Expert Systems With Applications 118 (2019) 201–208

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Subjects identification using EEG-recorded imagined speech

Luis Alfredo Moctezuma[a,*], Alejandro A. Torres-García[b], Luis Villaseñor-Pineda[b], Maya Carrillo[a]

[a] Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla (BUAP), Av. San Claudio #14, Puebla 72592, México
[b] Computer Science Department, Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), Luis Enrique Erro #1, Puebla 72840, México

ARTICLE INFO

Article history:
Received 6 December 2017
Revised 10 August 2018
Accepted 3 October 2018
Available online 5 October 2018

Keywords:
Subject identification
Electroencephalograms (EEG)
Imagined speech
Biometrics

ABSTRACT

Due to the problems presented in current traditional/biometric security systems, the interest to use new security systems, have been increasing. This paper explores the use of brain signals EEG-based during imagined speech in order to use it as a new biometric measure for Subjects identification and thus create a new biometric security system. The main contribution of this paper are two methods for feature extraction, first to improve the signal-to-noise ratio the Common Average Reference was applied. The first method was based on Discrete Wavelet Transform, and the second method was based on statistical features directly from the raw signal. The proposed methods were tested in a dataset of 27 Subjects who performed 33 repetitions of 5 imagined words in Spanish. The results show the feasibility of the task with accurate identification of the Subject, regardless of the imagined word used and using a commercial EEG system (EMOTIV EPOC). In addition, the scope of the method is displayed by decreasing the training data, as well as the number of active sensors for the identification task. Using the proposed method with future improvements and implementing it in a low-cost device can be a new and valuable biometric security system.

1. Introduction

Security systems are used by organizations in order to protect places or information for which privileges are needed or have access authorization, furthermore to deny unauthorized access to facilities, equipment or resources, and to protect against espionage, theft, or even for terrorist attacks. To achieve this, different safety measures have been proposed for a long time, ranging from the use of generic systems (security guards, closed-circuit television, smart cards, proximity readers and RFID) to the use of biometric identifiers (fingerprint, palmprint, retinal scan, etc.).

Biometric recognition refers to the automatic recognition of individuals based on their physiological and/or behavioral features (Jain, Ross, & Prabhakar, 2004). A biometric system is a pattern recognition system that operates by acquiring biometric data from subjects, extracting a feature set, and comparing this feature set against the template set in the database. The combination of biometric data systems and biometrics recognition/identification technologies creates biometric security systems. Biometric systems have been used in different low-cost devices, for example, in last years fingerprint/face-recognition were implemented in devices ranging from electronic boards to smartphones or computers.

Biometric systems are advantageous compared to generic systems, in fact, the key is more difficult to be stolen, compromised or duplicated. However, a biometric system is vulnerable to a variety of attacks aimed at undermining the integrity of the authentication process (Jain, Ross, & Uludag, 2005). For example, an intruder may obtain fraudulently the latent fingerprints of a user and later used to construct a digital or physical artifact of user's finger (Uludag & Jain, 2004). It is possible because authentication systems cannot discriminate between an intruder who fraudulently obtains the access privileges and authorized users.

In last years, systems have been used to capture brain signals in order to use them as a new way of recognizing subjects. Biometric systems based on the EEG are a low-cost non-invasive technique Jain et al. (2004). It should be noted that its main disadvantage is that the user must be in contact with the electrodes and the placement time is longer than other biometrics systems. However, EEG is a very promising modality that should be studied.

Electrophysiological sources refer to the neurological mechanisms or processes used by a user to generate brain signals (Bashashati, Fatourechi, Ward, & Birch, 2007). Among them, we find imagined or internal speech, which refers to the internal or

* Corresponding author.
E-mail addresses: luisalfredomoctezuma@gmail.com (L.A. Moctezuma), alejandro.torres@ccc.inaoe.mx (A.A. Torres-García), villasen@ccc.inaoe.mx (L. Villaseñor-Pineda), cmaya@cs.buap.mx (M. Carrillo).

https://doi.org/10.1016/j.eswa.2018.10.004
0957-4174/

imagined pronunciation of words but without uttering sounds or articulating gestures. Imagined speech, as an electrophysiological source, has advantages over others. Unlike other paradigms, imagined speech is consciously generated by the user, which leads to short periods of training.

In this research imagined speech from EEG signals is used as a biometric measurement for a subject identification system. This research used a dataset of EEG signals from 27 subjects captured while imagining 33 repetitions of five imagined words in Spanish, corresponding to the English words *up, down, left, right* and *select*.

Specifically, this work evaluates the following hypothesis: *"Using extracted features from EEG signals recorded during imagined speech, an automatic classifier is able to effectively recognize a specific subject"*. In addition, aiming to provide evidence about the robustness of imagined speech-based subject identification, we assessed either if a subject can be identified regardless of the imagined word or if a specific word is better for subject identification. Meanwhile, in this work, we analyzed the impact of using fewer both instances and channels.

The remaining sections are organized as follows: first, a set of related works is presented, then the proposed method is exposed, after that, the experiments are detailed. Finally, the conclusions and the future work looking for implementing an imagined-speech-based biometric security system are presented.

