UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI

Şcoala Doctorală de Electronică, Telecomunicații și Tehnologia Informației

Decizie nr. 972 din 08-12-2022

REZUMAT TEZĂ DE DOCTORAT

Ing. Ana-Luiza RUSNAC

Recunoașterea vorbirii imaginate prin analiza semnalelor EEG

Imaginary speech recognition by EEG signal analysis

COMISIA DE DOCTORAT

Prof. dr. ing. Ion MARGHESCU Univ. Politehnica din Bucureşti	Preşedinte
Prof. Dr. Ing. Ovidiu GRIGORE Univ. Politehnica din Bucureşti	Conducător de doctorat
Prof. dr. ing. Daniela TĂRNICERIU Univ. Tehnică „Gh. Asachi” din Iași	Referent
Prof. dr. ing. Ioan LIȚĂ Univ. din Pitești	Referent
Conf. dr. ing. Anamaria RĂDOI Univ. Politehnica din Bucureşti	Referent

BUCUREŞTI 2022

Cuprins

Chapter 1

Introduction

1.1 Presentation of the field of the doctoral thesis

In recent years, the brain-computer interface systems (BCI) have occupied an important role among the topics of interest of researchers all over the world. The system started from easier tasks, such as moving a wheelchair in the front-back and left-right directions, and tended towards more difficult targets, such as moving a prosthetic hand that achieves accurate movements of the phalanges or voice synthesis.

This field is in continuous growth due to the current trend to implement as many portable devices as possible with the aim of improving the quality of life of the users. An automatic imagined speech recognition device brings considerable value to patients suffering from conditions that affect the ability to communicate, such as brain attack, lock-down syndrome, etc., communication being a very important element in everyday life.

1.2 Scope of the doctoral thesis

This paper represents a study of the BCI systems capable of decoding signals acquired from the scalp during imaginary speech. Imaginary speech refers to imagining the thinking process of articulating the sound withouth the actual movement of the muscles involved in speech production. The developed imagined speech recognition systems presented in this thesys are non-invasive, using EEG signals, aiming to recognize in real-time a series of phonemes and words reaching the highest possible accuracy.

1.3 Content of the doctoral thesis

In the first part of the paper, ***Chapter 1***, general notions regarding the PhD thesis are presented, including a brief introduction to the intelligent imagined speech recognition systems together with the presentation of the field of study and its objectives.

Chapter 2 represents an introduction to the physyiology filed of the pronuntiation mechanisms, starting from the speech intention produced at the cortical level up to the speech articulation. In ***Chapter 3*** was made a short description of the different pronunciation mechanisms and the articulation method.

In ***Chapter 4*** we studied the state of the art of the imaginary speech recognition systems using cortical signals and their evolution was followed during the last years. The next two chapters, ***Chapter 5*** and ***Chapter 6***, focused on the description of the database used and its preprocessing method.

In ***Chapter 7*** are presented two methods of eye movement artefact removal: the first method is based on the implementation of an adaptive filter while the second method is based on the removal of the contaminated sources after decomposing the signal into principal components using Principal Component Analysys (PCA). A comparison study was made.

In ***Chapter 8*** we made a study of different pronunciation mechanisms based on (a) the articulation of the /iy/ phoneme (/iy/, /piy/, /tiy/, /diy/); (b) the articulation of the /uw/ phoneme; and (c) the articulation of the consonant /m/ and /n/, all phonemes being present in the Kara One database. To model the three pronunciation mechanisms, we used techniques widely applied in automatic speech recognition (ASR): the Mel-Cepstral coefficients (MFCC) and linear prediction algorithm (LPC). The comparative study was detailed in this chapter using Convolutional Neural Networks (CNN) in the classification stage.

Chapter 9 introduces the first system from the research process developed to recognize the phonemes presented in the Kara One database, based on MFCC coefficients and CNN neural network.

Chapter 10 introduces an analysis of the feature extraction methods: MFCC, cross-covariance in time-domain and cross-covariance in frequency domain using the unsupervised algorithm of Kohonen (self-organizing maps – SOM). In this chapter, the cross-covariance in frequency domain was first introduced for the imaginary speech recognition systems.

Chapter 11 followed to improve the automatic imaginary speech recognition. The design, implementation and analysis were made for all the phonemes and words from the Kara One database. The chapter presents a detailed analysis of the CNN architectures and hyperparameters used to achieve the best performances.

The next chapter, ***Chapter 12***, presents the proposed system of recognizing the eleven classes of the Kara One database. It uses the cross-covariance in frequency-domain in the feature extraction stage together with the CNNLSTM neural network for obtaining the best results of the study.

The last chapter, ***Chapter 13***, contains the final conclusions of the paper along with the original contributions made during the research study and the future prospects.

Chapter 2

Theoretical basis

2.1 Vocal signal production

Several studies regarding the vocal signal production [1], [2] confirmed the special importance of the temporal lobe in linguistic representation and understanding of concepts. Accordingly, the speech production mechanism starts from the temporal lobes and follows a route that allows the translation of thoughts into spoken words. The next element involved in the speech chain of articulation is the Broca area, which mainly plays a role in planning, initiating, and modifying the articulations needed in speech. In addition to Broca's area, the anterior insula participates in planning the positioning of the vocal tract elements for speech, the secondary motor area participates in the initiation of joint movements, and the facial primary motor cortex and the premotor cortex participate in the execution of the movements of the executive organs. The basal ganglia and the cerebellum are also involved in the completion of the speech act, which are activated to change the fundamental frequency, volume and rhythm of speech [2].

After planning and transmitting motor signals from the brain, they reach the effector organs. These organs are flexible, and their shape and size change depending on the signal received from the central nerves responsible for the articulation. Sound is formed starting from the lungs. The lungs provide the air force necessary to generate speech in acoustic form. Next, the air passes through the vocal tract, vocal cords, glottis, epiglottis, and other organs to reach further into the oral cavity in the form of an acoustic wave [3].

2.2 Producing, modeling and analyzing electrical activity of the brain

The nervous system can be divided into the central nervous system and the peripheral nervous system. The central nervous system is made up of the brain and spinal cord and has the role of interpreting sensory information and transmitting instructions to executive organs based on information from previous experiences [4].

The nervous tissue is made up of neurons, glial cells and endothelial cells. Its functional role is to receive and pass on electrical impulses that communicate messages regarding sensory, motor stimuli, or cognitive information.

Neurons are made up of body (or soma), axon and dendrites. The role of the axon is to transmit the received information unidirectionally to other nerve cells, while dendrites receive information from other neurons via synapses. In other words, the information is transmitted through the neuron starting from the dendrites, then is passing through the body of the neuron and finnaly reaching the axon. This exchange of information that takes place through the neuron can be recorded as an electrical signal by the electroencephalograph (EEG) [5].

2.3 Recording the neural activity of the brain

The transmission of information at the level of the nervous system is carried out using electrical impulses generated by (electro) biochemical processes. This activity can be recorded non-invasively, by positioning a set of electrodes on the surface of the scalp, method called electroencephalography (EEG) or invasively, which can be done using two methods: by positioning a set of electrodes directly on the surface of the cortex, method called electrocorticography (ECoG) or by inserting electrodes deep into the brain tissue, a method called stereoelectroencephalography (sEEG).

2.3.1 Electroencephalography (EEG)

Electroencephalography allows the observation of the brain electrical processes that take place on the surface of the cortex. Specifically, it is a measure of the electric field produced by a large number of simultaneously activated neurons as a function of time. The electrical activity measured on the surface of the scalp is the result of the excitation of tens of thousands of neurons present in the respective cortical region. An important element in the acquisition of EEG signals was the development of a system able to produce repeatability of the measurements. This is how the 10-20 system for electrode positioning was born [6]. The name 10-20 comes from the proportional distances, represented in percentages, relative to the cranial landmarks established as a standard.

