Article

Rethinking the Methods and Algorithms for Inner Speech Decoding - and Making them Reproducible

Foteini Simistira Liwicki [1,†]*, Vibha Gupta [1,†], Rajkumar Saini [1,†], Kanjar De [1,†] and Marcus Liwicki [1,†]

[1] firstname.lastname@ltu.se
* Correspondence: foteini.liwicki@ltu.se
† Embedded Intelligent Systems LAB, Machine Learning, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden.

Abstract: This study focuses on the automatic decoding of inner speech using noninvasive methods, such as electroencephalography (EEG)). While inner speech has been a research topic in philosophy and psychology for half a century, recent attempts have been made to decode nonvoiced spoken words by using various brain-computer interfaces. Although good accuracy has been reported in some cases (up to 90% for a binary task), results are often just above chance level (approximately 20% for 5 vowels). The main shortcomings of existing work are reproducibility and the availability of data and code. In this work, we investigate various methods (using deep Learning and standard classifiers) for the detection task of 5 vowels and 6 words on a publicly available EEG dataset. The main contribution of this work lies in its extensive experimentation and analysis. We analyse (1) subject dependent vs. subject-independent approaches, (2) the effect of different preprocessing steps (Independent Component Analysis (ICA), down-sampling and filtering), and (3) word classification (where we achieve state-of-the-art performance). Overall we achieve a performance accuracy of 35.20% when classifying the 5 vowels and 29.21% when classifying the 6 words, using our proposed iSpeech-Convolutional Neural Network (CNN) architecture and preprocessing methodology without downsampling. All of our code and processed data are publicly available to ensure reproducibility. As such, this work contributes to a deeper understanding and reproducibility of experiments in the area of inner speech detection.

Keywords: Brain-Computer Interface (BCI); Inner Speech; Electroencephalography (EEG); Deep Learning; Convolutional Neural Network (CNN); Independent Component Analysis; Supervised Learning

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

0. Introduction

This work focuses on inner speech decoding. Thought is strongly related to inner speech [1,2], through a voice being inside the brain that does not actual speak. Inner speech, although not audible, occurs when reading, writing, and even when idle (i.e., "mind-wandering" [3]). Moreover, inner speech follows the same pattern e.g., regional accents, as if the person is actually speaking aloud, for example [4].

While inner speech has been a research topic in the philosophy of psychology since the second half of the 20th century [5], with results showing that the part of the brain responsible for the generation of inner speech is the frontal gyri, including Broca's area, the supplementary motor area and the precentral gyrus, the automatic detection of inner speech has very recently become a research topic [6], [7]. However, a core challenge of this research is to go beyond single words and integrate other language domains (e.g., phonology and syntax) to reconstruct the entire speech stream.

In this work, we conduct extensive experiments using deep learning methods to decode 5 vowels and 6 words on a publicly available electroencephalography (EEG) dataset [8] - and achieve state-of-the-art performance on all tasks. Moreover, the main contributions of this work are as follows: (i) providing code for reproducing the reported results, (ii)

applying various methods for analysing the results (see Section 1), (iii) reporting results using raw and preprocessed data, and (iv) proposing a novel varying recognition architecture (iSpeech-CNN), which achieves state-of-the-art performance on word and vowel recognition.

0.1. State-of-the-art Literature

In [6], the authors reported an overall accuracy of 90% on the binary classification of vowels compared with consonants using Deep-Belief Networks (DBN) and the combination of all modalities (inner and outer speech), in a subject-independent approach. [8] reported a mean recognition rate of 22.32% in classifying 5 vowels and 18.58% in classifying 6 words using a Random Forest (RF) algorithm, with a subject-dependent approach. Using data from 6 subjects, [9] reported an average accuracy of 50.1% $\pm$ 3.5 for the 3 word classification problem and 66.2% $\pm$ 4.8 for a binary classification problem (long vs. short words), following a subject-dependent approach using a multi-class Relevance Vector Machine (MRVM). Martin et al. [10] attempted to detect single words from inner speech using electrocorticography (ECoG) recordings from inner and outer speech. This study included six word pairs and achieved a binary classification accuracy of only 58% (chance level 50 %) using a Support Vector Machine (SVM). Recent methods used a CNN with the "EEG-as-an-image" [11] and "EEG-as-raw-data" [7,12] inputs. In [11] magnetoencephalography (MEG) data from inner and outer speech and the information from jaw movement was used, and an average accuracy of 93% for the inner speech and 96% for the outer speech decoding of 5 phrases in a subject dependent approach using a CNN was reported. In [7], the authors used a CNN with transfer learning to analyse inner speech on the EEG dataset of [8] in a within-subject approach. In these experiments, the CNN was trained on the raw EEG data of all subjects but one. A subset of the remaining subject's data was used to fine-tune the CNN and the rest of the data was used for testing the CNN model. They authors reported an overall accuracy of 35.68% (5-fold cross-validation) for the 5 vowel classification task. [13] reported an average classification accuracy of 70.45 ± 19.19% for a binary word classification task using regularized linear discriminant analysis (RLDA). The EEGNet [14] is a CNN-based deep learning architecture for EEG signal analysis that includes a series of 2D convolutional layers, average pooling layers, and batch normalization layers with activations. Finally, there is a fully connected layer at the end of the network to classify the learned representations from preceding layers. The EEGNet serves as the backbone network in our model; however, the proposed model extends the EEGNet in similar manner to [7].

Table 1. Overview of the existing machine learning studies for inner speech using different modalities (2015-2020).

Study	Technology	Number of subjects	Number of classes	ML approach	Results
[6] - 2015	EEG, facial, audio	6	2 phonemes	DBN	90%
[8] - 2017	EEG, audio	15	5 vowels - 6 words	RF	22.32% - 18.58%
[9] - 2017	EEG	6	3-2 words	MRVM	50.1% ± 3.5 - 66.2% ± 4.8
[10] - 2018	ECoG, audio	5	2 (6) words	SVM	58%
[7] - 2019	EEG, audio	15	5 vowels	CNN	35.68%
[13] - 2019	Functional near-infrared spectroscopy (fNIRS), EEG	11	2 words	RLDA	70.45 ± 19.19%
[11] - 2020	MEG, audio, jaw movement	8	5 phrases	CNN	93%

Studies using invasive methods like ECoG [10] create a risk of infection and various ethical issues [15]. Therefore, these methods are typically only used for studies when the subjects undergoes brain surgery for other reasons. The methodology used in [7] is questionable due to the small test set (only 1/5 of one subject´s raw EEG data is used) and there is lack of regularization. Proper machine learning protocol studies would require a sufficiently large training set, an unbiased validation set and a test set which has not previously used. In many papers, it is unclear whether experiments were repeated. As Neural Networks (NN) models are stochastic, the accuracy varies from iteration to iteration and, therefore, needs to be averaged. Some studies stated average results with a very high

variance [13], signifying that the results are not stable and are more likely to be random. Other studies took advantage of outer speech and jaw movements [6,11].

