Technical Report

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

UNIVERSITY OF THE BASQUE COUNTRY
Department of Computer Science and Artificial Intelligence

A detailed investigation of classification methods for vowel speech imagery recognition

Roberto Santana

December 2013

San Sebastian, Spain
www.ccia-kzaa.ehu.es
hdl.handle.net/10810/4562

A detailed investigation of classification methods for vowel speech imagery recognition

Roberto Santana
email:roberto.santana@ehu.es
Intelligent Systems Group
Department of Computer Science and Artificial Intelligence
University of the Basque Country (UPV/EHU)

Abstract

Accurate and fast decoding of speech imagery from electroencephalographic (EEG) data could serve as a basis for a new generation of brain computer interfaces (BCIs), more portable and easier to use. However, decoding of speech imagery from EEG is a hard problem due to many factors. In this paper we focus on the analysis of the classification step of speech imagery decoding for a three-class vowel speech imagery recognition problem. We empirically show that different classification subtasks may require different classifiers for accurately decoding and obtain a classification accuracy that improves the best results previously published. We further investigate the relationship between the classifiers and different sets of features selected by the common spatial patterns method. Our results indicate that further improvement on BCIs based on speech imagery could be achieved by carefully selecting an appropriate combination of classifiers for the subtasks involved.

keywords: speech imagery, brain computer interface, classification methods.

1 Introduction

Brain computer interfaces (BCIs) translate brain electrical signals into commands without the need for motor intervention [18, 32]. BCIs were originally conceived for providing communication and control to people with severe muscular or neural handicaps [32], but have since then been also applied to applications oriented to healthy individuals [2]. An essential component of a BCI is a decoding algorithm that translates signals to actions. Machine learning methods are usually applied for this purpose [19]. Usually, a subject participates in a series of sessions in which brain recordings are collected and used to train a supervised classifier. In a second stage, the classifier is used online to decode the user's intention from the analysis of his brain recordings.

Most BCIs use the visual modality for control and feedback. However, there are situations where the user may lack the sense of sight or experience difficulties to watch a computer screen. In these situations, non-visual BCI systems can provide an alternative form of communication [30]. One of the most promising types of non-visual BCI systems are speech imagery interfaces where the task consists of decoding speech from neurological recordings. This approach has a number of benefits: it is a natural form of communication, particularly for mobile communication; it is also a more direct and efficient way to control a BCI device; and auditory and tactile feedback could be implemented as an alternative to more framework-dependent visual feedback.

Several studies have shown that it is possible to decode a variety of speech components from neural activity [5, 6, 20, 22]. However, although phoneme and word classification is possible, the conception of practical BCIs based on speech imagery is still a long-term goal. One of the obstacles is that classification accuracies can depend very much on the type of neurological recordings used. While invasive recordings taken from the cortex area show that classification accuracy well above chance level can be obtained, linguistic decoding from EEG recordings have not always been successful. Good results [6, 7, 12, 25, 27, 28, 30, 31], and also lower than chance or very low classification accuracies [4, 8, 23], have been reported.

There are many issues involved in the ability to accurately recover speech imagery from EEG data. For example, it has been hypothesized [23] that high accuracy decoding results for words presented in blocks may be due to temporal correlated artifacts that are detected by the classifiers. In general, further research is needed to elucidate the aspects that influence speech imagery recognition. Two of these aspects are the choice of the classifier and the way this choice is related to the method used for feature selection. Classification algorithms and feature selection strategies play a fundamental role in BCIs and a panoply of methods have been proposed for these purposes [15, 19]. It is expected that gains in vowel imagery decoding could be obtained from a better understanding of how the choice of the classifier is related to this particular type of mental task.

In this paper we present a detailed investigation of how the choice of the classifier and the method for feature selection influence the classification accuracy in the problem of vowel speech imagery recognition from EEG data. Ten different classification methods are applied to EEG data obtained from three subjects in three different tasks: imaginary speech of the English vowels /a/ and /u/, and a no action state as control [9]. We further extend the analysis of the classification algorithms by evaluating how is their behavior related to the features produced by different eigenvalues in the common spatial pattern (CSP) method [24].

The paper is organized as follows. In the next section, the experimental design is presented. Section 3 presents the classification problem, discuss a number of issues related to the CSP methods, and introduces the classification algorithms used in the comparisons. Related work is discussed in Section 5. Section 4 shows the results of the classification algorithms and discuss the implications of these results. Section 6 concludes the paper and present some lines for future research.

