Classification of Covert Vowels in Spanish and Dutch

What do brain signals say about inner speech?

Ioannis Kyriazis

TU Delft

Classification of Covert Vowels in Spanish and Dutch

What do brain signals say about inner speech?

Thesis report

by

Ioannis Kyriazis

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on August 28, 2023 at 12:00

Thesis committee:

Chair:	Prof. dr. ir. A. C. Schouten, TU Delft
Supervisors:	Prof. dr. ir. A. C. Schouten, TU Delft
	Dr. O. E. Scharenborg, TU Delft
External examiners:	Dr. ir. Ajay Seth, TU Delft
	Dr. ir. Yke Bauke Eisma, TU Delft
Place:	Faculty of Mechanical, Maritime & Materials Engineering, Delft
Project Duration:	February, 2023 - August, 2023
Student number:	5622921

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Classification of covert vowels in Spanish and Dutch: what do brain signals say about inner speech?

Ioannis Kyriazis
Department of Biomedical Engineering
Faculty of Mechanical, Maritime & Materials Engineering
Delft University of Technology

***Abstract*—Patients with neuromuscular diseases that are unable to speak, but whose cognitive ability has been maintained, can be benefited from Brain Computer Interfaces (BCIs). The decoding of inner (covert) speech from EEGs consists of one of the state of the art methods that aim to tackle this issue. High variability between subjects, as well as low signal to noise ratio (SNR) undermine the methods used, and introduce the need for computer assisted solutions. Thus, machine learning models as well as large amounts of recorded data are required to design effective algorithms and produce substantial results. In this study, covert vowel classification was performed in a systematic way, by making use of two openly shared databases from literature; the Coretto database, that contains EEG recordings of native Spanish speakers, and the DAIS dataset, which includes EEG recordings of native Dutch speakers. Six classifiers were initially selected to perform 5-class classification: a Random Forest (RF), a k Nearest Neighbours (kNN), a Gaussian Naive Bayers (GNB), a Deep Convolutional Neural Network (DCNN), a Shallow Convolutional Neural Network (SCNN) and a Long Short Term Memory Recurrent Neural Network (LSTM). The DCNN outperformed the other methods, with average intra-subject accuracies of 35% for Coretto and 39% for DAIS (chance level 20%). Afterwards, an Overt versus Covert trials experiment was implemented, to test the limits of overt speech decoding from EEGs. The overt result was slightly higher than covert, with an intra-subject average value of 37.8% for Coretto and 40.5% for DAIS (chance level 20%). Finally, binary classification was performed to identify those pairs of vowels that can be classified more efficiently. Vowels /a/ and /u/ seemed to perform better in average in both datasets (average of 64.8% for Coretto and 64.4% for DAIS with a chance accuracy of 50%). Future work should focus on identifying the useful parts of the EEG recordings, increasing the SNR and the resolution of the electrodes, and defining the most appropriate dictionaries of words/vowels for a BCI. Also, more studies should follow systematic ways of comparisons between datasets, to obtain less ambiguous insights and lead this field to improvements.**

***Index Terms*—brain computer interface (BCI), Dutch covert speech, Spanish covert speech, electroencephalography (EEG).**

I. Introduction

The decoding of covert speech from electroencephalograms (EEGs) is widely used to attempt the translation of thoughts into words. The application of this method could assist tremendously in cases of patients afflicted by neuromuscular disorders, whose cognitive ability still remains; such as those suffering from locked-in syndrome [1] [2]. Brain Computer Interfaces (BCIs) are able to provide the necessary human-machine interaction to achieve this objective [3].

Modern BCIs used for the purpose of covert speech decoding make use of the recorded brain activity from surface electrodes located around the healthy volunteer's head. The recorded EEGs are then processed and given as input to a computational setup. This method is low-cost, non-invasive and achieves high temporal resolution [4]. For those reasons, it is considered the state-of-the-art standard for such applications.

Decoding of covert speech is a complicated task, hindered by the limitations of the recording method itself as well as the variability of results per subject [5]. Multiple studies have attempted to provide high accuracy solutions, but results are often close to chance level [6]. In addition, the lack of a common protocol for result presentation introduces a lot of ambiguity on result interpretation, as it is unclear which dataset, processing method, or which classifier achieves the best results. Furthermore, even though similar brain cortex areas are activated during speech in healthy individuals, there is high variability per subject, which makes the generalization of insights more difficult.

Deep Learning (DL) Algorithms, and especially Convolutional Neural Networks (CNNs), are observed to produce the highest accuracy in literature [7]. However, unless the same architecture is tested for different groups of covert inputs, generalization is difficult. Furthermore, almost all DL algorithms suffer from overfitting, a factor that is also bothersome to judge, unless all the specific parameters of segmentation are known, and that is often omitted in literature.

The input prompts of any given covert speech dataset are also shown to be important for classification accuracy. Some articles assess that vowels achieve higher results than complete words [8], whereas others support the opposite view [9]. The number of prompts is also important, because a smaller number leads to higher chance level accuracy, meaning that even if the algorithm classifies randomly, a higher result might be obtained [6]. Thirdly, the language of the prompts is assumed to be of importance. Several studies choose to use English prompts, even though English does not consist the

native language of the subjects. Those studies usually state that the subjects have had sufficient English education [10]. However, even in the cases that the native language of the subjects is used, comparisons between articles with the same prompts, but in different language, are difficult to make [11].

In terms of classification methods, an additional level of ambiguity is added, with a multitude of used features and classifiers. Some of the most used features are: statistical ones (e.g. mean, standard deviation, variance, kurtosis, skewness, higher order moments, energy, zero-crossings), frequency features (e.g. Fourier Transform coefficients, Wavelet Decomposition coefficients, Spectral Densities, Gabor transform, Hilbert transform), coherence, Common Spatial Patterns, and covariance matrices [12], [13], [14], [15], [16], [17], [18]. Those techniques seem to be beneficial for some studies but without consistency, while other studies argue about the higher significance of separated frequency bands instead of just features (e.g. gamma band in [14], or theta band in [16]).

Machine Learning algorithms require large amounts of data. To this end, several datasets of covert speech EEG recordings have been introduced in literature, although just a few are openly shared (e.g. Coretto et al. [19], DaSalla [20], Nguyen et al. [21], DAIS dataset [22]). Typically, results obtained from different datasets are not easy to compare, because of differences in the acquisition setup and pre-processing steps. However, if treated correctly, comparisons between given datasets could be possible.

This research implements a systematic comparison between two openly shared datasets of covert vowels. The Coretto dataset contains native Spanish vowel recordings, while the DAIS dataset native Dutch vowel recordings. Comparisons between the two datasets were possible, as the vowels from one correspond to the other in a one to one way.

Any comparison between experimental setups is difficult, as recording protocols are often differentiated between articles. When it comes to dictionaries of similar prompts, only datasets that are used broadly in literature can be utilized, because there is a large pool of similar studies to compare results with. Comparisons between completely different datasets are also possible, but identical pre-processing of the data of each one cannot be made most of the times, because of the inherent characteristics of the datasets, or due to lack of comparison protocols. For example, an attempt for classification of vowels in English and Bengali was made [11], but it was never stated within the protocol that the goal was a comparison between the two to offer joined insights, nor it was mentioned if similar treatment was applied for the acquisition of the two datasets.

This study attempts to create a common ground for the generalization of results between the two covert vowel datasets. Using the same protocol when comparing datasets could prove to be useful, especially if future studies also follow a similar protocol to introduce or compare datasets.

II. Methods

A. Datasets

Two datasets were used in this study. The first one was the Coretto dataset [19], which contained EEG recordings of overt and covert vowels for 15 subjects, 8 male and 7 female, with an mean age of 25, whose native language was Spanish. Each trial was 4 seconds long. Six electrodes were used in total, placed on the F3, F4, C3, C4, P3, P4 locations of the 10-20 international system (figure 1). The sampling rate was 1024Hz. The number of trials per subject was variable, but ranged between 200 and 274. The vowel prompts were /a/, /e/, /i/, /o/ and /u/. The second dataset was the DAIS dataset [22], that contained EEG recordings of overt and covert vowels for 20 subjects, 6 male and 14 female (mean age of 24.6 years, range 23-26 years), whose native language was Dutch. Each trial was 2 seconds long. The number of electrodes was 62, placed according to the 10-20 international system (figure 1). A sampling frequency of 1024Hz was used. There were 20 trials per vowel, resulting in 100 trials per subject. The vowel prompts were /aa/, /ee/, /ie/, /oo/, /oe/, which correspond phonologically to the phonemes /a/, /e/, /i/, /o/ and /u/ respectively [23], [24], [25], [26]. The prompts of each dataset can be positioned on the vowel quadrilateral of each language (figure 2).

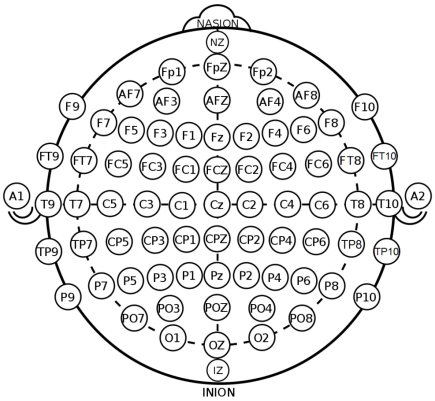


Fig. 1: 10-20 international system for EEG electrode placement. Each circle contains a letter that corresponds to head location (Frontal, Parietal, Central, Occipital, Temporal) and a number (odd numbers correspond to the left and even to the right hemisphere). The same 6 locations were selected for both datasets (F3, F4, C3, C4, P3, P4).

B. Pre-processing

Several pre-processing steps were implemented, to ensure similar conditions for both datasets and reinforce meaningful comparisons. First, only the data of right handed subjects were selected, as speech is a dominant hemisphere function, and the brain architecture of a subject due to handiness is considered to influence results [27]. Thus, subject 5 was excluded from Coretto, and subjects 9 and 13 from DAIS. Second, trials with artefacts were excluded from both datasets. In Coretto dataset, those trials were marked, so they were excluded explicitly. In

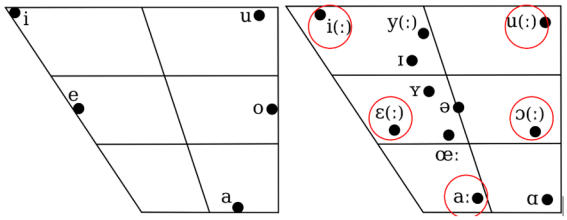


Fig. 2: Vowel quadrilaterals for Spanish (left) & Dutch (right) language. The vertical axis corresponds to mouth openness and the horizontal one corresponds to the position of tongue elevation. The red circles pinpoint the 5 vowels that correspond to those used in DAIS dataset.

DAIS dataset, subjects with marked trials were excluded (2, 7 and 17), similarly to the method followed in the initial article.

After the removal of trials with artefacts, further pre-processing was applied. The same 6 electrodes were chosen from DAIS dataset as in Coretto. Although a larger number of electrodes is considered more beneficial to cover adequately the speech related areas (e.g. Wernicke's and Broca's cortical areas), the low number of electrodes in Coretto introduced a hard upper limit. To this end, importance was given to making meaningful comparisons between the 2 datasets, which could only be possible if recordings were taken from the same brain regions. Tests were also performed with 62 and 16 electrodes only for DAIS dataset, to assess the possibilities of high electrode number classification(see Appendix C), but those results are omitted here. No downsampling was applied, but signals were filtered for both datasets in the range of 2 to 40Hz with 2 filters as in the original Coretto study [19]. Also, a robust scaler was applied per electrode, to ensure adequate normalization. The formula for the robust scaler is given in equation 1. The denominator consists of a subtraction between the 75th and the 25th percentile value of the data.

$$X_{scaled} = \frac{X - X_{median}}{X_{75} - X_{25}} \tag{1}$$

Then, the data was segmented. Windows with a duration of 250 ms (256 samples) were selected, with a 50% overlap, similar to literature . For both Coretto and DAIS, 2 seconds of recording were selected (the initial 2 for Coretto and the whole recording for DAIS). The reason for this was so that signals of the same duration would be chosen for all participants. Different parts of the 2 second windows were further tested, but no major differences were highlighted (see Appendix A), resulting in the choice of the whole 2 second period.

After segmentation, 34342 segments were obtained from Coretto and 21000 from DAIS. Those 2D representations were used as input for the Deep Learning Models. Since traditional Machine Learning (ML) classifiers were to be tested as well, further statistical features were calculated for every 2D segment too, to create 1D input vectors. The mean, standard deviation, median, kurtosis, skewness and third order moment were calculated per electrode per segment, resulting in 1D vectors of length 6x6 = 36. Mel coefficients were also tested, but because of poor results due to low resolution in the frequency domain, this method was not further implemented.