Most of the machine learning methods in the existing studies belong to the category of shallow (representation) learning, i.e., only a few layers of higher-level representations are learned or the classification is directly performed (typically with SVMs, linear classifiers, or 1-layer neural networks). Deep (representation) learning, although being already studied in the early second half of the previous century, has gained increasing popularity due to its increasing computational power, data availability, and method improvements [16, 17]. The potential of deep learning on functional magnetic resonance imaging (fMRI) data was shown in [18]. On segmentation benchmarks (e.g., MRBrainS[1]), all top state-of-the-art segmentation methods are based on deep learning. These methods include variants of (fully) CNNs, as well as faster implementations of Long Short-Term Memory Networks (LSTM) [19] networks. With combined CNN+LSTM implementations [20], the real-time application of those networks is whithin reach. It is important to note that deep learning (i.e., Generative Adversarial Networks (GAN)s) have also been applied to generate pseudovisual data from fMRI data [21]. However, it is questionable whether the generated data truly represent the human mind or if they are artefacts from the pretraining of the neural networks on large image datasets.

1. Materials and Methods

1.1. Dataset and Experimental Protocol

The current work uses a publicly available EEG dataset as described in [8]. This dataset includes recordings from 15 subjects using their inner and outer speech to pronounce 5 vowels *(/a/, /e/, /i /, /o/, /u/)* and 6 words *(arriba/up, abajo/down, derecha/right, izquierda/left, adelante/forward, and atr ás/backwards)*. A total of 3316 and 4025 imagined speech sample EEG recordings for vowels and words, respectively, are available in the dataset. An EEG with 6 electrodes was used in these recordings. More details can be found in [8].

Figure 1 shows the experimental design followed to collect the dataset in [8]. The experimental protocol consisted of a ready interval that was presented for 2 sec, followed by the stimulus (vowel or word) presented for 2 sec. The subjects were asked to use their inner or outer speech during the imagine interval to pronounce the stimulus. Finally, a rest interval of 4 seconds was presented, indicating that the subjects could move or blink their eyes before proceeding with the next stimulus. It is important to note that for the purpose of our study, only the inner speech part of the experiment is used.

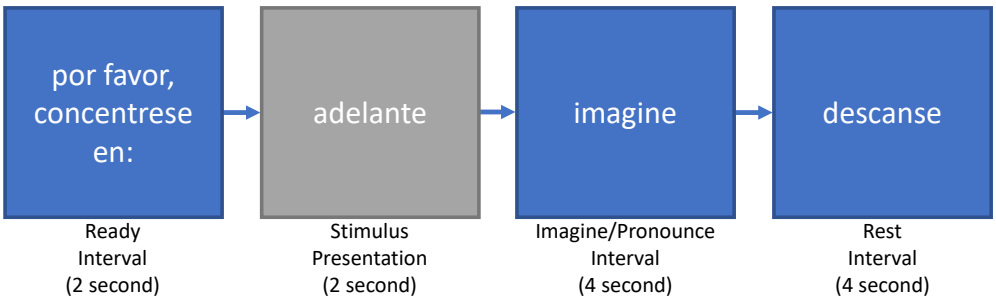

Figure 1. Experimental protocol [7].

1.2. Methods

The proposed framework uses a deep convolutional neural network (CNN) to extract representations from the input EEG signals. However, before applying the proposed CNN, the signals are preprocessed and then the CNN network is trained on the preprocessed signals. Separate networks are trained for vowels and words. Fig. 2 depicts the flow of the

[1] https://www.nitrc.org/projects/mrbrains/

proposed work. The proposed network is inspired by Cooney et al. [7]; they performed filtering, downsampling, and artefact removal before applying the CNN. However, we have noticed that downsampling degrades the recognition performance, see Section 3. Therefore, we do not downsample the signals in our experiments. The downsampling block is represented by a cross in Fig. 2 to indicate that this task is not included in our proposed system in comparison with [7]. The current work reports results on 3 different experimental approaches using preprocessed data and raw data. The 3 different approaches are discussed in detail in Sections 2.1.3 - 2.1.1. More information about the preproccessing techniques can be found in Section 1.3.

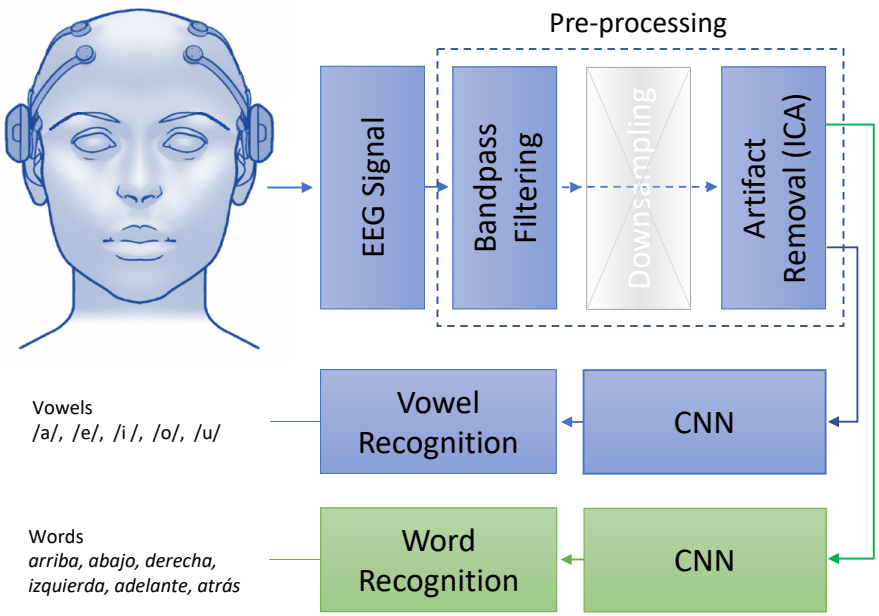

Figure 2. Proposed workflow.

1.3. Preprocessing

In the current work we apply the following preprocessing steps:
Filtering: A frequency between 2 Hz and 40 Hz is used for filtering [8].
Down-sampling: The filtered data are down-sampled to 128 HZ. The original frequency of the data is 1024 Hz.
Artefact removal: ICA is performed to compute mutually independent signal components. ICA facilitates the extraction of independent components from mixed signals by transforming a multivariate random signal. This approach effectively removes noise from the EEG data and is, therefore, an aid to classification. Given the small number of channels, we intact all the channels but used ICA [22] for artefact removal [2].

Figure A1 (see Appendix) depicts the preprocessed signal after applying ICA. This figure shows the vowel *a* for two subjects. From this figure, it can be noted that the subject´s model is not discriminative enough as overlapping is observed. The response from all electrodes' behaviour for all vowels for Subject-02 can be seen in Figure A2 (see Appendix). From this figure, it can be seen that all electrodes are adding information as they all differ in their characteristics.

