Recognition of grammatical classes of imagined speech words using a convolutional neural network and brain signals

Denise Alonso-Vázquez [1] **Tonatiuh Hernández-del-Toro** [2] **Omar Mendoza-Montoya** [1] **Ricardo Caraza** [3] **Hector R. Martinez** [3] **Carlos A. Reyes-García** [2] **Javier M. Antelis** [1]

Abstract

In this paper, we analyze in time domain the signals acquired with 32 electroencephalography (EEG) channels from 10 healthy participants obtained during the imagined speech task of words in Spanish. We performed a statistical test to determine the location in space and time of the differences produced by imagining words from two grammatical classes: decision adverbs and nouns. Based on the statistical test results and using the EEGNet convolutional neural network, we evaluated three different data window sizes for the classification of the two grammatical groups. In the larger window W1 (700ms), we obtained an accuracy of 60.1%, while in the smaller window W3 (200ms), the accuracy obtained was 69.5%. This work is a first approach for the decoding of imagined speech words that are intended to be implemented in a brain-machine interface focused on patients with amyotrophic lateral sclerosis.

1. Introduction

There are diseases in which motor neurons progressively die, and consequently, the ability to communicate through speech is partially or completely lost. An example is the neurological disorder amyotrophic lateral sclerosis (ALS), in which approximately 85% of patients experience symptoms of bulbar dysfunction such as decreased verbal communication and swallowing function (Lee et al., 2021). This limits the ability to speak and significantly affects the quality of life. Therefore, technological alternatives are needed to help these patients to recover communication, such as those offered by Brain-Machine Interfaces (BMI) (Rezeika et al., 2018). A BMI detects and quantifies features of brain signals that indicate user intent, translates these measurements in real-time into device commands, and provides concurrent feedback to the user (Wolpaw, 2013).

The most widely used paradigms in the development of BMIs are the evoked potentials, such as the P300 potential or the steady-state visual-evoked potential (SSVEP), and others with cognitive tasks, such as motor imagery (MI) (Rezeika et al., 2018). One of the limitations of these well-established traditional BMI paradigms is that they do not directly decode the speech-related response. Therefore, it is necessary to explore new paradigms that seek to remedy this limitation. Imagined or covert speech is the voluntary imagination of speaking without moving any articulator and, therefore, not making any sound (Panachakel & Ramakrishnan, 2021). Several previous works (Sarmiento et al., 2021; Cooney et al., 2020; García-Salinas et al., 2019; Nguyen et al., 2017; Sereshkeh et al., 2017) have investigated the decoding of imagined speech vowels, phonemes, and words from EEG signals as a potential paradigm for BMI systems. Some of the most widely used machine learning algorithms for decoding imagined speech are CNN, SVM, and LDA (Panachakel & Ramakrishnan, 2021).

In addition to word recognition, decoding grammatical classes is interesting for the development of BMIs that decode speech. Differences in processing between word classes have been studied for many years, and different grammatical classes have been shown to be processed in different brain regions; typically, this is mainly studied between nouns and verbs (Bierwisch, 1999). One way to study these differences is using the Event-Related Potential (ERP) waveform, which represents changes in voltage recorded at the scalp over time that reflect sensory, cognitive, affective, and motor processes elicited by a stimulus (Luck & Kappenman, 2011). In (Datta & Boulgouris, 2021), ten words in English are used, and they do a recognition by grammatical class between verbs and nouns, finding differences between the ERPs of each class of around 250ms. The ERP com-

*Equal contribution [1]Escuela de Ingeniería y Ciencias, Tecnologico de Monterrey, Monterrey, N.L., México [2]Departamento de Ciencias Computacionales, Instituto Nacional de Astrofisica Optica y Electronica, San Andrés Cholula, Puebla, México [3]Escuela de Medicina y Ciencias de la Salud, Tecnologico de Monterrey, Monterrey, N.L., México. Correspondence to: Denise Alonso-Vázquez <denise.alonso.v@tec.mx>, Javier M. Antelis <mauricio.antelis@tec.mx>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

ponent is a voltage change that reflects a specific neural or psychological process; the most used components in language are around 400ms (N400) and 600ms (P600) (Luck & Kappenman, 2011).

