Inner speech recognition through electroencephalographic signals

Francesca Gasparini[1,2,*,†], Elisa Cazzaniga[1,†] and Aurora Saibene[1,2,*,†]

[1]University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy
[2]NeuroMI, Milan Center for Neuroscience, Piazza dell'Ateneo Nuovo 1, 20126, Milano, Italy

Abstract

This work focuses on inner speech recognition starting from electroencephalographic (EEG) signals. Inner speech recognition is defined as the internalised process in which the person thinks in pure meanings, generally associated with an auditory imagery of own inner "voice". The decoding of the EEG into text should be understood as the classification of a limited number of words (commands) or the presence of phonemes (units of sound that make up words). Speech-related brain computer interfaces provide effective vocal communication strategies for controlling devices through speech commands interpreted from brain signals, improving the quality of life of people who have lost the capability to speak, by restoring communication with their environment. Two public inner speech datasets are analysed. Using this data, some classification models are studied and implemented starting from basic methods such as Support Vector Machines, to ensemble methods such as the eXtreme Gradient Boosting classifier up to the use of neural networks such as Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (BiLSTM). With the LSTM and BiLSTM models, generally not used in the literature of inner speech recognition, results in line with or superior to those present in the state-of-the-art are obtained.

Keywords

EEG, inner speech recognition, BCI

1. Introduction

Human speech production is a complex motor process that starts in the brain and ends with respiratory, laryngeal, and articulatory gestures for creating acoustic signals of verbal communication. Physiological measurements using specialised sensors and methods can be made at each level of speech processing, including the central and peripheral nervous systems, muscular action potentials, speech kinematics (tongue, lips, jaw), and sound pressure [1]. However, there are cases of subjects suffering from neurodegenerative diseases or motor disorders that prevent the normal activity of signal transmission from the brain to the peripheral areas. These subjects are prevented from communicating or carrying out certain actions.

Italian Workshop on Artificial Intelligence for Human Machine Interaction (AIxHMI 2022), December 02, 2022, Udine, Italy

[*]Corresponding author.

[†]These authors contributed equally.

francesca.gasparini@unimib.it (F. Gasparini); e.cazzaniga@campus.unimib.it (E. Cazzaniga); aurora.saibene@unimib.it (A. Saibene)

0000-0002-6279-6660 (F. Gasparini); 0000-0002-4405-8234 (A. Saibene)

Brain Computer Interfaces (BCIs) are promising technologies for improving the quality of life of people who have lost the capability to move or speak, by restoring communication with their environment. A BCI is a system that makes possible the interaction between an individual and a computer without using the brain normal output pathways of peripheral nerves and muscles. In particular, speech-related BCI technologies provide neuro-prosthetic help for people with speaking disabilities, neuro-muscular disorders and diseases. It can equip these users with a medium to communicate and express their thoughts, thereby improving the quality of rehabilitation and clinical neurology [2]. Speech-related paradigms, based on either silent, imagined or inner speech provide a more natural way for controlling external devices [3].

There are different types of brain-signal recording techniques that are mainly divided into invasive or non-invasive methods. The first ones involve implanting electrodes directly into the brain. They provide better spatial and temporal resolution, also increasing the quality of the signal obtained. However, invasive technologies have problems related to usability and the need for surgical intervention on the subject. This is why non-invasive techniques are increasingly used in BCI research. Among the non-invasive technologies, the electroencephalogram (EEG) is the most used method for measuring the electrical activity of the brain from the human scalp. It has an exceedingly high time resolution, it is simple to record and it is sufficiently inexpensive [4]. Over the years, EEG hardware technology has evolved and several wireless multichannel systems have emerged that deliver high quality EEG and physiological signals in a simpler, more convenient and comfortable design than the traditional, cumbersome systems.

Therefore, this paper focuses on inner speech recognition starting from EEG signals, where the basic definition of *inner speech* is [5] "the subjective experience of language in the absence of overt and audible articulation".

As suggested in [6], there is evidence from past neuroscience research that inner speech engages brain regions that are commonly associated with language comprehension and production [7]. This includes temporal, frontal and sensorimotor areas predominantly in the left hemisphere of the brain [7, 8]. Therefore, by monitoring these brain areas, it is theoretically possible to develop an inner speech BCI that classifies neural representations of imagined word [8].

We propose analyses on the literature available *Thinking Out Loud* and *Imagined Speech* datasets by using a Support Vector Machine (SVM) to provide a traditional machine learning model and then an ensemble approach based on eXtreme Gradient Boosting (XGBoost). Finally, we design two deep learning architectures based on Long Short Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models.

In Section 2, the studies in the field of inner speech are described. Section 3 presents the two publicly available datasets used for the analyses proposed in Section 4. In Section 5 the results obtained with our models are presented and discussed. Finally, in Section 6 some conclusions are proposed.