[2] https://github.com/pierreablin/picard/blob/master/matlab_octave/picard.m

1.4. *iSpeech*-CNN *Architecture*

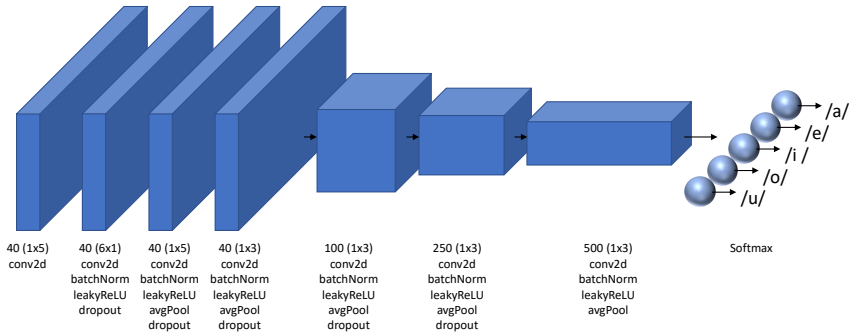

Figure 3. Proposed iSpeech-CNN architecture for imagined vowel recognition.

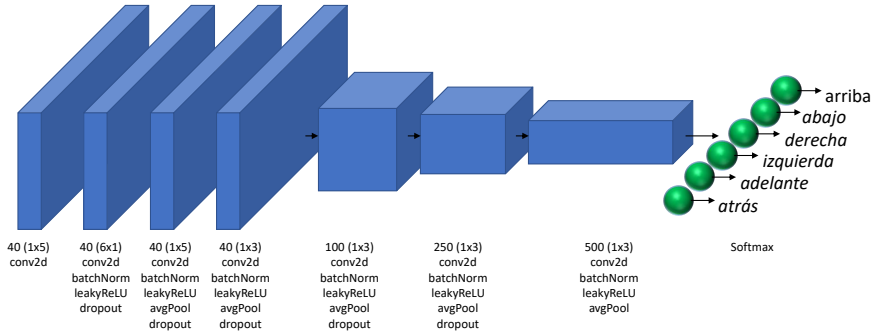

Figure 4. Proposed iSpeech-CNN architecture for imagined word recognition.

In this section, we introduce the proposed CNN-based iSpeech architecture. After extensive experiments on the existing CNN architectures for inner speech classification tasks, we determined that downsampling the signal has an effect on the accuracy of the classification and thus removed it from the proposed architecture. The iSpeech-CNN architecture for imagined vowel and word recognition are shown in Fig. 3 and Fig. 4, respectively. Unlike [7], after extensive experimentation, we observed that the number of filters has an effect on the overall performance of the system; 40 filters are used in the first four layers of both networks. The next three layers have 100, 250, and 500 filters, respectively. However, the filter sizes are different. Filters of sizes (1×5), (6×1), (1×5), (1×3), (1×3), (1×3), and (1×3) are used in the first, second, third, fourth, fifth, sixth, and seventh layers, respectively. The difference in the network architectures for vowels, and words is that the network for vowels (Fig. 3) has five units in its softmax layer (as there are five vowel classes in the dataset), and the network for words (Fig. 4) has six units in its softmax layer (as there are six word classes in the dataset).

We used Adam optimizer with a dropout of 0.0002 for the vowel classification and 0.0001 for the word classification. As the network is very small, dropping out more features will adversely affect the performance. The initial learning rate was fixed to 0.001 with a *piecewise* learning rate scheduler. Our network was trained for 60 epochs, and the best validation loss was chosen for the final network. The regularization was also fixed to a value of 0.001. Our proposed iSpeech-CNN architecture follows the same structure as [7] but with a different numbers of filters and training parameters and preprocessing.

2. Experimental Approaches and Performance Measures

This section describes the experimental approaches that have been utilized for the analysis of EEG data and the performance measures that quantify the obtained analysis.

2.1. Experimental Approaches

Three experimental approaches were used for analysis, and they are discussed in detail in the following subsections.

2.1.1. Subject-dependent/Within-subject Approach

Subject-dependent/within-subject classification is a baseline approach that is commonly used for the analysis of inner speech signals. In this approach, individual models are trained corresponding to each subject and for each subject, a separate model is created. The training, validation and testing sets will all have data from the same subject. This approach essentially measures how much an individual subject´s data changes (or varies) over time.

To divide the subject data into training, testing and validation datasets, a ratio of 80-10-10 is used. The training, validation and testing datasets contain all vowel/words category samples (five/six, respectively) in the mentioned ratio. To remove the bias towards the samples, five different trials are utilized. Furthermore, the mean accuracy and standard deviation are reported for all experimental approaches.

2.1.2. Subject-independent/Cross-subject Approach

The subject-dependent approach does not show generalization capability as it models a one subject at a time. Testing data contain only samples of the subject that is being modelled. Hence, to obtain some insights into the generalization ability of the EEG data, we apply the subject-independent (or cross-subject) approach. Using this approach, only one model will be trained for all subjects. This approach differs from the within-subject and leave-one-out approaches, where n models, which corresponds to the total number of subjects in the data, are trained. Testing contains samples of all the subjects under all subcategories (vowels/words).

To run this experiment, 80% of the samples of all the subjects are included in the training set, 10% in the validation set and remainder in the test set. We also ensure the class balancing, i.e., each class will have approximately the same number of samples of all vowel/word categories. The same experiment is repeated for five random trials, and the mean accuracy along with the standard deviation is reported.

2.1.3. Leave-one-out Approach

Compared to the subject-independent approach, the leave-one-out approach can be considered as less generalized (it includes data from n-1 subjects in training and is tested only with only one subject who was not included in the training). This approach helps to obtain a deeper analysis when there are fewer subjects or entities and shows how each individual subject affects the overall estimate of the rest of the subjects. Hence, this approach may provide more generalizable remarks than subject-specific models that depend on individual models.

In the leave-one-out approach, the data of each subject are tested using models that are trained using the data of all other subjects but one, i.e., $n - 1$ subjects out of total n will be used for the training model, and the rest will be used for testing. For example, Model-01 will be trained with data from subjects except Subject01 , and will be tested with Subject01 (see Table [A3-4]).

2.2. Performance Measures

The mean and standard deviation are used to report the performance of all the approaches. For the final results, the F-score are also given.

Mean: The mean is the average of a group of scores. The scores are totalled and then divided by the number of scores. The mean is sensitive to extreme scores when the population samples are small.

Standard deviation: In statistics, the standard deviation (SD) is a widely used measure of variability. It depicts the degree of deviation from the average (mean). A low SD implies

that the data points are close to the mean, whereas a high SD suggests that the data span a wide range of values.

F-score: The F-score is a measure of a model's accuracy that is calculated by combining the precision and recall of the model. It is calculated by the following formula:

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

where precision is the percentage of true positive examples among the positive examples classified by the model, and recall is the fraction of examples classified as positive, among the total number of positive examples.