C. Choice of Classifiers

Six classifiers were chosen to be trained on the covert speech recordings. Three of them were traditional ML classifiers, whereas the other three were DL models. All of them are used broadly in literature. This choice was made to provide comparisons between ML and DL classification performance.

1) ML classifiers: A Random Forest (RF), a k Nearest Neighbor (k-NN) and a Guassian Naive Bayes (GNB) classifier were chosen as traditional ML methods. For Random Forest, a maximum depth of 7 was chosen. A number of 7 neighbors was selected for the k-NN one. In this case, the 1D feature vectors were used as input. 75% of data were chosen for training and 25% for testing.

2) DL models: A Shallow Convolutional Neural Network (SCNN), a Deep Convolutional Neural Network (DCNN) and a Long Short Term Memory Recurrent Neural Network (LSTM) were chosen as the DL models. The strategy followed for defining the architectures of the SCNN and DCNN was inspired by previous studies that gave emphasis on layer selection for EEG classification problems [28] [29].

The SCNN consisted of 8 layers. Two Convolutional layers were responsible for spatial and temporal filtering, followed by an Average Pooling layer to reduce dimensionality and filter noise. A batch normalization layer was added, to ensure calibration of the data, and a Leaky Rectified Linear Unit (ReLU) was chosen as the activation function. Then, the 2D output was reorganized to be 1D via a Reshape layer. Finally, a Dense layer of size 5 and an Softmwax activation layer were included, to create the necessary output vector, and limit it between 0 and 1.

The DCNN consisted of 28 layers. Two initial convolutional layers performed spatial and temporal filtering of the data, in the same fashion as the SCNN. After those, max pooling, and batch normalization with Leaky ReLU as the activation function were applied. Then, three similar blocks were added, consisting of a convolutional layer, max pooling, batch normalization, Leaky ReLU and dropout layers. Following that, a fully connected layer was added, with Leaky ReLU as activation and dropout. At the end, a final dense layer of size 5 was introduced, followed by a final dropout layer. Lastly, a softmax activation function was used again, to make the output probabilistic and restrain it between 0 and 1. The filters of the convolutional layers were increasing in number the deeper they were in the network, to ensure that multiple high-level features would be identified within the data. Dropout was used to decrease overfitting, but the dropout rate in all cases was kept low (a value of 0.1), to also avoid underfitting, as data were limited. Regarding the activation functions, Leaky ReLU was chosen for the hidden layers, as it is considered more suitable for such classification problems [8]. However, Softmax was used for the final layer, to provide the necessary output in a probability range. As a note, max pooling was

preferred instead of average pooling in the DCNN case, so that low-level noise can be filtered faster.

The architecture for the LSTM was inspired in a similar way too, by articles that attempted to develop Recurrent Neural Network architectures [28] [30]. The LSTM classifier had 5 layers. First, three consecutive LSTM layers were defined, with sizes of 32, 16 and 8 units respectively. Then, a dense layer of size 5 was added followed by a Softmax activation function. Finally, a Reshape layer was responsible for giving the output vector the necessary size of (1,1,5) to avoid dimensionality coding errors.

The described architectures can be seen in figures 3a and 3b. The 2D segments of size 256x6 were used as input. In this case, both intra-subject classification and 5-fold Nested Cross Validation (NCV) were applied, to also test the feasibility of generalization between subjects. For intra-subject, 70% of data was used as input, 10% as validation and 20% as testing. For 5-fold NCV, 5 folds were defined per dataset. For each fold, the data of 2 subjects were used as validation data, and the data of 3 subjects as test data.

Regarding the hyperparameters, they were selected based on the requirements of each classifier. For the DCNN, the Adam optimizer was selected, with a learning rate of 0.01. categorical crossentropy for chosen for loss, 75 as the number of epochs, and 64 as the batch size. For the SCNN, the Adam optimizer with a learning rate of 0.001 was chosen, categorical crossentropy for loss, 75 as the number of epochs, and 64 as the batch size. Finally, for the LSTM, after a limited grid search, Adam with categorical crossentropy was used as well, with a learning rate of 0.01 and a batch size of 64. However, the number of epochs was reduced to 30, to avoid overfitting.

D. Evaluation Metrics

For comprehension of the results, the average accuracy plus standard deviation, range of accuracy, specificity and sensitivity of the classification objectives were selected as metrics. The confusion matrix of each training experiment was also extracted by interpolating the predicted with the true labels. From each confusion matrix, the specificity and sensitivity were calculated. Specificity is known as the True Negative Rate, in other words the trials that did not belong in a specific class, and were indeed classified as not belonging to that class. Specificity was calculated using equation (2). Sensitivity is referred to as the True Positive Rate, in other words the proportion of the data that were classified correctly per class. Sensitivity was obtained using equation (3). As a note, for the RF classifier, graphs that represented the importance of each feature in classification were also obtained (see Appendix D), but are not included in the coming results section.

$$Specificity = \frac{\text{True Negative}}{\text{True Negative + False Positive}} \quad (2)$$

$$Sensitivity = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad (3)$$

E. Overt versus Covert Trials

To test the potential of covert vowel classification, the accuracy of covert trials was compared to the accuracy of overt trials. Both datasets also contained several EEG recordings of overt trials, which were pre-processed in the same manner. After pre-processing 7840 segments (of 256 samples) were obtained for Coretto and 20902 (of 256 samples) for DAIS. Since overt speech is associated with the direct excitation of muscles and the effect is clearly distinguished by human ears, an assumption was made, that between the two cases, overt speech would result in higher accuracy, and the difference between the two would give an estimate of an upper limit for the covert case. For this experiment, the so far best performing ML method and the best performing DL method were chosen. Five-class classification was performed for each subject separately for the overt and covert case for both datasets. Then, the average for each dataset was calculated for both speech scenarios.

F. Binary classification of covert vowels

To test if certain vowels can be distinguished from one another easier from the EEG activity that produces them, intra-subject 2-class classification was also implemented for all the 10 possible pairs (/a/-/e/, /a/-/i/, /a/-/o/, /a/-/u/, /e/-/i/, /e/-/o/, /e/-/u/, /i/-/o/, /i/-/u/, /o/-/u/) for both datasets. In this case, the DCNN model was chosen as the classification method, under the assumption that DCNNs usually perform better in literature for this task [8], [15], [33]. The architecture was updated accordingly, to include a Dense layer of size 2 instead of 5 at the final stage of the classification. No further architectural changes were implemented during this step.

G. Additional Experiment

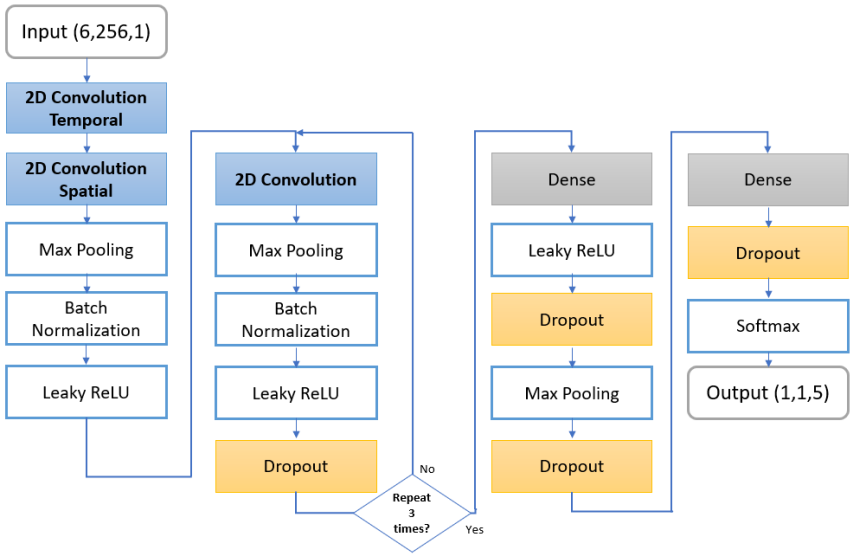
As a note, an additional experiment took place, testing the feasibility of subject identification through EEGs. The results are not discussed here, but the confusion matrices and model training history curves can be seen in Appendix B.

III. Results

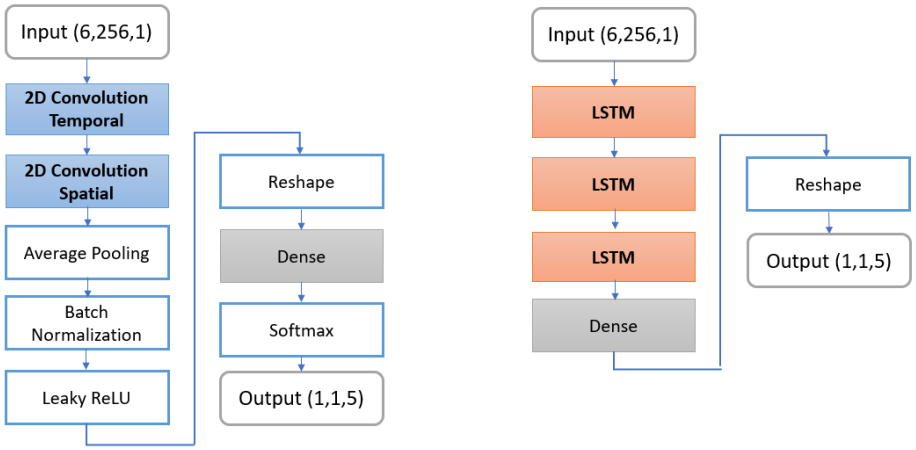
A. Classifier performance

The results of the 6 classifiers for the 5-class classification were compared with those of literature. Figure 4 shows several reported average accuracies from other articles as well as results of this study for the Coretto dataset by order of increasing magnitude. A comparison between the average performance of the 6 classifiers for both datasets can be seen in figure 5. The highest reported average accuracy for the ML methods was encountered for the RF classifier (24.3% for Coretto and 24.5% for DAIS). The highest reported average accuracy for the DL methods was observed for the DCNN classifier (36.0% for Coretto and 39.0% for DAIS). The remaining results can be seen in Table 1.

Confusion matrices as well as model accuracy history curves of the best performing method (the DCNN) are provided for 6 subjects of the Coretto and 6 subjects of the DAIS dataset. Those can be seen in figures 6 - 9. Some confusion matrices



(a)



(b)

Fig. 3: (a) The architecture of the DCNN. It consisted of 28 layers. Each type of layer is represented with a different color scheme. (b) The architecture of the SCNN (left). It consisted of 8 layers. The architecture of the LSTM (right). It consisted of 5 layers. Each type of layer is represented with a different color scheme.

for the DL methods can be seen in Appendix E, but are not discussed in this part.

As a note, confusion matrices depict the percentage of trials that were classified in every label, with the vertical axis being the true label and the horizontal one being the predicted label. In the ideal case, the diagonal of the matrix would contain percentages of 100%, meaning that all the predicted labels would be indeed the right ones. The rest of the spaces would be 0%. The subjects included span a range from fairly accurate to less successful predictions. The model accuracy history curves, on the other hand, show the accuracy of the model for the training and validation data for every single epoch of the training procedure. Typically, a positive slope is observed in the beginning stages of training, reaching a plateau towards its end.

In addition, the confusion matrix and model accuracy training curve is provided for one fold of the Coretto dataset (figures 10, 11). All the folds scored close to chance level during nested cross-validation (from 19.4% to 21.4%) and the resulting confusion matrices and model accuracy history

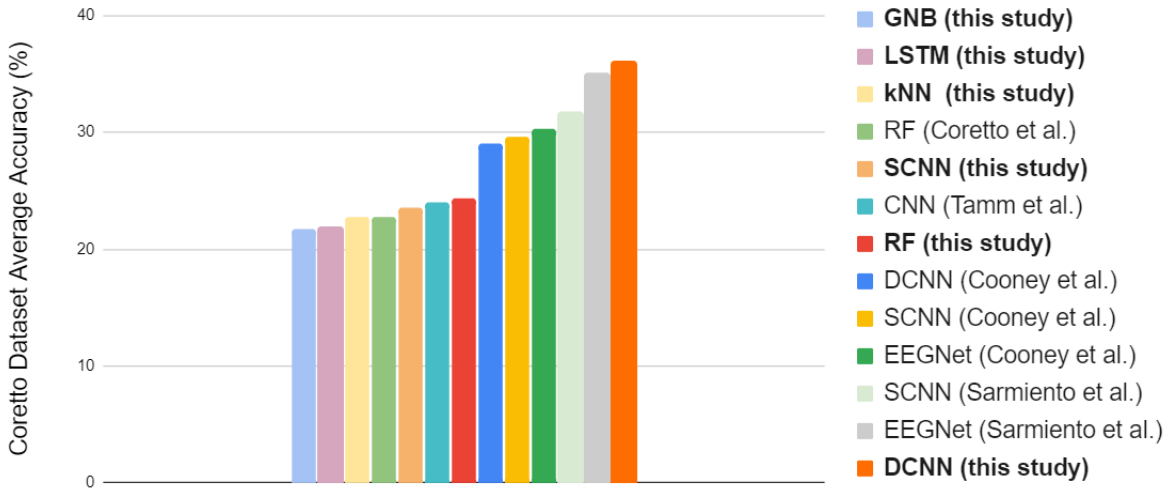


Fig. 4: Reported average performance of different classifiers from literature, including this study, by order of increasing accuracy (chance level 20%) [8], [15], [19], [29], [33].