2. Related works

Recently, several works have explored the use of brain signals recorded using surface electrodes, aiming for subject identification and authentication.

It is well-known that there are several challenges for Subjects identification task, it can be separated into; best neuro-paradigm, channels and localization and feature extraction technique. The number of channels and localization are hard-linked to the neuro-paradigm used, this is why in the state of the art there are several approaches and with different results. To stimuli the EEG-based brain signals have been used different activities, for example in the work presented by Ashby, Bhatia, Tenore, and Vogelstein (2011) they used the sensorimotor activity (Visual Counting and geometric figure Rotation), another approach using imagination of activities (mental composition of letters) was presented by Palaniappan (2006), visual stimulation of images approach was used by Ruiz-Blondet, Jin, and Laszlo (2016), among others similar approach can be found in Jayarathne, Cohen, and Amarakeerthi (2017); Marcos, Jess, Jaime, and Carlos (2014).

At first sight, the work presented by Jayarathne, Cohen, and Amarakeerthi (2016) for subjects identification while imagining four random digit numbers is interesting, however, the protocol design is not easy (neutralization, visual stimulation, and thinking stimulation phases) with a different duration of the phase (10, 5 and 10 s respectively). The idea of imagining four digit numbers is to use it as a standard bank password, but the work is limited to identify the subject, not the number imagined.

However, these before paradigms/stimulus present a problem: a long training period, since these neuro-paradigms are generated by the user in an unconscious way. In the before related work presented, in most cases the protocols for the signal acquisition were complicated, this is why this proposal is based on another paradigm: imagined or internal speech. This has the advantages of being a high-level task performed consciously and allows the subject to define a specific password.

In this sense, the most related work was presented by Brigham and Kumar (2010), they used EEG signals from 6 subjects who imagined the syllables /*ba*/ and /*ku*/ that according to the authors were selected since they contain no semantic meaning. The database they collected consisted of 20 trials per session, with a total of six sessions per subject and the EEG signal was captured from 128 channels with a sampling frequency of 1024 Hz. by Electrical Geodesics. For feature extraction, the authors have used the power spectral density (PSD) of each EEG signal and then for each dataset, *autoregressive (AR)* model coefficients were computed for each electrode's signal using the Burg method. For the classification step was used a linear *Support Vector Machine (SVM)* classifier and for these two syllables, they obtained 99.76% of accuracy. This experiment was also performed with *1-Nearest Neighbor (k-NN)* obtaining an accuracy of 99.41%. In addition, they considered the syllables separately, obtaining 99.49% and 99.43% respectively using *SVM* classifier.

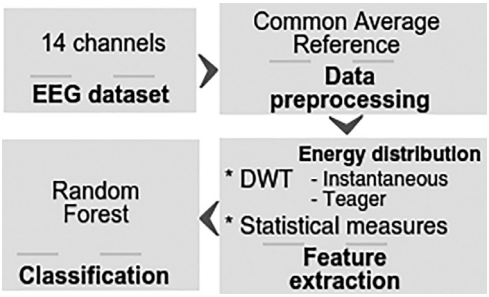

Fig. 1. Proposed method.

2.1. Discussion

Even though the work described by Brigham and Kumar (2010) is interesting; some factors as the impact of using words with a semantic meaning, fewer channels, non-clinical and cheap EEG device were not assessed. The use of words which have semantic meaning instead of syllables (Brigham & Kumar, 2010) could be useful for subject identification tasks. Also, it is not clear whether the use of fewer channels is an important factor looking for a real-life implementation of a biometric system based on EEG signals. In addition, it is not clear whether a subject identification system based on imagined speech could be as accurate as when a clinical EEG device is employed, using a non-clinical EEG headset. Finally, the question about how many repetitions of the words are the minimum necessary for subject identification has not been answered. Below, the proposed method looking for answering the questions before mentioned is described.

3. Proposed method

Basically, the proposed method consists of three main stages: data preprocessing, feature extraction and classifier construction. These steps are detailed in the sections below. It should be noted that to perform the classification was used the *random forest classifier* as it has shown the best results with imagined speech classification (Torres-García, Reyes-García, & Villaseñor-Pineda, 2011; Torres-García, Reyes-García, Villaseñor-Pineda, & Ramírez-Cortés, 2013).

The features were calculated using discrete wavelet transform (DWT) with biorthogonal 2.2 (bior2.2) as the mother function with 4 decomposition levels and each level was used to compute instantaneous and teager energy distribution in order to compare them. Other features based on statistical measures (Used first by Zhao & Rudzicz (2015)) were also computed directly from the signal. In general, the proposed method is shown in Fig. 1. Both EEG characterization methods were implemented looking for subjects identification regardless of the imagined word. With these representations, is obtained the same number of features in all instances,

Table 1
DWT with 4 decomposition levels, frequency ranges and related brain rhythms (taken a sampling frequency of 128 Hz).