Table 2. Classification accuracy

Study	Classifier	Vowels	Words
[8] - 2017	RF	22.32%±1.81	18.58%±1.47
[7] - 2019	CNN	32.75%±3.23	-
iSpeech-GRU	GRU	19.28% ±2.15	17.28% ±1.45
iSpeech-CNN (proposed)	CNN	35.20% ±3.99	29.21% ±3.12

3. Results and Discussion: Vowels (5 Categories)

The results estimated with the subject-specific approach are discussed first as this approach is common in most of EEG related papers. All code, raw data, and preprocessed data are provided on Github [3]. Related approaches are discussed in later subsections.

3.1. Subject-dependent/Within-subject Classification

In this section, we report the results when applying the subject-dependent approach. Figures 5, 6 and Table 3 show the results of our proposed iSpeech-CNN architecture. Table A1 and Table A2 show the results of the reference CNN architecture.

3.1.1. Ablation study - influence of downsampling

Table A1 shows the results with raw and downsampled data when used within the referenced CNN architecture framework. It is clearly observed from the Table A1 that downsample signal losses the information. In other words, the chosen sampling rate was not sufficient enough to retain the original information. Therefore, further results will be reported for both raw data and downsampled data, in order to obtain a better insight into the preprocessing (i.e., filtering and ICA) stage.

3.1.2. Ablation study - influence of preprocessing

Filtering and artefact removal plays an important role while analyzing the EEG signals. We applied the both bandpass filtering (see Section1), and picard (preconditioned ICA for real data) for artefact removal to obtain more informative signals. Table A2 shows the results of preprocessing when applied on the raw and downsampled data within the reference CNN architecture framework. The performance, i.e., the overall mean accuracy, decreased from 32.51% to 30.90%. The following points can be noted from Table A2: (1) Filtering and artefact removal highly influence the performance irrespective of raw and downsampled data. (2) The improved performance can also be observed with respect to each subject. A smaller standard deviation can also be seen. (3) The CNN framework generated higher performance than the handcrafted features, and the gated recurrent unit (GRU) (see Table 2).

[3] https://github.com/vibhasgupta/EEG_Matlab-Code/tree/master

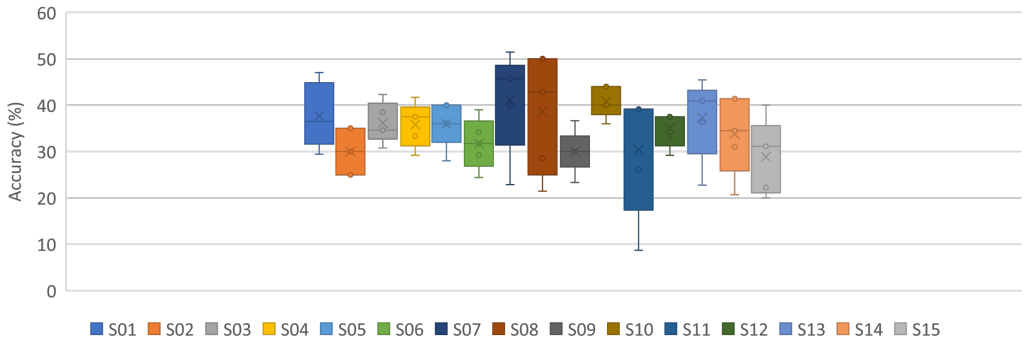

Figure 5. Subject dependent results for vowels with downsampling on preprocessed signals [iSpeech-CNN Architecture].

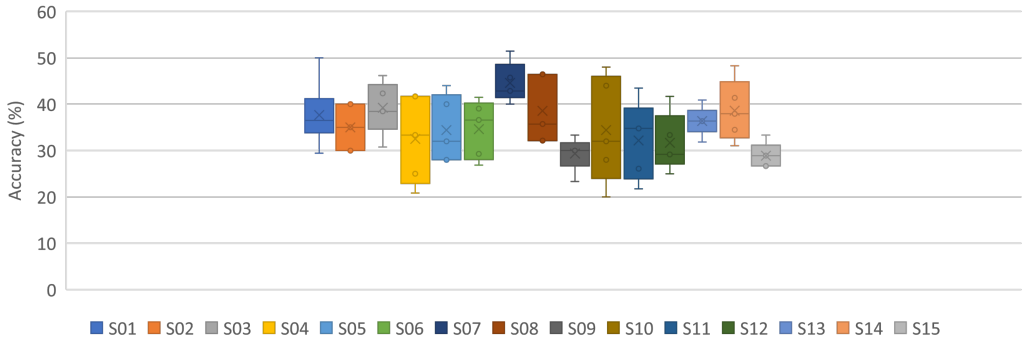

Figure 6. Subject dependent results for vowels without downsampling on preprocessed signals [iSpeech-CNN Architecture].

Table 3. Subject dependent results for vowels with and without downsampling on preprocessed signals [iSpeech-CNN Architecture].

Preprocessing (Filtering and Artefact Removal)						
	No Downsampling			Downsampling		
	Train	Validation	Test	Train	Validation	Test
Subj 01	86.33	48.80	37.65	81.86	33.60	37.65
Subj 02	98.00	38.00	35.00	84.33	31.00	30.00
Subj 03	85.76	38.00	39.23	98.00	37.00	36.15
Subj 04	97.09	46.00	32.50	97.45	41.00	35.83
Subj 05	91.41	44.00	34.40	94.59	38.00	36.00
Subj 06	89.87	41.00	34.63	95.33	38.00	31.71
Subj 07	98.34	39.00	44.57	98.21	39.00	41.14
Subj 08	80.24	41.00	38.57	96.82	38.00	38.57
Subj 09	96.65	32.80	29.33	92.22	35.20	30.00
Subj 10	97.41	44.00	34.40	87.53	37.00	40.80
Subj 11	95.43	41.00	32.17	98.06	31.00	30.43
Subj 12	91.54	40.00	31.67	83.43	34.00	35.00
Subj 13	82.74	46.00	36.36	85.60	46.00	37.27
Subj 14	80.23	43.20	38.62	78.06	36.80	33.79
Subj 15	89.60	31.00	28.89	90.80	38.00	28.89
Average/Mean	**90.71**	**40.92**	**35.20**	**90.82**	**36.91**	**34.88**
Standard Deviation	**6.24**	**4.65**	**3.99**	**6.60**	**3.66**	**3.83**

3.1.3. Ablation study - influence of architecture

Based on the CNN literature in the EEG paradigm [7] [23], adding more layers to the reference CNN architecture does not help to obtain improved performance. However, by changing the number of filters in the initial layers, some improvements can be observed. Based on the CNN literature for EEG signals, having a sufficient number of filters in the initial layers helps to obtain some improvement [7] [24]. Here, we choose three initial layers, unlike in natural images, in speech, initial layers are more specific to the task rather than the last few layers. The results with a changing number of filters in the initial layer within the iSpeech-CNN architecture are shown in Table 3. In the reference CNN architecture, this filter number was 20 for the initial three layers. However, we have changed this number to 40 (decided based on experimentation) in the iSpeech-CNN architecture. Table 3 clearly shows that changing the filter parameter gives higher performance than with the number of filters (compare to the reference architecture results in Tables A1 and A2 in the appendix). This improvement is observed with and without downsampled data and with respect to the subject (see Figure 5 and Figure 6). The standard deviation also decreases with these modifications (see Table 3).