2. Related works

Most studies on classification of inner speech focus on invasive methods such as electrocorticography [9] as they provide higher spatial resolution while fewer studies concerning inner speech classification using EEG data are available [10]. It is important for a BCI application to be non-invasive, accessible and easy to implement to be used by a large number of subjects.

Inner speech recognition is generally faced considering phonemes, in general vowels or syllables, such as /ba/ or /ku/, or simple words such as left, right, up and down, in subject-dependent approaches.

Preliminary works were conducted with very few participants and syllables by *D'Zmura et al.* [11], where EEG waveform envelopes have been adopted to recognise EEG patterns. Also *Brigham and Kumar* [12] and *Deng et al.* [13] considered the recognition of two syllables. In the first work, the accuracy obtained for the 7 subjects ranges from 46% to 88%. The authors preprocessed raw EEG data to reduce the effects of artifacts and noise, and applied k-Nearest Neighbor classifier to autoregressive coefficients extracted as features. Instead, *Deng et al.* used Hilbert spectra and linear discriminant analysis to recognise the two syllables imagined in three different rhythms, for a 6 classes task, with accuracy ranging from 19% to 22%.

Considering works where recognition of phonemes have been investigated, *Da Salla et al.* [14, 15], analysed the recognition of three tasks: /a/, /u/, and rest, obtaining from 68% to 79% of accuracy by using Common Spatial Pattern (CSP). On the same dataset, several other researchers have been tested different models obtaining promising results [16, 17].

Instead, *Kim et al.* [18] considered three vowels /a/, /i/ and /u/ and applied multivariate empirical mode decomposition and common spatial pattern for feature extraction together with linear discriminant analysis, reaching around 70% of accuracy.

Few representative studies that try to recognise imagined words using EEG data are reported in the literature. Given the complexity of the task, the number of terms considered is generally limited.

Suppes et al. [19] proposed an experiment in which five subjects performed the internal speech considering the following words: first, second, third, yes, no, right, and left for all subjects with the addition of to, too and hear for the last three subjects.

In the work performed by *Wang at al.* [20], eight Chinese subjects were required to read in mind two Chinese characters (that meant left and one). They were able to distinguish between the two characters and the rest state. Feature vectors of EEG signals were extracted using CSP, and then these vectors were classified with SVM. Accuracies between 73.65% and 95.76% were obtained when comparing between each of the imagined words and the rest state. A mean accuracy of 82.30% was achieved between the two words themselves.

Salama et al. [21] implemented different types of classifiers such as SVM, discriminant analysis, self-organising map, feed-forward back-propagation and a combination of them, to recognise two words (Yes and No). They used a single electrode EEG device to collect data from seven subjects and the accuracy obtained ranges from 57% to 59%.

In [22], *Mohanchandra at al.* constructed a one-against-all multiclass SVM classifier to discriminate five subvocalised words (water, help, thanks, food, and stop) and reported an accuracy ranging from 60% to 92%.

In the *González-Castañeda at al.* [23] analyses, some techniques of sonification and textification were applied, which allowed to characterise EEG signals as either an audio signal or a text document. Five imagined words (up, down, left, right) and 27 subjects were considered. The average accuracy rate using the EEG textified signals was 83.34%.

Using the data from six subjects, [24] reported an average accuracy of 50.10% for the three-short words (in, out and up) classification problem and 66.20% for the two long words classification problem (cooperate and independent), using a multi-class relevance vector machine. In order to evaluate the effect of the sound, three phonemes were used, namely /a/, /i/ and /u/ obtaining an accuracy of 49.0%. *Coretto et al.* [25], who collected one of the two datasets considered in this paper, reported a mean recognition rate of 22.32% in classifying five Spanish vowels and 18.58% in classifying six Spanish words using a Random Forest (RF) algorithm. Using the same dataset, in [26] 30.00% and 24.97% accuracies were obtained respectively for vowels and words using CNNs.

Recently, *van den Berg et al.* [6], working on the *Thinking Out Loud* dataset [3], also considered in our work, reported an average accuracy of 29.70% for a four-word classification task using a 2D CNN based on the EEGNet architecture [27].

3. Datasets

The testing of the proposed strategies is performed on two publicly available datasets, i.e., the *Thinking Out Loud* [3] and the *Imagined Speech* [25] datasets. In particular, the last dataset is used to check the validity of the resulting best approach for the *Thinking Out Loud* one.

3.1. Thinking Out Loud dataset

The first literature dataset chosen to conduct the subsequent analyses is the *Thinking Out Loud* [3] one, which is focused on an inner speech paradigm intended for the control of a BCI system through imagination of Spanish words.

The selected Spanish words are *arriba* (up), *abajo* (down), *derecha* (right), and *izquierda* (left). Notice that the words were presented randomly with a visual cue.