Most works related to imagined speech have been carried out with non-invasive acquisition methods (Panachakel & Ramakrishnan, 2021), resulting in a lower spatial resolution. Consequently, a gap exists between previous works and a BMI commanded by imagined speech. Distinguishing between grammatical groups using EEG signals contributes to decoding imagined speech words. Therefore, finding features that maximize the separability between classes is essential. In this article, we perform a hypothesis test to determine the temporal and spatial location of significant differences in EEG signals while performing the imagined speech task of adverbs with respect to decision nouns. Based on the results of the statistical tests, we use the EEGNet convolutional neural network to determine a specific time window in which these two grammatical classes are classified with the best performance.

2. Methods

2.1. Data description

Ten healthy (without any diagnosed psychological or neurological disorder) participants (S1-S10), Spanish language native speakers, 8 men and 2 women with a mean age of 24 years (std=4 years), voluntarily participated in this work. The signals of 32 active EEG electrodes (Ag/AgCl) distributed uniformly over the scalp were acquired according to the 10-20 system. The recording equipment used was the biosignal amplifier g.HIAMP (from g.tec), the ground electrode was placed on the AFz position while the reference was on the right ear lobe. The data was recorded at a sampling rate of 1200Hz, a Butterworth band-pass filter from 0.5Hz to 500Hz, and a Notch filter in 60Hz were applied.

The experiment consisted in the mental speech imagination of four words. The participants sat in front of an 18.5-inch computer monitor, which showed visual stimuli that guided and controlled the execution of the experiment. Participants were instructed that when a word appeared on the screen, they had to perform the imagined speech of that word, defined as follows: pronounce the word internally without making any sound or gesturing any movement.

A trial begins with a fixation cross presented during 3 seconds which instructed to pay attention by focusing the gaze on the cross symbol while relaxing. Later, one of the four words to imagine appears randomly, which instructed to perform the speech imagination task. The word is presented during three seconds. Finally, the image of a palm tree is presented on the screen, which indicated to relax from the experiment.

The four words in Spanish were *si*, *no*, *agua*, and *comida*, meaning *yes*, *no*, *water*, and *food*, respectively. We selected these words because we consider that they are helpful for the care of patients with communication limitations, in addition to the fact that they represent two grammatical groups: adverbs and nouns. Each participant performed forty speech imagination for each word, resulting in 160 trials.

2.2. EEG Data Preparation and Processing

The recorded EEG data was segmented in epochs from -1.5s to 1.5s with respect to the time instant where the word is presented on the screen. Afterwards, the EEG signals were band-pass filtered from 4Hz to 20Hz using a Butterworth-type eighth-order filter and sub-sampled to 256Hz. For each epoch, noisy channels were removed (maximum 4 channels per participant), and a threshold-based algorithm was used to remove noisy trials (based on the data distribution, it was considered a noisy trial if peak-to-peak voltage $\geq 150\mu$V and standard deviation $\geq 20\mu$V). Then, baseline correction was applied to each epoch using a pre-stimulus period of -400ms to -100ms. Finally, all epochs from the words *si* and *no* were grouped and labeled as *adverbs*, while all epochs from the words *agua* and *comida* were grouped and labeled as *nouns*. Therefore, there are two groups of 80 epochs each.

2.3. Statistical analysis

To determine the location in space and time in which the EEG signals are statistically different between the two grammatical classes, we implement a hypothesis test. The time segment used was from -0.2s to 1s, the data was subsampled at 80Hz, and we implemented the Wilcoxon rank sum test with α=0.05. Analysis was performed participant-specific to recognize potential features that can be used to decode imagined speech in real time for a BMI. Based on the results obtained in this analysis (see Section 3), three time windows will be proposed to be used in the classification of grammatical classes using EEGNet.