3.2. Subject-independent results

Having discussed the subject-specific results, in this section, the subject-independent results are discussed. Using this approach, data from all subjects are included in training, validation and testing. Table A4 shows the results for the subject-independent approach with and without downsampling. These results were compiled with filtering and ICA in both reference and modified CNN architectures.

From these results, it is noted that the obtained accuracies are random in nature. The modified CNN architecture parameters, do not help to abtain any improvements, and show random accuracy behaviour. In other words, it is difficult to achieve generalized performance with EEG signals. Based on the EEG literature, it has also been justified that models trained on data from one subject cannot be generalize to other subjects even though have been recorded using the same setup conditions.

Determining the optimal frequency subbands corresponding to each subject could be one possible direction that may be successful in such a scenario. We intent to explore such a direction in our future work.

3.3. Leave-one-out results

The leave-one-out approach is a variation of the subject independent approach. However, unlike the subject-independent approach, here the data of the testing subject for the leave-one-out approach are not included in the training. For example, in Figure 7, except $Subject01$, all other subjects were used in the training of $Model - 01$. Figure 7 and Table 4 show the results using the iSpeech-CNN architecture, while Table A3 shows the results using the reference architecture.

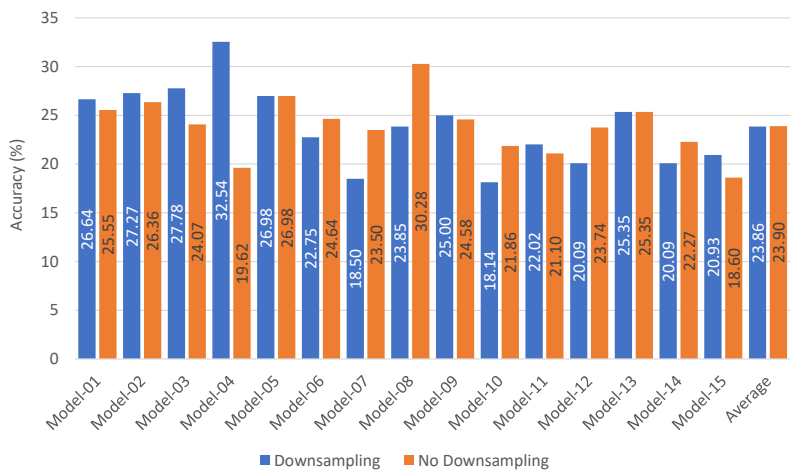

Figure 7. Leave-one-out results for vowels with and without downsampling on preprocessed signals [iSpeech-CNN Architecture].

Table 4. Leave-one-out results for vowels with and without downsampling on preprocessed signals [iSpeech-CNN Architecture].

With Preprocessing (Filtering and Artefact Removal)					
	No Downsampling			**Downsampling**	
	Validation	**Test**		**Validation**	**Test**
Model-01	46.65	26.64		38.23	25.55
Model-02	83.11	27.27		42.02	26.36
Model-03	74.32	27.78		43.16	24.07
Model-04	22.27	32.54		56.16	19.62
Model-05	64.79	26.98		56.24	26.98
Model-06	77.13	22.75		47.73	24.64
Model-07	47.21	18.50		40.76	23.50
Model-08	39.86	23.85		50.16	30.28
Model-09	82.38	25.00		46.36	24.58
Model-10	45.82	18.14		48.02	21.86
Model-11	52.42	22.02		54.23	21.10
Model-12	34.26	20.09		67.03	23.74
Model-13	46.05	25.35		50.08	25.35
Model-14	46.94	20.09		41.85	22.27
Model-15	51.69	20.93		46.31	18.60
Average/Mean	**54.33**	**23.86**		**48.56**	**23.90**
Standard Deviation	**18.13**	**4.02**		**7.51**	**2.97**

It can be noted that having fewer subjects in training (one less as compared to the subject-independent approach), shows slightly better behaviour than the subject-independent approach, where all subjects were included in the training. Moreover, changing the reference CNN parameters to our proposed iSpeech-CNN architecture also shows improved performance (see Figure 7 and Table 4).

The subject-independent and leave-one-out approaches both showed that generalizing the performance over all subjects is difficult in the EEG scenario. Hence, there is a need for the preprocessing stage, which can make the data more discriminative.

4. Results and Discussion: Words (6 Categories)

Having discussed all the approaches for the category of vowels, we noticed that only the subject-specific approach showed performance that was not random in nature and hence makes sense. Therefore, in this section, we only report results corresponding to the subject-specific approach for the word category.

This category contains six different commands (see Section 1.1). Table 5 and Figures 8 and 9 show the performance results for the classification of the 6 words, using the proposed iSpeech-CNN architecture. The performance results when using the reference architecture can be found in Appendix B. From these tables and figures, the same kind of behaviour as vowels is observed. The change in the number of filters in the initial layers affected the performance as shown in Table 5. The downsampling of data also affects the overall performance.

The iSpeech-CNN architecture shows better performance than handcrafted features such us real-time wavelet energy[8] and reference architecture (Appendix B).

Overall, we achieve a state-of-the-art performance of 29.21% when classifying the 6 words using our proposed iSpeech-CNN architecture and preprocessing methodology without downsampling.

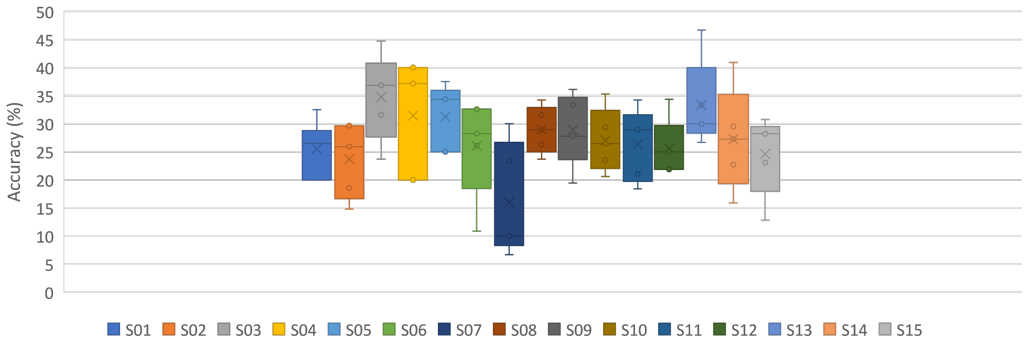

Figure 8. Subject-dependent results for words with downsampling on preprocessed signals [iSpeech-CNN Architecture].

