Using articulated speech EEG signals for imagined speech decoding

by

Chris Bras

Student Name	Student Number
Chris Bras	4394763

Instructor:	O. E. Scharenborg
Teaching Assistant:	T. B. Patel
Project Duration:	July, 2023 - April, 2024
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science, Delft

TUDelft

Summary

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

Contents

1

Introduction

Neurodegenerative disorders such as Amyotrophic Lateral Sclerosis (ALS) or conditions like locked-in syndrome frequently result in profound muscular impairment, rendering patients incapable of voluntary muscle movement and consequently unable to articulate speech [47]. This profound physical debilitation presents significant obstacles for individuals affected by these conditions when attempting to engage in effective communication with their external environment.

Brain-computer interfaces (BCIs) have emerged as a potential avenue to address this issue [47]. By analysing brain activity, BCI systems could facilitate communication solely based on the patient's thoughts. A promising approach in this context involves using imagined speech (*covert speech*), wherein an individual imagines to produce speech without any muscle movement nor audible or articulated (spoken) speech. By decoding and interpreting these neural signals, BCIs hold promise for enabling communication with patients affected by ALS and multiple sclerosis (MS), bypassing the physical limitations imposed by their conditions. Electroencephalography (EEG) is an non-invasive, low-cost technique widely used to capture the electrical signals generated by neural activity, which can be analyzed to understand both imagined and articulated speech [8, 5].

Research on imagined speech from EEG has focused on classifying small sets of stimuli, e.g., vowels (English [22], Dutch [28, 24], Japanese [54], Spanish [46]) and isolated words ("yes" and "no" [35], nine Russian words [53]). For the task of classifying EEG data of imagined speech, many different machine and deep learning techniques have been used, including, support vector machine [12], linear discriminant analysis [30], random forest [16], vanilla deep neural networks (DNNs) [42], and convolutional neural network (CNN) [14, 13]). However, deep learning models require large amounts of data to properly generalize for a given problem without having issues with overfitting [48] which often is not available for this type of EEG data. Moreover, different discriminative features extracted from the EEG signals have been used (e.g., wavelet domain features [25, 43] and common spatial patterns (CSP) [33, 55]). Nevertheless, no combination of classifier and features has proven to consistently achieve high decoding performances [13]; although Residual Network (ResNet) algorithms [53, 42] have been found to outperform other CNN algorithms on imagined speech classification tasks in both robustness and practicability.

Classifying imagined speech using EEG is however a challenging task, with classification results close to chance level [11, 16]. Being a non-invasive solution, the recording made using an EEG is not optimal, inducing noise from, for example, blinks or muscle movement in the data [18]. Moreover, there is a lack of imagined speech EEG data. Although different databases have been released[22, 24, 16, 38], they differ in the number of speakers, language, the absence or presence of articulated speech, and differences in the recording set-up, e.g., the number of channels used to record the EEG signals (e.g., 6 for Coretto et al.[16] and 62 for DAIS [24]). DaSalla et al. [22] includes 3 subjects that each performed 3 tasks for imagined speech, creating a dataset of /a/ and /u/ vowels in the English language with a no action state task as a control. Dekker et al. introduces the Delft Articulated and Imagined Speech (DAIS) [24] that will be used in this thesis. The DAIS dataset includes both imagined and articulated speech

EEG data and audio data for the articulated speech for a total of 20 subjects. 15 prompts are included for each subject for both imagined and articulated speech, consisting of 5 Dutch vowels and 10 Dutch words. Coretto et al. [16] contains 5 words containing the vowels /a/ /e/ /i/ /o/ and /u/ in the Spanish language for 15 subjects of imagined speech EEG data. Each subject in the dataset repeated each word 50 times in a random order. Nguyen et al. [38] introduces a dataset of 15 subjects of imagined speech EEG data.

On top of the difference between available datasets, EEG signals vary a lot, especially between different subjects [26]. This makes it difficult for models to generalize and validate their performance. A further difficulty is the lack of an accurate ground truth: it is not easily verifiable if the subject performed the imagined task correctly, as is the case with articulated speech. A roundabout way of testing whether participants complied with the task of imagining speech is to visually investigate whether structural differences exist between the event-related potentials (ERPs) of the EEG signals for rest and imagined speech or run a classification task predicting whether an EEG signal came from the rest state or imagined speech [24]. Besides these difficulties, the prompts used to gather the data from the subjects also influences the results greatly, making different datasets hard to compare. Vowels for example can be found easier to classify than complete words [15].

Research suggests that imagined speech production can be seen as interrupted articulated speech production without the actual muscle movement required for producing sound [36]. In other words, the onset of imagined and articulated speech is comparable, meaning that the part before the sound production of articulated speech is similar to imagined speech.

Figure 1.1 shows the averaged event related potentials (ERP) for participant #12 from the DAIS dataset for rest (top panel), imagined (middle panel) and articulated (bottom panel) speech [24]. Comparing the three ERP signals shows clear differences between the EEG data of the rest segment, imagined speech, and articulated speech. For rest (top panel) only background EEG activity was found. For both imagined and articulated speech activity is observed around 0.25 - 0.3 seconds, which is followed by a broad peak/trough (depending on the channel) starting at 500 ms for articulated speech. This corresponds to start of the articulated speech and is therefore associated with the movement of the articulators.

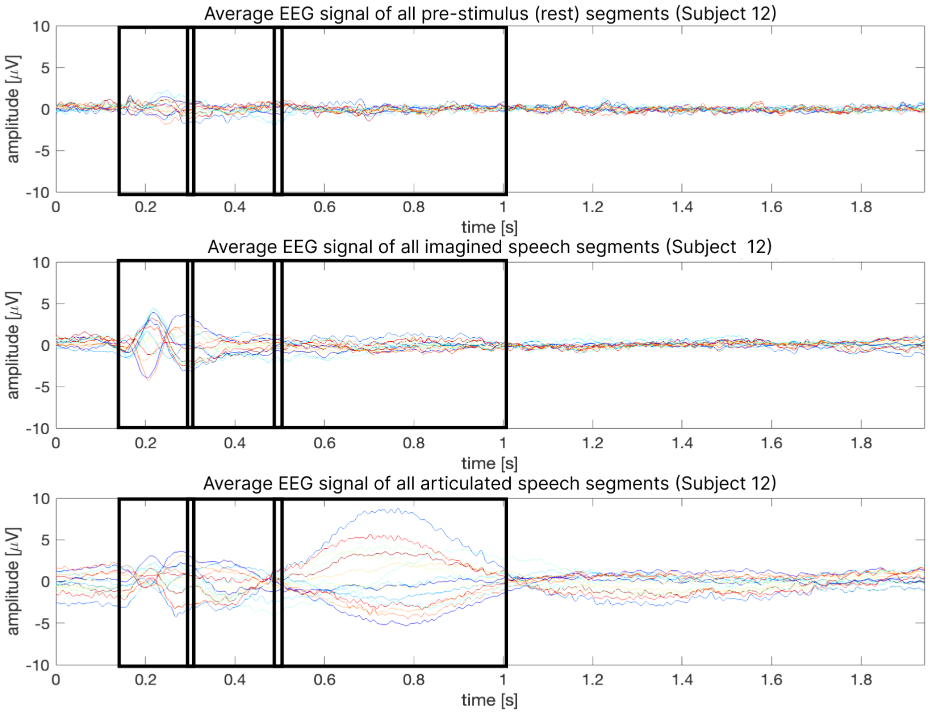

Figure 1.1: Averaged event related potentials (ERP) for Participant 12 from the DAIS dataset for rest (top panel), imagined (middle panel) and articulated (bottom panel) speech[24].

In this work, it is investigated whether EEG data captured from articulated speech can be used to improve classification accuracy and generalization of EEG data captured from imagined speech in two ways: It is investigated whether articulated speech EEG signals can be used to predict the end point of the imagined speech in the EEG signal and whether using the articulated speech EEG as additional training data improves speaker-independent imagined vowel classification.

The overall aim for this thesis is to explore the viability of using articulated speech EEG data for improving classification results on imagined speech EEG vowel data in the Dutch language. For this, multiple questions need to be answered first.

- What model architecture works best for this use case?
- When removing the sound production part of articulated speech, is it similar enough to imagined speech to be used together in a vowel EEG data classifier?
- Is it possible to mark the start of speech within an EEG signal, using only the EEG signal as a guide?

The approach to answer these questions consists of 2 main parts. First a model needs to be created that can mark the start of speech, or more precisely, the end of ”pre-speech”. Pre-speech is considered to be the part of speech that occurs before audible speech is produced. In this work a model is presented that marks the start of speech within an EEG signal. The model is trained on articulated speech EEG data, using the accompanying audio data as a guide during training.