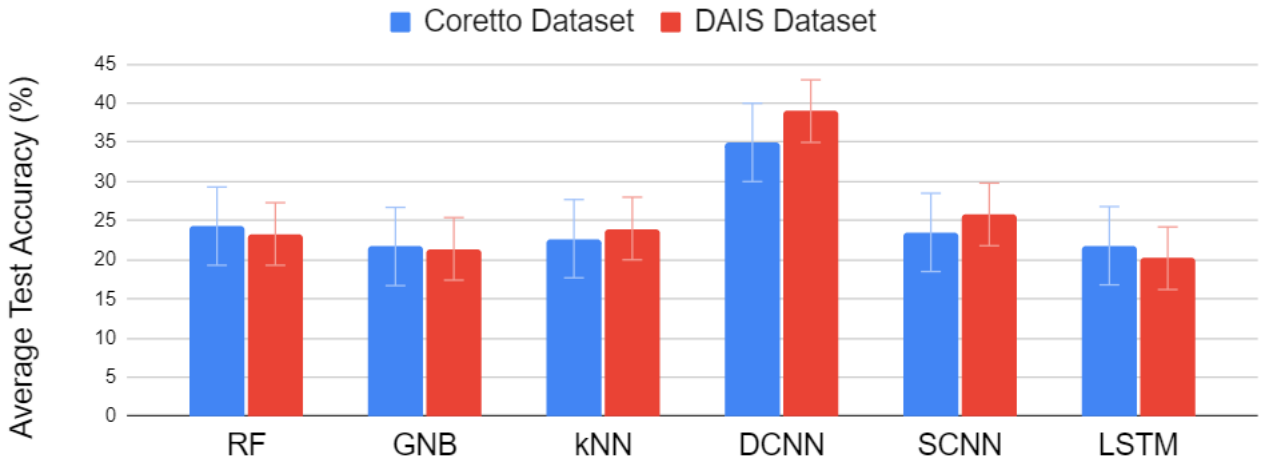


Fig. 5: Average intra-subject classification accuracy for the 6 classifiers (confidence level 95%, chance level 20%).

curves were almost identical. The specificity and sensitivity values were also identical to the average accuracy per subject.

B. Overt versus covert trials accuracy

The average accuracy for the two datasets in the case of overt versus covert speech can be seen in figures 12 and 13. The DCNN scored an average accuracy of 40.5% for DAIS and an average accuracy of 37.8% for Coretto, whereas the RF classifier scored an average of 25.3% for DAIS and an average of 29.4% for the two datasets respectively. In all cases, the overt speech training scored slightly higher than the covert equivalent. The best recorded accuracy during this experiment was 50.0%, and it was achieved from subject 3 of DAIS. Despite the minor fluctuations in accuracy achieved by different subjects, both datasets reached similar scores, as can be seen by the range of results in table 2. In addition, the values of the standard deviation were lower in the case of the RF and higher in the case of the DCNN for the same data.

C. Binary classification results

The ten binary combinations of vowels were classified per subject and per dataset and the results can be seen in figures 14 and 15. For DAIS, the lowest pairs were /a/ versus /i/ and /a/ versus /e/ with accuracies of 57.9% and 58.2% respectively. The rest of the pairs followed a slightly increasing trend starting from /i/ versus /o/ and ending with /i/ versus /u/. Finally, the best performing pair was /a/ versus /u/, with an average accuracy of 64.4%. Standard deviation ranged from 5.8% (/a/ - /u/) to 8.0% (/a/ - /i/).

Regarding the behavior of Coretto dataset, the combinations /o/ versus /u/ and /e/ versus /o/ performed the worst, scoring 60.4% and 60.0% respectively. A similar increasing trend was observed for the rest of the pairs. Finally, the best performing

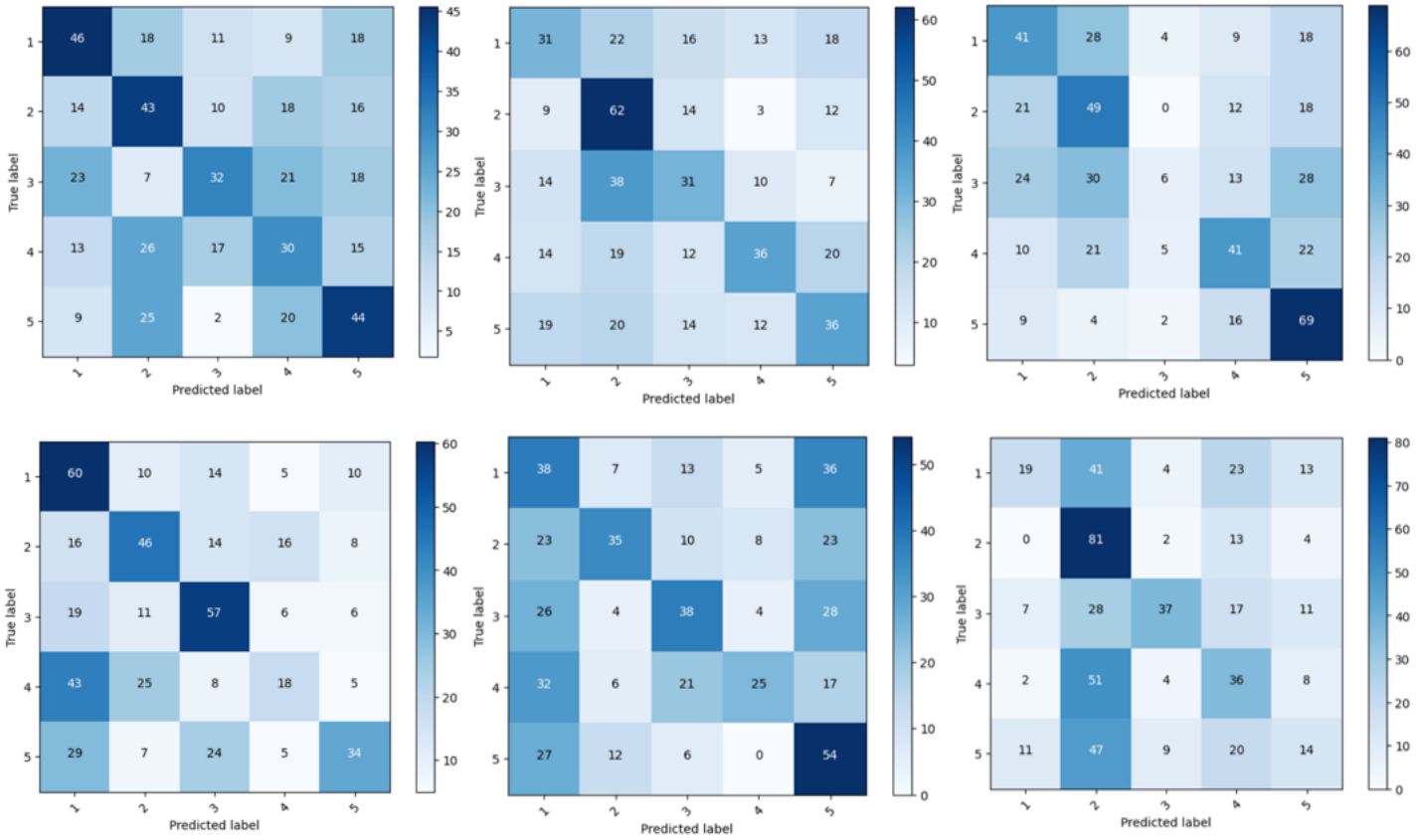


Fig. 6: DAIS Subject normalized confusion matrices for DCNN. Upper left: subject 15, upper middle: subject 8, upper right: subject 5, bottom left: subject 12, bottom middle: subject 14, bottom right: subject 19. Numbers correspond to the 5 vowels: 1 - /a/, 2 - /e/, 3 - /i/, 4 - /o/, 5 - /u/.

TABLE I: Comparison of mean, std and range for all classifiers for both datasets.

Classifier	Dataset	Metrics (%)			
		mean	std	min	max
RF	Coretto	24.3	1.9	21.3	28.6
	DAIS	24.5	2.5	18.7	27.5
GNB	Coretto	21.7	1.7	18.9	24.5
	DAIS	21.4	1.6	19.1	24.5
kNN	Coretto	22.7	1.9	20.0	26.7
	DAIS	24.0	1.8	21.5	27.8
DCNN	Coretto	36.0	5.5	21.3	40.0
	DAIS	39.0	2.6	34.1	43.3
SCNN	Coretto	23.5	2.3	20.2	28.5
	DAIS	25.8	2.6	20.0	30.4
LSTM	Coretto	21.8	1.9	20.1	25.6
	DAIS	20.2	2.7	16.7	25.7

pair was /a/ - /u/, with an average accuracy of 64.8%. Standard deviation ranged from 3.2% (/e/ - /u/) to 6.0% (/a/ - /o/).

Overall, the standard deviation values were lower for Coretto dataset than DAIS. Also, intra-subject accuracies were between 44.7% for subject 1 of DAIS dataset (/o/ - /u/) and 76.3% for subject 15 of DAIS dataset (again /o/ - /u/). Finally, the sensitivity and specificity values were almost identical as the average classification accuracy per subject per vowel pair (with a maximum difference of 5%). The average accuracies and standard deviations in arithmetical form can be seen in Table 3.

TABLE II: Comparison of mean, std and range for the overt versus covert case.

Classifier	Dataset	Speech type	Metrics (%)			
			mean	std	min	max
RF	Coretto	Overt	29.4	4.1	22.3	35.3
		Covert	23.3	2.5	18.7	27.5
	DAIS	Overt	25.3	1.4	23.2	29.6
		Covert	24.3	1.9	21.3	28.6
DCNN	Coretto	Overt	37.8	4.6	26.6	45.9
		Covert	36.0	5.5	21.3	40.0
	DAIS	Overt	40.5	6.4	28.5	50.0
		Covert	39.0	2.6	34.1	43.3