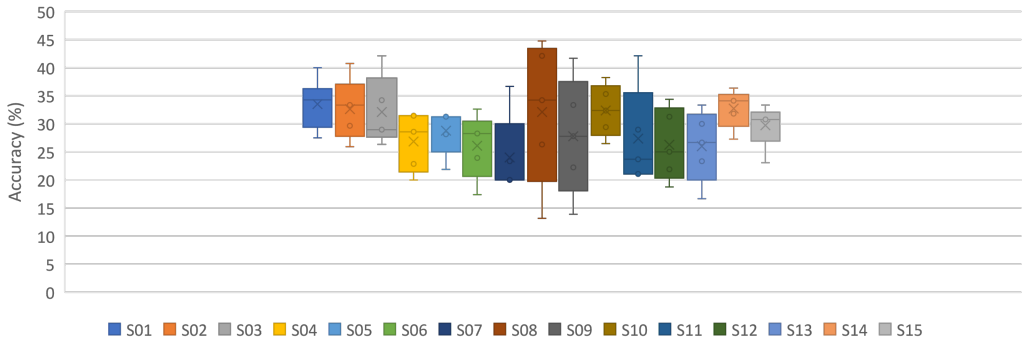

Figure 9. Subject-dependent results for words without downsampling on preprocessed signals [iSpeech-CNN Architecture].

Table 5. Subject-dependent results for words with and without downsampling on preprocessed signals [iSpeech-CNN Architecture].

Preprocessing (Filtering and Artefact Removal)						
	No Downsampling			Downsampling		
	Train	Validation	Test	Train	Validation	Test
Subj 01	97.95	34.67	33.50	77.35	30.00	25.50
Subj 02	92.86	34.17	32.59	89.05	28.33	23.70
Subj 03	93.05	32.00	32.11	76.67	28.67	34.74
Subj 04	94.51	35.00	26.86	95.59	29.17	31.43
Subj 05	95.38	34.17	28.75	82.47	26.67	31.25
Subj 06	96.99	38.33	26.09	67.20	30.83	26.09
Subj 07	87.47	33.33	24.00	80.71	30.00	16.00
Subj 08	98.33	35.00	32.11	83.70	30.00	28.95
Subj 09	98.86	37.50	27.78	70.00	27.50	28.89
Subj 10	92.38	32.50	32.35	93.52	34.17	27.06
Subj 11	97.43	36.67	27.37	87.90	35.33	26.32
Subj 12	92.86	37.33	26.25	69.05	32.00	25.62
Subj 13	84.57	31.67	26.00	68.29	31.67	33.33
Subj 14	96.38	37.33	32.73	66.38	30.67	27.27
Subj 15	94.69	33.33	29.74	86.67	31.67	24.62
Average/Mean	**94.25**	**34.87**	**29.21**	**79.64**	**30.45**	**27.38**
Standard Deviation	**4.00**	**2.14**	**3.12**	**9.51**	**2.25**	**4.37**

Table 6. Precision and F-score (with respect to Tables 3- 5) for vowel and word classification [iSpeech-CNN Architecture].

Vowel (iSpeech-CNN)			
	Precision	**Weighted F score**	**F-score**
No Downsampling	34.85	41.12	28.45
Downsampling	34.62	38.99	30.02
Cooney et al.[7] (Downsampling)	33.00	-	33.17
Words (iSpeech-CNN)			
	Precision	**Weighted F-score**	**F-score**
No Downsampling	29.04	36.18	21.84
Downsampling	26.84	31.94	21.50

5. Performance Comparison and related discussion

In this section, we compare our results on the vowels and words dataset with existing work and discuss related findings. Based on the reported performances in the Table 2, it is clearly noted that the CNN performs better than the handcrafted features for both datasets.

The precision, weighted F-score, and F-score for our proposed iSpeech-CNN in comparison with the reported results of Cooney et. al.[7] are shown in Table 6. From this table, we can note that our proposed system results in a higher precision but a lower F-score compared to the model in [7]. Hence, the reproducibility of the results reported in [7] is questionable.

Our proposed CNN architecture and preprocessing methodology outperform the existing work in both categories. Hence, it is concluded that the correct selection of preprocessing methods and the number of filters in the CNN, greatly add to the performance. The elaborated results for each category and with each approach have been added to Appendices A and B.

6. Conclusions

This study explores the effectiveness of preprocessing steps and the correct selection of filters in the initial layers of the CNN in the context of both vowel and word classification. The classification results are reported on a publicly available inner speech dataset of 5 vowels and 6 words [8]. Based on the obtained accuracies, it is found that such a direction of exploration truly adds to the performance. We report state-of-the-art classification performance for vowels and words with mean accuracies of 35.20% and 29.21% respectively, without downsampling the original data. Mean accuracies of 34.88% and 27.38% have been reported for vowels and words, respectively, with downsampling. Furthermore, the proposed CNN code in this study is available to the public to ensure reproducibility of the research results and to promote open research. Our proposed iSpeech-CNN architecture and preprossessing methodology are the same for both datasets (vowels and words). Evaluating our system in other publicly available datasets is part of our future work.

In the future, we will address the issues related to the selection of the downsampling rate and the selection of the optimal frequency subband with respect to subjects. In the results section, we have already discussed about their importance in the EEG paradigm.

Funding: This research received no external funding by the Grants for excellent research projects proposals of SRT.ai 2022

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. The results on vowels

Table A1. Subject-dependent results for the vowels using raw and downsampled data [Reference Architecture].

	Raw				**Downsampled**		
	Train	**Validation**	**Test**		**Train**	**Validation**	**Test**
Subj 01	91.07	36.00	35.29		79.07	39.20	26.47
Subj 02	92.44	29.00	17.00		72.00	24.00	16.00
Subj 03	83.65	34.00	27.69		86.94	26.00	19.23
Subj 04	89.94	38.00	33.33		89.33	30.00	31.67
Subj 05	86.59	16.00	21.60		72.82	19.00	17.60
Subj 06	91.73	27.00	25.85		80.00	29.00	27.80
Subj 07	95.17	31.00	26.86		85.66	33.00	28.00
Subj 08	88.12	29.00	21.43		83.06	24.00	21.43
Subj 09	83.89	35.20	30.00		82.38	30.40	26.00
Subj 10	88.35	29.00	20.80		63.88	26.00	19.20
Subj 11	80.91	25.00	22.61		86.06	23.00	16.52
Subj 12	88.57	29.00	26.67		96.23	37.00	30.83
Subj 13	80.00	38.00	29.09		93.94	36.00	31.82
Subj 14	79.66	27.20	22.76		73.49	27.20	23.45
Subj 15	83.20	23.00	20.00		91.73	28.00	20.89
Average/Mean	**86.88**	**29.76**	**25.29**		**82.44**	**28.79**	**23.79**
Standard Deviation	**4.81**	**5.96**	**5.14**		**8.73**	**5.44**	**5.35**

Table A2. Subject-dependent results for vowels with and without downsampling on preprocessed data [Reference Architecture].