The second part then consists of combining articulated and imagined speech EEG data. This is done in multiple ways, first the data is naively combined without any extra pre-processing steps. This is then compared to combining the data after isolating the pre-speech part in both the imagined and articulated EEG data.

Before the main experiments can be run, a shorter pilot experiment is run in which multiple different model types are compared. The models compared are a support vector machine (SVM) [17], random forest (RF) [29], K-nearest Neighbour (KNN), long short-term memory model (LSTM) and a convolutional neural network (CNN). The best model from this pilot experiment is then used for the main experiments.

In all main experiments, performance is compared on different numbers of channels, to investigate whether there is an influence of the number and location of the electrodes from which the EEG signals are collected on imagined and articulated speech EEG classification. Muscle movement could influence the classification of articulated speech, but is not present in imagined speech. Taking smaller subsets of EEG electrodes, focused more on the regions of the brain that are active during imagined speech production, like Wernickes’s region [9], can limit the influence of muscle movement on classification performance.

2
Background

This chapter will focus on information and techniques used throughout the rest of this thesis. Terminology and standard methods and frameworks are explained here.

2.1. Vowel Articulation

Vowels are articulated by opening up the vocal tract. Which vowel is produced depends on 3 main parameters: the rounding of the lips, height of the tongue and position of the tongue from front to back in the mouth. Besides these 3, vowels can further be separated based on length and whether a single mouth position is used to articulate them (monophtongs) or 2 mouth positions are used (diphthongs). To schematically display this, the vowel quadrant is introduced. Figure 2.1 shows this quadrant for the dutch language. In here, the vowel extremes are put on the corners of the quadrant. The Horizontal axis represents the height of the tongue, the vertical axis represents the amount that the mouth is opened for that vowel.

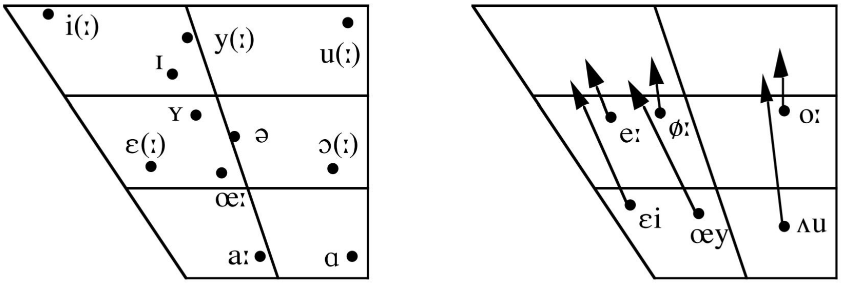

Figure 2.1: Vowel quadrant for the Dutch language, left shows the steady-state and right shows the diphtongs [27].

2.2. EEG Data

Brain activity can be recorded using Electroencephalography (EEG) in a non-invasive way [20]. This is done by placing electrodes on a subject's head. The placement of these electrodes is commonly done according to the international standardized 10-20 system. This system defines two axes, from the nasion (bone above the nose) to the inion (bone lump on the back of the head) and the lateral (left to right) axis. The positions are coded by an alphanumerical combination. Even numbers indicate placement on the right brain half, odd numbers indicate placement on the left brain half. Letters are used to denote the placement from front to back on the brain: Fp (pre-frontal), F (frontal), C (central), P (parietal) and O (occipital). Figure 2.2 shows the schematic layout of this system.

Areas of interest for speech are Broca's area and Wernicke's area [9]. Broca's area lies around the frontal lobe . This area is associated with the motor functions of speech production and articulation,

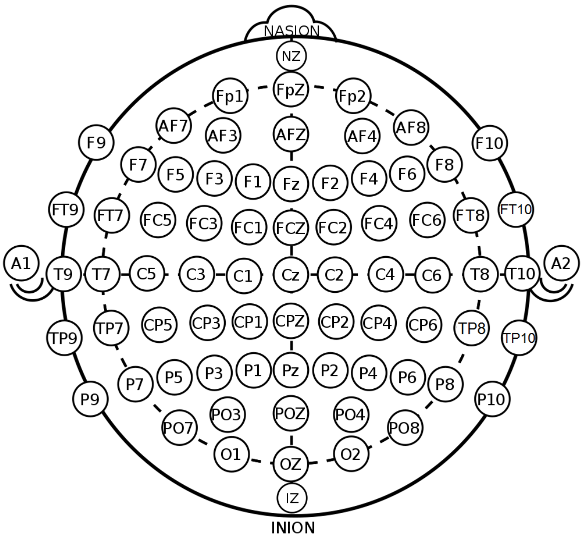

Figure 2.2: Electrode layout of international standardized 10-20 64-channel EEG capture [50]

as well as the correct use of words for both spoken and written language. Wernicke's area lies in the posterior superior temporal lobe and is connected to Broca's area. This area is associated with language processing. Figure 2.3 shows these locations highlighted in the brain.

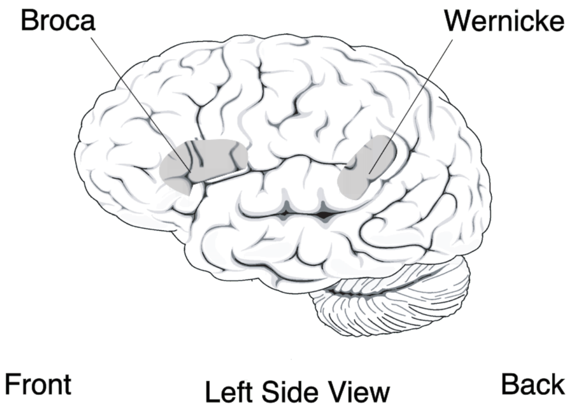

Figure 2.3: Location of Broca's area and Wernicke's area within the brain. [1]

2.2.1. Brainwaves

Brainwaves are oscillating voltages within the brain and are widely classed as 5 different types [6]: Delta, Theta, Alpha, Beta and Gamma. An overview is given of the frequency bands and their relation to speech production and perception [34]:

- **Theta band**: between 4 and 8 Hz. These waves become active when the phonemic restoration effect occurs within in the brain [32]. This is when the brain tries to fill in the gaps if a phoneme is not heard properly due to for example noise. Theta waves can also help with identifying consonants [40].
- **Alpha band**: between 8 and 12 Hz. Associated with speech perception. Differs between imagined and articulated speech. Alpha waves are stronger in articulated speech than they are in imagined speech [31].
- **Beta band**: between 12 and 35 Hz. Associated with muscle movement for the articulated part of speech [7].
- **Gamma band**: from 35 Hz and up. Associated with both articulated and imagined speech. Different

locations of the brain are active in this region for articulated and imagined speech however [44]. Figure 2.4 shows an example of these different brainwaves.

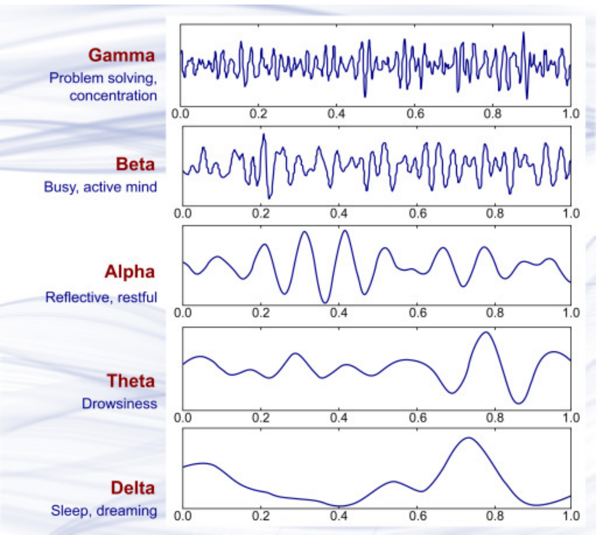

Figure 2.4: Example of the different types of brainwave [6]

2.2.2. Preprocessing and Data Cleaning

After EEG signals have been recorded, they need to be pre processed to remove any signals and noise that are not wanted within the data. This consists of three main parts: Signal filtering, removal of artifacts and channel selection.

Artifact Removal and Filtering

Artifacts are part of the signal that are not directly related to the source of interest, in this case brain activity. Examples of artifacts are electromagnetic interference, power line hum, oscillations from the instrumentation used. These types of electrical artifacts are effectively removed using the correct combination of high pass, low pass and notch/band pass filters as stated earlier.