Ten (four females) healthy right-handed subjects with mean ± std age 34 ± 10, without any hearing or speech loss, nor any previous BCI experience, participated in the experiment, which consisted of three experimental conditions, i.e., inner/pronounced speech and visualised condition. During the *inner speech* condition the participant was asked to imagine his/her own voice, repeating the corresponding word. Instead, the participant was asked to repeatedly pronounce aloud the word corresponding to each visual cue during the *pronounced speech* condition. Finally, during the *visualised condition*, the participant was asked to focus on moving a circle presented at the center of the screen. The direction of the movement was provided by a visual cue.

Each subject participated in 3 consecutive sessions (200 words/session), separated by a break. Each session consisted of a baseline recording (15s), the *pronounced speech* run, two *inner speech* runs, and two *visualised condition* runs. Each run was constituted by a series of trials containing the different experimental tasks.

Notice that in this paper we consider only the *inner speech* condition and that the number of trials for each of its classes varied from subject to subject, however the number of trials for each class were balanced for the same subject. The minimum (maximum) number of trials for a pronounced speech class was 25 (30), the minimum (maximum) number of trials for an inner speech class was 45 (60). Please, refer to Table 4 of the original work by *Nieto et al.* [3] for further details on the dataset.

Figure 1 shows the organisation of each *inner speech* trial, under investigation in this paper. A white circle was shown in the center of the screen and the subject was asked to stare at it without blinking. Subsequently, a white triangle was shown pointing in one of the four directions corresponding to the chosen Spanish words. When the triangle disappeared and the white circle was presented again, the subject had to perform the indicated task. The task execution had to be stopped when the white circle turned blue. The subject was asked to control eye blinking until the circle disappeared. Finally, to evaluate the participants' attention, the subjects were asked to indicate the last inner speech and visualised conditions after a random number of trials. The subject answered using keyboard arrows and feedback was displayed.

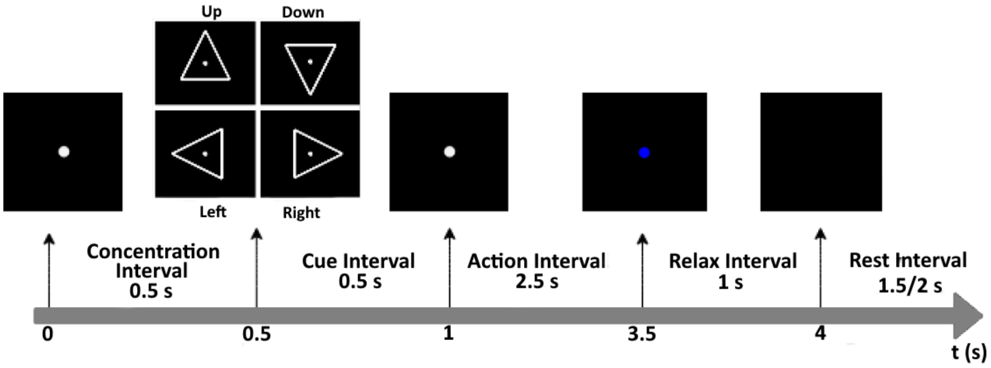

Figure 1: Trial workflow reported following the *Thinking Out Loud* dataset original paper [3].

The data acquisition was performed using 128 active EEG wet electrodes and 8 external active EOG/EMG wet electrodes. The resolution was of 24 bits resolution and 1024 Hz sampling rate applied.

The EEG signals of the *Thinking Out Loud* dataset were preprocessed by its authors. The preprocessing included a band pass filter between 0.5-100 Hz, a notch filter at 50 Hz and downsampling to 254 Hz.

3.2. Imagined Speech dataset

The *Imagine Speech* dataset [25] was chosen to confirm the validity of the model obtaining the best results on the *Thinking Out Loud* dataset. In fact, similar experimental conditions are presented considering Spanish words and also vowels.

In fact, the vowels /a/, /e/, /i /, /o/ and /u/ have been selected due to their acoustic stationarity, simplicity and lack of meaning by themselves. While the Spanish words *arriba* (up), *abajo* (down), *derecha* (right), *izquierda* (left), *adelante* (forward), and *atras* (backward) were chosen as possible BCI commands to control the movements of an external device.

Fifteen (seven females) healthy subjects with mean age of 25 years old, without any hearing or speech loss, participated in the experiment. Only one of the subjects reported to be left-handed, while the rest were right-handed.