2.3.2 Electrocorticography (ECoG)

ECoG is an invasive method of acquiring brain activity directly from the surface of the scalp in the operating room. This method acquires signals in a manner similar to EEG, with the mention that attenuation given by the skull and scalp is eliminated.

2.3.3 Stereoelectroencephalography (sEEG)

The discovery of ECoG led to the discovery of the possibility of using stereotactically positioned depth electrodes. They are especially used for the precise determination of the epileptic focus.

Chapter 3

Pronunciation mechanisms

Language is a communication system made up of articulated sounds, specific to people, through which they communicate. The way these sounds are articulated to express language-specific elements represents the pronunciation mechanism.

Sounds are generally divided into two basic categories: segmental and suprasegmental sounds. Segmental sounds include consonants and vowels while suprasegmental sounds are described by a series of phonetic parameters such as: tonality (fundamental frequency), intonation and accent.

Vowels are sounds pronounced without major obstructions of the vocal tract, so that the air leaving the lungs passes through the phonatory mechanism quite easily during speech. Example of vowel sounds: /iy/, /uw/, /ah/, /oh/.

Consonants, unlike vowels, involve obstructions or constrictions of the vocal tract, its elements changing their position to restrict air according to the desired utterance. For example, in the articulation of the consonant /p/ the lips are closed preventing the air to came out during speech [7].

3.1 The pronunciation of vowels

The pronunciation of vowels requires a greater opening of the vocal tract, unlike pronouncing consonants. There are two primary elements involved in the mechanism of speech: the shape and position of the tongue and the shape of the lips.

Next, the three mechanisms of vowel articulation will be analyzed in detail***: the opening*** (degree of opening of the oral cavity), ***the place of articulation*** and ***the roundness*** of the lips.

The opening or the degree of opening of the oral cavity describe the opening of the lips when the vowel is being articulated. This element also provides information about the frequency of the vowels', specifically, closed vowels have a higher frequency while open vowels have a lower frequency.

The place of articulation refers to the positioning of the tongue at the time of pronouncing the vowel. This can be anterior (/æ/) or posterior (/ɑ/). To differentiate between the two types of utterances, one can try the articulation of the English words pan (/pæn/) and palm (/pɑlm/). In this case, during the pronunciation of the word pan, the

tongue is in the front part of the oral cavity, unlike the palm, where the tongue is positioned in the back part of the oral cavity for the pronunciation of the vowel.

The roundness describes how the lips are positioned in a rounded shape or not during the pronunciation of the vowel.

3.2 The pronunciation of consonants

Consonants are sounds that are created by partially or totally restricting the passage of air through the vocal tract. This restriction is achieved by moving at least one component towards the other so that they touch or are very close. The moving part is called the active component and is represented by the lower articulators, while the fixed part is called the passive component and is given by the upper articulators.

Place of articulation

In pronouncing consonants, a very important element is the **place of articulation**. It represents the combination of an active articulator with a passive one.

The phonation

In addition to their role as an articulator, the vocal cords are also used to control the passage of air through the vocal tract. There are cases in which the vocal cords are positioned in a specific manner so the passage of air through the glottis allows them to vibrate.

When the vibration of the vocal cords is obtained at the time of pronouncing a consonant or vowel, the speech is called ***voiced*** speech, or otherwise, when no vibration of the vocal cords occurs, it is called ***unvoiced*** speech.

Manners of articulation

Consonants can also be classified by the manner of articulation, which refers to how air moves through the vocal tract, based on the size and shape of the constriction between the articulators.

The physiological processes involved in articulation are complex and all these processes are controlled by the brain. The way the muscles are commanded to utter each mechanism leads to complex brain activity during the execution of the movements. This is why the electrical activity of the brain recorded by the EEG signals during utterances can be considered to contain information about these processes.

Chapter 4

State of the art

Brain-Computer Interface (BCI) is a computer-based system that measures the neuronal activity of the central nervous system (CNS) and decodes it into a command capable of replacing, restoring, enhancing or supplementing the natural motor function of the CNS thereby modifying the primary interaction between the CNS and the external environment [8].

The most popular BCI systems are represented by the systems in which data acquisition is performed non-invasively, with the help of the surface electroencephalograph. These devices allow the measurement of neural activity by amplifying the potential differences between the electrodes placed on the scalp and the electric field emitted by the pyramidal neurons of the cerebral cortex [9].

For better spatial and temporal resolution, electrocorticographic (ECoG) signals are also used in BCI applications. These signals offer a very high SNR (signal-to-noise ratio). However, the big disadvantage of this method of acquisition is represented by its invasive nature because it requires direct contact between the electrodes and the cerebral cortex.

BCI for imaginary speech recognition is a system that acquires signals from the brain, processes the signals and encodes them further into speech synthesis, commands that actuate various devices or text. The principle underlying these types of BCI starts from the idea that in order to produce a word, the brain must transmit specific information to the motor elements of the vocal tract such as the tongue, jaw, lips, larynx, etc., in the same way that the brain sends signals for movement of other motor elements such as hand or foot.

In 2014, in study [10] the researchers aimed to differentiate the vowels „a”, „e”,„i”,„o” and „u” using EEG signals. They computed the mean, variance, standard deviation, and average power signal for differentiating the given five vowels. Signal classification was performed using a Multilayer Perceptron (MLP) neural network for each subject and the following results were obtained: 44% accuracy for the first subject and 32% for subjects 2 and 3.

Vowel recognition from EEG signals was further studied in 2016 in Colombia by Diego A. Rojas, Olga L. Ramos and Jorge E. Saby in study [11]. They used EMOTIV Epoch for signal acquisition, a simple and easy-to-wear device. Signals were acquired during the utterance of two vowels: "a" and "e". For feature extraction and selection, researchers used Symbolic Aggregate Approximation (SAA) together with Support Vector

Machine (SVM) algorithm for classification. The results obtained exceeded 84% for the differentiation between the vowels "a", "e" and the neutral signal, in which no vowel was spoken. The results of this study encourage the decoding of signals from EEG recordings, but it is necessary to mention that the differentiation was made only between two vowels, and the accuracy is expected to decrease significantly when introducing more vowels/consonants.

The significant development of this field in recent years has led to the encouragement of researchers to get involved in the creation of BCI systems for imaginary speech recognition by making available several databases for these applications. It is known that data acquisition represents a great challenge in this field regarding several points of view: professional equipment is needed, specialized knowledge is required for the correct positioning of the electrodes, and signals are difficult to acquire because of the special context that must be created in which the subject can focus specifically on the application, he must be completely rested, and he must take regular breaks for recovery.

There are currently several public access databases that can be used for this application. One of the open-source databases is provided by Chuong H. Nguyen, George K. Karavas and Panagiotis Artemiadis in study [12]. It contains signals acquired from 15 healthy subjects (11 men and 4 women) during the utterance of three groups of words: short words ("in", "out" and "up"), long words ("cooperative" and "independent") and vowels ("a", "i" and "u"). Each word was spoken 100 times in one recording step.

Another popular database in this field was developed at the University of Toronto by researchers Shunan Zhao and Frank Rudzicz [13]. This database contains signals collected from 12 subjects (8 males and 4 females) during the utterance of 7 phonemes ("iy", "uw", "piy", "tiy", "diy", "m" and " n") and 4 short words ("pat", "pot", "knew" and "gnaw"). The set of words and phonemes was uttered 12 times by each subject, reaching a total of 144 utterances for each phoneme and word.

In a recent study on imagined speech recognition conducted in Russia [14] the largest database was obtained by acquiring signals from 268 subjects for eight different Russian words: "forward", "back", "up", "down", "help", "take", "stop" and "release". Following the study, the researchers argued that it is more feasible to create a subject-dependent system that exhibits higher accuracy in comparison to developing a generalized system using signals acquired from a large number of different subjects.