Another type of artifact often come across in the EEG data is high voltage spikes due to for example muscle movement while blinking, swallowing or eye movement of the subject. These are harder to filter as they are transient, not present in every data point and often of significantly higher magnitude that the signals of interest. In this case it can be considered to leave out the data points containing these artifacts altogether.

Channel Reconstruction

The last EEG data pre-processing step to be discussed in this section is channel selection and bad channel detection. When working with EEG data it is often best to use the most amount of channels available unless you are targeting a specific part of the brain, for example when focusing on parts of the brain responsible for speech production. However sometimes during recording errors can occur in certain channels. It is important to detect and handle these errors before doing any further analysis. Detecting bad channels can often be done by simply looking at the raw waveform. Figure 2.5 shows an example of EEG data containing a bad channel. Channel 53 in this case contains no discernible signal, only noise.

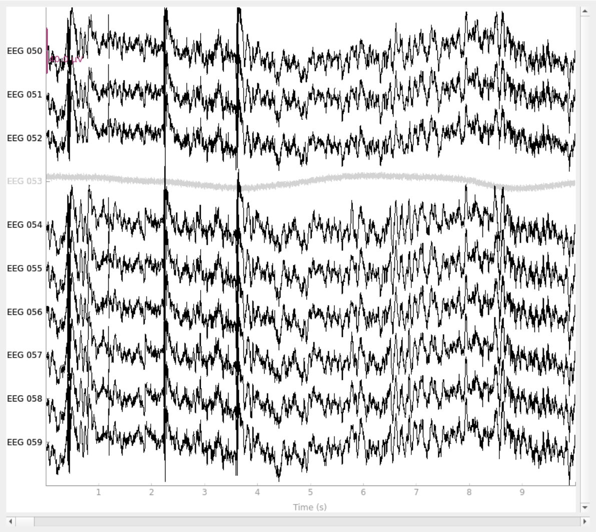

Figure 2.5: Example of EEG data containing a bad channel, in this case channel 53.[2]

In some cases it is sufficient to drop bad channels from the dataset. However, this is not always wanted. For example in cross-subject analyses you want to have the same data dimensions for each subject. In this case a single bad channel for one subject would mean you have to drop that channel for every other subject as well, unnecessarily wasting clean data. In this case interpolating the bad channel could be a better option. Figure 2.6 shows EEG data with a bad channel in red, before interpolation on the left side, where the signal is almost flat except for some noise, and after interpolation on the right side where the signal is more similar to the signal from the other EEG channels.

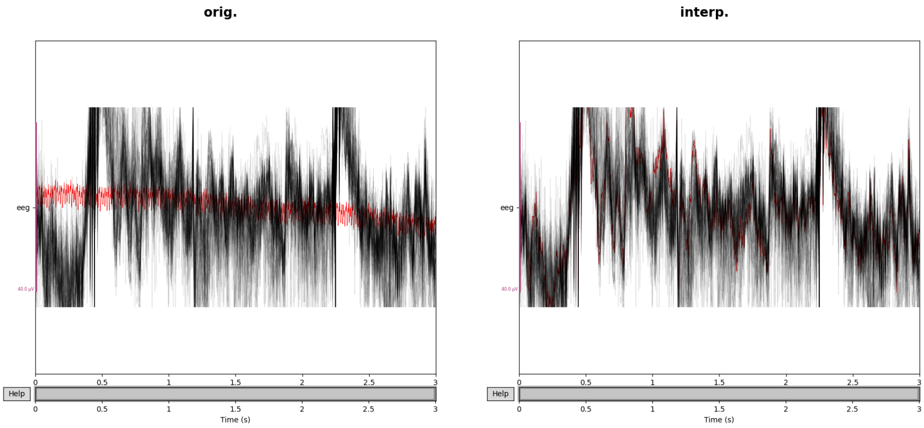

Figure 2.6: Example of EEG data containing a bad channel (red) before and after interpolation of the bad channel.[2]

2.3. Machine Learning Models

In this thesis several models are tried and used, both classical machine learning (ML) and deep learning (DL) models. This section will go over how these models function.

2.3.1. Support Vector Machine

A Support Vector Machine (SVM) works in concept by assuming that the input vector can be non-linearly mapped to a high dimensional feature space, in which then a linear decision boundary can be created [17]. This allows data that normally cannot be separated in a linear way to still be categorized. Figure 2.7 shows how an SVM transforms the data such that it can be categorized using a linear decision boundary. The data shown is only separable using a non linear curve, but after the SVM performs its transform, the problem space is transformed in such a way that now the data is separable in a linear way.

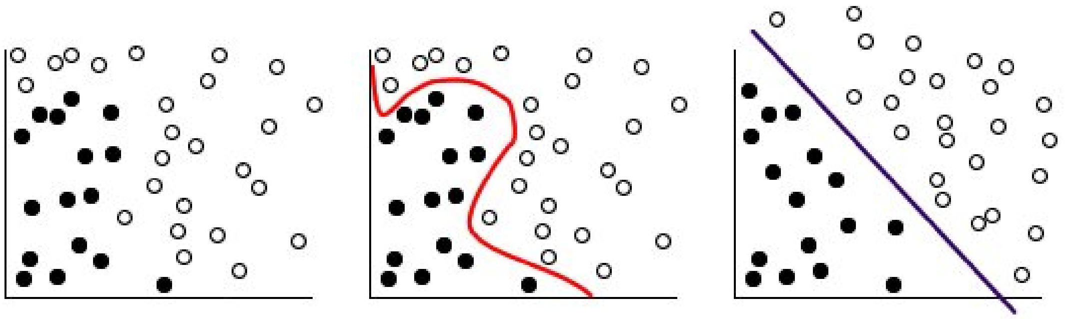

Figure 2.7: Example of how an SVM creates a linear decision boundary in non-linearly separable data [3]. Left image shows the two classes of data, which in the middle image are categorisable by a curve as a decision boundary. The right image shows the data after the SVM applied its transform, now the data is linearly separable.

SVM models can use different types of kernel functions: linear, nonlinear, polynomial, radial basis function (RBF) and sigmoid. This kernel is what performs the transform. Most commonly RBF is used as a kernel function [51].

2.3.2. Random Forest

The Random Forest (RF) algorithm is a type of decision tree algorithm. An RF model creates multiple decision trees within arbitrary subspaces of the input data [29]. This divides the feature space of the data over multiple trees, making each tree focus on its own small subset of features of the data. The idea is that these trees in different subspaces perform their classification in different ways from each other, based on different features, so that when they are combined they will complement each other. This way the complete forest can have higher accuracy and generalization than any single tree in the forest could achieve. Classification using random forest models is done by letting each smaller decision tree make its own prediction, which are then used to determine a majority vote, combining the output of each tree in one single output.

2.3.3. k-Nearest Neighbour

K-Nearest Neighbour is the simplest machine learning algorithm to be discussed. It works by the assumption that samples from the same category can be found close to each other in the input space. Classification is done by a majority vote. This works by comparing a to be classified sample (from for example the test set) to a number of already known samples (from for example the training set). The model will look at the label of the k number of closest neighbours to the input sample. The label that is most prevalent, in other words the label for which the majority of samples ”voted”, is then also assumed to be the category of this new sample. Figure 2.8 shows this process. The number of neighbours to consider, K, is the only hyper parameter to choose for this model. A smaller k will make a model more likely to over fit and produce unstable decision boundaries, while a larger k can make the model lose too much detail, but creates smoother decision boundaries.

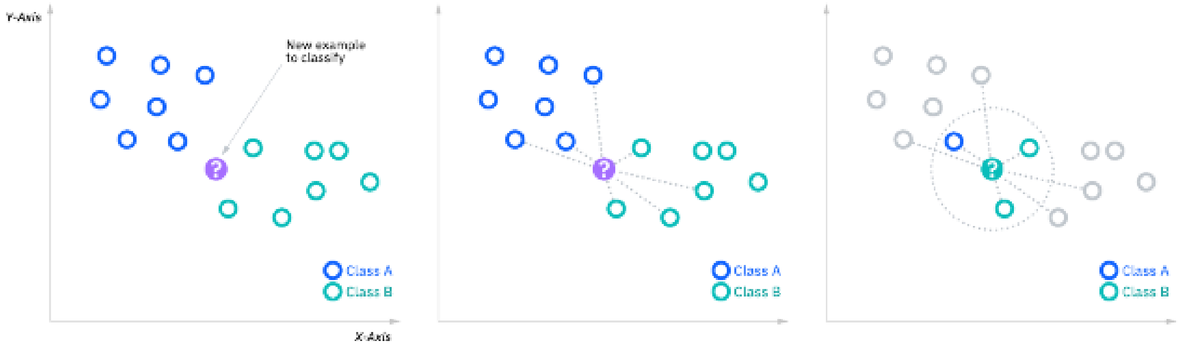

Figure 2.8: Example of a sample being classified by a KNN model. A k-number of known points closest to the new sample is considered. The label that is most common between these will be the label of the new sample [4].