Level	Frequency range (Hz)	Brain rhythm
D1	32-64	*Gamma*
D2	16–32	*Beta (16–30 Hz) and Gamma (30–32 Hz)*
D3	8–16	*Alpha (8–12 Hz) and Beta (12–16 Hz)*
D4	4–8	*Theta*
A4	0–4	*Delta*

thus removing the dependence on the duration of the imagined words.

With the proposed method can be implemented a higher security level, this is because the same imagined word can be used for subject identification and then a second step will find the imagined word and the action which should have (i.e give access, lock the system, etc.), like a two-step verification system.

3.1. Data preprocessing

At this stage, the EEG signals were processed using common average reference (CAR) method. This method improves the signal-to-noise ratio from the EEG signal by removing the common information in all electrodes simultaneously recorded. CAR can be computed using the following formula:

$$V_i^{CAR} = V_i^{ER} - \frac{1}{n}\sum_{j=1}^{n} V_j^{ER} \tag{1}$$

Where V_i^{ER} is the potential between the i-th electrode and the reference, and n is the number of electrodes.

3.2. Feature extraction

This stage is based on statistical measures and energy coefficients that were computed for each decomposition level of the DWT biorthogonal 2.2 (bior2.2). In both cases, the features for each channel were computed and concatenated, thus obtaining a single feature vector per instance.

3.2.1. Features based on DWT

EEG signals are usually non-stationary because they change rapidly over time and patterns of brain activity contain information related to specific variations over time. This is why a representation should be used to consider this behavior.

As a first option, DWT with bior2.2 as the mother function was applied for each channel. Applying this transform with a decomposition level j gives a structure with vectors of approximation CA_j and detail CD_j coefficients: $[CA_j, CD_j, CD_{j-1}, \ldots, CD_1]$. Also, it is possible to determine the maximum number of decomposition levels to be computed using DWT based on the size of the epochs, which for this dataset was 4 levels. Table 1 shows the frequencies for each level of decomposition.

However, the wavelet coefficients for each decomposition level will vary depending on the pronunciation window length of the imagined word (between imagined words from the same subject and between imagined words from different subjects). To deal with this problem the energy distribution instantaneous (IWE) and teager (TWE) were compute (Didiot, Illina, Fohr, & Mella, 2010). Once the energy coefficients were computed is possible to have the same number of features per instance, below are the formulas for these energy distributions.

- Instantaneous energy: this energy coefficient is a ratio from the amplitude of the signal.

$$f_j = log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}(w_j(r))^2\right) \tag{2}$$

- Teager energy: reflects variations in both amplitude and frequency of the signal and it is a robust parameter for speech recognition as it attenuated auditory noise.

$$f_j = log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j-1}|(w_j(r))^2 - w_j(r-1) * w_j(r+1)|\right) \tag{3}$$

Now instead of having a vector of values for each decomposition level, a single value per each one is obtained. This is for each decomposition level and the process is repeated for each channel. Then the energy coefficients from each channel were concatenated in order to have a features vector to represent the EEG signal.

3.2.2. Features based on statistical values

Another way for feature extraction from the EEG signal was motivated by Zhao and Rudzicz (2015). In such work, the signal was characterized by computing a set of 15 statistical values (9 values and 6 combinations of them). In the present work, only statistical values (StV) were considered without combinations between them, thus obtaining a set of 9 values per channel. After computing of statistical features per channel, the values were concatenated to have a single feature vector per instance.

- StV: mean, maximum, minimum, standard deviation, variance, kurtosis, skewness, sum, median.

3.3. Classification

Once features vectors were obtained for each instance of the EEG signal, for automatic classification *random forest* was used with the implementation in weka 3.6.14 using the default parameters. This classifier was selected because of the results reported in works related to imagined speech classification using EEG signals (Torres-García et al., 2011; Torres-García et al., 2013). *Random forest* is a combination of predictive trees such that each tree depends on the value of a random vector sampled independently and with the same distribution for all forest trees (Breiman, 2001).

Random forest has two main parameters: the number of trees and the number of predictors to be used in each partition of each tree. One of the RF's advantages is its low sensitivity to these parameters so that the default values usually produce good results (Liaw & Wiener, 2002). In addition, the classifier performance was evaluated with the accuracy using 10-folds cross-validation.

4. Experiments and results

This section aims to show the results by experimenting with two sets of different features. Below, the dataset used is detailed, and later the results obtained when using different types of features. In addition, the results obtained when using a reduced set of channels are shown.