EEG signals were recorded under two conditions: *imagined speech* and *pronounced speech*. During *imagined speech*, the subjects had to imagine pronouncing the word without moving

muscles or producing sounds.
Target stimuli were presented in a sequence comprised of four intervals of predefined duration (Figure 2). During the ready interval (2 s), the subject was informed that the rest interval finished and a new cue would be displayed soon. Afterwards, the target word was presented, both visually and acoustically, during the stimulus presentation interval (2 s). In the Imagine/Pronounce stage an image represented the requested task (either imagined or pronounced speech). In this stage the subject had to imagine the pronunciation or pronounce the word given as a cue. In the case that the word was a vowel, the subject had to perform the task during the complete 4 seconds of this interval duration, while if the word was a command, a sequence of three audible clicks indicated when to imagine or pronounce the target word. Finally, during the rest interval (4 s) the subject was allowed to move, swallow or blink.

Even for the Imagined Speech dataset, the number of trials for each class and condition varied from subject to subject, however the number of trials for each class were sufficiently balanced for the same subject, with variations of most 4 trials. The minimum (maximum) number of trials for a pronounced speech class was 7 (14), the minimum (maximum) number of trials for an inner speech class was 39 (51). Please, refer to Table 4 of the original work by *Coretto et al.* [25] for further details on the dataset.

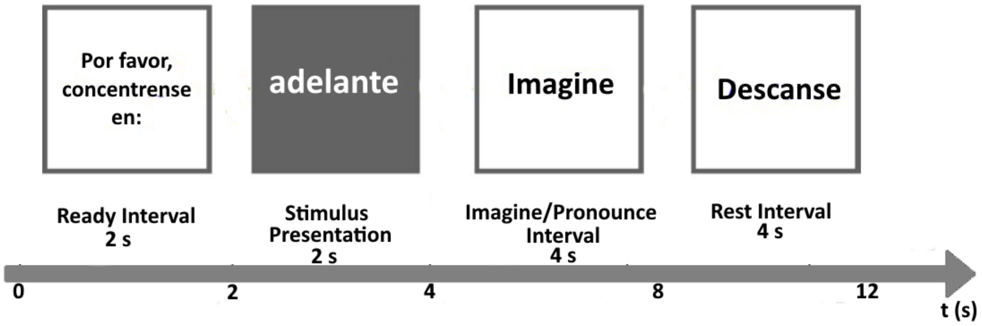
Por favor, concentrense en:
adelante
Imagine
Descanse
Ready Interval 2 s
Stimulus Presentation 2 s
Imagine/Pronounce Interval 4 s
Rest Interval 4 s
0
2
4
8
12
t (s)

Figure 2: Sequence time course for the presentation of one stimulus, in this particular case for the word *adelante* and under the *imagined speech* condition. Graphic inspired from the original dataset paper [25].

EEG signals were recorded using Ag-AgCl cup electrodes, attached to the scalp according to the 10-20 international system and with conductive paste. No electrode cap was used. F3, F4, C3, C4, P3, and P4 were chosen as active electrodes, while reference and ground electrodes were placed on the left and right mastoids. The EEG signals were acquired with 1024 Hz sampling rate and 16 bit resolution.

The EEG signals of the *Imagined Speech* dataset were preprocessed by its authors. The preprocessing included a band pass filter between 2-40 Hz.

4. Proposed approaches

Some classification models are studied and implemented starting from basic methods such as SVM, to ensemble methods such as the XGBoost classifier up to the use of neural networks such as LSTM and BiLSTM.

4.1. Machine Learning approaches

Since inner speech recognition is a very complicated task, a simpler preliminary analysis was performed on the *Thinking out Loud* dataset. Therefore, a binary classification between the resting state and the action interval was performed. The first 1.5 s of the rest interval and the 2.5 s of the action interval were considered for each trial (Figure 1). Subsequently, the multiclass classification was considered on the four words (*left*, *right*, *up* and *down*) of the same dataset. The action intervals of 2.5 s were used for each trial (Figure 1).

The following ML analyses were performed for both classification tasks.
Power Spectral Density (PSD) was used as a feature extraction technique before proceeding with the classification. PSD was calculated using Welch's method and based on relative power in specific frequency bands: alpha (8-13 Hz), beta (13-30 Hz) and gamma (30-100 Hz).
The models were trained and tested on each subject individually, using K-fold cross validation. In this study, the data was split into four folds, resulting in a number of trials ranging from 237 to 285 trials (depending on the subject) in the test set and from 713 to 855 trials in the training set in binary classification, while in multiclass classification from 118 to 142 trials in the test set and from 357 to 428 trials in the training set. The SVM and XGBoost classifiers were trained on PSD feature vector with dimension $(n\_epochs, n\_channels * n\_band\_freqs)$ for each subject. Since the number of features is too high compared to the number of trials of each subject, three possible solutions were analysed:

- to apply Principal Component Analysis (PCA);
- to extract the most important features identified with the XGBoost classifier, both considering the subjects individually and making an intersection of the most important features in common to all subjects;
- to choose a subset of meaningful electrodes, since the neural correlates of inner speech processing are reported to be mainly present in the left hemisphere (see Section 1).