To determine how close 2 different samples are to each other, a distance metric is needed. This metric gets smaller the closer 2 samples are alike. For example for samples lying in n dimensional space, the Euclidean distance between them can be used given by the formula

$$d_{Euclidean}(i, j) = \sqrt{\sum_{t=1}^{n} (i_t - j_t)^2}$$

Where i and j are the samples to be compared. Other commonly used distance metrics are the Minkowski distance, of which Euclidean distance is a special case where $q = 2$, given by

$$d_{Minkowski}(i, j) = \sqrt[q]{\sum_{t=1}^{n} (i_t - j_t)^q}$$

and the Manhattan distance, which is a special case of Minkowski distance where $q = 1$, given by

$$d_{Manhattan}(i, j) = \sum_{t=1}^{n} (i_t - j_t)$$

When using a KNN model for EEG however, these metrics can not be used. The reason for this is that not just simple samples are compared, but complete time series are compared to each other. Instead of the aforementioned distance metrics, another type of similarity measure is used, Dynamic Time Warping.

Dynamic Time Warping

Dynamic time warping (DTW) [45] is a measurement of similarity between two signals, even if they do not perfectly sync up. This can be used as a distance metric for a KNN model when working with time series data, which can then be used to compare samples to each other during majority voting. With

DTW, the time indices of the two time series are aligned such that the Euclidean distance between the signals is minimized. This becomes the optimization problem

$$DTW(X,Y) = \min_{\pi \in A(X,Y)} \left(\sum_{(i,j)\in\pi} d(X_i, Y_j)^q \right)^{\frac{1}{q}}$$

Here $A(X,Y)$ is the set of all possible paths between signals X and Y. π is the alignment path for these signals. A path should always start with both start indices of the signals, and should always end with both end indices of the signals. All indices of both signals should appear at least once in a path.

2.3.4. Long Short-Term Memory Model

The Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN). By using gated memory cells, it can be used to learn information over longer time scales compared to simple recurrent neural networks. LSTM models are therefor useful for time series classification tasks, such as speech recognition [41].

An LSTM model is made up of multiple LSTM cells. Within each cell, there are 3 gates: the input, output and forget gate. These gates control how information is passed through the cell. For the gates a sigmoid $\sigma(x)$ activation function is often used, which is defined as follows:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

The activation of the gates determines which information of the previous time steps is kept in the cell state and which is forgotten. This is what makes an LSTM model stand out from a standard Recurrent Neural Network (RNN) model and what allows for the long term connections within the data to be remembered by the model.

Figure 2.9 shows an overview of an LSTM memory cell. In here, x denotes the input to the layer, h the hidden state and c the cell state.

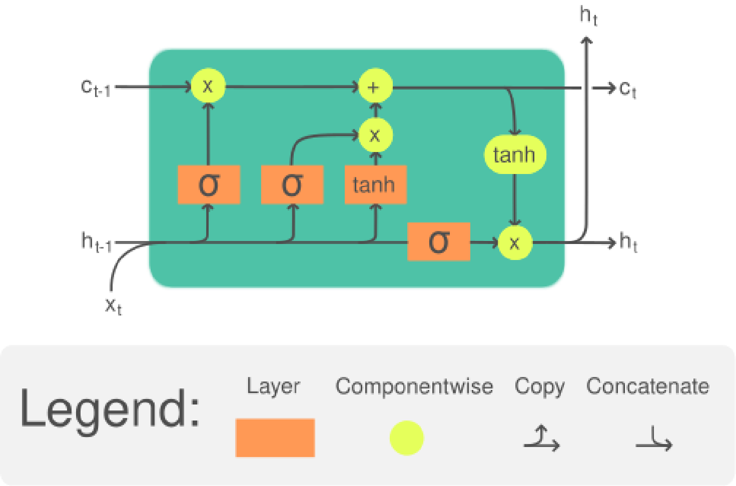

Figure 2.9: Memory cell of an LSTM recurrent neural network model. [10].

Listing 2.1 shows how everything is put together in equation form for a forward pass of an LSTM model. Here t indicates the current time step, i_t, o_t and f_t represent the activation vectors for the input, output and forget gates respectively. h_t represents the hidden state for time step t. c'_t and c_t represent the cell input activation and cell state vectors respectively. $\odot$ is the element wise multiplication of 2 vectors.

$W \in \mathbb{R}^{h \times d}$, $U \in \mathbb{R}^{h \times h}$ and b are the weight matrices and bias vector that are learned during training of the model. The size depends on the input vector size (d) of and the size of the hidden state for the model (h).

$$\begin{aligned}
&i_t = \sigma(W_i x_t + U_i h_{t-1} + bi) \\
&o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
&f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
&c'_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
&c_t = f_t \odot c_{t-1} + i_t \odot c'_t \\
&h_t = o_t \odot \tanh(c_t)
\end{aligned} \tag{2.1}$$

2.3.5. Convolutional Neural Network

A convolutional neural network (CNN) consists of 3 main layer types. A convolution layer and a pooling layer that focus on feature extraction and a fully connected layer that is used to make predictions from these features. This section will cover each of these layers.

Convolutional Layer

The convolutional layer does most of the heavy lifting of a CNN model. A convolutional layer consists of 3 components: the input data, a filter (also known as a kernel) and a feature map. When using time series EEG data as an input, the input to the convolutional layer has 2 dimensions: the length in time steps of a sample and the number of features per time step.

A kernel is one of the trainable parameters in a CNN model. A kernel can be considered a feature detector. Features in this context are completely abstract and learned by the model during training. During convolution, the kernel moves across the input signal looking for a certain feature within its receptive field. The output of a convolutional layer is the convolution product of the input signal with the kernel. A convolutional layer can consist of many of these kernels, which can result in a higher dimensionality on the output than the input signal.

There are 3 main hyper parameters that can be tuned for a convolutional layer. First, the kernel size is an important hyper parameter. This determines the window size of the kernel, essentially dictating the time frame in which a kernel looks for features. Second, the kernel count determines how many different kernels are used independently on the input signal. The last main hyper parameter is stride. Stride in essence is the distance, in number of time steps, that a kernel will jump over to the next time frame that it considers. Higher strides result in a smaller output vector of the convolutional layer.

Pooling Layer

The pooling layer is a down sampling layer and most often found between subsequent convolutional layers. As stated before, using multiple kernels can result in a higher dimensionality on the output of a convolutional layer than the input. The pooling layer is then used to bring this down again.

A pooling layer works in a similar way to a convolutional layer. It also runs a kernel across the input signal, but instead of containing weights, the kernel applies an aggregate function to the input signal. Most commonly used aggregate functions are max pooling and average pooling. Max pooling takes the highest value of each time step to send to the output. Average pooling calculates the average value within its receptive field and puts that into the output array.

Fully Connected Layer

A fully connected layer is added to a CNN to perform the actual task of classification or regression on the features that are extracted by the convolutional and pooling layers. This layer is the same as used in classical neural networks. For classification, the number of output neurons correspond to the number of classes to classify. A softmax activation is used on this layer for classification. For regression problems, 1 output neuron is used.

3 Methodology

This chapter will focus on the methods used to combine articulated and imagined speech EEG data. First a detailed overview is given of the dataset used. After that, that the general pre-processing steps are discussed that are then applied to the data for each experiment. The first experiment to be discussed is the pilot experiment, in which multiple models are compared to each other on articulated speech EEG data vowel classification. From this experiment, the best model is chosen and used for all remaining experiments. After this, a method is introduced to separate the pre-speech section from the rest of the EEG data using a model trained on the articulated EEG data. Finally the experimental setup is introduced. Here the remaining comparison experiments are explained, that explore the usage of articulated speech in classification of imagined speech data.

3.1. Database

This thesis uses the Delft Articulated and Imagined Speech (DAIS) dataset [21], which consists of EEG signals of imagined and articulated Dutch and speech from 20 native Dutch subjects, 6 male and 14 female [24]. The subjects were asked to imagine and articulate speech of 15 prompts: five vowels (/a:, e:, i, o:, u/ where ":" indicates long vowels) and 10 words. The 5 vowels constitute the different corners of the Dutch vowel quadrant. The 10 words are 5 Dutch word-pairs that are also words when read backwards (with english translations provided), each vowel is part of one word pair:

- *taal* and *laat* for /a:/ (*"language"* and *"late"*)
- *leeg* and *geel* for /e:/ (*"empty"* and *"yellow"*)
- *niet* and *tien* for /i/ (*"not"* and *"ten"*)
- *toon* and *noot* for /o:/ (*"tone"* and *"note"*)
- *soep* and *poes* for /u/ (*"soup"* and *"cat"*)

Figure 3.1 shows the Dutch vowel quadrant with the available vowels marked in red.