2.4. EEGNet

EEGNet is a compact convolutional neural network architecture for EEG-based BMIs, which can be used for different BMI paradigms and trained with a very limited amount of data (Lawhern et al., 2018).

2.4.1. Network architecture and hyperparameters

The EEGNet architecture, see Figure 1, is composed of three blocks: the first consists of a two-step convolutional sequence. It starts with a temporal convolution to learn frequency filters. For the first step of the block, we used

$F1 = 16$, where $F1$ is the number of temporary filters with a kernel size of $(1, 128)$, that is, half of the sampling rate. The second step is a depthwise convolution to learn spatial filters for each temporal filter to extract frequency-specific spatial filters efficiently. The size is $(C, 1)$, with C defined as the number of channels; therefore, $C = 32$. The depth parameter D, set in $D = 10$, which is the number of spatial filters to learn within each temporal convolution. This part of the first block (the combination of temporal filtering with spatial filtering) is inspired by the Filter-Bank Common Spatial Pattern (FBCSP) algorithm (Ang et al., 2012). Subsequently, Batch normalization is applied between the dimension of the feature maps before applying the exponential linear unit (ELU) nonlinearity and the Dropout layer to regularize the model, with a probability set at 0.85. The block ends with an average pooling layer of size $(1, 4)$ to reduce the sampling rate of the signal to 64Hz.

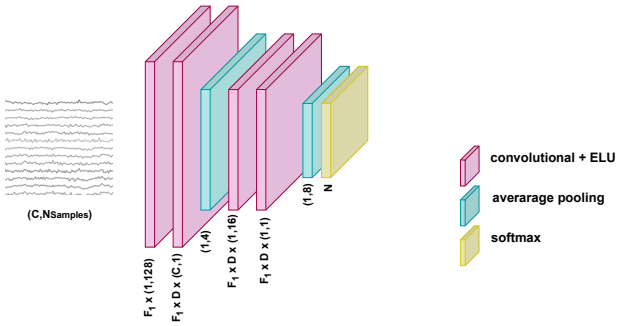

Figure 1. Architecture of the convolutional neural network EEGNet.

The second block is composed of a separable convolution formed by the combination of a deep convolution (size $(1, 16)$) and a point wise convolution, with $F2 = F1 * D$ where $F2$ is the number of point filters to learn. As in the previous block, batch normalization is applied to the entire dimension of the feature maps obtained from the separable convolution. Then, the exponential linear unit (ELU) nonlinearity activation layer, and the Dropout layer to regularize the model, with a probability set at 0.85 are used. Those layers are followed by average pooling, set to a size of $(1, 8)$ to reduce dimension.

The third block corresponds to the classification made with the softmax function explained in Section 2.4.2. Detailed network information is available at (Lawhern et al., 2018).

2.4.2. Network training, classification, and validation procedure

The input vector to the EEGNet has dimension $\mathbf{x} \in \mathbb{R}^{(N_{Channels} \cdot N_{Samples})}$, where $N_{Channels} = 32$ is the number of EEG channels and $N_{Samples}$ is the number of samples in each time-window. $N_{Samples}$ will depend on the result of the statistical analysis obtained in Section 2.3 and the size of the windows will be defined in Section 3. The training was carried out intra-subject, which means that one model was made per each subject. A softmax layer with N units is used in the classification block, where N is the number of classes; therefore, N=2. The softmax function $g_n(T)$ assigns a probability to each class based on the values of the vector of outputs T (Hastie et al., 2009), expressed as follows

$$g_n(T) = \frac{e^{T_n}}{\sum_{l=1}^{N} e^{T_l}}. \quad (1)$$

We use Google Colab GPUs to train the network developed in Tensorflow using Keras API. The training was performed for each subject, and the number of epochs was 300. The Adam optimizer and the categorical cross-entropy loss function were used to fit the model. A cross-validation of five folds was performed for each participant. To evaluate the performance of the model, we use classification accuracy.