The analyses carried out in the binary classification showed a difference between the action interval and the rest interval. It was therefore verified whether this difference could be associated with one or more time windows, in order to identify a particularly significant area in which the inner speech activity could be encoded. The action interval has been split into 0.5 s wide sliding windows with a 50% overlap. For each window of the action interval, a binary SVM model was trained, considering all the resting state and using the features extracted with XGBoost. The idea was to identify the best window for binary classification and then use it in multiclass classification. Again a subject-based approach was used.

The performed analyses do not justify the choice of one interval over another and this suggests continuing to consider the entire interval in the following tests carried out with deep learning methods.

4.2. Deep Learning approaches

The DL models were trained and tested for the multiclass classification task on each subject individually, using nested cross-validation.
Three different types of analysis were performed in order to obtain the input data to train the models:

- the PSD using the Welch's method was calculated and the most important features were extracted using the XGBoost features importance vector;
- the raw data considering all the channels were used;
- the raw data considering only the channels associated with the most important features extracted using the XGBoost were used.

LSTM and BiLSTM networks were trained for each type of input data. The architecture and parameter choices were performed iteratively by a trial and error process, focusing on the accuracy and loss trend.

The *Imagine Speech* dataset [25] was chosen to confirm the validity of the model that obtained the best results on the *Thinking Out Loud* dataset.

5. Results and discussion

5.1. Machine Learning Models results

In the binary classification task, among the various tests performed using the PCA, the best results for SVM were obtained without PCA and an accuracy of 79% is obtained, while with XGBoost an accuracy of 81% is obtained using PCA, explaining 99% of the variance. Using the most important features extracted with XGBoost considering each subject individually, the SVM performances improve up to 80% accuracy with a gain of 0.9.

In the multiclass classification task, the results obtained with SVM and XGBoost are very similar, respectively 26.20% and 27.90% accuracy. In this case, using the most important features extracted with XGBoost both considering each subject individually and in common to all subjects, the results are approximately the same. This means that the subjects have common characteristics relevant for classification. Furthermore, there are no particular differences when using all channels or only those of the left hemisphere.

The characteristics extracted in common to all subjects were analysed and are highlighted in Figure 3. Considering all the channels, the most affected areas are the occipital one, probably involved due to the visual signals presented on the screen, the temporal and the frontal ones. Both the right and left hemispheres are involved. Since, even using only the electrodes identified in the left hemisphere, the performance still remains good, this could mean that there are electrodes in the left channels that compensate for the absence of the right ones. The frequency band most involved is alpha, usually associated with intense mental activity. These features were used later for some analyses carried out with the DL models.

5.2. Deep Learning Models results

This paragraph summarises the performances obtained with the different deep learning models using the *Thinking out loud dataset.* In Table 1 the results are shown averaging over all subjects.

In general, with BiLSTM the performances increase rather than using LSTM. This is due to the fact that BiLSTM is able to capture the sequential dependencies between data in both directions. The greatest improvement is obtained using raw data from all channels as input.

Looking at the average performance subject by subject, we have a repeating trend. Using both different models and different types of input data, the data of some subjects are classified better

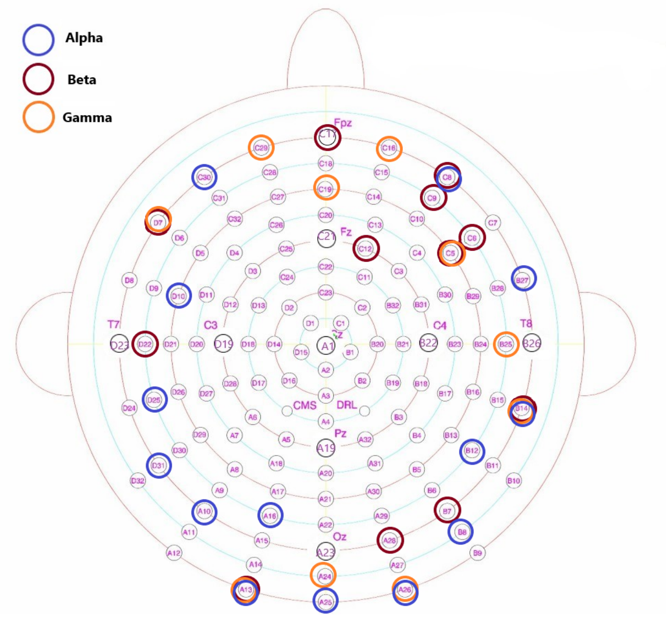

Figure 3: Features in common to all subjects in the multiclass task identified with XGBoost using all channels and gain = 0.95.

Table 1
Deep learning models comparison.