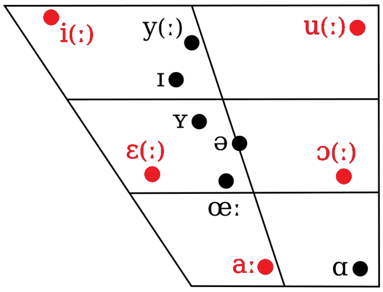

Figure 3.1: Dutch vowel quadrant with the vowels available in the DAIS dataset marked in red. The vertical axis represents how open the mouth is to produce the sound and the horizontal axis represents the position of the tongue elevation.

The EEG signals were recorded over 62 channels, placed according to the standard 10-20 international system [39] using the TMSi SAGA 64+ and with a BrainWave EEG Cap at a sampling frequency of 1024 Hz and the TMSi SAGA interface for MATLAB. The SAGA docking station was located outside the sound-attenuating room. Impedances were kept below $50k\Omega$. The audio is sampled at 44.1 kHz.

Each participant completed 20 runs of 15 trials, one for every prompt (i.e., the 15 Dutch vowels and words), where a trial consisted of 4 successive segments: rest, reading of the prompt, imagining to produce the prompt, and articulating the prompt. Each run was followed by (another) 2s rest. Each EEG recording is divided into 2s segments, for each prompt [24].

3.2. Data Pre-processing: Filtering

First, the EEG data was band-pass filtered (Second order Butterworth filter) using a high-pass filter at 1 Hz and a low-pass filter at 40 Hz to limit any electrical noise, such as power line noise, present in the signal. Artifacts such as blinks and muscle movement are removed where possible by using the low-pass filter. If it is not possible to remove the artifact from the EEG data frame, the whole frame is discarded.

Lastly, when dropping channels from the EEG data for the different experiments that are discussed in 3.6.1, the channels must be re-referenced to each other by subtracting the average of all remaining channels combined from each channel.

3.3. Pilot Experiments: Model Selection

A number of pilot experiments were done to determine the best model type to use for all further experiments. In these pilot experiments we compared support vector machine (SVM), K-nearest neighbour (KNN) and random forest (RF) machine learning models as well as Long short-term memory (LSTM) and Convolutional Neural Networks (CNN) deep learning models on articulated speech vowel classification. The deep learning (DL) models, (LSTM and CNN) can be used directly on the pre-processed EEG signals. The machine learning (ML) models (SVM, KNN and RF) need some adaptation to work with the time series data from the EEG channels. Performance for all models is measured in accuracy as correct predictions divided by total number of test samples. After all filtering steps as discussed in section 3.2 a total of 356 training samples and 83 test samples remain.

Time Series K-Nearest Neighbour

A KNN works by comparing a distance value between multiple samples. Dynamic Time Warping (DTW)[45] is used as a distance measurement for the KNN algorithm. With this distance measurement, the KNN algorithm can be applied as normal: k-number of samples from the training set are compared to the given sample. The label for the unknown sample is then determined by a majority vote from the k-nearest samples found in the training set. For the pilot experiment a k of 7 is used. This number is chosen such that no tie can occur, and every class could be represented at least once in the majority vote.

Time Series Random Forest

The time series variant of the random forest algorithm works by extracting features from a given window. The features extracted per window are the mean, standard deviation and the slope. The total number of features for this random forest is then 3 times the number of windows that fit the data. A standard random forest tree is then trained with these features as input data. Since every channel in the EEG data needs to be considered as a separate feature, the model needs to be adapted to work with multivariate data. To do this, one of these trees is made for each feature (EEG channel) in the data. Each feature tree then comes up with a classification result, considering only the feature (channel) they are assigned. The overall classification result for the model is then determined by a majority vote from all these feature trees. For example, consider a dataset consisting of 6 EEG channels. In this case 6 feature trees are created that all focus on a single channel. 4 of the 6 trees then classify their channel with label ”A”, while the other 2 classify their channel as label ”B”. The overall model will then output ”A” as the label for the overall sample, as label ”A” has gotten 4 votes versus only 2 votes for sample ”B”.

Time Series SVM

The adaption of the SVM model for time series is comparable to that of the KNN model. A different type of internal metric needs to be used in order to compare the different samples. For this experiment, global alignment kernels (GAK)[19] are used, which is a form of DTW, as used for the KNN.

3.4. Final Model Architecture

Based on the pilot experiments described in section 3.3, the best-performing model is chosen for the experiments reported here. The model used is a CNN model with an input size of 2048 timesteps with a variable number of features. The feature count per timestep depends on the number of EEG channels used, as each channel is considered to be a separate feature. The model consists of three repeated convolutional layers followed by global average pooling and finally a fully connected prediction layer, similar to commonly used convolutional models for time series classification [49]. Each convolutional layer has batch normalization and a dropout of 25% to prevent overfitting with rectified linear activation at each layer.

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

3.5. Predicting the End Point of Imagined Speech

To separate the imagined speech from possible noise in the EEG frame, a model is made to detect where imagined speech ends. While there is no ground-truth for the endpoint in imagined speech available, for articulated speech, it is known when speech starts, and the time stamps of the speech are aligned with the EEG signal of the articulated speech. It is assumed that the starting point of articulated speech is the end of the preparations to speak, and then is taken as the end point of the imagined speech.

To determine at which timestamp speech starts for each data sample, SileroVAD, a neural network based voice activity detection (VAD) algorithm [52], is applied to the speech files provided in the DAIS dataset. From this, a timestamp is extracted at which speech starts in the articulated speech. The period in the EEG signal prior to this point we refer to as “pre-speech”. A CNN model was trained with articulated speech EEG signal segments as input and the timestamps obtained from the VAD as target, transforming this into a regression problem. The model trained for 150 epochs.

Figure 3.2 shows an example of a segment of articulated speech (bottom left) and the accompanying EEG signal (top left). The blue vertical line marks the onset of speech as provided by the VAD. The top right panel shows the EEG signal after removing the EEG signals after the speech onset, i.e., it only shows the “pre-speech” EEG signal. Here the difference can be seen between the pre-speech only and full EEG inputs to the models.

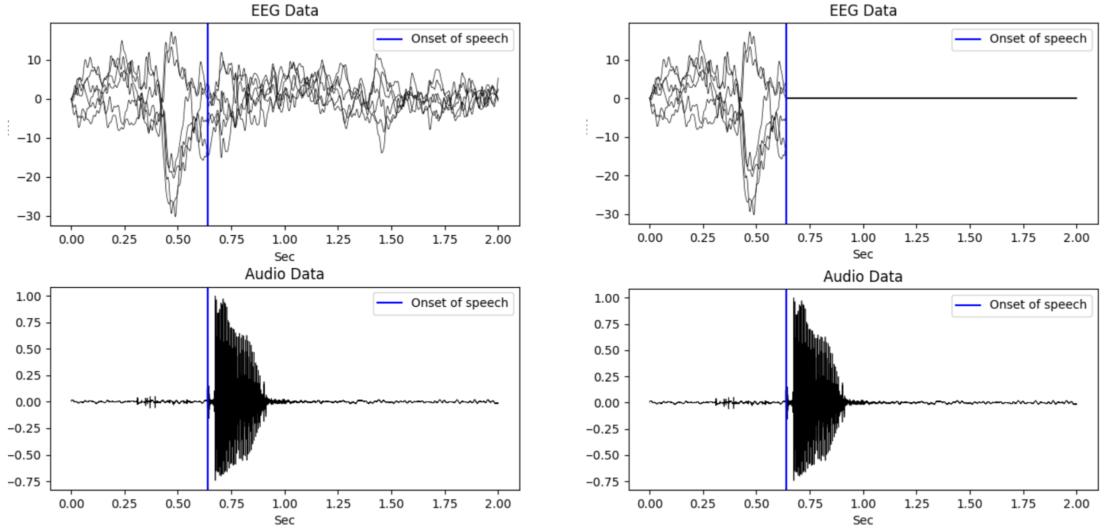

Figure 3.2: Left panels: EEG segment of articulated speech (top) and the acoustic signal (bottom); Right panels: EEG signal of the "pre-speech", after cutting the segment at the start of speech (vertical line).