4.1. Dataset

For the following experiments, a dataset of EEG signals from 27 subjects captured using EMOTIV EPOC while imagining 33 repetitions of five Spanish imagined words, corresponding to the English words *up, down, left, right* and *select* is used. Each repetition of the imagined words was separated by a state of rest, as shown in Fig. 2 and described by Torres-García et al. (2013). EEG signals were recorded from 14 high-resolution channels (AF3, F7, F3, FC5,

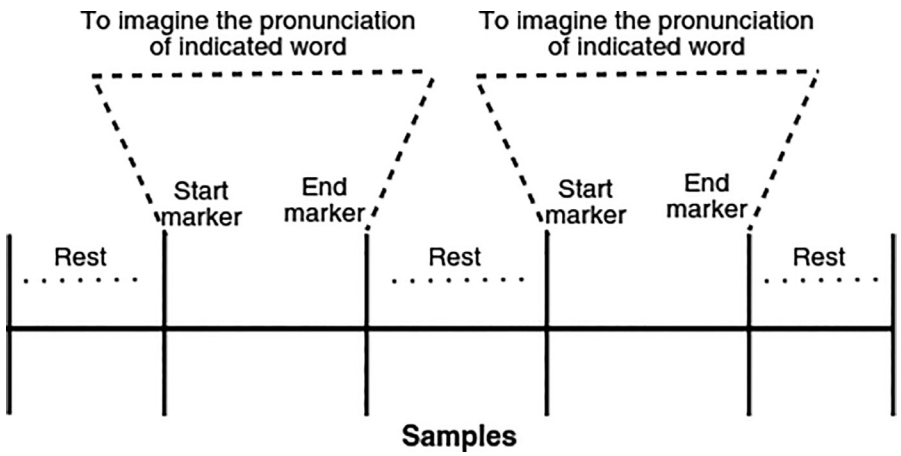

Fig. 2. Protocol design for EEG signal acquisition using EMOTIV EPOC.

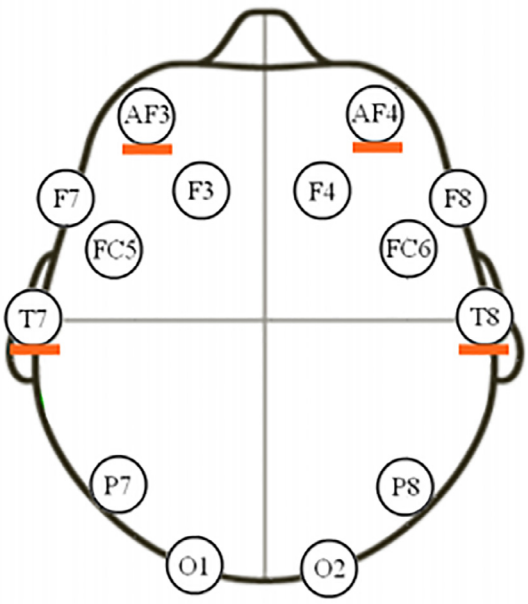

Fig. 3. Location of 14 electrodes according to international 10–20 system for EEG.

T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4) with a 128 Hz sample frequency which were placed according to the international 10–20 system.

In Fig. 3 the location of the 14 channels of EMOTIV EPOC according to the international system 10–20 is shown. The channels marked with an orange line are the common channels with EMOTIV INSIGHT.

4.2. Subject identification regardless of the imagined word

In the first place, an experiment was carried out to the aim of checking if there is different information between the subjects that allows classification to be performed of the EEG signals using imagined speech. For this the 5 imagined words were considered as a single class, labeling each of 165 instances by each subject, corresponding to the 33 repetitions of the 5 imagined words, with a subject ID.

This experiment was performed with 5 and 27 subjects using StV and the IWE and TWE distribution based on the DWT. The aim of using 5 subjects first and then 27, is to verify if, in small populations, the accuracy is higher compared to consider more subjects.

Table 2
Accuracy and standard deviation obtained for 5 y 27 subjects groups after 10-fold cross-validation with *Random Forest* for subject identification regardless of the imagined word.

Feat.	5 subjects	27 subjects
StV	94% ± 3	85% ± 5
IWE	**97%** ± 2	**95%** ± 4
TWE	96% ± 3	93% ± 5

The results obtained in the classification step with 10-fold cross-validation with *random forest* are shown in Table 2.

In Table 2 can be observed that when using DWT with instantaneous energy distribution the best results are obtained (97% and 95% for 5 and 27 subjects respectively). These results suggest that it is possible to identify subjects regardless of the imagined word. In other words, subjects can use different words and the method can still distinguish them, in this case, 5 different words were tested.

Something that is also important is that with the statistical values StV similar results are achieved and this type of characterization has the advantage that it does not need any transformation of domain or definition of additional parameters. As shown in the proposed method, when using StV, 9 values per channel are computed, obtaining 126 characteristics for the 14 channels. When the energy distribution IWE and TWE are calculated, 5 values for each channel are obtained, in total 70 features with the 14 channels per instance. The above described should be considered to determine which is the appropriate way to represent the EEG signal because, on the one hand, the StV are computationally cheaper but at the stage of classification, there are almost double values to compare with respect to IWE and TWE.

4.3. Subjects identification using a specific imagined word

The previous experiment shows that if the imagined words of each subject are considered as a single class, the subjects can be distinguished. If words are considered separately, is still possible to distinguish between subjects? From the dataset, is there a word that works better for this task? To give evidence of this, the following experiment was carried out, which consists in testing the classification with *random forest* with 5 and 27 subjects to show that the task is complex and that in larger and larger populations, accuracy is affected but not for all words. The last idea shows the possibility about if there are words that are better for this task.