Chapter 5

Databases

At the moment, data acquisition is still a challenge for researchers in this field. It is known that data acquisition represents a great challenge in this field regarding several points of view: professional equipment is needed, specialized knowledge is required for the correct positioning of the electrodes, and signals are difficult to acquire because of the special context that must be created in which the subject can focus specifically on the application, he must be completely rested, and he must take regular breaks for recovery.

5.1 Kara One Database (KODB)

Kara One database was developed by a research crew at the University of Toronto [13] in 2015. This database contains signals collected from four women and eight men, with an average age of 27.4 years. All participants who took part in the study are right-handed, have higher education, have no visual, auditory, or motor problems and have no history of neurological problems or drug use.

Subjects were instructed to follow the installed monitor and to stay still. A recording session lasted between 30 and 40 minutes, during which one of the 7 phonemes used for recognition could be seen on the screen: "iy", "uw", "piy", "tiy", "diy", "m", "n" or one of the words: "pat", "pot", "knew", "gnaw".

Each experiment consisted of four successive stages: (1) a 5s rest period, during which participants were instructed to relax and not think about anything; (2) a stimulus period, in which a text containing a phoneme, or a word appeared on the screen, together with an auditory stimulus corresponding to the stimulus on the screen. After the appearance of the auditory stimulus, followed a period of 2 seconds in which the subject was instructed to move his joints in the position necessary to start the pronunciation of the visual stimulus; (3) a 5s period in which each participant was instructed to imagine saying the word; (4) a period in which the subject spoke the word aloud, and the Kinect sensor recorded both the vocal signal and the facial features.

Each visual stimulus was presented 12 times, resulting in 132 trials. Finally, 4 of the 12 subjects were removed from the study because they had detached electrodes, and two of the subjects fell asleep during the recordings.

Chapter 6

Preprocessing the KODB

The purpose of the research done by the author of this paper was to identify words based on EEG signals, words spoken during imagined speech. To achieve the proposed goal, for the carried out studies made in this work only the 5s signals corresponding to the mental utterance of phonemes and words from the Kara One Database were segmented for further use. In order to eliminate transitions from one state to another, the first and last 0.5s were further removed from the 5s of signal, finally obtaining for each stimulus a 4s EEG signal.

The obtained signals were further visually analyzed by an expert. In the first stage of visual analysis, six of the 14 recording sessions were found to have very high noise or unattached ground wires, giving a signal that could not be used in imaginary speech recognition. For this reason, all these subjects were removed from the study.

Afterwards, in this stage a visual analysis of all EEG signals corresponding to the imagined speech followed and the recordings containing loud noises, generally due to subject movement, were removed from the study.

Following this process of visual analysis of the signals, we finally obtained the cleared database containing a total of 993 signals that were further used during the study.

Finally, all remained signals from the Kara One database were filtered using a 60Hz Notch filter to remove the power line artifacts.

Chapter 7

Ocular artefacts removal [15]

EEG signals are low-amplitude signals, having a range between 5 and 200 µV [16]. Due to this low-amplitude these signals can easily be contaminated by other biological signals such as electrocardiographic signals, electromyographic signals, eye movement signals, etc. In this chapter, we aimed to compare two methods for ocular artefacts removal: the adaptive filter and the PCA method, because ocular artifacts have the greatest influence on EEG signals.

7.1 Adaptive filter

An often used method by the researchers to improve the quality of EEG signals consists of implementing adaptive filters. The great advantage of this filtering method consists in adapting the coefficients progressively, taking into account the statistics of the signal at each time. In the present study, an adaptive filter of size 400 coefficients was used using root mean square error as training algorithm. Since this algorithm is one that performs well over time, the 4 seconds of signal of each spoken recording from the KODB database was expanded to 60s by successively concatenating the same recording multiple times.

7.2 PCA

The algorithm aims to find a matrix of coefficients containing the uncorrelated sources in the signal. The first step of the algorithm was to compute the covariance matrix of the recorded EEG signal with respect to the features. Next, after computing the covariance matrix, the eigenvectors and eigenvalues of the matrix will be easily obtained. Finally, the resulted vectors are corresponding to the principal components of the signal.

7.3 Results

After applying the adaptive filter, an improvement in the EEG signal qualities can be observed by attenuating the ocular component in the frontal channels. **Figure 7.3** highlights the best this result.

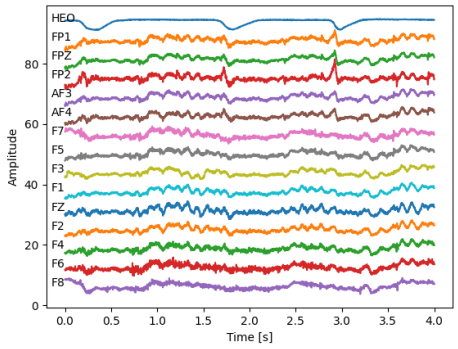

Figure 7.3 *EEG signal after applying adaptive filter – channels FP1-F8 and HEO*

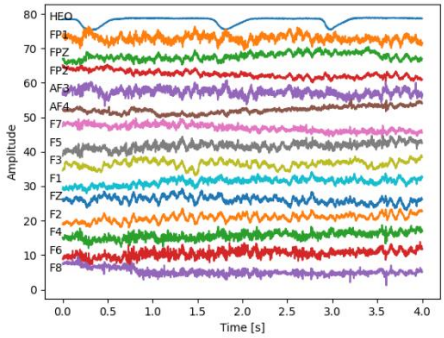

Figure 7.5 *EEG signals after applying the first two PCA components – channels FP1-F8 and HEO*

The second method of removing ocular artifacts was based on the computation of principal components. After obtaining the principal components, we observed that the first two components were very similar to the HEO signal. We further eliminated these components and reconstructed the signal. The results are presented in **Figure 7.5**. As can be seen, the ocular artefact were eliminated along with the first two principal components.

Frontal channel signals are most affected by these eye movements. The correlations of these channels with the HEO signal are very high, reaching values up to 0.8. After applying the adaptive filter, this correlation value decreased to approx. 0.1, and using the PCA filter at 0.2. Next, a quantitative analysis based on inter-class correlation was performed. It could be seen that for PCA, correlations between records of the same class increased for most phonemes and words, while the adaptive filter decorrelated these records.

7.4 Conclusions

This chapter aimed to compare two different methods of ocular artifacts removal: adaptive filtering and PCA for the further use of filtered signals in imagined speech recognition applications. Following this study, we concluded that both methods attenuated the ocular artifacts by decorrelating signals with HEO. Initially, it started from a very high correlation between the signals acquired from the frontal electrodes and the HEO, having values exceeding 0.8 and reaching values of approximately 0.1 for the adaptive filter and approximately 0.2 after filtering with PCA. It was observed that the signals after being filtered using the PCA method showed a higher inter-class correlation, unlike the signals obtained after applying the adaptive filter, where a decorrelation appeared between them.

Chapter 8

Pronunciation mechanisms recognition system [19]

In this chapter we proposed a BCI system for recognizing phonemes from the KODB database grouped into three categories: C1, phonemes containing the vowel /iy/, C2, the phoneme /uw/ and C3, the consonants /m/ and /n/. To achieve the proposed goal, we used the KODB preprocessed method described in ***Chapter 6***. In the feature extraction stage, we compaired four types of features based on the computation of the MFCC and LPC coefficients. The four methods analyzed in this chapter were: (1) MFCC coefficients: MFCC (size 62x62); (2) LPC coefficients: LPC (size 62x62); (3) concatenation of MFCC and LPC into a matrix of size 62x36 over which the covariance was computed: MFCC+LPC V1 (size 62x62); (4) concatenation of MFCC and LPC into a three-dimensional matrix: MFCC + LPC V2 (size 62x62x2).