3. Results

Figure 2 shows the number of participants in each channel-time pair that presented significant differences between the two grammatical classes. The darker color represents a higher concentration of subjects with significant differences at that time and in that channel. These results show that between 200ms and 300ms, most participants have differences in the EEG signal between the imagined pronunciation of adverbs and nouns in most channels. In the literature, it has been proposed that two brain areas are mainly related to

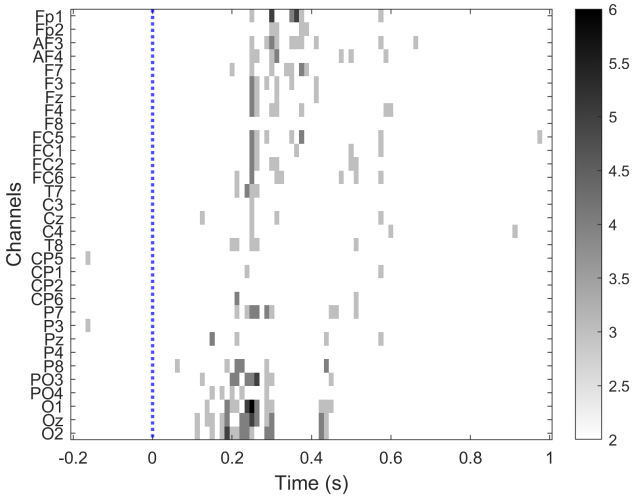

Figure 2. Results of the Wilcoxon hypothesis test with α=0.05, a darker color indicates a greater accumulation of subjects with the same significant differences

language processing: Broca's area, associated with the language production system, and Wernicke's area, associated with word recognition and the association of words with other information (Ardila et al., 2016). According to the positions of the electroencephalography channels used in this study, Broca's area is covered by: F7, F3, F4, F8, FC5, FC1, FC2, and FC6. The channels that cover Wernicke's area are: T7, C3, C4, T8, CP5, CP1, CP2, CP6, P7, P3, P4, and P8. Between 100ms and 200ms, there are differences in the parietal-occipital and occipital regions, that is, in the PO3, PO4, O1, Oz, and O2 channels. Around 300ms and 400ms, there are differences in the channels related to Broca's area. Finally, between 400ms and 600ms, the differences spread out spatially and temporally.

In Section 1, it was mentioned that around 250ms, 400ms, and 600ms post-stimulus are time segments of interest in language processing. To study the entire time window from the stimulus appearance and including the activity around 600ms, the W1 window size was chosen, which ranges from 0 to 700ms. The second window, W2, takes the time segment from when significant differences begin to be observed in 100ms up to 550ms. The third window, W3, takes the time segment in which the greatest number of significant differences are concentrated in most of the channels, that is, from 175ms to 375ms. Therefore, $N_{SamplesW1} = 180$, $N_{SamplesW2} = 116$, and $N_{SamplesW3} = 52$ are the number of samples in each time-window.

Figure 3 shows the results obtained by evaluating three different time windows in the classification task of two grammatical classes using EEGNet. According to the boxplots, the data distribution in W1 and W2 is similar; however, their means and medians are located at different accuracy values. Using the W1 time window that contains the information of all the potentials evoked by the stimulus, an average accuracy of 60.1% was obtained. W2 consisted of reducing the W1 time window by 36% fewer data, by eliminating the first 100ms of the signal after the stimulus appeared and limiting it to 550ms. In this case, W2 obtained an average accuracy of 64.4%. For the W3 time window, the information was further limited only to data around 200ms and 300ms, representing only 29% of the original W1 data, obtaining an average accuracy of 69.5%. Although there are outliers in the boxplot that contains the results of W3, the variability of the data is less than the results of W1 and W2.