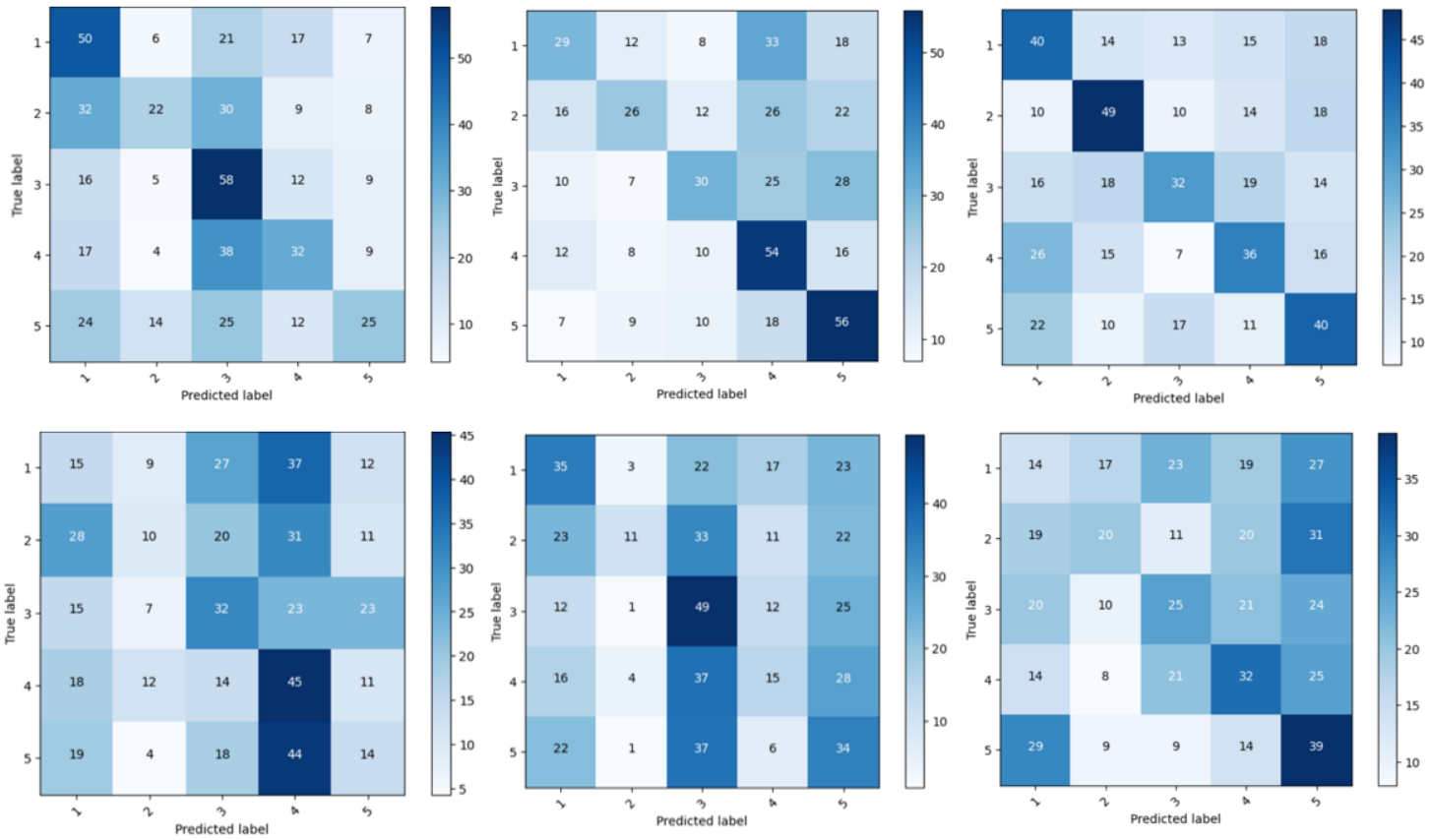


Fig. 7: Coretto Subject normalized confusion matrices for DCNN. Upper left: subject 1, upper middle: subject 2, upper right: subject 8, bottom left: subject 6, bottom middle: subject 7, bottom right: subject 12. Numbers correspond to the 5 vowels: 1 - /a/, 2 - /e/, 3 - /i/, 4 - /o/, 5 - /u/.

IV. Discussion

A. 5-class Classification

None of the classifiers for covert vowels achieved average accuracy levels higher than 40%. This suggests that, despite the different methods used, the SNR of the data was significantly low. Among the traditional ML methods, the RF classifier performed slightly better overall (about 1% difference with kNN), but still limited below 25%, given a chance level of 20%. On the other hand, between the DL models, the Convolutional Neural Networks managed to surpass the 25% limit. More specifically, the SCNN achieved an average accuracy of 25.8% for the DAIS dataset, and the DCNN, which outperformed all the other methods by far, scored 36.0% for Coretto and 39.0% for DAIS respectively.

The CNNs, in general, focus on identifying spatial and temporal features in their initial layers, whereas in the primary focus of the other classifiers is the temporal flow of information. Given the level of accuracy that was achieved, this might suggest that, to solve the problem, a spatial component is also necessary. Also, the hidden layers of the DCNN were responsible for the acquisition of high level features, that might be able to detect small important discrepancies in the data.

The possibility of overfitting is also necessary to be assessed. However, many dropout layers were used, especially towards the final layers of the model, where the dense connections easily result to the emergence of this issue. The percentage of dropout was chosen to be kept at 0.1, given the small size of the datasets. A larger ratio could result in less successful training conditions. Secondly, by looking at figures 8 and 9, it is visible that the training accuracy reaches a plateau towards the very end of the training for most subjects - but not earlier - while the validation accuracy keeps increasing at a low but steady rate. This suggests that the model is found at the early stages of overfitting, but not deep enough to be considered biased. The difference between training and validation accuracy is substantially large, which is evidence pinpointing to the data having a low SNR. As a note, accuracy history curves are almost never displayed in articles - according to the author's extent of knowledge - so comparisons about overfitting are difficult to be made between articles.

The model accuracy history curves highlight another fact. While the training accuracy follows a pretty smooth trend in most cases, the same is not true for validation accuracy. Instead, rough fluctuations can be observed, which suggest

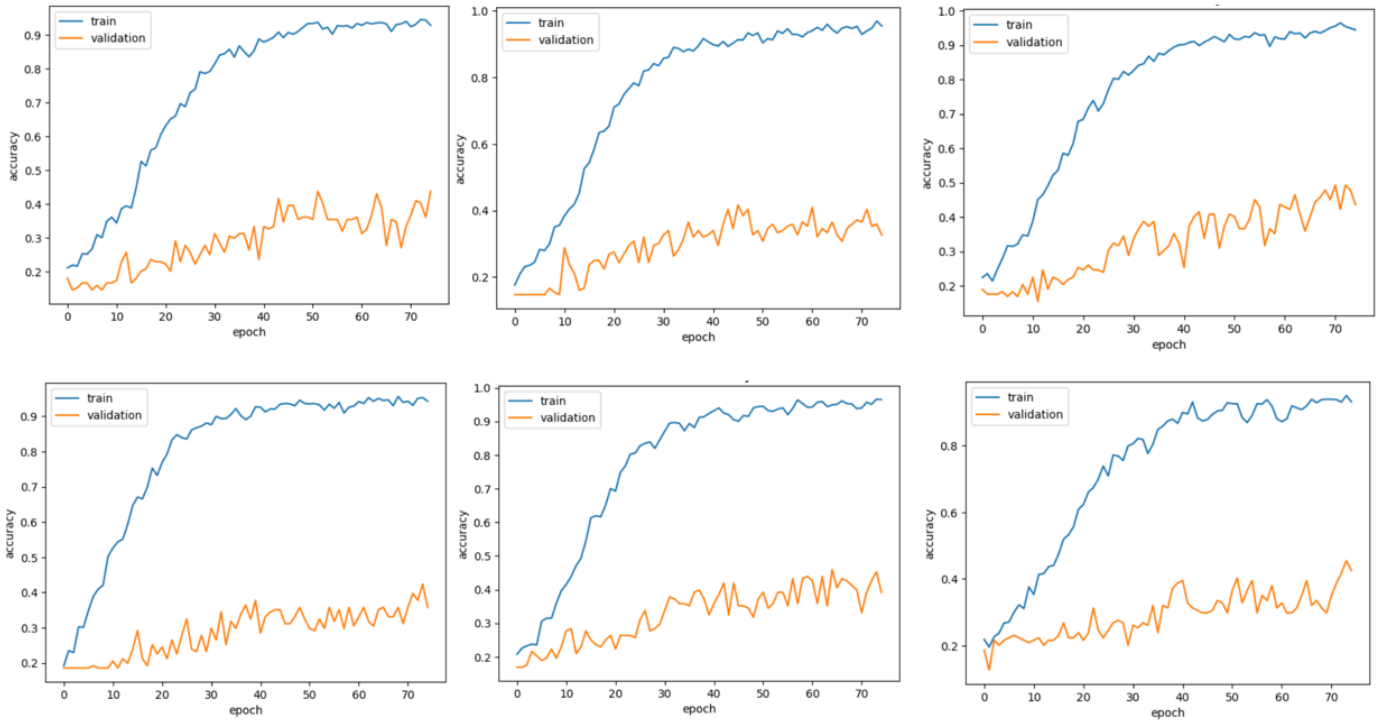


Fig. 8: DAIS Subject model accuracy history curves for DCNN. Upper left: subject 15, upper middle: subject 8, upper right: subject 5, bottom left: subject 12, bottom middle: subject 14, bottom right: subject 19.

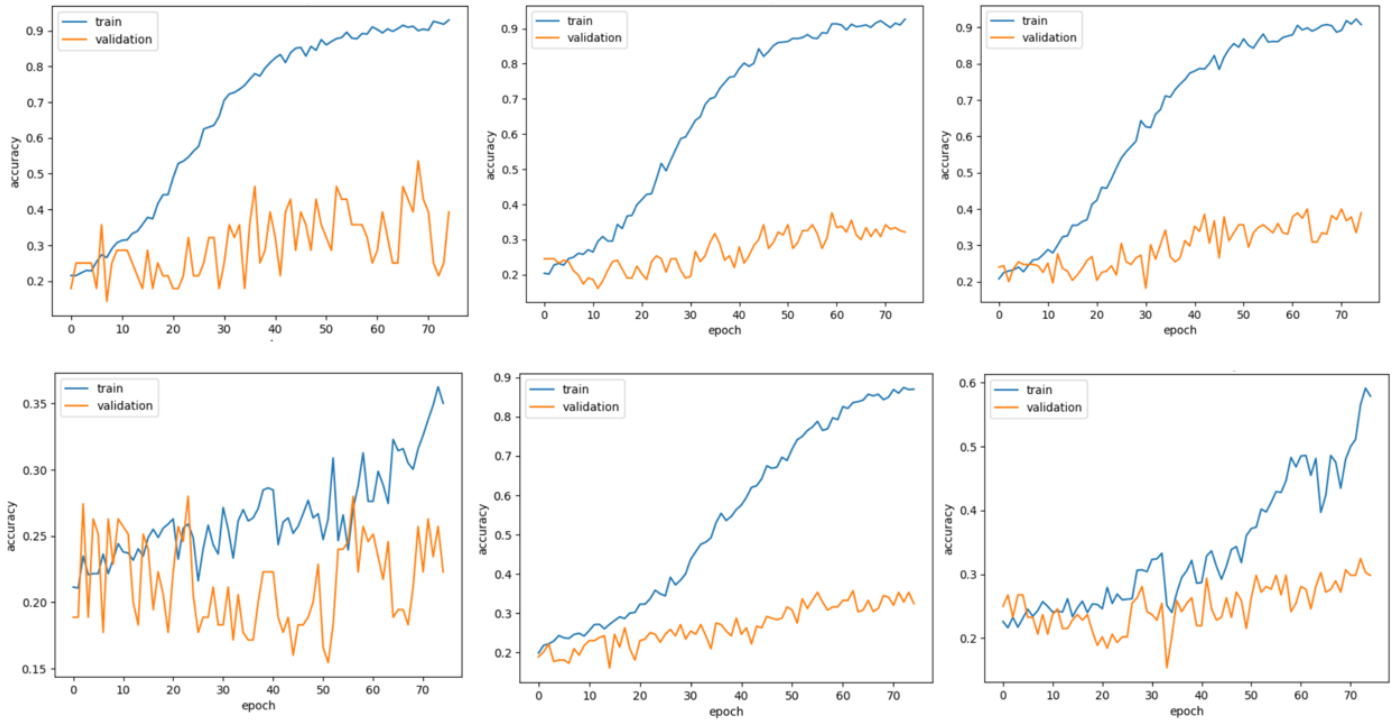


Fig. 9: Coretto Subject model accuracy history curves for DCNN. Upper left: subject 1, upper middle: subject 2, upper right: subject 8, bottom left: subject 6, bottom middle: subject 7, bottom right: subject 12.

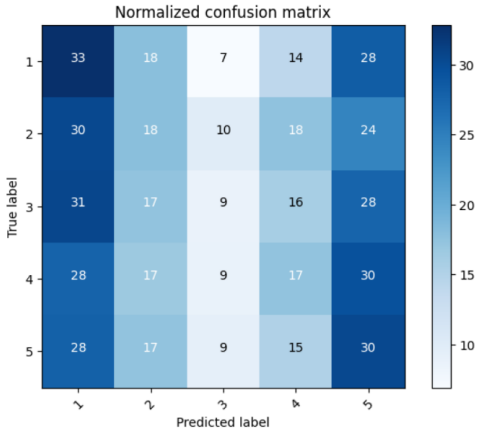


Fig. 10: Coretto Dataset, Fold 2, normalized confusion matrix. Numbers correspond to the 5 vowels: 1 - /a/, 2 - /e/, 3 - /i/, 4 - /o/, 5 - /u/.

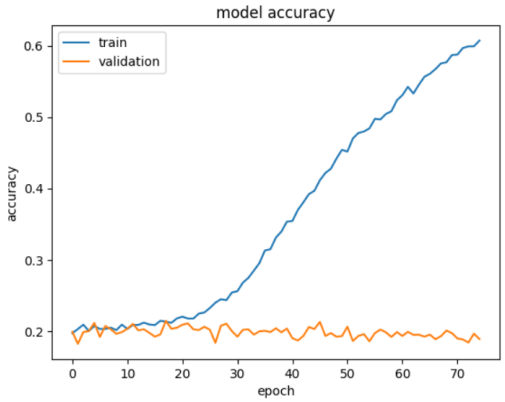


Fig. 11: Model accuracy history curve for Fold 2 of Coretto dataset.

that chunks of trials are classified randomly every time, despite the average accuracy increasing as epochs go by. Since the classification was intra-subject, meaning that no data from multiple subjects was used for training, those randomly classified segments either represent the cases where SNR is really low, or contain data that is not relevant to the classification problem. This introduces the question: which exact parts of an EEG recording contain important information for classification? An additional experiment took place, where the 2 second windows for each dataset were separated in 4 500ms subwindows and training occurred for each category (see Appendix A).