3.6. Experimental Setup

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

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

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

3.6.1. EEG Channel Selection

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

- ***Channel Set 6***: 6 channels {F3, F4, C3, C4, P3, P4}. Following the Coretto database [16], we use 6 channels, which is the lowest number used in any database of imagined speech.
- ***Channel Set 8***: 8 channels {Fz, C3, Cz, C4, Pz, PO7, Oz, PO8} as used in [5]. These channels are chosen as they are close to Broca's and Wernicke's regions of the brain, which is assumed to produce good quality imagined speech-based EEG signals [5].
- ***Channel Set 16***: 16 channels {F7, F5, FT7, FC5, FC3, FC1, T7, C5, C3, Cz, C4, TP7, CP5, CP3, P5, P3} that are located on specific areas of the cortex that are known to be involved in language processing. These 16 channels were also used in the validation study reported in [24].

- ***Channel Set 62***: All available channels are used.

Each channel subset is run 5 times for each combination of train and test data (articulated, imagined and combined data) as discussed in Section 3.6.3.

3.6.2. Predicting end point of imagined speech

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

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

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

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

3.7. T-SNE Analysis

After the experiments described above are run, some more insight in to what the model is doing is needed to help understand the results. For this t-SNE analysis [37] is done on the layers within a given model. First an analysis is done on a model trained and tested with articulated speech EEG data containing all channels. This is then compared to a model trained and tested on imagined speech data containing all channels, to see how the 2 types of data differ from each other when used in a model. The training and testing procedure is the same as for the other experiments. The same pre-processing steps are applied and the same datasets are used as for the articulated and imagined speech experiments as described in section 3.6.3. For legibility, a smaller subset of subjects was used. Only subject 1, 3 4 5 and 6 are considered. A total of 63 samples are used for imagined speech analysis and 99 samples are used for articulated speech.

4

Experimental Results

4.1. Initial Model Comparison

Table 4.1 shows the accuracies together with the standard deviation of the models used in the model exploration experiment when classifying articulated speech vowels using all channels on full EEG frames. It can be seen that from the basic ML (KNN, RF and SVM) models, the KNN model scores the best on average, although a high standard deviation is observed. The SVM scores just slightly above chance level of 20 %. The RF model scores comparable to the LSTM model, but the LSTM model has a lower standard deviation. The CNN model scores highest overall with an average of 49% and a standard deviation of 2.8%. This makes it the best choice to use for all other experiments.

Table 4.1: The accuracy and $Stdev$ for articulated speech vowel classification task of the 5 models tested in the pilot experiment.

Model	**KNN**	**SVM**	**RF**	**LSTM**	**CNN**
Accuracy (%)	32.2	23.3	28.4	28.1	49.0
std dev (%)	7.5	3.8	3.3	1.8	2.8

4.2. Predicting the End point of Imagined Speech

Table 4.2 shows the MSE and standard deviation ($Stdev$) of the difference (in milliseconds) between the ground-truth timestamp of start time in the articulated speech EEG signal and the predicted timestamp for the four models with different channel sets. First, with increasing number of channels, the MSE and standard deviation reduce, although the smallest standard deviation was found for the model with only 6 channels. Importantly, all models show relatively good prediction results, with a maximum MSE of only 5.65 ms, which is a lot less than the duration of a single vowel sound, which is around 300 ms in this dataset. This indicates that the start of speech can be predicted from the EEG signal of articulated speech within a relatively small margin.

Table 4.2: The MSE and $Stdev$ of the predicted start of the articulated speech (in ms).

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

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

Tables 4.3 and 4.4 show the classification results of the different models on the EEG of articulated and imagined speech in terms of accuracy together with the standard deviation for each channel subset when the full EEG signal is used (Table 4.3) and when only the pre-speech is used (Table 4.4) for the four different channel sets. The results are grouped by the type of training and test data used (articulated, imagined or a combination) and the number of channels.

The results for experiment 1 show a baseline of 53.8% accuracy for speaker-independent articulated Dutch vowel classification when all available channels and the full EEG segments are used. This can be viewed as an upper bound for the imagined vowel classification task. This performance drops to 43.9% when only the pre-speech part of the segments is used. The information related to articulation in the EEG signal is thus needed for improved classification of articulated vowels. Interestingly however, this is still much higher than any of the imagined speech results, suggesting that there is a non trivial difference between the pre-speech part of articulated speech and imagined speech.

The results for experiment 2 show a baseline of 24.8% accuracy for speaker-independent imagined Dutch vowel classification when all available channels and the full EEG segments are used. This increases to 27% accuracy when only using the pre-speech part of the EEG segment, although this still falls within 1 standard deviation from the baseline. Chance level in all cases is 20%.

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

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

Table 4.3: Accuracy (in %) of the classification experiments using the full EEG frames, grouped by channel sets used.

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

Table 4.4: Accuracy (%) of the classification experiments using the pre-speech, grouped by channel sets used.

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

4.4. T-SNE Analysis

Figure 4.1 shows the t-SNE analysis result for articulated speech. The left image shows the data points grouped by classification targets, with red for /a:/, green for /o:/, blue for /e:/, yellow for /i/ and orange for /u/. The right image shows the same data points, but grouped by speaker. Figure 4.2 shows the same for imagined speech.

From figure 4.1 it can be seen that clustering happens for the different classification targets on the articulated speech model. This clustering is not present when looking at figure 4.2 for the imagined speech model. Interestingly however, when labeling the points per speaker, it can be seen that the data for each speaker clusters together for the imagined speech model. This is less obvious for the articulated speech model, suggesting that the imagined speech model internally models the difference between speakers instead of the difference between classification targets, while the articulated speech model does the opposite.

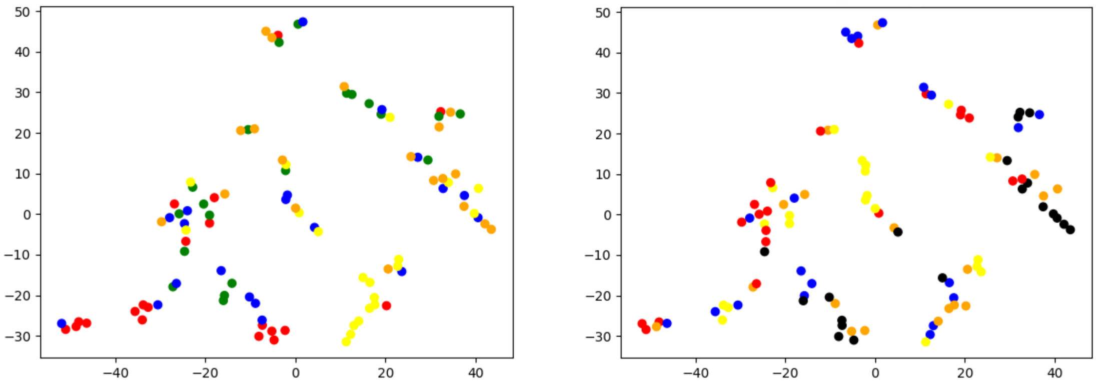

Figure 4.1: Left panel: t-SNE result of articulated speech model, grouped by classification target. Right panel: t-SNE result of articulated speech model, grouped by speaker.

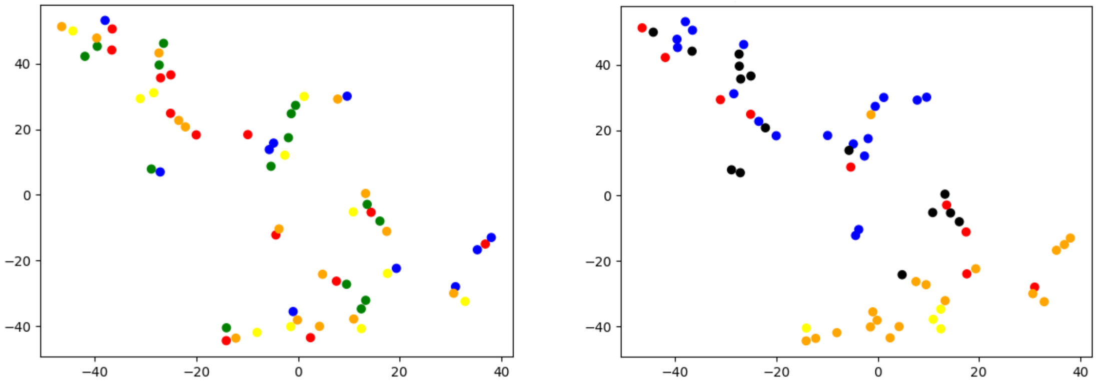

Figure 4.2: Left panel: t-SNE result of imagined speech model, grouped by classification target. Right panel: t-SNE result of imagined speech model, grouped by speaker.