8.1 MFCC coefficients

The input signals, corresponding to the EEG signals acquired during imagined speech, are transformed using FFT in the frequency domain. Over the signal spectrum, a bank of triangular filters are applied whose bandwidths are computed using the Mel scale. For each triangular filter, the spectral energy is obtained as the sum of the squared samples. Finally, the MFCC coefficients will be computed by converting the logarithm of the previously calculated coefficients from the Mel bands to the time domain using iFFT.

8.2 LPC

The linear prediction algorithm is a well-known technique in automatic speech recognition because provides important information both in the time domain and in the frequency domain [20]. The widespread use of the LPC algorithm in the recognition of speech signals is based on the ability of this method to extract the essential information from the signal and to provide a small number of parameters that describe the configuration of the vocal tract during speech [21]. In the developed work, 18 LPC coefficients were computed for decoding EEG information of seven imagined phonemes grouped into three classes.

8.3 Data augmentation

After clustering the phonemes into the desired three classes, the problem of imbalanced number of vectors appeard. To solve this problem, a data augmentation step was introduced. In this step vectors from clusters C2 and C3 were artificially generated to finally obtain 200 observations from each class, while for cluster C1 vectors were randomly removed to reach the same number equal to 200 observations from this category as well. Artificial generation of input data for C2 and C3 clusters was performed using a Gaussian distribution of each feature.

8.4 Classification

For this study, a CNN neural network containing two 2D convolutional layers with the number of filters 64 and 32 respectively and three fully connected layers of size 32, 16 and 3 neurons in the last layer equivalent to the three clusters was used in the classification stage. A training batch normalization layer was inserted after each convolutional layer. The activation function used was the hyperbolic tangent (tanh) for all layers except the last layer, where the activation function used was softmax.

8.5 Results

Following the obtained results, it can be argued that compared to LPC, the MFCC coefficients provided a better system performance reaching the average value of 0.39 accuracy, being more suitable for the classification of imaginary speech. It can also be seen that better results were obtained by combining the two features in a two-channel three-dimensional matrix (MFCC+LPC V2) compared to the MFCC + LPC V1 method. Using an i7-3537U processor with 6GB RAM and 2.5GHz clock frequency, we were able to achieve an average for phoneme recognition using MFCC features of 5.28s and 105.53s for LPC.

8.6 Conclusions

Following the results obtained, we concluded that the MFCC features provide a better understanding of imaginary speech, providing the best accuracy results (0.39) compared to the other features. We also observed that following the concatenation of the two types of features, MFCC and LPC, in a three-dimensional matrix the results were improved, obtaining an average value of 0.38.

Chapter 9

Phoneme recognition using MFCC and CNN [23]

In this chapter, we aimed to differentiate seven phonemes acquired during imaginary speech of KODB. A generalized subject-independent intelligent system based on the computation of 18 MFCC coefficients in the feature extraction stage and a CNN neural network in the classification stage was designed. The preprocessing stage followed the procedure detailed in ***Chapter 6***. In addition to these preprocessing steps, a bandpass filter with a bandwidth between 0.5 and 100 Hz was also introduced in this study. In this stage of the study the channels from the occipital area were removed, being located in the area of the visual cortex. In addition, channels in the frontal area were also removed because these channels are generally heavily affected by the eye movements. Thus, after removing the channels, the features extracted for 45 of the 62 channels were further used. These features were passed on to a CNN neural network for phoneme classification.

9.1 Results

The purpose of this chapter was to differentiate the seven phonemes acquired during the development of the Kara One database. For each signal in the database, MFCC coefficients were computed over the 45 channels resulting a matrix of size [45 x 18] for each utterance. The extracted features were passed through a CNN neural network having three 2D convolutional layers followed by max-pooling and two fully connected layers. Using this architecture, the best results achieved were 24.19% accuracy for the test set.

9.2 Conclusions

In this chapter, the development of an intelligent phoneme recognition system was pursued. The developed system is a generalized system relative to the subjects in the database. It was based on the computation of MFCC coefficients in the feature extraction stage, together with a CNN neural network in the classification stage.

The study showed that EEG signals contains hidden information regarding imagined speech, and features commonly used in automatic speech recognition, such as MFCC, contains essential markers regarding imagined speech as well.

Using 18 MFCC coefficients and a CNN neural network having three convolutional layers of size 64, 32 and 32, each of them followed by a max-pooling layer, and two fully connected layers of size 16 and 7 respectively, an accuracy on the test set of 24.19% was obtained.

The obtained results show that the use of the CNN neural network improves the performance of the system, by comparison with the study [24], in which the maximum accuracy reached only 20.80% using the same database and features based on MFCC coefficients, but using an SVM network in the classification stage. It should be noted that the study is specific to each subject, but all phonemes and words from the KODB database were used in the analysis, ultimately differentiating 11 classes.

After analyzing the confusion matrix, we could see that the system was able to best recognize /iy/, /uw/, /m/ and /n/, which have different pronunciation mechanisms. Most of the phonemes that could not be recognized, such as /tiy/ and /diy/, were confused with the phoneme /iy/, having similar pronunciation mechanisms. An item in the confusion matrix that was also of interest was the network error between the phonemes /piy/ and /m/. Analyzing in detail the pronunciation mechanism of the two phonemes that were confused, it can be argued that this is due to the fact that both consonants, /p/ and /m/, are pronounced through transient lip closures.

Preprocessed signals using the PCA filter to remove ocular artifacts led to a poorer response of the neural network, achieving an accuracy of 14.05%. Following the two classification studies of preprocessed EEG signals using the PCA method that led to a poorer system performance we can conclude that the removal of components leads to a decrease in hidden information in the EEG signals.

Chapter 10

Imaginary speech analysis and classification using SOM

The purpose of using SOM neural network was to perform a comparative analysis of several types of features in order to observe their behavior relative to the classification of imagined speech phonemes and words from the database. The features analyzed were: (1) MFCC coefficients; (2) Cross-covariance in time-domain; (3) Cross-covariance in frequency-domain. The study carried out in this chapter also aimed to analyze the response of the SOM network after introduceing in the preprocessing stage the eye movement artefact removal using the PCA algorithm. In the second stage of the study, this network was used to classify the input data by creating two-dimensional classification maps based on the response of the majority of winning neurons. The preprocessed signals were segmented into non-overlapping 0.25s windows, and 50% of the windows were randomly distributed in the training set, while the remaining 50% were distributed in the test set.

10.1 Cross-covariance in time-domain

Let cross covariance between two channels, c1 and c2, be described by:

$$Cov(X^{c1}(t), X^{c2}(t)) = E\big[[\,X^{c1}(t) - E(X^{c1}(t)][\,X^{c2}(t) - E(X^{c2}(t)]\big], \quad (10.1)$$

where $X^{c1}(t)$ is the acquired EEG signal for the c1 channel, $X^{c2}(t)$ is the acquired EEG signal for the c2 signal and $E[X^{\mathrm{ch}}(t)]$ is the average of the chanel ch (which can be c1 or c2).

10.2 Cross-covariance in frequency-domain

The FFT transform of a channel can be described by:

$$FX^{ch}(f) = \sum_{t=0}^{n-1} X_t^{ch} e^{-\frac{j2\pi ft}{n}} \quad (10.6)$$

The cross-covariance in frequency domain is computed as follows:

$$Cov(FX^{c1}(t), \mathrm{F}X^{c2}(t)) = E\big[[\,FX^{c1}(t) - E(\mathrm{F}X^{c1}(t)][\,FX^{c2}(t) - E(\mathrm{F}X^{c2}(t)]\big], \quad (10.2)$$

10.3 The representation of EEG signals feature extraction based on MFCC and SOM

After the visual analysis of the resulted feature space of the SOM network we observed that the MFCC features, in this case, do not provide a good separability of the classes. The winning majority neurons were distributed over the entire surface of the SOM map. The results could not be improved even after increasing the input space from (31, 31) to (62, 62). Also, no major differences can be observed between the unprocessed signals and the processed signals using the PCA method.