Input Type	LSTM Accuracy	BiLSTM Accuracy
Most important features	30.40%	31.30%
Raw data (all channels)	27.20%	36.10%
Raw data (channels most important features)	26.70%	33.10%

than others. Specifically, subjects 4 and 5 achieve an average performance which is always slightly lower. This could probably be related to the data acquisition phase of these volunteers. Maybe the placement of the electrodes was not perfect or their data is noisier. Analysing the results of attention monitoring there are no differences with the other subjects, so the lack of attention in carrying out the task should not be the cause of the lower performance.

The best model network is composed of a BiLSTM layer followed by two dense layer (with ReLU activation function) and a dense output layer (with softmax activation function). Two dropout layers were used to reduce overfitting. SGD was used as optimisation method and categorical cross entropy as loss function.

We remind that the architecture and parameter choice was performed iteratively by a trial and error process, focusing on the accuracy and loss trend.

Figure 4 shows the results obtained for each subject. All the subjects achieve an accuracy above randomness (represented by the red line - 25%) and the mean accuracy considering all the subjects is 36.10% (± 0.054 std). Table 2 summarises the results of our best model for each subject.

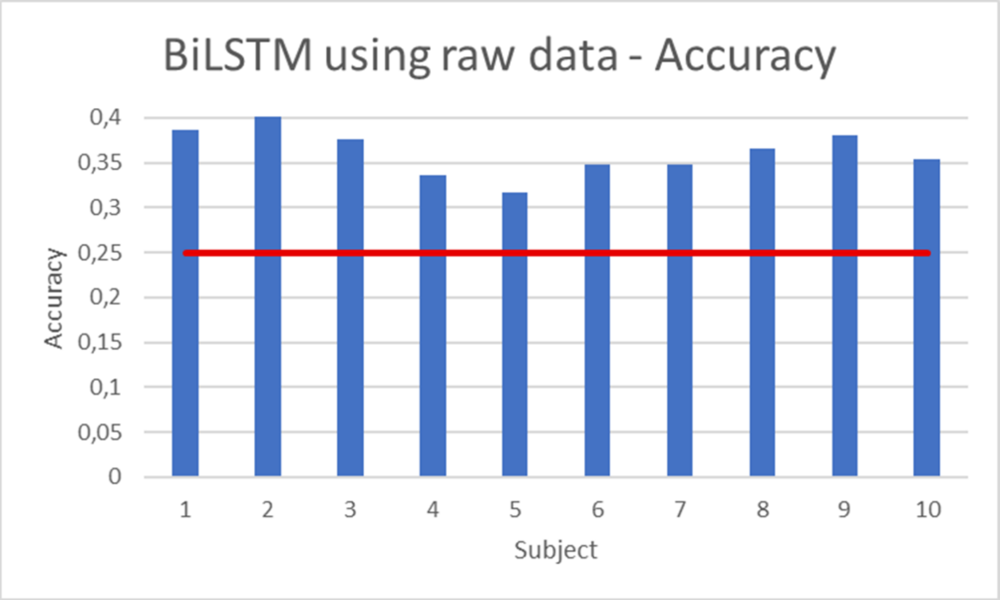

Figure 4: Accuracy of the BiLSTM network for multiclass classification using the *Thinking out loud* raw data. The red line represents the chance level (25%).

Table 2
BiLSTM performance for each subject on the 4-class inner speech classification task using the *Thinking out loud* raw data.

Subject	**Accuracy** ± **std**	**Precision**	**Recall**	**F1-score**
Sub 1	38.60 ± 0.050%	40.37%	38.93%	37.87%
Sub 2	40.17 ± 0.029%	40.06%	40.43%	39.56%
Sub 3	37.60 ± 0.055%	38.25%	37.50%	35.50%
Sub 4	33.67 ± 0.072%	34.44%	33.69%	32.81%
Sub 5	31.67 ± 0.028%	32.13%	31.94%	31.06%
Sub 6	34.81 ± 0.057%	37.69%	33.00%	33.75%
Sub 7	34.83 ± 0.062%	36.00%	34.94%	34.31%
Sub 8	36.60 ± 0.058%	37.69%	36.88%	34.88%
Sub 9	38.00 ± 0.089%	38.19%	37.75%	37.31%
Sub 10	35.33 ± 0.043%	34.50%	34.94%	34.38%
Average	36.12 ± 0.054%	36.93%	36.00%	35.14%

Instead, Table 3 shows a comparison of our proposed approaches and works in the literature using the *Thinking out loud* dataset.

The *Imagined Speech* dataset was chosen to confirm the validity of the model that obtained the best results on the *Thinking Out Loud* dataset. Figure 5 shows the results obtained for each subject. All the subjects achieve an accuracy above chance (represented by the red line - 16.67%). The mean accuracy considering all the subjects is 25.10% (± 0.045 std).