5 Discussion and Conclusion

The aim of this thesis is to use articulated speech EEG data to improve classification results on imagined speech EEG data. First, multiple classification models are compared to each other, to pick the best one for further experiments. This resulted in the use of a CNN model. After that, an onset of articulated speech detection model is created which is then used to mark the end of imagined speech in the imagined speech EEG segments. This is done to try and separate useful information in the EEG signal from noise by removing anything that is not considered to be articulated speech. The experiment investigating the viability of such a model showed potential for a speech detection model based on EEG signals instead of audio signals when used on articulated speech EEG data in a speaker-independent way. The EEG signal before articulation onset is referred to as “pre-speech”. One assumption made in this thesis, is that this model can then be used to mark the point in the EEG data where imagined speech ends. Due to the lack of ground truth data, this assumption however cannot be validated. Future research could investigate the performance of this approach in a speaker-dependent scenario. Using a speaker-dependant model could more accurately capture the useful information in the imagined speech EEG data, as the between-speaker variability in the EEG signals is removed completely this way.

In a second experiment, it is investigated whether combining articulated and imagined speech EEG without extra processing steps improves imagined speech vowel classification. This does not improve the classification results, and in most cases even deteriorates the model performance. These results indicate that articulated and imagined speech can not be used together without extra processing steps. A possibility for this is the presence of motor functions for speech in articulated speech that are not present in imagined speech. A model can use these extra signals to help classification in articulated speech, which then does not translate into imagined speech classification.

After this, an experiment was done to investigate if using only the pre-speech part of articulated speech EEG data makes it more compatible with the imagined speech EEG data for use in training a classifier. The result from this experiment again shows no improvement to the classification results. This further suggests that articulated speech cannot be used to increase the amount of data available for imagined speech classification.

Overall the results on speaker-independent imagined vowel classification are in line with other research where the focus is on speaker-independent models [23]. From the results of experiment 3 in Tables 4.3 and 4.4, it can be seen that there is usable information in the pre-speech part of articulated speech EEG data as all results are above chance level. Nevertheless, in general using pre-speech only did not improve vowel classification of imagined nor articulated speech. However, there are two interesting findings. For the all-channels model trained on imagined vowel EEG and tested for imagined vowel classification, using only the pre-speech gave an improvement over using the full EEG, suggesting there is information in the full signal that is unnecessary and reduces performance. This is in line with the second interesting finding: the 16-channel model trained on imagined speech outperforms the all-channels model on imagined speech vowel classification. Also here, using less data improved imagined vowel classification. Future research should focus on investigating what information is necessary for improved imagined vowel classification and what information is better removed from the EEG signal.

The lack of an improvement when using pre-speech EEG is not in line with the assumptions made in this thesis based on the literature: the difference between imagined speech EEG and the phase before speech in articulated speech EEG is too different to be used together to train a classifier.This can be seen from the classification results of full EEG frames compared to pre-speech only frames in articulated speech. While the pre-speech only frames score lower than the full frames, suggesting some data is lost when removing the articulation part of speech from the EEG frame, they still score much higher than any imagined speech classification models in this thesis, further accentuating the non trivial difference between articulated and imagined speech. Although, of course, it cannot be excluded that the speech onset detection in EEG algorithm used in this thesis is not working well enough for imagined speech. However, due to a lack of ground truth, this cannot be verified.

When looking at the t-SNE analysis plots, one thing that is noticeable for the articulated speech model is that some vowels cluster together more than others. For example, /a:/ and /u/ cluster far away from each other, while /o:/ and /u/ cluster together more. This suggests that /a:/ and /u/ are easier to distinguish from each other by a classifier than /o:/ and /u/ are. This also corresponds with their relationship to each other on the dutch vowel quadrant, as /o:/ and /u/ lay on the same side and /a:/ and /u/ lay on opposite sides, with regards to tongue position. This indicates that the position of the tongue is encoded somewhere in the articulated speech EEG data.

Another thing to notice from the t-SNE plots is the lack of clustering for vowels in imagined speech, but the presence of clustering in subjects. For articulated speech this is the opposite. This further confirms the high variability between subjects in imagined speech data, and indicates further research should focus on speaker dependant models.

To conclude, the data scarcity problem in imagined speech EEG classification cannot easily be solved by adding more data from articulated speech EEG. Instead, research focus should lie on investigating what information in the EEG signal to use for imagined speech classification, and better ways to generalize the EEG signals across speakers. A good place to start is comparing the pre-speech part of articulated speech to imagined speech in a speaker-dependant way, to get a better idea of how these signals exactly differ from each other.

References

[1] URL: https://www.nidcd.nih.gov/health/aphasia.

[2] URL: https://mne.tools/dev/auto_tutorials/preprocessing/15_handling_bad_channels.html.

[3] URL: https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works#:~:text=SVM%20works%20by%20mapping%20data,be%20drawn%20as%20a%20hyperplane..

[4] URL: https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20non,used%20in%20machine%20learning%20today..

[5] Mokhles M. Abdulghani, Wilbur L. Walters, and Khalid H. Abed. "Imagined Speech Classification Using EEG and Deep Learning". In: *Bioengineering* 10.6 (2023). ISSN: 2306-5354. DOI: 10.3390/bioengineering10060649. URL: https://www.mdpi.com/2306-5354/10/6/649.

[6] Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra. "Chapter 2 - Technological Basics of EEG Recording and Operation of Apparatus". In: *Introduction to EEG- and Speech-Based Emotion Recognition*. Ed. by Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra. Academic Press, 2016, pp. 19–50. ISBN: 978-0-12-804490-2. DOI: https://doi.org/10.1016/B978-0-12-804490-2.00002-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128044902000026.

[7] Andrew Bowers et al. "Suppression of the µ Rhythm during Speech and Non-Speech Discrimination Revealed by Independent Component Analysis: Implications for Sensorimotor Integration in Speech Processing". In: *PLOS ONE* 8.8 (Aug. 2013), pp. 1–17. DOI: 10.1371/journal.pone.0072024. URL: https://doi.org/10.1371/journal.pone.0072024.

[8] Jonathan S. Brumberg et al. "Brain–computer interfaces for speech communication". In: *Speech Communication* 52.4 (2010). Silent Speech Interfaces, pp. 367–379. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2010.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0167639310000105.

[9] Edward F. Chang, Kunal P. Raygor, and Mitchel S. Berger. "Contemporary model of language organization: an overview for neurosurgeons". In: *Journal of Neurosurgery* 122.2 (Feb. 2015), pp. 250–261. ISSN: 1933-0693. DOI: 10.3171/2014.10.jns132647. URL: http://dx.doi.org/10.3171/2014.10.JNS132647.

[10] Guillaume Chevalier. *Schematic of the Long-Short Term Memory cell, a component of recurrent neural networks*. https://upload.wikimedia.org/wikipedia/commons/9/93/LSTM_Cell.svg.

[11] Etienne Combrisson and Karim Jerbi. "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy". In: *Journal of Neuroscience Methods* 250 (2015). Cutting-edge EEG Methods, pp. 126–136. ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2015.01.010. URL: https://www.sciencedirect.com/science/article/pii/S0165027015000114.

[12] Ciaran Cooney, Rafaella Folli, and Damien Coyle. "Mel Frequency Cepstral Coefficients Enhance Imagined Speech Decoding Accuracy from EEG". In: *2018 29th Irish Signals and Systems Conference (ISSC)*. 2018, pp. 1–7.

[13] Ciaran Cooney, Raffaella Folli, and Damien Coyle. "Optimizing Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019, pp. 1311–1316.

[14] Ciaran Cooney et al. "Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG". In: *Sensors* 20.16 (Aug. 2020), p. 4629. ISSN: 1424-8220. DOI: 10.3390/s20164629. URL: http://dx.doi.org/10.3390/s20164629.

[15] Ciaran Cooney et al. “Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG”. In: *Sensors* 20.16 (Aug. 2020), p. 4629. ISSN: 1424-8220. DOI: 10.3390/s20164629. URL: http://dx.doi.org/10.3390/s20164629.

[16] Germán A Pressel Coretto, Iván E Gareis, and H Leonardo 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] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/bf00994018. URL: http://dx.doi.org/10.1007/BF00994018.

[18] R.J. Croft and R.J. Barry. “Removal of ocular artifact from the EEG: a review”. In: *Neurophysiologie Clinique/Clinical Neurophysiology* 30.1 (2000), pp. 5–19. ISSN: 0987-7053. DOI: https://doi.org/10.1016/S0987-7053(00)00055-1. URL: https://www.sciencedirect.com/science/article/pii/S0987705300000551.