4. Conclusions

In this work, we recognized two grammatical classes: decision adverbs and nouns, from electroencephalography signals recorded during imagined speech task. We implemented a statistical test to determine the spatial and temporal location of the significant differences between the two grammatical groups. Based on the results obtained in the statistical test, we evaluated three different time window sizes: W1 contains the first 700ms of the post-stimulus signal, W2 reduces the original window size by 36%, and W3 by 71%. We use the EEGNet convolutional neural network and evaluate its classification performance in the three time windows. Using the shorter W3 time window, we obtained an average accuracy of 69.5%, while using the original W1 time window, we obtained 60.1%. These results indicate that between 200ms and 300ms post-stimulus, the language process takes different paths depending on the grammatical class, and these differences are separable using a CNN classification model like EEGNet.

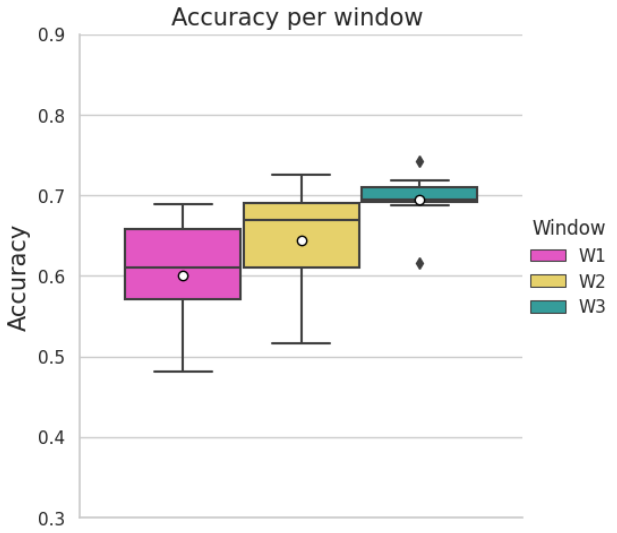

Figure 3. Classification accuracies for recognition of grammatical classes in three time windows. The line in the box is the median, and the empty circle is the average.

References

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., and Zhang, H. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in neuroscience*, 6:39, 2012.

Ardila, A., Bernal, B., and Rosselli, M. How localized are language brain areas? a review of brodmann areas involvement in oral language. *Archives of Clinical Neuropsychology*, 31(1):112–122, 2016.

Bierwisch, M. Words in the brain are not just labelled concepts. *Behavioral and Brain Sciences*, 22(2):280–282, 1999.

Cooney, C., Korik, A., Folli, R., and Coyle, D. Evaluation of hyperparameter optimization in machine and deep learn-

ing methods for decoding imagined speech eeg. *Sensors*, 20(16):4629, 2020.

Datta, S. and Boulgouris, N. V. Recognition of grammatical class of imagined words from eeg signals using convolutional neural network. *Neurocomputing*, 465:301–309, 2021.

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

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15 (5):056013, 2018.

Lee, J., Madhavan, A., Krajewski, E., and Lingenfelter, S. Assessment of dysarthria and dysphagia in patients with amyotrophic lateral sclerosis: Review of the current evidence. *Muscle & Nerve*, 64(5):520–531, 2021.

Luck, S. J. and Kappenman, E. S. *The Oxford handbook of event-related potential components*. Oxford university press, 2011.

Nguyen, C. H., Karavas, G. K., and Artemiadis, P. Inferring imagined speech using eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering*, 15(1):016002, 2017.

Panachakel, J. T. and Ramakrishnan, A. G. Decoding covert speech from eeg-a comprehensive review. *Frontiers in Neuroscience*, 15:392, 2021.

Rezeika, A., Benda, M., Stawicki, P., Gembler, F., Saboor, A., and Volosyak, I. Brain–computer interface spellers: A review. *Brain sciences*, 8(4):57, 2018.

Sarmiento, L. C., Villamizar, S., López, O., Collazos, A. C., Sarmiento, J., and Rodríguez, J. B. Recognition of eeg signals from imagined vowels using deep learning methods. *Sensors*, 21(19):6503, 2021.

Sereshkeh, A. R., Trott, R., Bricout, A., and Chau, T. Eeg classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2292–2300, 2017.

Wolpaw, J. R. Brain–computer interfaces. In *Handbook of clinical neurology*, volume 110. Elsevier, 2013.