For this, the 33 repetitions of each imagined word were used separately and the experiment was repeated for the 5 words from

Table 3
Accuracy and standard deviation obtained after 10-fold cross-validation with the *random forest classifier* using different feature extraction techniques for each imagined words.

Imagined word	Feat.	5 subjects	27 subjects
Up	StV	99% ± 2	90% ± 6
	IWE	99% ± 1	96% ± 6
	TWE	99% ± 3	92% ± 9
Down	StV	94% ± 5	87% ± 7
	IWE	99% ± 1	**98%** ± 5
	TWE	97% ± 4	96% ± 4
Left	StV	96% ± 3	87% ± 9
	IWE	99% ± 2	96% ± 4
	TWE	99% ± 1	94% ± 7
Right	StV	95% ± 5	86% ± 7
	IWE	99% ± 2	96% ± 3
	TWE	98% ± 4	95% ± 5
Select	StV	97% ± 4	86% ± 8
	IWE	96% ± 2	96% ± 5
	TWE	98% ± 2	94% ± 6

Table 4
Accuracy and standard deviation obtained after 10-fold cross-validation with the *random forest classifier* using fewer and fewer instances.

% Ins	Accuracy
100	95% ± 4
75	94% ± 4
50	94% ± 4
25	92% ± 6
15	90% ± 8
10	86% ± 15

Table 5
Accuracy and standard deviation obtained after 10-fold cross-validation with the *random forest classifier* using fewer and fewer instances for each word by separate.

% Ins	Up	Down	Left	Right	Select
100	96% ± 6	98% ± 5	96% ± 4	96% ± 3	96% ± 5
75	95% ± 6	96% ± 5	95% ± 4	96% ± 4	96% ± 5
50	95% ± 7	96% ± 5	95% ± 6	96% ± 5	95% ± 6
25	95% ± 8	96% ± 7	92% ± 11	91% ± 12	92% ± 12
15	91% ± 12	87% ± 19	90% ± 16	86% ± 23	89% ± 15
10	75% ± 25	76% ± 30	70% ± 33	73% ± 31	84% ± 23

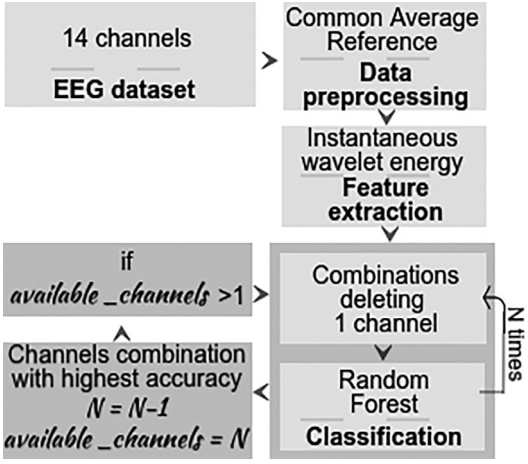

Fig. 4. Proposed method to delete channels.

the dataset of EEG signals. The classification for each of the imagined words was done with the three feature extraction ways used in the previous experiment (StV, IWE, TWE) in order to make a comparison. The results obtained in the classification step with 10-fold cross-validation with *random forest* are shown in Table 3.

Table 3 again shows that when using fewer subjects, the accuracy is greater. When using the imagined word Down, the best accuracy is obtained for 5 and 27 subjects using IWE, 99%, and 98% respectively. For this reason, it could be said that the word that is best for the task of subjects identification is the imagined word Down. However, when using the other imagined words, the results are not very different and in all cases, they are above 96 of accuracy with IWE.

4.4. Use of fewer instances

To incorporate new subjects into a dataset in order to have samples with which to compare in future records in a security system, a set of instances of the same imagined word must be recorded. However, performing the 33 instances that were performed in the database used in this work would be impractical. That is why it is necessary to analyze the impact of using fewer instances and thus decrease the process of adding new subjects.

For this experiment, all words were considered as a single class and then the words were analyzed separately, the 14 channels of the dataset were used and the number of instances was decreased in order to analyze if it is possible to identify the subjects with fewer instances compared to using the 33 of the dataset.

The Table 4 shows the accuracy and standard deviation obtained when performing 10-fold cross-validation using the *random forest classifier*. It shows the accuracy obtained with 100% of instances (33 instances per imagined word) and then with 75%, 50%, 25%, 15% and 10% of instances, the instances always were the first part of the set. The results give evidence that it is possible to distinguish between subjects with 15% of instances, with 5 instances per imagined word, and to obtain similar results with a standard deviation of 8%.

When performing the same experiment considering the imagined words separately, similar results are obtained to use all instances (see Table 5), with the 33 instances per imagined word, 25% of the instances are 8 instances, with which similar results are obtained with respect to using all instances. When using 15% of instances the accuracy decreases and the standard deviation increases, this is because the number of instances is 5 per class (subject) and the classes are 27.