[19] Marco Cuturi. “Fast global alignment kernels”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML’11. Bellevue, Washington, USA: Omnipress, 2011, pp. 929–936. ISBN: 9781450306195.

[20] Lopez-Bernal D et al. “A State-of-the-Art Review of EEG-Based Imagined Speech Decoding”. In: *Hum Neurosci* (2022). DOI: 10.3389/fnhum.2022.867281.

[21] “DAIS: The Delft Database of EEG Recordings of Dutch Articulated and Imagined Speech”. English. In: *Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. United States: IEEE, 2023. ISBN: 978-1-7281-6328-4. DOI: 10.1109/ICASSP49357.2023.10096145.

[22] Charles S. DaSalla et al. “Single-trial classification of vowel speech imagery using common spatial patterns”. In: *Neural Networks* 22.9 (2009). Brain-Machine Interface, pp. 1334–1339. ISSN: 0893-6080.

[23] Debadatta Dash et al. “Towards a Speaker Independent Speech-BCI Using Speaker Adaptation”. In: *Interspeech 2019*. ISCA, Sept. 2019.

[24] Bo Dekker, Alfred C Schouten, and Odette Scharenborg. “DAIS: The delft database of EEG recordings of dutch articulated and imagined speech”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, June 2023.

[25] Jesús S. García-Salinas et al. “Tensor Decomposition for Imagined Speech Discrimination in EEG”. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 239–249. ISBN: 9783030044978.

[26] Erin Gibson et al. “EEG variability: Task-driven or subject-driven signal of interest?” In: *NeuroImage* 252 (2022), p. 119034. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2022.119034. URL: https://www.sciencedirect.com/science/article/pii/S105381192200163X.

[27] Carlos Gussenhoven. “Dutch”. In: *Journal of the International Phonetic Association* 22.1–2 (June 1992), pp. 45–47. ISSN: 1475-3502. DOI: 10.1017/s002510030000459x. URL: http://dx.doi.org/10.1017/S002510030000459X.

[28] Lars Hausfeld et al. “Pattern analysis of EEG responses to speech and voice: Influence of feature grouping”. In: *NeuroImage* 59.4 (2012), pp. 3641–3651. ISSN: 1053-8119.

[29] Tin Kam Ho. “Random Decision Forest”. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (1995).

[30] Amir Jahangiri, David Achanccaray, and Francisco Sepulveda. “A Novel EEG-Based Four-Class Linguistic BCI”. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 3050–3053. DOI: 10.1109/EMBC.2019.8856644.

[31] David Jenson et al. “Temporal dynamics of sensorimotor integration in speech perception and production: independent component analysis of EEG data”. In: *Frontiers in Psychology* 5 (2014). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.00656. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2014.00656.

[32] Anne Kösem and Virginie van Wassenhove. “Distinct contributions of low- and high-frequency neural oscillations to speech comprehension”. In: *Language, Cognition and Neuroscience* 32.5 (Oct. 2016), pp. 536–544. ISSN: 2327-3801. DOI: 10.1080/23273798.2016.1238495. URL: http://dx.doi.org/10.1080/23273798.2016.1238495.

[33] Seo-Hyun Lee, Minji Lee, and Seong-Whan Lee. “Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.12 (2020), pp. 2647–2659. DOI: 10.1109/TNSRE.2020.3040289.

[34] Diego Lopez-Bernal et al. “A State-of-the-Art Review of EEG-Based Imagined Speech Decoding”. In: *Frontiers in Human Neuroscience* 16 (Apr. 2022). ISSN: 1662-5161. DOI: 10.3389/fnhum.2022.867281. URL: http://dx.doi.org/10.3389/fnhum.2022.867281.

[35] M A Lopez-Gordo et al. “An auditory brain–computer interface evoked by natural speech”. In: *Journal of Neural Engineering* 9.3 (May 2012), p. 036013. ISSN: 1741-2552.

[36] Palmer E. D. Rosen H. J. Ojemann J. G. Buckner R. L. Kelley W. M. and Petersen S. E. “An event-related fMRI study of overt and covert word stem completion.” In: *NeuroImage, 14(1 Pt 1)* (2001), pp. 182–193. DOI: https://doi.org/10.1006/nimg.2001.0779.

[37] Laurens van der Maaten and Geoffrey Hinton. “Viualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.

[38] Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. “Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features”. In: *Journal of Neural Engineering* 15.1 (Nov. 2017), p. 016002. ISSN: 1741-2552.

[39] Ernst Niedermeyer and F H Lopes Da Silva, eds. *Electroencephalography*. en. 5th ed. Philadelphia, PA: Lippincott Williams and Wilkins, Nov. 2004.

[40] Sanne ten Oever and Alexander T. Sack. “Oscillatory phase shapes syllable perception”. In: *Proceedings of the National Academy of Sciences* 112.52 (Dec. 2015), pp. 15833–15837. ISSN: 1091-6490. DOI: 10.1073/pnas.1517519112. URL: http://dx.doi.org/10.1073/pnas.1517519112.

[41] Jane Oruh, Serestina Viriri, and Adekanmi Adegun. “Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition”. In: *IEEE Access* 10 (2022), pp. 30069–30079. DOI: 10.1109/ACCESS.2022.3159339.

[42] Jerrin Thomas Panachakel, A.G. Ramakrishnan, and T.V. Ananthapadmanabha. “Decoding Imagined Speech using Wavelet Features and Deep Neural Networks”. In: *2019 IEEE 16th India Council International Conference (INDICON)*. 2019, pp. 1–4. DOI: 10.1109/INDICON47234.2019.9028925.

[43] Dipti Pawar and Sudhir Dhage. “Multiclass covert speech classification using extreme learning machine”. In: *Biomedical Engineering Letters* 10.2 (Mar. 2020), pp. 217–226. ISSN: 2093-985X.

[44] Xiaomei Pei et al. “Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition”. In: *NeuroImage* 54.4 (2011), pp. 2960–2972. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2010.10.029. URL: https://www.sciencedirect.com/science/article/pii/S1053811910013212.

[45] H. Sakoe and S. Chiba. “Dynamic programming algorithm optimization for spoken word recognition”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (1978), pp. 43–49. DOI: 10.1109/TASSP.1978.1163055.

[46] Luis Carlos Sarmiento et al. “Recognition of EEG Signals from Imagined Vowels Using Deep Learning Methods”. In: *Sensors* 21.19 (Sept. 2021), p. 6503. ISSN: 1424-8220.

[47] Eric W. Sellers, David B. Ryan, and Christopher K. Hauser. “Noninvasive brain-computer interface enables communication after brainstem stroke”. In: *Science Translational Medicine* 6.257 (Oct. 2014). ISSN: 1946-6242. DOI: 10.1126/scitranslmed.3007801. URL: http://dx.doi.org/10.1126/scitranslmed.3007801.

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

[49] Aboozar Taherkhani, Georgina Cosma, and T. M. McGinnity. “A Deep Convolutional Neural Network for Time Series Classification with Intermediate Targets”. In: *SN Computer Science* 4.6 (Oct. 2023). ISSN: 2661-8907. DOI: `10.1007/s42979-023-02159-4`. URL: `http://dx.doi.org/10.1007/s42979-023-02159-4`.

[50] Bang-Bei Tang et al. “The effect of odor exposure time on olfactory cognitive processing: An ERP study”. In: *Journal of integrative neuroscience* 18 (Mar. 2019), pp. 87–93. DOI: `10.31083/j.jin.2019.01.103`.

[51] DataFlair Team. *Kernel functions-introduction to SVM Kernel and Examples*. Mar. 2021. URL: `https://data-flair.training/blogs/svm-kernel-functions`.

[52] Silero Team. *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. `https://github.com/snakers4/silero-vad`. 2021.

[53] Darya Vorontsova et al. “Silent EEG-Speech Recognition Using Convolutional and Recurrent Neural Network with 85% Accuracy of 9 Words Classification”. In: *Sensors* 21.20 (Oct. 2021), p. 6744. ISSN: 1424-8220.

[54] Natsue Yoshimura et al. “Decoding of Covert Vowel Articulation Using Electroencephalography Cortical Currents”. In: *Frontiers in Neuroscience* 10 (May 2016). ISSN: 1662-453X.

[55] Yongsheng Zhao, Ying Liu, and Yunlong Gao. “Analysis and classification of speech imagery EEG based on Chinese initials”. In: *J. Beijing Inst. Tech* 30 (2021), pp. 44–51.