The confusion matrices span a range of types. For both datasets, several subjects produce the diagonal form of the matrix. However, most of the times, one or two vowels are classified the best, with the rest hardly rising above chance level. In addition, a specific behavior emerges, where really high values are located in a single column of the matrix (e.g. subject 14 of DAIS in figure 6). This shows that the network becomes biased in favor of the vowel that is classified the best,

TABLE III: Comparison of mean, std and range for all 2-class classification vowel pairs for both datasets. The DCNN was used as the classification method.

Vowel pair	Dataset	Metrics (%)			
		mean	std	min	max
/a/-/e/	Coretto	62.8	4.3	47.0	69.7
	DAIS	58.2	7.1	56.2	71.4
/a/-/i/	Coretto	62.4	5.3	49.8	71.2
	DAIS	57.9	8.0	45.2	72.8
/a/-/o/	Coretto	63.0	6.0	47.7	69.2
	DAIS	61.9	6.6	49.9	72.0
/a/-/u/	Coretto	64.8	3.8	55.1	70.5
	DAIS	64.4	5.8	48.6	73.0
/e/-/i/	Coretto	63.1	3.5	56.9	70.0
	DAIS	61.5	5.9	50.0	74.0
/e/-/o/	Coretto	60.0	4.8	51.8	69.0
	DAIS	62.3	6.4	50.5	70.6
/e/-/u/	Coretto	62.0	3.2	55.3	69.2
	DAIS	62.2	6.3	55.1	74.8
/i/-/o/	Coretto	61.1	6.3	48.5	73.5
	DAIS	60.0	6.6	47.4	70.2
/i/-/u/	Coretto	62.5	7.1	57.8	70.2
	DAIS	62.8	3.4	46.2	72.1
/o/-/u/	Coretto	61.1	7.6	55.5	66,7
	DAIS	60.4	3.5	44.7	76.3

showing an inclination to label data from different prompts as that one. Also, there are subjects whose confusion matrices follow a random pattern (e.g. subject 6 of Coretto in figure 7). This suggests that several trials from different labels look similar. For example, in the case of subject 6, /o/ and /u/ seem to be confused with one another. Finally, in the case of nested cross-validation, almost all elements of the matrices are close to chance levels, with the column behavior emerging again (figure 10). In this case, no pattern can be learned, as variability between subjects is high. This is also evident from the model accuracy history curve (figure 11), where validation accuracy never rises above the chance level throughout the training session (even though the training accuracy increases).

Overall, the average accuracy results were consistent between the two datasets, with only minor differences, especially visible in the SCNN and DCNN experiments. Since only 6 electrodes were used for the experiments, two observations can be made. First, given the same number of electrodes (and the same prompts - 5 vowels) both datasets seem to perform similarly. Second, even with a small number of electrodes, accuracy levels up to 40% could be obtained. Those numbers are only slightly lower than other studies that used more electrodes, that were even targeting the cortical areas responsible for speech processing (e.g. [33]). In this case, the electrodes were distributed evenly, not focusing on spatial resolution, but rather covering as much volume as possible.

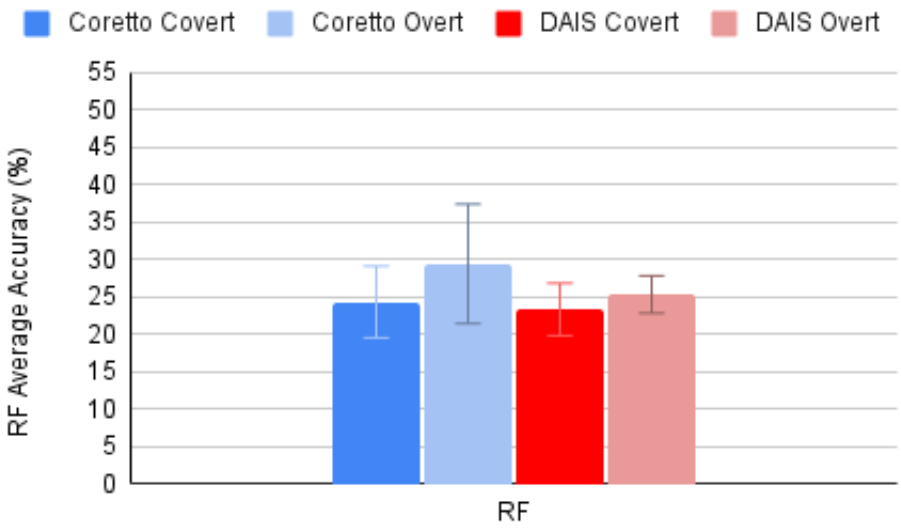


Fig. 12: Average intra-subject classification accuracy for the RF classifier for covert & overt trials, for both datasets (confidence level 95%, chance level 20%).

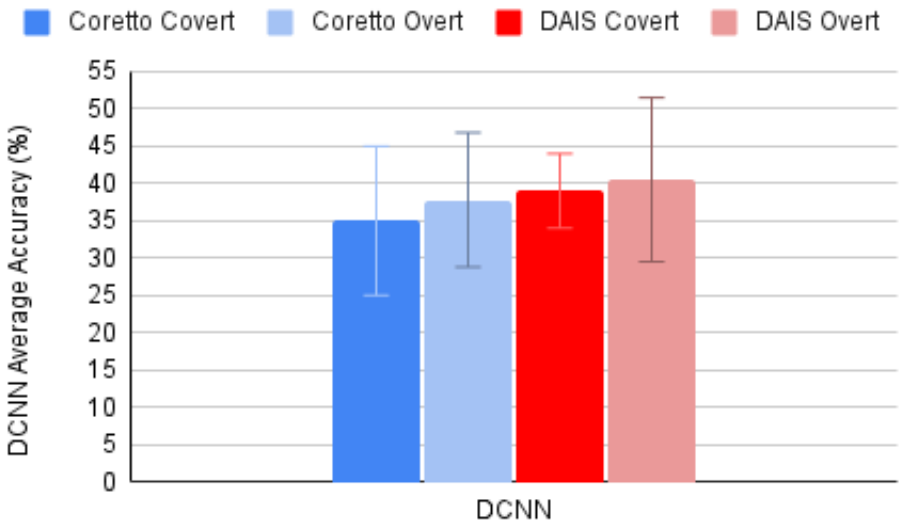


Fig. 13: Average intra-subject classification accuracy for the DCNN classifier for covert & overt trials, for both datasets (confidence level 95%, chance level 20%).

B. Overt vs. Covert vowel classification

Classification of overt trials resulted in higher accuracy than that of the covert ones. However, this increase was small, suggesting that, even in the case of loud speech, where it is clear which vowel the subject is pronouncing at a given point in time, decoding speech from the EEGs is challenging. Possible reasons for this might be the low SNR of the data, the small number of electrodes used, or the fact that all vowels require the activation of similar neural paths, the location of which lies really close together in the cortical brain areas, making differences hard to distinguish. Furthermore, the fact that a subject is able to hear their own voice during articulation, might create a feedback loop, responsible for sound correction through the whole duration of speech [34]. The same condition is not satisfied when subjects perform covert speech.

In addition, the gap between the two classifiers (RF and DCNN) remained large, even in the case of overt trials. The advantage of the DCNN for calculation of spatial features was evident. Overall, RF results showed smaller variance than DCNN. Also, in the case of the RF classifier, Coretto dataset achieved higher results than DAIS. The opposite is true, however, for the DCNN model, where DAIS scored better than

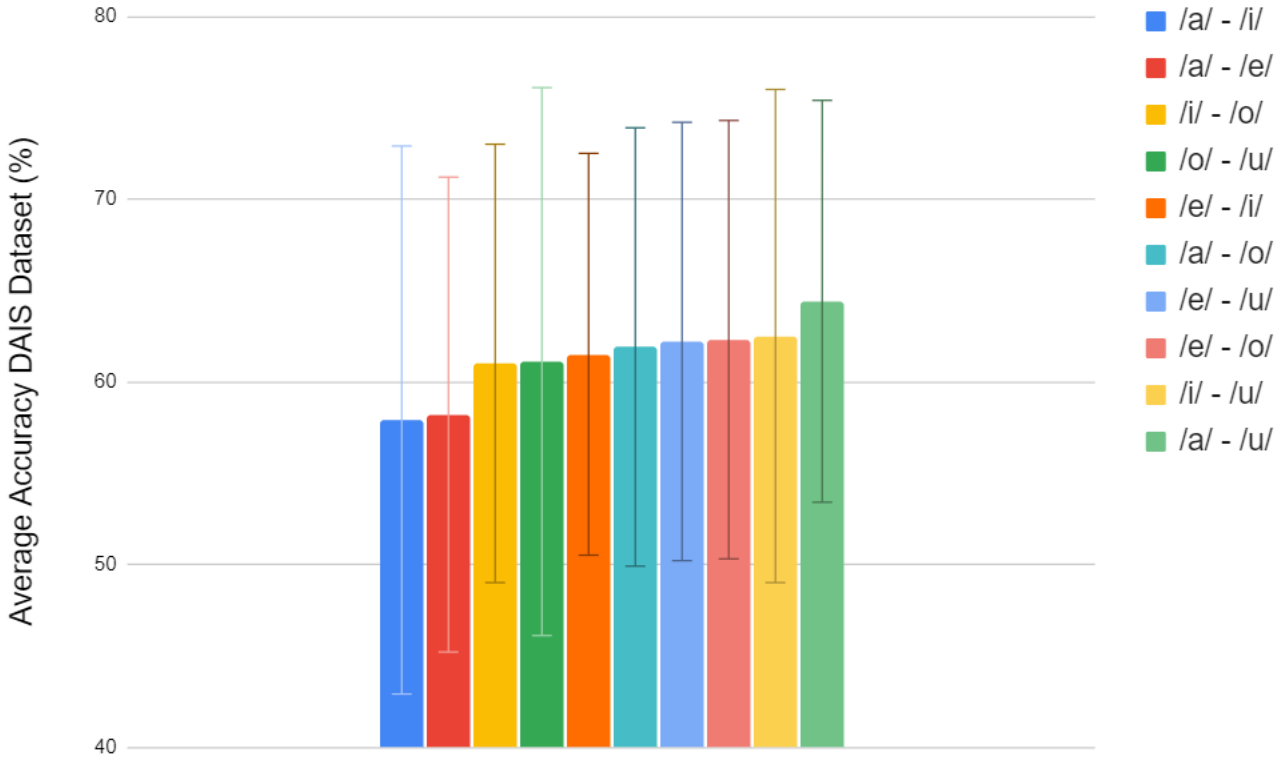


Fig. 14: Average intra-subject classification accuracy for the DCNN classifier for covert & overt trials, for the DAIS dataset (confidence level 95%, chance level 50%).

