch with the researchers mentioned below if you have any questions.

Study background

People who have lost the ability to speak and who cannot or can no longer use sign language due to severe neuromuscular disease (e.g., severely paralyzed people or patients of locked-in syndrome) are strongly impaired in the communication with the external world. However, as their cognitive abilities are preserved, the neural signals of these patients during covert speech might be used to offer a way of communicating. Covert speech is imagining speaking without moving any of the articulators or making any sound. A non-invasive way of measuring neural activity is electroencephalography (EEG).

However, measuring neural activity leads to a large amount of data. This makes it difficult to distinguish the specific neural signals related to the covert speech from the background signals. Machine learning algorithms allow us to address tasks that are too difficult to solve using programs designed by humans. As these algorithms can learn from data, new datasets containing different types of covert speech are required to make use of recent advances in the field of machine learning.

Study goal

The goal of this study is (1) to develop a database with healthy subjects combining covert and produced speech for Dutch vowels and words recorded with EEG and audio, and (2) to train a machine learning algorithm using the developed database to decode covert and produced speech.

What does participation involve?

During the study you will be seated before a computer monitor and a microphone. You will be instructed to look at the screen and move as little as possible. Individual prompts will appear on the screen one-at-a-time. The study will consist of multiple trials. Each trial consists of four successive states:

1) A rest state, in which you can relax and clear you mind.
2) Stimulus state, where the specific prompt appears on the screen.
3) Covert speech state, in which you imagine speaking the prompts without moving.
4) Produced speech state, in which you speak the prompt aloud.

Brain activity will be measured by electroencephalography (EEG), which is a non-invasive method to measure brain activity. To measure EEG, you will be asked to wear a cap throughout the experiment in which measurement electrodes are integrated. The risks associated with the study are small. Recording EEG is routine research and clinical procedures which are performed daily without known harmful effects or significant risks. To have a good conductance between skin and electrodes, each electrode will have some conducting gel. At the end of the experiment, we will remove the gel as much as we can, but some remaining gel will have to be washed out, in a shower at the faculty or at home.

The study takes place at the Delft University of Technology. The total experiment takes about 2 hours including set-up and removal of measurement equipment.

Participation preparation
We ask participants to withhold from taking caffeinated drinks, like coffee, two to three hours before the experiment since this might influence resting brain activity. Additionally, we would like to ask you to wash your hair in the morning or the day before and not use any hair products after washing until the experiment so there are no remnants of hair products negatively impacting conductivity.

Participation is voluntary!
Your participation in the study is voluntary. If you agree on participating in the study, you have the right to withdraw any time, even during the study. There is no need to have a legitimate reason to do so. If you agree to participate in the study, you will be provided with an informed consent form for you to sign.

Confidentiality
We will treat your personal details and data confidentially. People not authorised to access your details will not be able to do so. The recorded date will be pseudonymized by storing personal details and recorded data in different places using a key to link the two. Both are stored in a secure storage environment at the TU Delft.

The results will be published in a Master thesis report. To ensure that the data cannot be traced back to you, the audio recording and EEG data will be cut into small fragments corresponding to single words or vowels. The de-identified and pre-processed recorded EEG data and audio recordings will be archived in a data repository so it can be used for future research.

If you have any complaints regarding confidentiality of your data, please contact the TU Delft Data Protection Officer (Erik van Leeuwen) via privacy-tud@tudelft.nl.

Summary
Participating in this study is voluntary. Summarized, when you decide to participate:

- You are willing to participate in research during which EEG and audio will be recorded while you perform a simple speech task.
- You adhere to the asked preparations on the day before and the day of the experiment.
- You agree with the use of your data for purposes of the study and future research.
- You understand we cannot provide individual study results.

For more information, feel free to contact one of the researchers mentioned below.
Thank you in advance for considering participation in our study.

Bo Dekker (first point of contact)
MSc. Student Biomedical Engineering
[redacted]

Dr. Ir. Alfred C. Schouten
Associate Professor
[redacted]

Neuromechanics & Motor Control Laboratory

Informed Consent Form

Concerning a study on decoding covert speech using electroencephalography (EEG).

Participant number: ____________

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
GENERAL AGREEMENT		
I have read and understood the study information dated 24/05/2022, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	□	□
I understand that taking part in the study involves data recording with an EEG cap and audio recording while performing a speech task. I can request for my data to be removed up to one week after the experiment has taken place.	□	□
POTENTIAL RISKS OF PARTICIPATING		
I understand that taking part in the study also involves collecting specific personally identifiable information (PII; name and email) and associated personally identifiable research data (PIRD; age, gender, and hand dominance) with the potential risk of my identity being revealed.	□	□
I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach: • All data is stored at a secure storage environment at the TU Delft. • Directly identifiable PII is stored at a different place than the recorded data and will be destroyed after the study. • The recorded data is saved under the participant number. • The audio and EEG data will be cut into small fragments corresponding to single words or vowels.	□	□
I understand that personal information collected about me that can identify me, such as my name and email, will not be shared beyond the study team.	□	□
(LONGTERM) DATA STORAGE, ACCESS, AND REUSE		
I give permission for the audio and pre-processed EEG data that I provide to be archived in a data repository so it can be used for future research. The audio and EEG data will be pseudonymized and cut into small fragments corresponding to single words and vowels. No raw EEG data will be shared.	□	□
I understand that there is a possibility that my voice is recognized from the audio fragments (i.e., single words or vowels).	□	□

Signatures

____________________	____________________	________
Name of participant	Signature	Date

____
Email of participant

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Bo Dekker	____________________	________
Researcher name	Signature	Date

Study contact details for further information:
Bo Dekker
[redacted]