Preprocessing (Filtering and Artefact Removal)						
	No Downsampling				Downsampling	
	Train	Validation	Test	Train	Validation	Test
Subj 01	96.65	38.40	38.82	91.44	43.20	38.82
Subj 02	92.78	40.00	25.00	99.33	32.00	27.00
Subj 03	99.88	38.00	34.62	99.41	36.00	32.31
Subj 04	99.15	43.00	32.50	97.45	42.00	30.83
Subj 05	89.29	37.00	26.40	93.41	34.00	35.20
Subj 06	95.47	38.00	37.56	98.80	38.00	32.20
Subj 07	98.62	36.00	27.43	81.52	28.00	28.00
Subj 08	97.06	45.00	38.57	98.59	39.00	32.14
Subj 09	97.41	36.00	35.33	97.41	32.00	30.00
Subj 10	96.24	35.00	37.60	98.47	35.00	28.80
Subj 11	88.69	30.00	31.30	91.66	35.00	33.91
Subj 12	86.86	35.00	30.83	91.31	28.00	24.17
Subj 13	99.54	40.00	33.64	95.89	46.00	31.82
Subj 14	87.43	39.20	35.86	84.11	40.00	33.79
Subj 15	91.87	33.00	22.22	99.73	30.00	24.44
Average/Mean	**94.46**	**37.57**	**32.51**	**94.57**	**35.88**	**30.9**
Standard Deviation	**4.44**	**3.62**	**5.05**	**5.48**	**5.28**	**3.83**

Table A3. Leave-one-out results for vowels with and without downsampling on preprocessed data [Reference Architecture].

Preprocessing (Filtering and Artefact Removal)				
	No Downsampling		Downsampling	
	Validation	Test	Validation	Test
Model-01	71.43	21.17	23.08	18.61
Model-02	88.99	20.91	40.41	21.36
Model-03	86.00	25.93	57.48	19.91
Model-04	87.77	24.40	55.20	22.01
Model-05	69.33	19.53	58.72	24.65
Model-06	55.36	14.69	60.84	23.70
Model-07	60.30	26.00	36.36	30.50
Model-08	83.99	28.44	41.51	21.10
Model-09	45.45	26.25	44.05	24.58
Model-10	73.72	20.00	41.99	17.67
Model-11	51.45	25.23	68.88	24.31
Model-12	91.90	22.37	55.34	20.55
Model-13	63.63	20.28	61.47	22.12
Model-14	78.62	23.14	51.31	21.83
Model-15	84.42	25.58	24.28	22.33
Average/Mean	**72.82**	**22.93**	**48.06**	**22.35**
Standard Deviation	**14.84**	**3.55**	**13.10**	**2.95**

Table A4. Subject-independent/Cross-subject results for vowels with and without downsampling on preprocessed data (Filtering and Artefact Removal) [Reference Architecture; iSpeech-CNN Architecture].

With preprocessing; Reference architecture parameters						
	No Downsampling			Downsampling		
	Train	Validation	Test	Train	Validation	Test
Trial 1	72.45	20.95	22.27	62.61	20.32	17.63
Trial 2	76.73	22.22	19.03	58.75	20.63	24.36
Trial 3	64.82	22.54	19.26	50.58	20.32	20.65
Trial 4	67.55	18.10	19.95	57.04	18.41	20.19
Trial 5	60.78	20.32	22.04	47.78	22.86	23.90
Mean/Average	**68.47**	**20.83**	**20.51**	**55.35**	**20.51**	**21.35**
Standard Deviation	**5.61**	**1.59**	**1.38**	**5.43**	**1.42**	**2.50**
With preprocessing; iSpeech-CNN Architecture parameters						
	no Downsampling			Downsampling		
	Train	Validation	Test	Train	Validation	Test
Trial 1	57.98	20.63	20.42	55.10	21.90	18.33
Trial 2	68.79	21.27	19.72	46.46	23.17	22.04
Trial 3	54.63	21.59	21.11	44.44	18.73	22.27
Trial 4	37.70	17.46	20.42	24.36	21.27	21.35
Trial 5	86.15	24.13	20.19	57.20	22.86	20.42
Average/Mean	**61.05**	**21.02**	**20.37**	**45.51**	**21.59**	**20.88**
Standard Deviation	**16.04**	**2.14**	**0.45**	**11.64**	**1.58**	**1.43**

Appendix B. The results on words

Table A5. Subject-dependent results for words with and without downsampling on preprocessed data [Reference Architecture].

Preprocessing (Filtering and Artefact Removal)						
	No Downsampling			With Downsampling		
	Train	Validation	Test	Train	Validation	Test
Subj 1	90.60	31.33	30.00	83.16	32.67	28.50
Subj 2	96.57	30.83	23.70	79.62	30.00	22.22
Subj 3	90.19	32.00	31.58	67.33	30.00	29.47
Subj 4	92.75	36.67	25.14	75.49	33.33	25.71
Subj 5	87.53	41.67	28.13	69.89	30.83	31.25
Subj 6	83.12	39.17	25.65	71.72	29.17	23.04
Subj 7	92.22	26.67	25.33	79.39	35.00	21.33
Subj 8	85.37	38.33	30.53	86.48	34.17	23.16
Subj 9	92.86	32.50	25.56	82.57	30.00	28.89
Subj 10	89.05	31.67	29.41	92.00	30.00	30.59
Subj 11	93.14	31.33	27.37	95.24	32.00	22.11
Subj 12	88.00	33.33	25.62	82.00	34.67	32.50
Subj 13	96.76	28.33	33.33	86.10	39.17	25.33
Subj 14	86.10	35.33	28.64	68.95	36.00	28.18
Subj 15	96.98	27.50	29.23	91.77	28.33	27.69
Average/Mean	**90.75**	**33.11**	**27.95**	**80.78**	**32.36**	**26.66**
Standard Deviation	**4.27**	**4.36**	**2.76**	**8.48**	**2.91**	**3.53**

Appendix C. Dataset samples

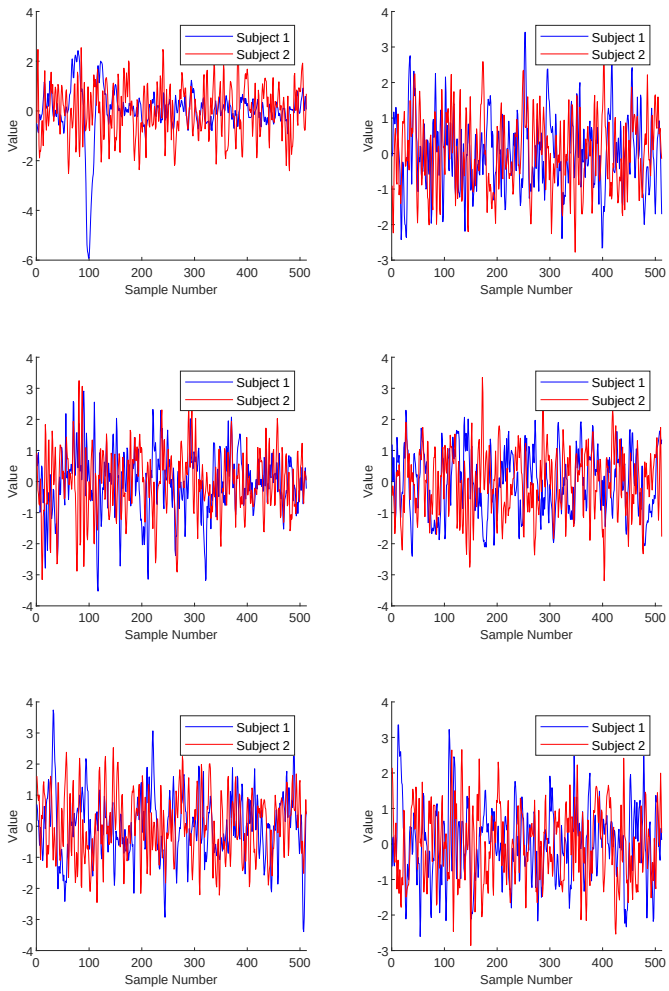

Figure A1. Example of preprocessed signals for all electrodes (after ICA) for the vowel /a/ for Subject01 and Subject02.