10.4 The representation of EEG signals feature extraction based on cross-covariance in time-domain and SOM

The qualitative results observed in this sub-chapter showed that the Kohonen neural network fails to separate the eleven classes of the database in the output space, the winning neurons being overlapped over the entire surface of the resulting map. According to the obtained maps, we can support the fact that there are no major differences between the representation of the feature space using the unprocessed signals versus the processed ones. The resulting classes were further distributed over the entire surface of the map.

10.5 The representation of EEG signals feature extraction based on cross-covariance in frequency-domain and SOM

The analysis carried out in this chapter showed that there are phoneme grouping areas and word grouping areas, but regions specific to each class cannot be identified. One can observe the dispersion of the classes on the entire map of the neural network, without the possibility of specific delimitation of the regions corresponding to the different classes. Increasing the output space did not significantly improve the mapping of the input data.

10.6 Classification of the input data using SOM

The best results were obtained after 100,000 iterations using a SOM neural network of size (31, 31) together with cross-covariance in the frequency-domain using signals processed using the PCA method. The accuracy reached a value of 28.49%.

10.7 Conclusions

It was observed from the qualitative analysis that the features computed in the frequency-domain presented better mapping in terms of class differentiation, but there were no clearly differentiated areas for each phoneme or word in the database for these features either. Regarding the classification, it can be considered that the input data corresponding to the cross-covariance in the frequency domain provided a better classification than those in the time-domain, raising the accuracy to a value of 0.25. By processing the signals using the PCA method, the system performance increased to 0.28.

Chapter 11

Word and phoneme recognition system from the KODB database using CNN [25]

This study was performed on eight different subjects and was designed as a subject's shared system. One of the goals was to compare two different types of feature extraction: time-domain and frequency-domain cross-covariance. Another direction we turned our attention during the development of the study was testing different analysis window lengths: 0.25s, 0.5s and 1s. In the second part of the work, we focused on testing different architectures of the CNN network used to classify the extracted features to determine which one best fits our application.

11.1 Signal classification

In the research carried out in this paper, we tested different architectures of CNN networks with the aim of finding the architecture that provides the best performance while also taking into account complexity, memory and runtime. We started with a low-complexity architecture, one convolutional layer and one fully connected layer (without the output layer), and increased the complexity to three convolutional layers and one fully connected layer with a larger number of filters and neurons. At this point, we considered that system performance does not improve, instead memory and runtime will be affected.

11.2 System performance metrics

To evaluate the performance of the system, a series of metrics such as accuracy, balanced accuracy, precision and sensitivity were computed in order to provide quantitative information regarding the degree of recognition of phonemes and words from the database.

11.3 Results

11.3.1 Analysing activation functions: Tanh and ReLU

The results obtained on the test set using different CNN network architectures and different activation functions for convolutional layers: hyperbolic tangent versus Rectified Linear Unit (ReLU) showed that the ReLU activation function provides significantly better results, raising the maximum accuracy from 0.3169 (tanh) to 0.3758 (ReLU). The results were obtained using the cross-covariance in the time domain over a 0.25s window as feature extraction method.

11.3.2 Feature extraction study: Time vs Frecquency

Next, the study aimed to compare the differences between the features computed in the time-domain and in the frequency-domain. A study of different CNN architectures shows that using two convolutional layers with filter numbers 64 and 128 connected to a fully connected layer of size 64 neurons works best for the features computed in the frequency-domain achieving the performance of 37% accuracy. In the time-domain, the best results were obtained using less complex architectures, with the best system performance captured by a single convolutional layer network having 64 filters and a fully connected layer with 64 neurons.

11.3.3 Window length analysis: 0.25, 0.5 and 1s

The next step was to test the network with different analysis window lengths applied to the input data: 0.25s, 0.5s and 1s. After the made study, we observed that the results obtained for the analysis window of 0.25s are the best, reaching an accuracy of 37%.

11.3.4 Mean filter comparison: B0, B3 and B5

Another study pursued at this stage of the work focused on applying an averaging filter on the spectrum before computing the covariance matrix on channels having kernels of different sizes: three and five samples. The results of the study showed no improvement in network accuracy for any of the kernel sizes. The maximum value when using the filter with the kernel equal to three samples being 0.2886, and for the kernel of five samples 0.2863.

11.3.5 System performance metrics

For a better understanding of the recorded results as well as the performance of the system, a series of new metrics were introduced: balanced accuracy, precision and sensitivity, computed according to the paper [26]. The results showed that there is no significant signal imbalance in the database.

11.3.6 Complexity and memory metrics

Using an AMD Ryzen 7 4800HS CPU system with 16 GB of RAM and 2.9 GHz clock frequency, an average recognition time of an input vector is 1.8×10^{-3}s. The time was obtained starting from the feature extraction stage until the decision making. Time was estimated using cross-covariance in frequency-domain over a 0.25s window fed into a CNN network having the C64-128/D64 architecture.

11.3.7 Comparison between raw signals and processed signals using PCA for eye movement artefact removal

After analyzing the results we observed that when using the PCA analysis for eye movement artefact removal the performance of the CNN network drops from 0.37 to 0.35, a different result than the one obtained using the SOM neural network.

11.4 Discussion

The comparison between the two methods of feature extraction, the cross-covariance in time and in frequency domain, showed that when using the frequency-domain the accuracy increased by approximately 16% reaching a value of 0.37 compared to 0.21, a value obtained when using time-domain features . Another element that this study pursued was the comparison of different analysis window sizes in order to observe the signal statistics for different time intervals. Following this comparison we concluded that the best analysis window is 0.25s. The accuracy obtained for this window length is significantly higher, reaching a value of 0.37, compared to 0.29 obtained after using a window of 1s. The difference between the performance obtained for a window of 0.25s and 0.5s is not significant, it drops by only 1%. The final study was based on testing different CNN architectures to observe the performance of the system and to model its final characteristics. We concluded that for features extracted in the frequency-domain (the features that also provided the best system performance) the best architecture used contains only two 2D convolutional layers having 64 and 128 connected filters with a fully connected layer containing 64 neurons.

11.5 Conclusions

The best results were obtained using the cross-covariance in the frequency-domain using an analysis window of 0.25s. The best performance of the system was achieved using a CNN with two 2D convolutional layers having 64 and 128 filters and a fully connected layer having the number of neurons equal to 64. Using these system characteristics, the maximum accuracy of the system reached a value of 37%. We have also shown that a smaller analysis window provides a better understanding of imagined speech. Finally, we can argue that the proposed system can be implemented on a low-cost portable device with limited resources to make decisions about the imagined pronunciation of phonemes or words.

Chapter 12

Word and phoneme recognition system from the KODB database using CNNLSTM [27]

In this chapter, the behavior of the imagined speech recognition system was tested using in the classification stage a Convolutional Neural Network that includes in the convolutional layers recurrent cells of the Long-Short Term Memory (CNNLSTM).

This study also highlighted the fact that using only signals acquired from the anatomical areas recognized for their implications in speech production: Broca's area, the primary motor cortex and the secondary motor cortex, approximately 93% of the information obtained from all electrodes is preserved.

12.1 Feature computation

The features computed for system development were chosen according to the results of the previous study. Acordingly, in this study, only the cross-covariance in frequency-domain for 0.25s segments was pursued. To provide the time variation needed for the CNNLSTM network, the 0.25s segments were in turn divided into 0.1s windows with 50% overlap.