4.5. Using fewer channels

According to the previous experiments, hereafter only IWE is considered for feature extraction and only with 27 subjects since the best results were obtained with IWE and the objective of using 5 subjects was already shown previously. As was presented in the introduction, the question arises whether there are channels that provide more information for subjects identification. This is done in order to have less data with which to work and make use of increasingly simple and accessible devices.

In order to analyze which channels provide more information the general proposed method is modified by adding a greedy algorithm, which can be observed in Fig. 4. First, preprocessing with CAR is applied and then extract features with IWE based on DWT. Next, the channel vector with 0 or 1 take the corresponding channel feature vector for join all channels and make the feature vector for each instance.

The channels were represented by a size vector 14, one position per channel. The available channels were initially marked with the number 1 in the vector, then all possible combinations are made by eliminating a channel, to perform the classification using 10-fold cross-validation with *random forest* and select the channel vector with which the highest accuracy is obtained. The deleted chan-

Table 6
Accuracy and standard deviation obtained after 10-fold cross-validation with the *random forest classifier* for Subject level analysis using fewer and fewer channels.

No. channels	Accuracy	No. channels	Accuracy
14	95% ± 4	7	90% ± 8
13	94% ± 4	6	88% ± 9
12	94% ± 4	5	85% ± 11
11	94% ± 5	4	80% ± 11
10	93% ± 5	3	73% ± 13
9	92% ± 5	2	59% ± 17
8	92% ± 6	1	29% ± 17

nel is marked with the number 0 and the process is repeated to eliminate another channel, this is done until having a single channel. Thus, it is possible to eliminate the channel that provides less information for classification, one by one and analyze the impact of using fewer and fewer channels for subjects identification.

The method described was used to perform the experiments below.

4.5.1. Subject level analysis

In order to analyze if there are channels that provide more information or more important to the classification of subjects, the method described in Fig. 4 is applied to the configuration of the first experiment. In this experiment all words were considered as a single class, thus having 165 instances per subject. The experiment was performed with the 27 subjects, with the IWE features and in the classification step with 10-fold cross-validation. The best results obtained are shown in Table 6[1].

In Table 6, the results are shown when using 14 channels and then the results obtained by eliminating in each iteration a channel until the classification with a single channel. It is shown that according to how they are eliminated, the accuracy decreases and the standard deviation increases. Using less than 50% of channels, the accuracy is similar compared to the use of all channels.

This seems to indicate that subjects identification can be performed with devices that are more accessible and with fewer channels to facilitate and speed up the identification process. Fig. 5 shows the order in which the channels were removed, which gives an idea of the importance of the channels. In the figure, the channel 14 was the first on be deleted and then 13, etc. It can be noted that some of the channels that are the last ones to delete are the ones closer to the Broca's and Wenicke's area, that according to the literature, these areas have the responsibility for generating language and the semantic relationships.

4.5.2. Subject-word level analysis, all cases

In Table 7 the results of channel elimination are shown when using the imagined words separately. So in this experiment, like the second, each subject contains 33 instances corresponding to the repetitions of each imagined word. After applying the method from Fig. 4, the results obtained show that the accuracy in the classification by imagined word is greater compared with the accuracy obtained with all imagined words as a single class[2].

[1] Currently there are devices with fewer channel that EMOTIV EPOC that was used to capture the EEG signal used in this work. For example, the EMOTIV INSIGHT device has 5 channels (AF3, AF4, T7, T8 and Pz) set according to the international system 10–20 and the sampling frequency is 128 Hz.

From the 5 channels of this device, in this dataset, there are 4 channels in common, as it can be seen in Fig. 3. To give an approximation and to compare the results obtained the experiment is proposed was also done with all imagined words as a single class for the 27 subjects. The accuracy when the common channels were used was 76% ± 14.

[2] The same experiment was performed using the 4 common channels with EMOTIV INSIGHT considering the words separately and the results that were obtained

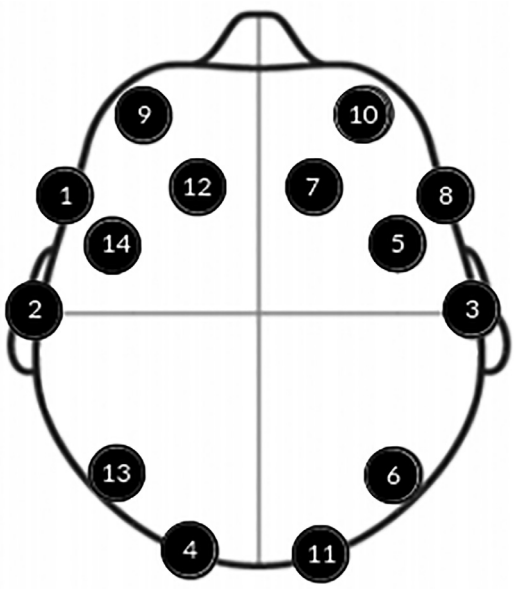

Fig. 5. Order in which channels were deleted for subject level analysis.

Table 7
Accuracy and standard deviation obtained after 10-fold cross-validation with the *random forest classifier* for Subject-Word level analysis using fewer and fewer channels.