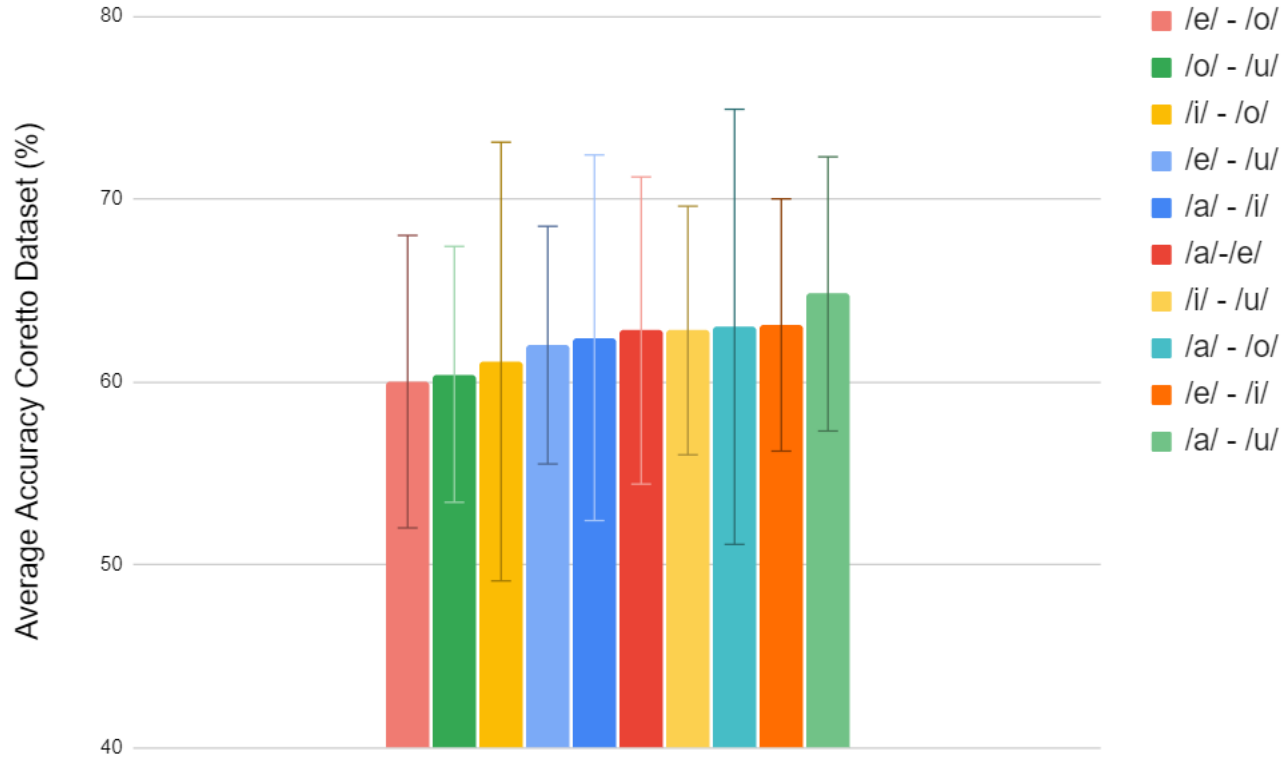


Fig. 15: Average intra-subject classification accuracy for the DCNN classifier for covert & overt trials, for the Coretto dataset (confidence level 95%, chance level 50%).

Coretto. Additionally, the variance of results between subjects was overall higher when using the DCNN. Even in the case of overt speech, however, average results did not exceed 45% with only a single subject of DAIS dataset crossing the 50% mark. Again, the two datasets behaved in a similar fashion.

C. 2-class Classification

Regarding the classification of vowel pairs, results were significantly higher than in the case of 5-class classification. However, this is mainly because of the smaller number of classes, as the chance level now increased to 50%. Coretto and DAIS datasets performed similarly again. Almost all the average accuracies fluctuated around 60%, with significant variances. Again, the variance of Coretto dataset is smaller than that of DAIS. Second, several vowel-pairs are observed to have a similar place in the order of increasing accuracy. The classification between /a/ and /u/, for example, seems to be performing the best in both cases. On the other hand, /o/ versus /u/ and /i/ versus /o/ seem to be performing poorly in both datasets.

These observations might be explained, if the vowel quadrilateral for each language is taken into consideration. Vowels that tend to be placed further away in the quadrilateral are also far away in terms of the position of the articulators (e.g. how open the mouth is). This could explain why /a/ is classified more accurately when it is grouped with /u/, as the two vowels are found the far away in the quadrilateral. This observation is linked to literature, as another broadly used and openly shared dataset - the DaSalla dataset - contains recordings of /a/,/u/ and rest [20]. Comments have been made before regarding the higher degree of variability between /a/ and /u/ [20].

If this observation is valid for one vowel combination, all that remains is to be extended for all vowel combinations. However, no significant evidence exists or stems from this study, regarding the relation of EEGs from the vowel pairs to anatomy and physiology. As the average values for each pair are really close together, and the variances really large, this observation remains an assumption.

Overall, the accuracy of the DCNN in this experiment, is again comparable to literature. For example, in [18] an average accuracy of 50% was achieved for an SVM-R, and an average accuracy of 65% was scored for an Extreme Learning Machine (ELM). In [13], all methods used scored average accuracy between 50% and 70%. In those articles, a vowel versus the resting state results in higher accuracy than the training between two different vowels. This often results in scores around 90%, but was not implemented in this study.

D. Comparison to literature

Results obtained from this study lead to several important points regarding previous published articles. First, as seen from figure 5, the accuracy levels of all classifiers are on par with other studies. Traditional ML solutions produced results really close to chance levels. This showcases that the statistical features chosen for this study show high spatiotemporal variability between subjects, as well as for a single subject between segments. The DL models, on the other hand, achieve a performance of around 30% for 5-class classification, which can be considered statistically significant. Among the studies that have made use of neural networks for the Coretto Dataset, several cases have been tested. For example, the best choice of hyperparameters [8], the possibility of transfer learning [15], or the use of a powerful network made for image recognition called EEGNet [33].

The last one of those studies also introduced a binary pair CNN, an architecture that classifies vowels based on the majority outcome after a series of binary classifications, in order to extract the most possible label, scoring around 60%. This is the only reported result for Coretto dataset with such a high accuracy. However, there is ambiguity regarding the chance level in this case, as the network only needs to choose between 2 vowels every time, and in some cases it is forced to even choose between 2 wrong answers (e.g. a true label of /a/ needs to be classified as /e/ or /i/). In the same study, when a second dataset was introduced, results were significantly higher, ranging from 50% to 90%. This brings about a point regarding the efficiency of a dataset, as the reasons why such differences were observed between the two are unknown.

Leaving behind the Coretto dataset, other studies have performed multiclass covert vowel classification, making sole use of their own recorded datasets. In those cases, all the subjects were chosen to be right-handed. In [30], Japanese vowels were classified, and accuracies above 60% were reported. In [13], Japanese vowels were classified again, this time with adaptive collection, bringing about similar results. However, the amount of test trials per subject was only 2. In [11], classification on English and Bengali vowels was performed, using a stacked autoencoder, again achieving similar results. There were 5 trials per subject. In this case, however, inter-subject training is implied, which is notorious for results close to chance level. In addition, the features used as input were kurtosis, band power, entropy, peak to peak amplitude, Hjorth parameters and relative wavelet energy. These statistical features have been used in other studies as well, but due to the EEGs being non-stationary signals, results are often close to chance accuracy. However in that study, the reported accuracy was around 70%. It is unknown, whether the native language of the subjects, or the differences between the methods used in the aforementioned studies are responsible for the differences in results.

E. Future work

Future work could entail the extension of this protocol to more covert vowel datasets, so that an assessment can be made regarding the generalization of the results. Also, next steps of development should focus on pinpointing the most important characteristics of the EEGs, or to try and increase the signal to noise ratio through filtering. Additionally, different placements and numbers of electrodes should be tested, to assess if locations next to the cortical areas close to speech result consistently in higher accuracy, or if just a bigger volume coverage can be as successful. After all, due to limitations,

the number of electrodes in this study was small and their location was not focused around those areas, and yet results were comparable to literature.

When it comes to the EEGs themselves, the useful part of the whole recording in time is crucial to be identified. There is no substantial evidence, except some arguments when the event related potentials begin, of which part is associated with covert speech, and which is just noise. Additionally, experiments need to be made, to comprehend why multi subject classification performs close to chance level. Even though a similar brain structure is defined in every human, there is still uncertainty regarding the extent of that similarity. Therefore, a higher level of understanding of the brain function is necessary to improve this field.

Since different pairs of vowels seem to be classified with slightly different levels of accuracy, the effect of different dictionaries, or prompts, that will be used in a specific BCI is important to be studied. To do that, however, most other parameters, like identity of subjects, language, or electrode number, must be kept constant, to decrease uncertainty. Datasets have been compared before, but due to constant parameters, the outcome is always ambiguous (e.g. [Lee] 5-class versus 6-class classification).

Future work should also entail the extension of datasets to contain information from patients, since the purpose of the BCIs is to assist those, and not healthy subjects. Then, similarities and differences between the two groups can also be researched.

Finally, although a systematic comparison between 2 covert vowel datasets is useful, it needs to be extended to more, so that the same hypotheses can be tested repeatedly. Only then the most efficient methods for covert speech classification will be determined.

V. Conclusion

In this study, a systematic comparison between two openly shared covert vowel datasets was performed. The methods used suggested similarities through several experiments, from 5-class and 2-class classification to the comparison of accuracy between overt and covert trials. The DCNN model was observed to be the best performing method for covert speech decoding, compared to other DL and ML classifiers.

All in all, EEG covert speech classification is a method that could provide a solution to the communication problems of people with neuromuscular disorders. However, current methods and available data only go that far. Deep learning models - and especially DCNNs - are a step towards the right direction, and their use in the future will be crucial in tackling this complicated objective. Results, as demonstrated by both the 5-class as well as the 2-class classification, show that statistically significant and above chance level accuracy is possible. Also, several observations indicate that the obtained data are not just noise, but rather meaningful information affected by noise. If the SNR of the data is also improved in the future, EEG classification could reach its full potential.

References

[1] M. Das, J. Anosike, K. Asuncion. Locked-in Syndrome. [Updated 2022 Dec 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK559026/.

[2] P. Francesca, C. Antonio, S.Simona, S. Marco. (2017). Commentary: Embodied Medicine: Mens Sana in Corpore Virtuale Sano. Frontiers in Human Neuroscience. 11. 10.3389/fnhum.2017.00381.

[3] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, F. H. Guenther. (2010). Brain-Computer Interfaces for Speech Communication. Speech communication, 52(4), 367–379. https://doi.org/10.1016/j.specom.2010.01.001

[4] X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T. P. Jung, T. Lin. (2021). EEG-Based Brain-Computer Interfaces (BCIs): A Survey of Recent Studies on Signal Sensing Technologies and Computational Intelligence Approaches and Their Applications. IEEE/ACM transactions on computational biology and bioinformatics, 18(5), 1645–1666. https://doi.org/10.1109/TCBB.2021.3052811

[5] E. Gibson, N. J. Lobaugh, S. Joordens, A. R. McIntosh. (2022). EEG variability: Task-driven or subject-driven signal of interest?, NeuroImage, Volume 252, 119034, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2022.119034.

[6] E. Combrisson, K. Jerbi. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. Journal of neuroscience methods, 250, 126–136. https://doi.org/10.1016/j.jneumeth.2015.01.010.

[7] J. J. Bird, D. R. Faria, L. J. Manso, P. P. S. Ayrosa, A. Ekárt. (2021). A study on CNN image classification of EEG signals represented in 2D and 3D. Journal of neural engineering, 18(2), 10.1088/1741-2552/abda0c. https://doi.org/10.1088/1741-2552/abda0c.

[8] C. Cooney, A. Korik, R. Folli, D. Coyle. (2020). Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG. Sensors; 20(16):4629. https://doi.org/10.3390/s20164629.

[9] J. T. Panachakel, A. G. Ramakrishnan. (2021). Decoding Covert Speech From EEG-A Comprehensive Review. Frontiers in neuroscience, 15, 642251. https://doi.org/10.3389/fnins.2021.642251

[10] S. H. Lee, M. Lee and S. W. Lee. 2020. "Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 12, pp. 2647-2659, doi: 10.1109/TNSRE.2020.3040289.

[11] R. Ghosh, N. Sinha, S. Phadikar. (2022). Classification of Silent Speech in English and Bengali Languages Using Stacked Autoencoder. SN COMPUT. SCI. 3, 389. https://doi.org/10.1007/s42979-022-01274-y.

[12] M. Matsumoto, J. Hori. (2013). Classification of silent speech using adaptive collection. 2013 IEEE Symposium on Computational Intelligence in Rehabilitation and Assistive Technologies (CIRAT), 5-12.

[13] T. Morooka, K. Ishizuka, N. Kobayashi. (2018). Electroencephalographic Analysis of Auditory Imagination to Realize Silent Speech BCI. 683-686. 10.1109/GCCE.2018.8574677.

[14] D. Sikdar, R. Roy, M. Mahadevappa. (2017). Multifractal Analysis of Electroencephalogram for Human Speech Modalities. 10.1109/NER.2017.8008432.

[15] C. Cooney, R. Folli, D. Coyle. (2019). Optimizing Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG. 1311-1316. 10.1109/SMC.2019.8914246.

[16] R. A. Sree, A. Kavitha. (2017). Vowel classification from imagined speech using sub-band EEG frequencies and deep belief networks. 1-4. 10.1109/ICSCN.2017.8085710.

[17] S. Chengaiyan, A. S. Retnapandian, K. Anandan. (2019). Identification of vowels in consonant-vowel-consonant words from speech imagery based EEG signals. Cogn Neurodyn. 2020 Feb;14(1):1-19. doi: 10.1007/s11571-019-09558-5. PMID: 32015764; PMCID: PMC6974026.

[18] B. Min, J. Kim, H. Park, B. Lee. (2016). Vowel Imagery Decoding toward Silent Speech BCI Using Extreme Learning Machine with Electroencephalogram. Biomed Res Int. 2016;2016:2618265. doi: 10.1155/2016/2618265. PMID: 28097128; PMCID: PMC5206788.