Next, the computed features were qualitatively investigated using LDA to reduce the number to a two-dimensional space so that their visual inspection could be performed. After the visual and quantitative analysis of the previously mentioned features, we concluded that the ones computed using the cross-covariance in the frequency-domain best partitioned the feature space with respect to the imagined speech, and only these were used further in the study.

12.2 CNNLSTM classification

The neural network architecture is also based on the results of studies previously obtained in ***Chapter 11***. The best results obtained after testing several architectures and hyperparameters were obtained using two convolutional layers of size 64 and 128 connected to a fully connected layer containing 64 neurons, the output layer having 11

neurons corresponding to the number of classes. Finally the neural network was trained using the Adam optimizer with a learning rate of 0.0001 and using the cross-entropy error function.

12.3 Results

12.3.1 CNNLSTM vs CNN

This study aims to highlight the advantages of using the CNNLSTM neural network for recognizing spoken phonemes and words during imagined speech using cross-covariance in frequency-domain in the feature extraction stage. The results obtained at this stage showed an improvement over the CNN network, with accuracy increasing from 37% to 43%.

12.3.2 Brain regions analysis

The next study carried out in this chapter evaluates the performance of the system relative to reducing the number of electrodes so that only the signals taken from certain cranial regions are analyzed.

The regions were initially selected based on the major regions defined by the electrodes position: Frontal, Central and Occipital. Then the electrodes corresponding to the anatomical areas involved in speech production, starting from the conceptualization and planning of joint movements to the initiation and coordination of the neurons involved in the transmission of the electrical stimulus sent to the effectors. The results obtained for each area of the brain and the combinations made between these areas analyzed in the study showed that the best recognition rate has the electrodes positioned in the anatomical regions specific to speech, reaching a value of 0.4027.

12.3.3 Memory, computation and time execution study

This section of the chapter focused on studying the complexity and memory of the proposed system. The complexity of an intelligent system is generally given by the neural network. In the present case, the maximum complexity is given by the second convolutional layer and is of the form O(4x4x(2(N+3)2x64)xlog(2(N+3)2x64)x128). In terms of execution time, measured using an AMD Ryzen 7 4800HS processor with 16GB of RAM and 2.9 GHz clock frequency, the average time execution value of the vectors in the data set is 81.9ms.

12.3.4 Raw signal analysis vs processed signal using PCA

An analysis of the system performances using raw signals and processed signals using PCA for eye movement artefact removal was introduced in this chapter. The obtained results provide information similar to those in the previous chapter, where the CNN network was used for classification. The decrease in system performance when using processed signals can be explained by the same process in which deep learning neural networks use their own filters to extract from the signal the essential information for classification and to remove artifacts or elements that can disrupt the learning process.

12.4 Discussions

The major advantages of the LSTM neural network is the long-term memorization of input features. This ability has a significant value when analyzing non-stationary time-invariant signals such as EEG signals. This time-space connections helped the neural network to raise the accuracy from 0.37 to 0.43 using similar architectures and parameters. The average confusion matrix for the 4-folds shows that there is a fairly clear distinction between phonemes and words, they are very rarely confused with each other, the network confusion being between phonemes and between words.

The final goal of the developed system was to obtain the highest possible accuracy with the possibility of a real-time implementation using a portable device with limited resources. Next we studied the behavior of the system using a reduced number of electrodes located in specific areas. This helps with device portability and reduces development resources, but has the downside of a slight decrease in the system accuracy. By reducing the number of channels, the accuracy also decreased, which was an anticipated phenomenon. However, using electrodes from the specific anatomical areas involved in speech production, the system's accuracy reached a value of 0.40, a decrease of only 3% compared to using all channels. This means that 93% of the information of imagined speech is concentrated in these channels and only 7% of the information is distributed to the parietal and occipital regions.

An important aspect to consider when developing an imagined speech recognition system is the complexity and memory used. In general, the biggest consumer of resources is the neural network. The largest number of computational operations is given by the second layer of the CNNLSTM network and is the order of approx. O (6.3×10^{-9}). However, the execution time to make a decision about an input stimulus by the network is below 100ms, even using all channels in the computational process. These values indicate that the system can still be deployed in real time. Regarding the memory used, the system presents a limitation as a minimum of 2GB is required to retain only the weights of the neural network. This is due to the long-term memory of the network used.

12.5 Conclusions

This chapter showed an improvement in system performance when using the CNNLSTM neural network compared to the CNN neural network. Accuracy increased from 37% to 43% when using the CNNLSTM with no changes in the preprocessing or feature extraction chain. The developed system aimed to consider in the design the possibility of a real-time implementation on a portable device with limited resources. Therefore, a study was also carried out in terms of reducing the number of electrodes in the system. We concluded from the study that 93% of the information is concentrated in the anatomical regions specific to speech production, obtaining an accuracy of 40% for the use of 29 electrodes, compared to 62, which was their original number.

Chapter 13

Conclusions

In this paper, the development of an intelligent system for automatic recognition of imaginary speech was pursued. In order to achieve the proposed goal, a study of the pronunciation mechanisms of the utterance was made, starting from the intention of the articulation that occurs at the cortical level to the transmission of the electrical impulse to the effector organs involved in the utterance process.

This thesis used for the systems development the Kara One database acquired during the collaboration of the University of Toronto with Toronto Rehabilitation Center. The database contains signals acquired during the imagined speech of seven phonemes and four words. Next, the signals were analyzed and preprocessed in order to increase their quality. Preprocessing consisted of visually analyzing them by an expert and removing epochs containing noisy signals or electrodes with poor connectivity. The main artifacts of the EEG signals are given by the eye movement, because the electrical activity of the muscles at the level of the eyes is recorded, which has a higher amplitude than the EEG signal. This is why two methods of eye movement artefacts removal were implemented. The first method consisted of filtering them using an adaptive filter, and the second method was based on removing the signal sources containing these motions by separating them into principal components using the PCA algorithm.

The next study carried out looked at the possibility of recognizing three different types of phonetic mechanisms: (a) pronunciation of the phoneme /iy/ (/iy/, /piy/, /tiy/, /diy/) (b) pronunciation of the phoneme /uw/ (/uw/) and (c) pronunciation of consonants (/m/ and /n/). This study showed that there are descriptive makers for different pronunciation mechanisms when analyzing EEG signals. Going further, a system was created to differentiate all phonemes (seven classes) in the database. At this point the occipital channels were removed in order not to influence the response of the system, taking into account that the stimulus was applied visually. In the feature extraction stage, the CNN neural network was used to consider the spatial connections of the electrodes in the classification.

In the next chapter, the SOM neural network was used to represent the features of the data set in the two-dimensional space with the aim of changing the feature space into a space with greater separability between classes. The study is a comparative one between

features: (1) MFCC, (2) cross-covariance in the time-domain, and (3) cross-covariance in the frequency-domain.

The study made in ***Chapter 11*** focused on the classification of all phonemes and words from the KODB database. Several objectives were pursued in this chapter, including: (a) the influence of CNN hyperparameters; (b) testing different network architectures; (c) the impact of different activation functions used for CNN layers; (d) different features capable of decoding the hidden information in EEG signals by computing the time-domain and frequency-domain cross-covariance; (e) different analysis window sizes for feature extraction methods; (f) applying an averaging filter having the kernel of three (B3) and five (B5) samples applied to the signal spectrum. The latter study included the implementation of an intelligent phoneme and word recognizer using frequency-domain cross-covariance and the CNNLSTM convolutional neural network. During the development of the system, the performance of the system was also studied when analyzing different cranial regions: frontal, central and occipital for their left (S) and right (D) emispheres, as well as combinations between these regions. The anatomical regions involved in the speech production process were also selected. Finally, the analysis of system complexity and memory was also pursued for the possibility of implementation on a portable device with limited resources.

13.1 Results

13.1.1 Chaper 7: Eye movement artefact removal

This chapter aimed to remove the ocular artifacts from EEG signals to improve their quality. To achieve the proposed goal, we tested two different filtering methods, one based on adaptive filter and the other on principal components analysis.