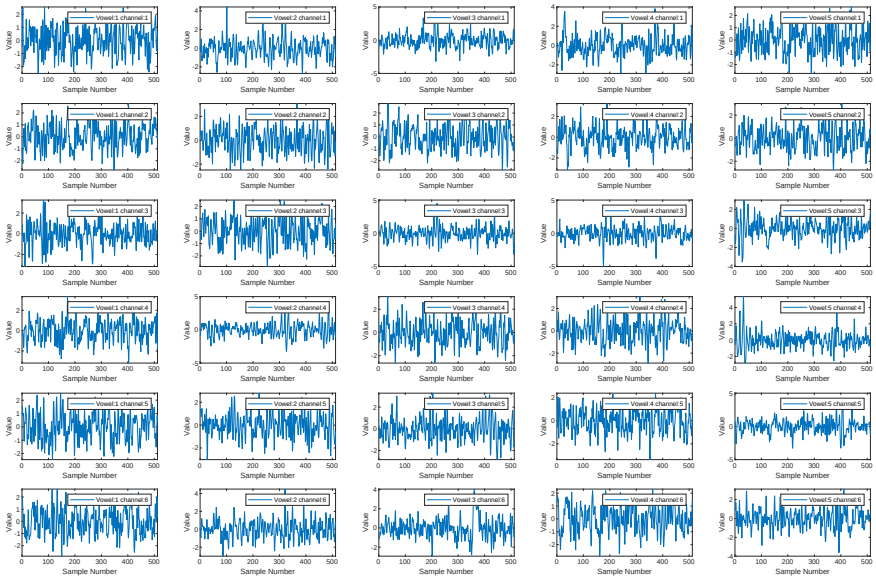

Figure A2. Example of preprocessed signals (after ICA) for all vowels and all electrodes for Subject02.

References

1. Alderson-Day, B.; Fernyhough, C. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin* **2015**, *141*, 931.
2. Whitford, T.J.; et al. Neurophysiological evidence of efference copies to inner speech. *Elife* **2017**, *6*, e28197.
3. Smallwood, J.; Schooler, J.W. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology* **2015**, *66*, 487–518.
4. Filik, R.; Barber, E. Inner speech during silent reading reflects the reader's regional accent. *PloS one* **2011**, *6*, e25782.
5. Langland-Hassan, P.; Vicente, A. *Inner speech: New voices*; Oxford University Press, USA, 2018.
6. Zhao, S.; Rudzicz, F. Classifying phonological categories in imagined and articulated speech. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 992–996.
7. Cooney, C.; Folli, R.; Coyle, D. Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019, pp. 1311–1316.
8. Coretto, G.A.P.; Gareis, I.E.; Rufiner, H.L. Open access database of EEG signals recorded during imagined speech. 12th International Symposium on Medical Information Processing and Analysis, 2017, Vol. 10160, p. 1016002.
9. Nguyen, C.H.; Karavas, G.K.; Artemiadis, P. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of neural engineering* **2017**, *15*, 016002.
10. Martin, S.; Iturrate, I.; Millán, J.d.R.; Knight, R.T.; Pasley, B.N. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience* **2018**, *12*, 422.
11. Dash, D.; et al. Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Frontiers in neuroscience* **2020**, *14*.
12. Schirrmeister, R.T.; et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* **2017**, *38*, 5391–5420.
13. Rezazadeh Sereshkeh, A.; et al. Development of a ternary hybrid fNIRS-EEG brain–computer interface based on imagined speech. *Brain-Computer Interfaces* **2019**, *6*, 128–140.
14. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* **2018**, *15*, 056013.
15. Glannon, W. Ethical issues with brain-computer interfaces. *Frontiers in systems neuroscience* **2014**, *8*, 136.
16. LeCun, Y.; et al. Deep learning. *nature* **2015**, *521*, 436–444.
17. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks* **2015**, *61*, 85–117.
18. Plis, S.M.; et al. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience* **2014**, *8*, 229.
19. Stollenga, M.F.; et al. Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation. NIPS, 2015.
20. Breuel, T.M. High performance text recognition using a hybrid convolutional-lstm implementation. 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE, 2017, Vol. 1, pp. 11–16.
21. Seeliger, K.; et al. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage* **2018**, *181*, 775–785.

22. Ablin, P.; Cardoso, J.F.; Gramfort, A. Faster independent component analysis by preconditioning with Hessian approximations. *IEEE Transactions on Signal Processing* **2018**, *66*, 4040–4049.
23. Cheng, J.; Zou, Q.; Zhao, Y. ECG signal classification based on deep CNN and BiLSTM. *BMC medical informatics and decision making* **2021**, *21*, 1–12.
24. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT press, 2016.

Cover Letter

We hereby submit our study titled *Rethinking the Methods and Algorithms for Inner Speech Decoding - and Making them Reproducible* to the NeuroSci journal. Our study focuses on the automatic decoding of inner speech using noninvasive methods, i.e., EEG). While inner speech has been a research topic in philosophy and psychology for half a century, recent work tries to decode nonvoiced spoken words by using various brain-computer interfaces. Although good accuracy has been reported in some cases (up to 90% for a binary task), often, the results are just above chance level (approximately 20% for 5 vowels). The main shortcoming of existing work is reproducibility and availability of data and code. In this work, we investigate various methods for the detection task of 5 vowels and 6 words on a publicly available EEG dataset. The main contribution of this work lies in the extensive experimentation and analysis. We analyse (1) the subject-dependent vs. subject-independent approach, (2) the effect of different preprocessing steps (ICA, downsampling, filtering), and (3) word classification (where we achieve state-of-the-art performance). Overall, we achieve a state-of-the-art performance of 35.20% when classifying the 5 vowels and 29.21% when classifying the 6 words, using our proposed iSpeech-CNN architecture and preprocessing methodology. Our reported results in vowel and word inner speech decoding outperform previous studies on the same dataset. [7] reports an overall accuracy of 32.75% in vowel decoding, and [8] reports an overall accuracy of 18.58% for the word dataset. All our code and preprocessed data are publicly available to ensure reproducibility. As such, this work contributes to a deeper understanding and reproducible experiments in the area of inner speech detection. Therefore we believe that our research work contributes with novel findings to the area of cognitive neuroscience.

We confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal. All authors have approved the manuscript and agree with its submission to NeuroSci.