[19] G. Coretto, I. Gareis, H. Rufiner. (2017). Open access database of EEG signals recorded during imagined speech. 1016002. 10.1117/12.2255697.

[20] C. S. DaSalla, H. Kambara, M. Sato, Y. Koike, (2009). Single-trial classification of vowel speech imagery using common spatial patterns, Neural Networks, Volume 22, Issue 9, Pages 1334-1339, ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2009.05.008.

[21] C. H. Nguyen, G. K. Karavas, P. Artemiadis. (2018). Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. Journal of neural engineering, 15(1), 016002. https://doi.org/10.1088/1741-2552/aa8235

[22] B. Dekker, A. C. Schouten and O. Scharenborg. (2023). "DAIS: The Delft Database of EEG Recordings of Dutch Articulated and Imagined Speech," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096145.

[23] C. Gussenhoven. (1992). The Dutch vowel system. Language, 68(3), 525-563.

[24] G. Booij. (1995). The Phonology of Dutch. New York: Oxford University Press.

[25] C. Gussenhoven, V. J. van Heuven. (2002). Phonetic similarity between vowels: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, 111(5), 2845-2854

[26] J. J. Bolger, V. J. van Heuven. (1999). The acoustics of Dutch vowels: Contextual effects and speaker variability. Speech Communication, 29(2-4), 151-174.

[27] S. Knecht, B. Dräger, M. Deppe, L. Bobe, H. Lohmann, A. Flöel, E. B. Ringelstein, H. Henningsen. (2000). Handedness and hemispheric language dominance in healthy humans. Brain : a journal of neurology, 123 Pt 12, 2512–2518. https://doi.org/10.1093/brain/123.12.2512.

[28] R. T. Schirrmeister, J. T. Springenberg, L. D. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human brain mapping, 38(11), 5391–5420. https://doi.org/10.1002/hbm.23730.

[29] M. O. Tamm, Y. Muhammad, N. Muhammad. (2020). Classification of Vowels from Imagined Speech with Convolutional Neural Networks. Computers; 9(2):46. https://doi.org/10.3390/computers9020046.

[30] N. Kobayashi, T. Morooka. (2021). Application of High-accuracy Silent Speech BCI to Biometrics using Deep Learning. 2021 9th International Winter Conference on Brain-Computer Interface (BCI), 1-6.

[31] T. Cho, M. E. Beckman, J. Edwards. (2002). Phonetic similarity in German and English: Evidence from vowel systems. Journal of Phonetics, 30(2), 255-276.

[32] D. Y. Lee, M. Lee, S. W. Lee. (2021). Decoding Imagined Speech Based on Deep Metric Learning for Intuitive BCI Communication. IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society. PP. 10.1109/TNSRE.2021.3096874.

[33] L. C. Sarmiento, S. Villamizar, O. López, A. C. Collazos, J. Sarmiento, J. B. Rodríguez. (2021). Recognition of EEG Signals from Imagined Vowels Using Deep Learning Methods. Sensors (Basel);21(19):6503. doi: 10.3390/s21196503. PMID: 34640824; PMCID: PMC8512781.

[34] M. Maslowski, A. S. Meyer, H. R. Bosker. (2018). Listening to yourself is special: Evidence from global speech rate tracking. PloS one, 13(9), e0203571. https://doi.org/10.1371/journal.pone.0203571.

[35] G. Dunn, Self Reflected Gallery. (Aug 10, 2023). Accessed here: https://www.gregadunn.com/self-reflected/self-reflected-gallery/. Thesis cover.

APPENDIX

A. *Testing different time windows*

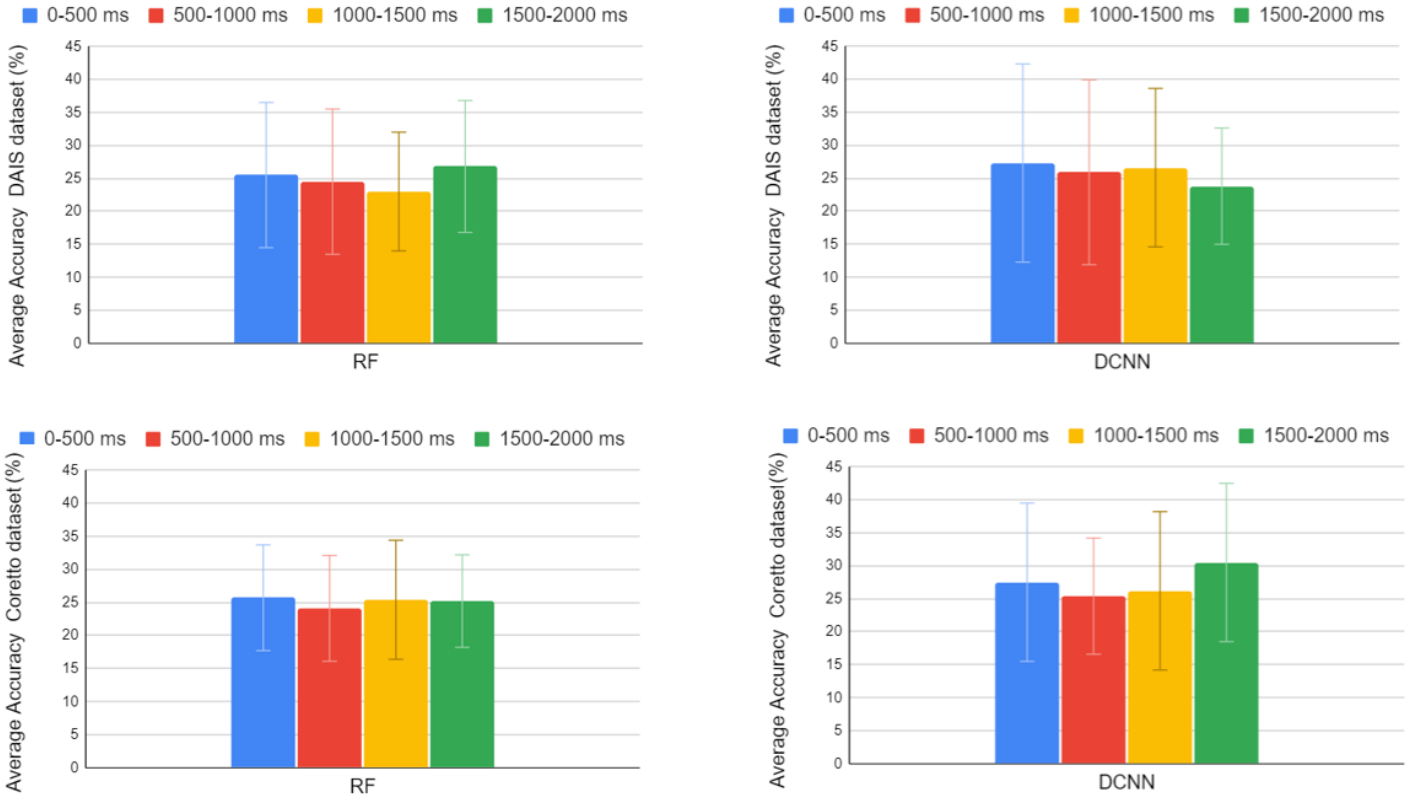


Fig. 16: Average intra-subject accuracy when training took place with a subset of data, for different parts of the initial 2 second window. DAIS dataset with an RF classifier (upper left), DAIS dataset with a DCNN classifier (upper right), Coretto dataset with an RF classifier (bottom left), Coretto dataset with a DCNN classifier (bottom right). Chance level is 20%.

B. Classification per subject

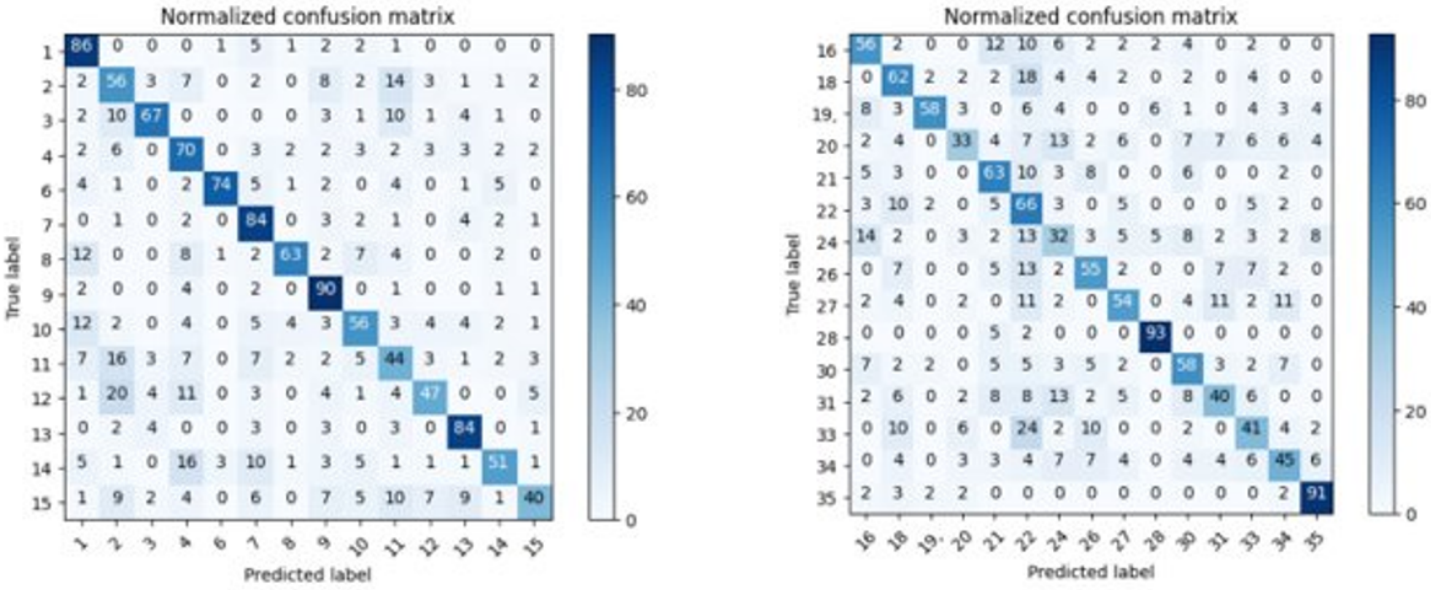


Fig. 17: Normalized confusion matrices of multi-class classification to identify the subject from the EEG signals. All data from the 5 vowels were used for this experiment. Subjects 1 through 15 correspond to Coretto (left), whereas subject 16 to 35 to DAIS dataset (right).

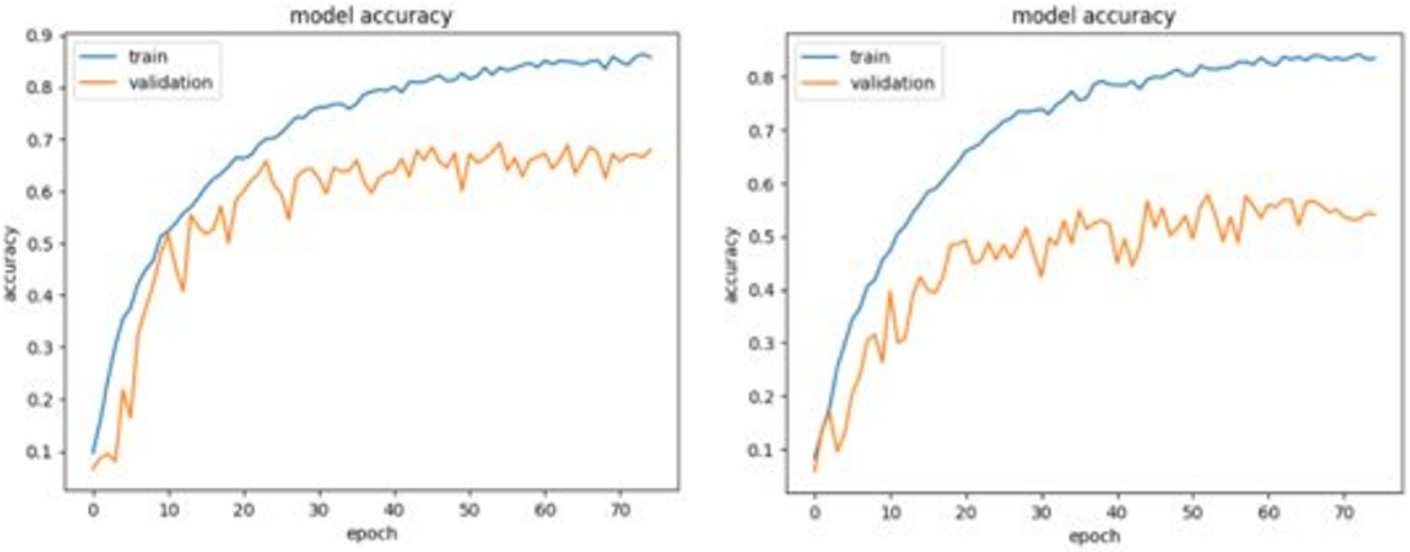


Fig. 18: Model accuracy history curves for multi-class classification for subject identification. The blue lines corresponds to training accuracy and the orange ones to validation accuracy. Coretto dataset to the left, DAIS dataset to the right.