Ch	Up	Down	Left	Right	Select
14	96% ± 6	98% ± 5	96% ± 4	96% ± 3	96% ± 4.5
13	96% ± 6	97% ± 3	96% ± 4	96% ± 3	96% ± 4
12	95% ± 8	97% ± 4	96% ± 4	95% ± 4	96% ± 4
11	95% ± 6	97% ± 3	95% ± 4	95% ± 5	95% ± 5
10	95% ± 6	96% ± 4	95% ± 4	95% ± 5	95% ± 5
9	94% ± 8	95% ± 5	95% ± 4	94% ± 5	95% ± 5
8	93% ± 8	95% ± 6	94% ± 4	93% ± 6	94% ± 6
7	93% ± 8	94% ± 5	93% ± 4	93% ± 5	93% ± 6
6	92% ± 10	92% ± 7	92% ± 6	92% ± 5	92% ± 7
5	91% ± 12	90% ± 8.5	88% ± 9	88% ± 7	90% ± 7
4	89% ± 13	87% ± 8	83% ± 12	84% ± 10	86% ± 10
3	81% ± 18	80% ± 10	75% ± 17	78% ± 10	81% ± 13
2	70% ± 19	67% ± 15	62% ± 20	62% ± 14	66% ± 17
1	46% ± 21	45% ± 18	41% ± 19	40% ± 21	37% ± 19

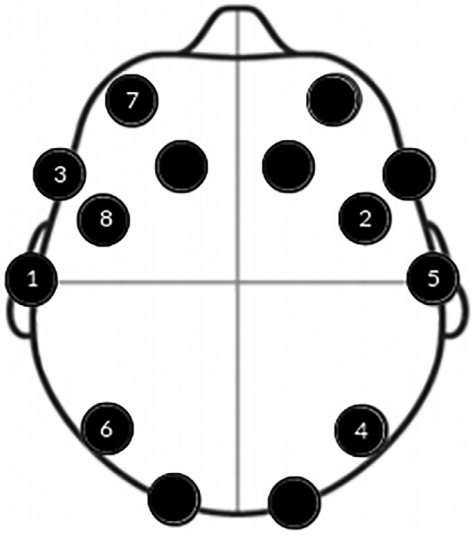

Fig. 6. Order in which channels were deleted for subject-Word level analysis.

In Table 7 it can be seen that when using less than 50% of channels, the accuracy is similar to using the 14 channels. In addition, the results obtained in this experiment show that with only 4 or 5 channels an accuracy of 83% to 91% can be obtained, which is of great interest since there are some commercial devices with this number of channels.

Below are shown in Fig. 6 the last 8 channels to be deleted and which are common in the 5 imagined words. This gives an idea of which channels provide the most information needed for subjects identification and that when using only these channels, accuracy rates similar to using the 14 channels can be obtained.

5. Conclusions and future work

Currently, some works have aimed to develop EEG-based biometric systems which allow the authentication of a given subject from his/her brain activity. Several EEG neuro paradigms have been assessed although we focused our work on imagined speech, which especially has the flexibility to be used for either setting or modifying a given password. The last one is of special interest and the main advantage over previous works because it allows robustness against an intrusion situation, allowing the setting of a new password based on brain signals.

The main contribution of this work was the evaluation of the robustness of the imagined speech for subject identification. Specifically, we found that the system can support the identification of different passwords keeping the same performance. In addition, the subject identification could be achieved without accuracy decreasing using fewer channels and instances or even using a different kind of features. In the following paragraphs, we detail the main conclusions for each experiment assessed.

As to the feature extraction, two ways of feature extraction were presented, the first one based on the discrete wavelet transform and the second one using statistical values. The results obtained give evidence that by calculating the instantaneous energy distribution, the EEG signal seems to be better characterized for subjects identification using *random forest* with cross-validation with 10 folds. However, it should be noted that with IWE, TWE, and StV, similar accuracy rates are obtained that suggest the feasibility of the task.

Whereas, as to the channel selection process, it was proposed a method that uses a greedy algorithm to delete channels, this method deletes a channel by each iteration, remaining only with the channels with which the highest accuracy rate was obtained. When applying this algorithm to the tasks of subjects identification by imagined word and considering all the imagined words together, an idea of the channels that contribute with more information for the classification stage is obtained. Some of the channels that are the last ones to delete are the ones closer to the Broca's and Wenicke's area, that according to the literature, are the areas that are responsible for generating language and semantic relationships. In general, the experiments performed show that subjects identification can be performed even with few channels. With less than 50% of the channels, similar accuracy rate are obtained that when using the 14 channels. A potential application of our method along with 3d printers would be to personalize EEG devices for covering the selected channels for each subject.

On the other hand, as future work, we propose to use the method for eliminating channels presented in Torres-García, Reyes-García, Villaseñor-Pineda, and García-Aguilar (2016). The objective to apply that method is to compare if the two methods delete the same or similar channels, and then select the common best channels for subject identification task. Other activities are both to evaluate other EEG devices and other EEG-based imagined speech datasets. The first activity aims to verify our outcomes using devices with either more or fewer channels. Whereas, the second one could evaluate the application of our method in another imagined speech database such as karaone dataset[3], which has been recently published.