We concluded that both methods attenuate ocular artifacts by decorrelating signals with HEO. It could be observed that initially it started from a very high correlation between the signals acquired from the frontal electrodes and the HEO, having values exceeding 0.8 and reaching values of approximately 0.1 for the adaptive filter and approximately 0.2 after filtering with PCA. We further observed that the signals after being filtered using the PCA method showed a higher correlation when looking at the relationship between the records of the same class, in contrast to the signals obtained after applying the adaptive filter, where a decorrelation appeared between them.

13.1.2 Chapter 8: Pronunciation mechanism recognition system

This study aimed to confirm the possibility of differentiating three different pronunciation mechanisms by analyzing only EEG signals acquired during imagined speech. We used in this stage features commonly used in automatic speech recognition: MFCC and LPC coefficients.

The obtained results showed that MFCC features provide a better understanding of imagined speech, providing higher accuracy and precision results than LPC. We also be

observed during the study that using the concatenation of the two types of features, MFCC and LPC, in a three-dimensional matrix the results were improved, the accuracy reaching a value of 0.38. This study also showed that the features used in automatic speech recognition can differentiate between different pronunciation mechanisms when aiming to classify signals acquired during silent speech.

13.1.3 Chapter 9: Phoneme recognition using MFCC and CNN

The study presented in this chapter aimed to test the MFCC coefficients for an automated imagined speech recognition system. These coefficients were combined with a CNN neural network able to find spatial links between the computed coefficients for each channel. The best results achieved were 24.19% accuracy for the test set.

13.1.4 Chapter 10: Imaginary speech analysing and classifictaion using SOM

This chapter aimed to analyze three types of features: MFCC, cross-covariance in time-domain and cross-covariance in frequency-domain using the unsupervised SOM neural network to map the feature space into a two-dimensional space aiming to obtain a space transformation that improve data separability. We observed in this study that features computed in the frequency domain showed better mapping in terms of differentiation between classes, but there were no clear distinct areas for each phoneme or word in the database for these features either. In contrast, in the frequency domain one could see groupings of phonemes and words in different areas of the map, making a better differentiation between these two classes.

In terms of classification, the same conclusion can be drawn, namely the cross-covariance in frequency-domain inputs offered better results than the time-domain, raising the accuracy to a value of 0.25. Applying the PCA technique to eliminate the ocular artifacts, the accuracy reached a value of 0.28.

13.1.5 Chapter 11: Words and phonemes recognition system from the KODB using CNN

The study carried out in this chapter aimed to analyze the EEG signals for the recognition of imagined speech of seven phonemes and four words. To achieve the proposed goal, an intelligent subject-shared system was developed using the processing chain applied to signals from the KODB database. In the feature extraction stage, the results obtained after computing the cross-channel covariance in time-domain and in frequency-domain were compared. An analysis of different window lengths: 0.25, 0.5 and 1s was also performed to find the window where the signal becomes quasi-stationary or nearly quasi-stationary, but which also contains information about the utterance. Finally, in the classification stage, multiple CNN architectures were tested to see which one provides the best performance.

The best results were obtained using the cross-covariance in frequency-domain using an analysis window of 0.25s. The best performance of the system was achieved using a CNN with two 2D convolutional layers having 64 and 128 filters and a fully connected

layer having the number of neurons equal to 64. Using these system characteristics, an accuracy of 37% was achieved, a significant improvement compared to using MFCC coefficients, where the maximum accuracy recorded was 20.80% using SVM in the classification step [24] and 24.19% using a CNN classifier [23].

13.1.6 Chapter 12: Words and phonemes recognition system from the KODB using CNNLSTM

This chapter aimed to develop a shared subject system for the recognition of seven phonemes and four words acquired during imagined speech. The current chapter showed an improvement in system performance for using the CNNLSTM network compared to the CNN. Accuracy increased from 37% to 43% when using the CNNLSTM network with no changes in the preprocessing or feature extraction chain. The developed system aimed to consider in the design the possibility of real-time implementation on a portable device with limited resources. Therefore, a study was also carried out in terms of reducing the number of electrodes in the system. The study concluded that 93% of the information is concentrated in anatomical regions specific to speech production, achieving an accuracy of 40% when using 29 electrodes compared to 62, which was the original number of channels.

13.2 Original contributions

The original contributions made during the development of the doctoral thesis are:

- The implementation of algorithms used to eliminate ocular artifacts with applications in imaginary speech recognition, detailed in the paper (C2). Two types of algorithms were comparatively tested: adaptive filter and PCA. The PCA method provided better results, decorrelating the HEO signal from affected frontal channels (FP1, FP2, FPZ) and improving inter-class correlations of the signals.
- Highlighting the presence of descriptive markers for different pronunciation mechanisms when analyzing EEG signals in the study (C1). The pronunciation mechanisms analyzed were: the phoneme /iy/ (containing the vowels: /iy/, /piy/ /tiy/ and /diy/), the phoneme /uw/ and consonants (/m/ and /n/).
- Implementation of a seven-phoneme recognition system from the Kara One database in the paper (C5). The study uses the MFCC coefficients using the mel-scale transformation equation adapted for EEG frequencies, in the feature extraction stage. By combining the Mel-Cepstral coefficients with the CNN neural network, the system exceeded the results presented in the literature.
- Study on the influence of hyperparameters and different CNN neural network architectures in imagined speech recognition (J1).
- Study of window length in the realization of an imagined speech recognition system (J1). Different windows length were analyzed: 0.25s, 0.5s and 1s. This analysis

showed that the window size affects the performance of the system, playing an important role in finding the optimal size for imaginary phonemes and words.

- Introduction of cross-covariance in frequency-domain in the field of automatic recognition of imagined speech from EEG signals (J1). This feature extraction method has been shown to significantly improve system performance.
- Development of an intelligent imagined speech recognition system that can be implemented in real time on a low-cost portable device (J1), (J2). this was evidenced by compting its complexity, memory and execution time.
- Analysis of different cranial regions in the development of the imaginary speech recognition system (J2). This analysis showed that the anatomical regions involved in the process of speech preparation and execution provide the best system performance results.

13.3 List of original work

In this section of the paper are presented all the papers published during the development of this thesis.

Papers published in international scientific journals:

(J1) **A.L. Rusnac**, O. Grigore, *"CNN Architectures and Feature Extraction Methods for EEG Imaginary Speech Recognition"*, Sensors, 22(13), p. 4679, 2022, WOS:000822119000001, Articol Q2, Factor de Impact: 3.847, eISSN:1424-8220, DOI: 10.3390/s22134679

(J2) **A.L.Rusnac**, O. Grigore, "*Imaginary speech recognition using a convolutional network with long-short term memory*", Applied Science, 2022, WOS:000887061100001, Articol Q2, Factor de Impact: 2.838, DOI: 10.3390/app122211873

Papers in the volumes of international scientific events (conferences, symposia) indexed by ISI:

(C1) **A. L. Rusnac**, O. Grigore, *"Convolutional Neural Network applied in EEG imagined phoneme recognition system"*, 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), MAR 25-27, 2021, București, ROMÂNIA, pp: 1-4, ISSN: 1843-8571, ISBN: 978-1-6654-1878-2, DOI: 10.1109/ATEE52255.2021.9425217, WOS:000676164800094

(C2) **A. L. Rusnac**, O. Grigore, *"EEG Preprocessing Methods for BCI Imagined Speech Signals"*, 9th IEEE International Conference on e-Health and Bioengineering (EHB), NOI 18-19, 2021, Grigore T Popa Univ Med & Pharmacy,

ELECTR NETWORK, pp: 1-4, ISSN: 2575-5137, ISBN: 978-1-6654-4000-4, DOI: 10.1109/EHB52898.2021.9657563, WOS:000802227900027