C. Different number of electrodes

TABLE IV: Classification accuracy for the subjects of DAIS dataset, for different numbers of electrodes (62 - all the electrodes of the EEG cap used, 16 - electrodes covering the Wernicke and Broca's areas, 6 - electrode locations used in the Coretto dataset.

Subject	**Number of Electrodes**		
	62	**16**	**6**
1	32.4	25.0	34.1
3	34.8	34.9	43.3
4	28.8	38.1	37.7
5	44.3	36.2	40.4
6	39.3	33.6	40.7
8	33.6	31.5	39.7
10	42.0	33.8	36.1
11	43.4	27.5	40.0
12	41.7	37.1	43.3
14	34.8	34.1	37.8
15	33.3	38.4	39.0
16	28.7	30.3	41.0
18	33.1	22.6	36.6
19	31.5	36.7	535.1
20	41.3	33.1	40.1
Average	$\mathbf{34.2 \pm 9.7}\%$	$\mathbf{32.9 \pm 4.6}\%$	$\mathbf{39.0 \pm 2.6}\%$

D. Order of feature importance - Random Forest

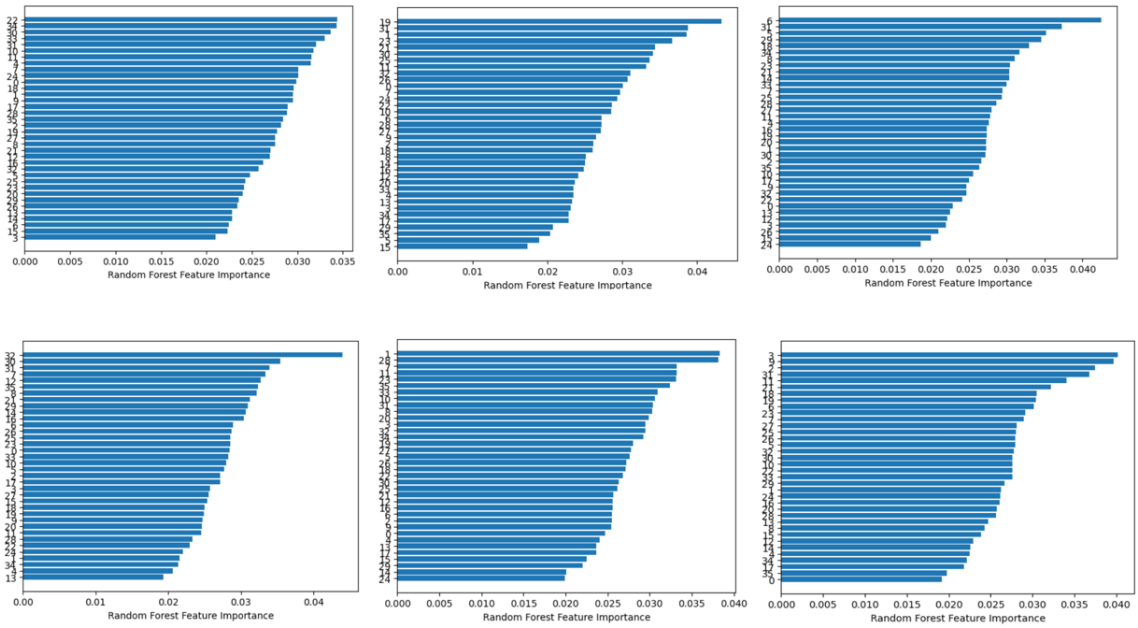


Fig. 19: Feature importance curves obtained via the Random Forest classifier for DAIS dataset. Upper left: subject 15, upper middle: subject 8, upper right: subject 5, bottom left: subject 12, bottom middle: subject 14, bottom right: subject 19. There are 6 features per electrode (6 electrodes in total), leading to 36 features. The sum of the importance of all features equals to 1.

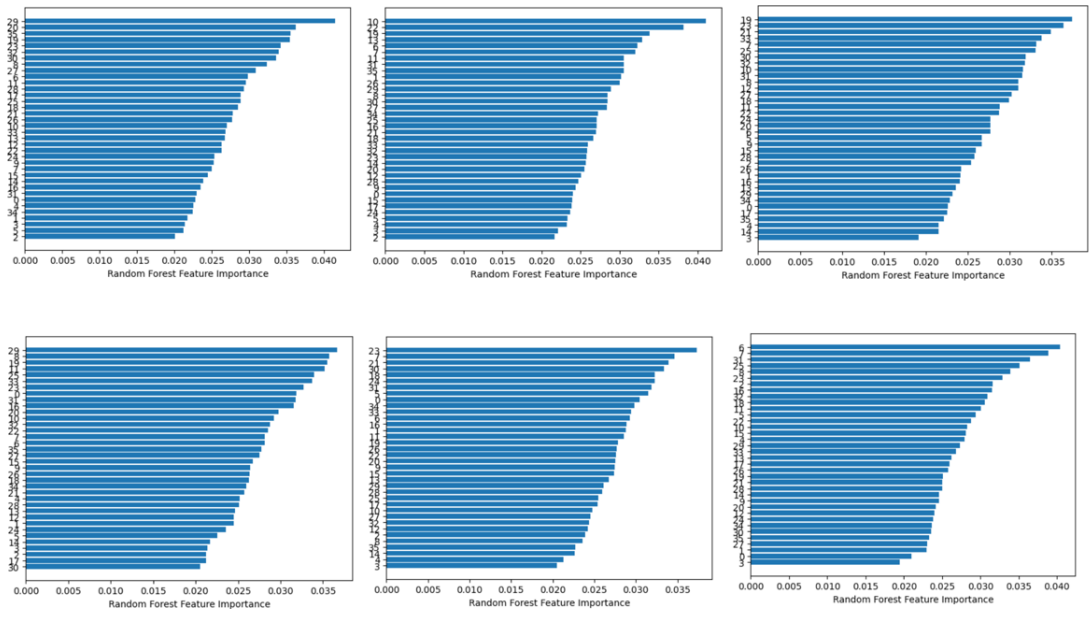


Fig. 20: Feature importance curves obtained via the Random Forest classifier for Coretto dataset.. Upper left: subject 1, upper middle: subject 2, upper right: subject 8, bottom left: subject 6, bottom middle: subject 7, bottom right: subject 12.

E. Confusion matrices of other methods

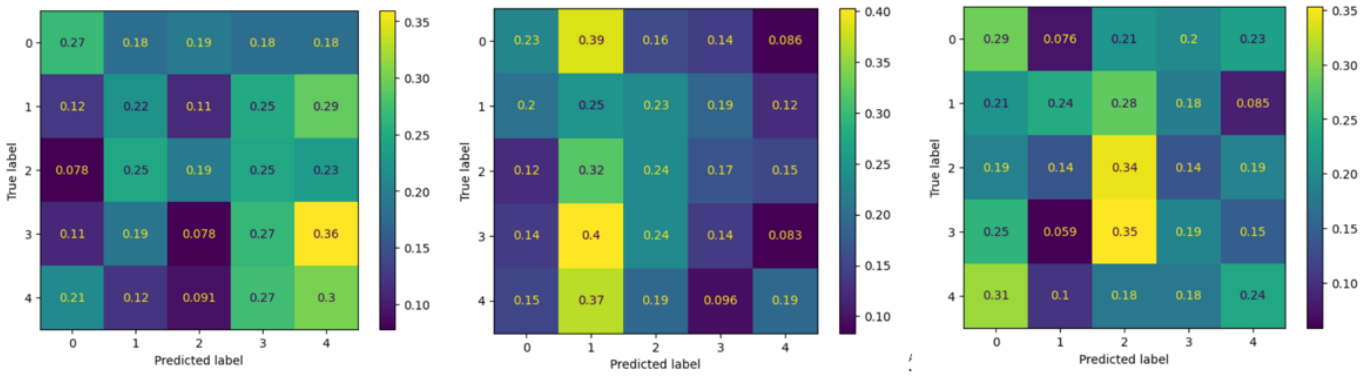


Fig. 21: DAIS Subject normalized confusion matrices for the RF classifier. Upper left: subject 15, upper middle: subject 8, upper right: subject 5. Numbers correspond to the 5 vowels: 1 - /a/, 2 - /e/, 3 - /i/, 4 - /o/, 5 - /u/.

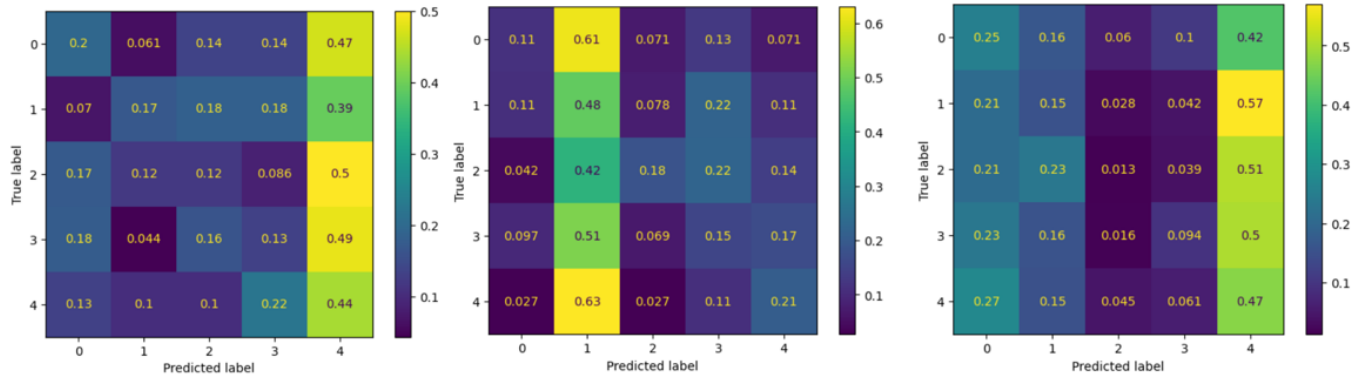


Fig. 22: DAIS Subject normalized confusion matrices for the NB classifier. Upper left: subject 15, upper middle: subject 8, upper right: subject 5. Numbers correspond to the 5 vowels: 1 - /a/, 2 - /e/, 3 - /i/, 4 - /o/, 5 - /u/.

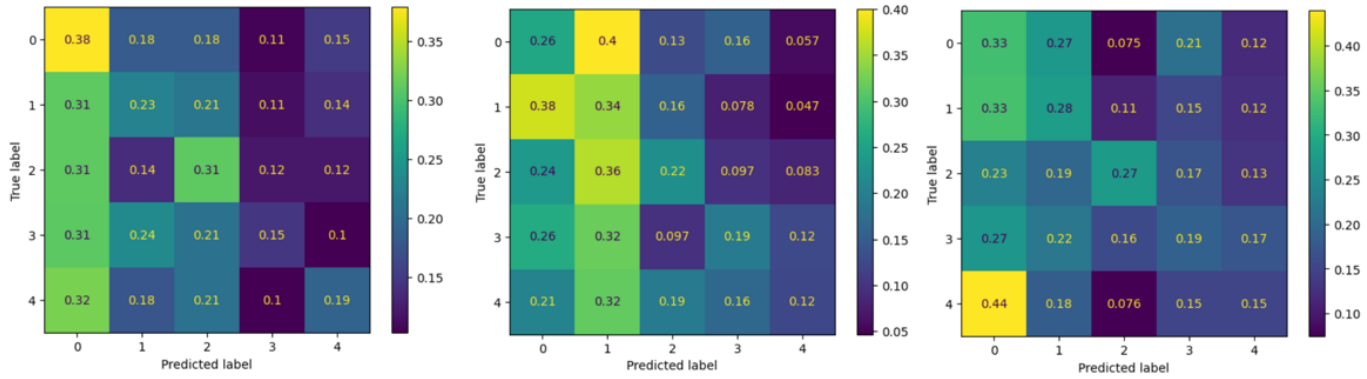


Fig. 23: DAIS Subject normalized confusion matrices for the k-NN classifier. Upper left: subject 15, upper middle: subject 8, upper right: subject 5. Numbers correspond to the 5 vowels: 1 - /a/, 2 - /e/, 3 - /i/, 4 - /o/, 5 - /u/.