Table 4 shows a comparison of our proposed approach and works in the literature using the *Imagined Speech* dataset.

Table 3
Comparison of our ML and DL models with results in the literature using the *Thinking Out Loud* dataset.

Classifier	Input Data	Accuracy
SVM	PSD Features (channels left hemisphere) + PCA (0.99)	26.20%
XGBoost	PSD Features + PCA (0.99)	27.90%
LSTM	most important features	30.40%
	Raw data (all channels)	27.20%
	Raw data (channels most important features)	26.70%
BiLSTM	most important features	31.30%
	Raw data (all channels)	**36.10%**
	Raw data (channels most important features)	33.10%
EEGNet	Raw Data (channels left hemisphere)	29.67% [6]

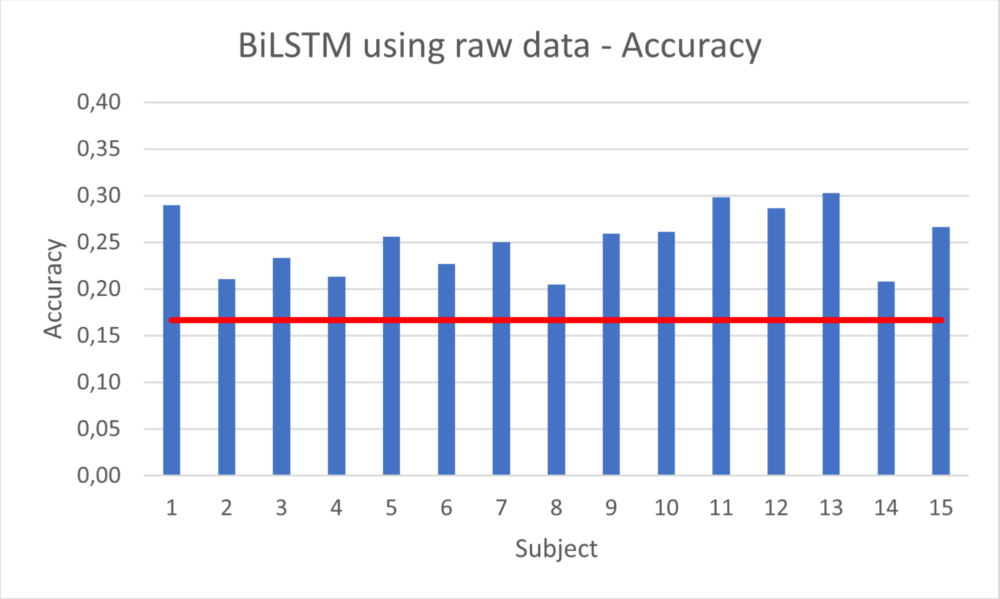

Figure 5: Accuracy of the BiLSTM network for multiclass classification using the *Imagined Speech* raw data. The red line represents the chance level (16.7%).

6. Conclusions

Inner speech recognition decoding EEG signal is still an open field of research. Few datasets are available in the literature and the classification performance, even if above chance, is still very low. The results obtained with this work confirm that the adoption of a BiLSTM architecture increases the performance of classification with respect to those in the state-of-the-art. In particular, the model designed for the *Thinking Out Loud* dataset was tested on the *Imagined Speech* one, acquired using a similar experimental protocol but with lesser electrodes, and thus confirming the validity of our proposal. The best classifier is obtained considering raw data of all channels, denoting that a deeper analysis on the most significant features should be performed, recalling that inner speech recognition should be considered in BCI applications

Table 4
Comparison of our DL model with results in the literature using the *Imagined Speech* dataset.

Classifier	Input Data	Accuracy
BiLSTM	Raw data (all channels)	**25.10%**
RF	Relative Wavelet Energy (RWE)	18.58% [25]
EEGNet	Raw Data (all channels)	24.97% [26]

where classification should be performed in real time. In future works, a parameter optimisation will be performed considering a subject-based approach, to further increase classification performance.

Another future development of this work would be to test the API proposed by [28] on the datasets analysed in the present paper to provide an insight on speech activity recognition and to extend it to multiclass identification of words in real-time.

Finally, the need of more numerous and less noisy datasets is crucial for further development in this field of research.
In fact, the studies on speech related tasks may benefit from a collection of a larger pool of data or the introduction of data augmentation approaches. However, it will be of fundamental importance to provide an unbiased data augmentation that may be based on data driven approaches like the empirical mode decomposition described in [29].

References

[1] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, J. S. Brumberg, Biosignal-Based Spoken Communication: A Survey, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (2017) 2257–2271.

[2] P. Saha, M. Abdul-Mageed, S. Fels, Speak your mind! towards imagined speech recognition with hierarchical deep learning, arXiv preprint arXiv:1904.05746 (2019).