(C3) **A. L. Rusnac**, O. Grigore, *"Development of an Intelligent Seizure Prediction System"*, 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE), MAR 28-30, 2019, București, ROMANIA, pp: 1-5, ISSN: 1843-8571, ISBN: 978-1-4799-7514-3, WOS:000475904500102

(C4) **A. L. Rusnac**, O. Grigore, *"Intelligent Seizure Prediction System Based on Spectral Entropy"*, 14th International Symposium on Signals, Circuits and Systems (ISSCS), IUL 11-12, 2019, Iași, ROMANIA, pp: 1-4, ISBN: 978-1-7281-3896-1, WOS:000503459500070

Papers in the volumes of international scientific events (conferences, symposia) indexed by BDI:

(C5) A. L. Rusnac, O. Grigore, *"Generalized Brain Computer Interface System for EEG Imaginary Speech Recognition"*, 24th International Conference on Circuits, Systems, Communications and Computers (CSCC), IUL 19-22, 2020, Chania, GRECIA, p. 184-188, eISBN: 978-1-7281-6503-5, ISBN: 978-1-7281-6504-2, DOI: 10.1109/CSCC49995.2020.00040

Scientific research reports:

(R1) **A. L. Rusnac,** Coordonator: O. Grigore, "Studiu asupra metodelor actuale de recunoaștere a discursului imaginativ din semnale EEG"

(R2) **A. L. Rusnac,** Coordonator: O. Grigore, "Analiza utilizării entropiei spectrale a semnalelor EEG în aplicații de recunoaștere a cuvintelor rostite imaginativ"

References

[1] J. T. Crinion, M. A. Lambon-Ralph, E. A. Warburton, D. Howard, and R. J. S. Wise, "Temporal lobe regions engaged during normal speech comprehension," *Brain*, vol. 126, no. 5, pp. 1193–1201, May 2003, doi: 10.1093/brain/awg104.

[2] G. Mobus, "Part 4. The Neuroscience of Sapience," in *A THEORY OF SAPIENCE: Using Systems Science to Understand the Nature of Wisdom and the Human Mind*, Millennium Alliance for Humanity and the Biosphere, 1992, pp. 192–258. Accessed: Feb. 08, 2022. [Online]. Available: https://mahb.stanford.edu/library-item/theory-sapience-using-systems-science-understand-nature-wisdom-human-mind/

[3] H. C. Mahendru, "Quick review of human speech production mechanism," *International Journal of Engineering Research and Development*, vol. 9, no. 10, pp. 48–54, 2014.

[4] E. N. Marieb, S. J. Mitchell, L. A. Smith, and P. Z. Zao, *Human anatomy & physiology laboratory manual*, Eleventh edition (cat version). Boston: Pearson, 2014.

[5] T. H. Wideman *et al.*, "Brain, Tissue," in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds. New York, NY: Springer New York, 2013, pp. 262–263. doi: 10.1007/978-1-4419-1005-9_1105.

[6] M. Teplan, "FUNDAMENTALS OF EEG MEASUREMENT," 2002.

[7] H. Rogers, *The sounds of language: an introduction to phonetics*. Harrow, England ; New York: Longman, 2000.

[8] N. J. Hill and J. R. Wolpaw, "Brain–Computer Interface☆," in *Reference Module in Biomedical Sciences*, Elsevier, 2016, p. B978012801238399322X. doi: 10.1016/B978-0-12-801238-3.99322-X.

[9] M. E. Saab, "Basic Concepts of Surface Electroencephalography and Signal Processing as Applied to the Practice of Biofeedback," 2009.

[10] K. Ravi, R. Rajkumar, M. M. Raj, and S. S. Devi, "Imagined Speech Classification using EEG," *Advances in Biomedical Science and Engineering*, vol. 1, pp. 20–32, Dec. 2014.

[11] D. A. Rojas, O. L. Ramos, and J. E. Saby, "Recognition of Spanish Vowels through Imagined Speech by using Spectral Analysis and SVM," *J. Inf. Hiding Multim. Signal Process.*, vol. 7, pp. 889–897, 2016.

[12] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," *J. Neural Eng.*, vol. 15, no. 1, p. 016002, Feb. 2018, doi: 10.1088/1741-2552/aa8235.

[13] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 992–996. doi: 10.1109/ICASSP.2015.7178118.

[14] D. Vorontsova *et al.*, "Silent EEG-Speech Recognition Using Convolutional and Recurrent Neural Network with 85% Accuracy of 9 Words Classification," *Sensors*, vol. 21, no. 20, p. 6744, Oct. 2021, doi: 10.3390/s21206744.

[15] A.-L. Rusnac and O. Grigore, "EEG Preprocessing Methods for BCI Imagined Speech Signals," in *2021 International Conference on e-Health and Bioengineering (EHB)*, Iasi, Romania, Nov. 2021, pp. 1–4. doi: 10.1109/EHB52898.2021.9657563.
[16] R. Srinivasan and P. L. Nunez, "Electroencephalography," in *Encyclopedia of Human Behavior 2nd Edition*, vol. 2, Academic Press, 2012, pp. 15–23.
[17] R. Kher and R. Gandhi, "Adaptive filtering based artifact removal from electroencephalogram (EEG) signals," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, Melmaruvathur, Tamilnadu, India, Apr. 2016, pp. 0561–0564. doi: 10.1109/ICCSP.2016.7754202.
[18] L. Tan and J. Jiang, "Chapter 9: Adaptive filters and applications," in *Digital signal processing: Fundamentals and applications*, New Mexico: Academic Press, 2018, pp. 421–462.
[19] A.-L. Rusnac and O. Grigore, "Convolutional Neural Network applied in EEG imagined phoneme recognition system," in *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, Bucharest, Romania, Mar. 2021, pp. 1–4. doi: 10.1109/ATEE52255.2021.9425217.
[20] M. M. Azmy Gad, "Classification of mental tasks using support vector machine based on linear predictive coding and new mother wavelet transform," in *2015 International Conference on Biomedical Engineering and Computational Technologies (SIBIRCON)*, Novosibirsk, Russia, Oct. 2015, pp. 156–159. doi: 10.1109/SIBIRCON.2015.7361873.
[21] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, Feb. 1988, doi: 10.1109/45.1890.
[22] Y. Padmasai, K. SubbaRao, V. Malini, and C. R. Rao, "Linear Prediction Modelling for the Analysis of the Epileptic EEG," in *2010 International Conference on Advances in Computer Engineering*, Bangalore, Karnataka, India, Jun. 2010, pp. 6–9. doi: 10.1109/ACE.2010.20.
[23] A.-L. Rusnac and O. Grigore, "Generalized Brain Computer Interface System for EEG Imaginary Speech Recognition," in *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, Chania, Greece, Jul. 2020, pp. 184–188. doi: 10.1109/CSCC49995.2020.00040.
[24] C. Cooney, R. Folli, and D. Coyle, "Mel Frequency Cepstral Coefficients Enhance Imagined Speech Decoding Accuracy from EEG," in *2018 29th Irish Signals and Systems Conference (ISSC)*, Belfast, Jun. 2018, pp. 1–7. doi: 10.1109/ISSC.2018.8585291.
[25] A.-L. Rusnac and O. Grigore, "CNN Architectures and Feature Extraction Methods for EEG Imaginary Speech Recognition," *Sensors*, vol. 22, no. 13, p. 4679, Jun. 2022, doi: 10.3390/s22134679.
[26] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview." arXiv, Aug. 13, 2020. Accessed: May 24, 2022. [Online]. Available: http://arxiv.org/abs/2008.05756
[27] A.-L. Rusnac and O. Grigore, "Imaginary speech recognition using a convolutional network with long-short term memory," *Applied Sciences*, 2022.