Also, the use of others transforms like Hilbert-Huang Transform (HHT), could be explored. This because HHT does not depend on a predefined mother function.

Finally, the main drawback of our work is to be based on EEG signals which are sensitive to artifacts and non-stationary signals. However, we took into account these problems to minimize them but our work could be improved when better and fast artifact removal methods are proposed.

Conflict of interest

This work is part of a patent.

Acknowledgements

This work was done under partial support of CONACYT (scholarship #591475) and the project No. Ref. 2016-01-2228. The authors are grateful for the support of the *Red Temática en Tecnologías del Lenguaje/CONACYT* for the collaboration mechanisms to carry out this research.

References

Ashby, C., Bhatia, A., Tenore, F., & Vogelstein, J. (2011). Low-cost electroencephalogram (eeg) based authentication. *International IEEE/EMBS Conference on Neural Engineering*, 442–445. doi:10.1109/NER.2011.5910581.

Bashashati, A., Fatourechi, M., Ward, R. K., & Birch, G. E. (2007). A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering, 4*(2), R32. doi:10.1088/1741-2560/4/2/R03.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. doi:10.1023/A:1010933404324.

Brigham, K., & Kumar, B. V. (2010). Subject identification from electroencephalogram (eeg) signals during imagined speech. *Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, 4*, 1–8. doi:10.1109/BTAS.2010.5634515.

Didiot, E., Illina, I., Fohr, D., & Mella, O. (2010). A wavelet-based parameterization for speech/music discrimination. *Computer Speech & Language, 24*, 341–357. doi:10.1016/j.csl.2009.05.003.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(1), 4–20. doi:10.1109/TCSVT.2003.818349.

Jain, A. K., Ross, A., & Uludag, U. (2005). Biometric template security: Challenges and solutions. In *Signal processing conference, 2005 13th European* (pp. 1–4). IEEE.

Jayarathne, I., Cohen, M., & Amarakeerthi, S. (2016). Brainid: Development of an eeg-based biometric authentication system. In *Information technology, electronics and mobile communication conference (iemcon), 2016 ieee 7th annual* (pp. 1–6). IEEE.

Jayarathne, I., Cohen, M., & Amarakeerthi, S. (2017). Survey of eeg-based biometric authentication. In *Awareness science and technology (icast), 2017 ieee 8th international conference on* (pp. 324–329). IEEE.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R news, 2*(3), 18–22.

Marcos, D. P.-B., Jess, B. A., Jaime, R. T.-R., & Carlos, M. T. (2014). Electroencephalogram subject identification: A review. *Expert Systems with Applications, 41*, 6537–6554. doi:10.1016/j.eswa.2014.05.013.

Palaniappan, R. (2006). Electroencephalogram signals from imagined activities: A novel biometric identifier for a small population. *International Conference on Intelligent Data Engineering and Automated Learning, 4224*, 604–611. doi:10.1007/11875581_73.

Ruiz-Blondet, M. V., Jin, Z., & Laszlo, S. (2016). Cerebre: A novel method for very high accuracy event-related potential biometric identification. *IEEE Transactions on Information Forensics and Security, 11*, 1618–1629. doi:10.1109/TIFS.2016.2543524.

were $83\% \pm 14$, $81\% \pm 12$, $79\% \pm 14$, $79\% \pm 13$, $82\% \pm 12$ respectively. As it can be observed that the accuracy is smaller compared to the results when using 4 channels in Experiment 4.5.2 (89%, 87%, 83%, 84%, 86% respectively), it is also important to note that the channels used are not the same ones that the proposed method selects (see Fig. 6) by eliminating the channels that provide less information.

[3] Computational Linguistics, University of Toronto, Department of Computer Science: http://www.cs.toronto.edu/~complingweb/data/karaOne/karaOne.html.

Torres-García, A., Reyes-García, C., & Villaseñor-Pineda (2011). Hacia la clasificación de habla no pronunciada mediante electroencefalogramas (eeg). In *Xxxiv congreso nacional de ingeniería biomédica, ixtapa-zihuatanejo, guerrero, mexico*.

Torres-García, A., Reyes-García, C., Villaseñor-Pineda, L., & Ramírez-Cortés, J. (2013). Análisis de señales electroencefalográficas para la clasificación de habla imaginada. *Revista mexicana de ingeniería biomédica, 34*(1), 23–39.

Torres-García, A. A., Reyes-García, C. A., Villaseñor-Pineda, L., & García-Aguilar, G. (2016). Implementing a fuzzy inference system in a multi-objective eeg channel selection model for imagined speech classification. *Expert Systems with Applications, 59*, 1–12. doi:10.1016/j.eswa.2016.04.011.

Uludag, U., & Jain, A. K. (2004). Attacks on biometric systems: a case study in fingerprints. In *Proceedings of spie: 5306* (pp. 622–633). doi:10.1117/12.530907.

Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 1–5. doi:10.1109/ICASSP.2015.7178118.