[3] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, R. Spies, Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition, Scientific Data 9 (2022) 1–17.

[4] X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T.-P. Jung, C.-T. Lin, 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 (2021) 1645–1666.

[5] B. Alderson-Day, C. Fernyhough, Inner speech: development, cognitive functions, phenomenology, and neurobiology., Psychological bulletin 141 (2015) 931.

[6] B. van den Berg, S. van Donkelaar, M. Alimardani, Inner Speech Classification using EEG Signals: A Deep Learning Approach, in: 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), IEEE, 2021, pp. 1–4.

[7] E. Amit, C. Hoeflin, N. Hamzah, E. Fedorenko, An asymmetrical relationship between verbal and visual thinking: Converging evidence from behavior and fMRI, NeuroImage 152 (2017) 619–627.

[8] F. Bocquelet, T. Hueber, L. Girin, S. Chabardès, B. Yvert, Key considerations in designing a speech brain-computer interface, Journal of Physiology-Paris 110 (2016) 392–401.

[9] S. Martin, I. Iturrate, J. d. R. Millán, R. T. Knight, B. N. Pasley, Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis, Frontiers in neuroscience 12 (2018) 422.

[10] J. T. Panachakel, A. G. Ramakrishnan, Decoding covert speech from EEG-a comprehensive review, Frontiers in Neuroscience (2021) 392.

[11] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, R. Srinivasan, Toward EEG sensing of imagined speech, in: International Conference on Human-Computer Interaction, Springer, 2009, pp. 40–48.

[12] K. Brigham, B. V. Kumar, Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy, in: 2010 4th International Conference on Bioinformatics and Biomedical Engineering, IEEE, 2010, pp. 1–4.

[13] S. Deng, R. Srinivasan, T. Lappas, M. D'Zmura, EEG classification of imagined syllable rhythm using Hilbert spectrum methods, Journal of neural engineering 7 (2010) 046006.

[14] C. S. DaSalla, H. Kambara, M. Sato, Y. Koike, Single-trial classification of vowel speech imagery using common spatial patterns, Neural networks 22 (2009) 1334–1339.

[15] C. S. DaSalla, H. Kambara, Y. Koike, M. Sato, Spatial filtering and single-trial classification of EEG during vowel speech imagery, in: Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology, 2009, pp. 1–4.

[16] B. M. Idrees, O. Farooq, Vowel classification using wavelet decomposition during speech imagery, in: 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2016, pp. 636–640.

[17] A. Riaz, S. Akhtar, S. Iftikhar, A. A. Khan, A. Salman, Inter comparison of classification techniques for vowel speech imagery using EEG sensors, in: The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), IEEE, 2014, pp. 712–717.

[18] J. Kim, S.-K. Lee, B. Lee, EEG classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition, Journal of neural engineering 11 (2014) 036010.

[19] P. Suppes, Z.-L. Lu, B. Han, Brain wave recognition of words, Proceedings of the National Academy of Sciences 94 (1997) 14965–14969.

[20] L. Wang, X. Zhang, X. Zhong, Y. Zhang, Analysis and classification of speech imagery EEG for BCI, Biomedical signal processing and control 8 (2013) 901–908.

[21] M. Salama, L. ElSherif, H. Lashin, T. Gamal, Recognition of unspoken words using electrode electroencephalograhic signals, in: The Sixth International Conference on Advanced Cognitive Technologies and Applications, Citeseer, 2014, pp. 51–5.

[22] K. Mohanchandra, S. Saha, A communication paradigm using subvocalized speech: translating brain signals into speech, Augmented Human Research 1 (2016) 1–14.

[23] E. F. González-Castañeda, A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, Sonification and textification: Proposing methods for classifying unspoken words from EEG signals, Biomedical Signal Processing and Control 37 (2017) 82–91.

[24] C. H. Nguyen, G. K. Karavas, P. Artemiadis, Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features, Journal of neural engineering 15 (2017) 016002.

[25] G. A. P. Coretto, I. E. Gareis, H. L. Rufiner, Open access database of EEG signals recorded during imagined speech, in: 12th International Symposium on Medical Information Processing and Analysis, volume 10160, SPIE, 2017, p. 1016002.

[26] C. Cooney, A. Korik, R. Folli, D. Coyle, Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG, Sensors 20 (2020) 4629.

[27] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, Journal of neural engineering 15 (2018) 056013.

[28] L. A. Moctezuma, M. M. Molinas Cabrera, Towards an API for EEG-based imagined speech classification, in: ITISE 2018-International Conference on Time Series and Forecasting, 2018.

[29] J. Dinarès-Ferran, R. Ortner, C. Guger, J. Solé-Casals, A new method to generate artificial frames using the empirical mode decomposition for an EEG-based motor imagery BCI, Frontiers in neuroscience 12 (2018) 308.