2 Experimental design and data acquisition

Three subjects participated in the experiment. Each subject was instructed to perform three mental tasks: imaginary speech of the English vowels /a/ and /u/, and a no action state as control. A randomly selected visual cue was displayed on a computer monitor placed in front of the subject. Vowel /a/ was represented with an image of an open mouth, vowel /u/ with an image of rounded lips, and control with a continuation of the fixation cross. Subjects were instructed to perform and maintain the appropriate task until the visual cue disappeared 2 s later. Each epoch had a duration of 3 s, 1 s of pre-stimulus and 2 s of stimulus. 50 trials were performed for each task, resulting in a total of 150 trials per subject.

EEG was recorded using a BioSemi ActiveTwo system (BioSemi B.V., Amsterdam, Netherlands) with $64 + 8$ active electrodes and a sampling rate of 2048 Hz. Data was downsampled in software to 256 Hz. The 8 extra electrodes were used for: EEG reference (1), measure vertical and horizontal electrooculography (2), and detect unwanted mouth electromyography (2). The remaining 3 extra electrodes were unused. Trials were inspected for movement artifact and only 4 trials in one subject (S2) needed to be rejected and repeated in an extra session.

3 Classification problem and classifiers

The general classification problem consists of decoding, from the EEG recordings of a subject, one of the three possible classes (vowels /a/ and /u/, and a no action state as control). However, we approach the problem as in [9], solving three different binary classification that consists in distinguishing for every possible pair of classes. Each binary problem is called a task and the three possible tasks (F1,F2,F3) are solved for each possible subject (S1,S2,S3). As an initial step CSPs are computed from the EEG signals and features constructed using these patterns are then used for classification.

3.1 Feature selection and common spatial patterns

The goal of the CSP method is to construct a number of distinctive time-series whose variances contain the most discriminative information between the classes [24]. The raw EEG data of a single trial is represented as an $N \times T$ matrix E, where N is the number of channels and T is the number of samples per channel. The normalized spatial covariance of the EEG can be obtained from:

$$C = \frac{EE'}{trace(EE')} \tag{1}$$

where $'$ denotes the transpose operator and $trace(x)$ is the sum of the diagonal elements of x. For each of the two distributions to be separated (i.e., vowels /a/ and /u/), the spatial covariance $\bar{C}_d, \in [a, u]$ is calculated by averaging over the trials of each group. The composite spatial covariance is given as

$$C_c = \bar{C}_a + \bar{C}_u. \tag{2}$$

C_c can be factored as $C_c = U_c \lambda_c U_c'$, where U_c is the matrix of eigenvectors and λ_c is the diagonal matrix of eigenvalues. Note that from now on the eigenvalues are assumed to be sorted in descending order. The whitening transformation

$$P = \sqrt{\lambda_c^{-1}} U_c' \tag{3}$$

equalizes the variances PC_cP' in the space spanned by U_c, i.e., all eigenvalues PC_cP' of are equal to one. If $\bar{C}_1$ and $\bar{C}_r$ are transformed as $S_a = PC_cP'$ and $S_u = PC_cP'$ then S_a and S_u share common eigenvectors B.

The projection of whitened EEG onto the first and last eigenvectors in B will give feature vectors that are optimal for discriminating two populations of EEG in the least squares sense. With the projection matrix $W = (B'P')$, the decomposition (mapping) of a trial is given as $Z = WE$. The columns of W^{-1} are the CSPs.

In our experiments we used the same data investigated in [9] where the two fist and the two last CSPs were used. Each CSP produces a vector of 128 features.

3.2 Classifiers

We select ten classifiers that differ according to their functioning principles, search strategies, and efficiency considerations. Previously, only support vector machines (SVMs) [29] had been applied to this data [9]. The classifiers selected, as implemented in the scikit-learn software [21] programmed in Python language, were:

- Regularized logistic regression with norm l1 (Ll1) [33]
- Regularized logistic regression with norm l2 (Ll2) [33]
- Linear discriminant analysis (LDA) [13]
- k-nearest neighbor classifier (KNN) algorithm [1] with $k = 3$ and using the Euclidean distance
- Gaussian naive Bayes classifier (GNB)
- Gradient boosting (GB) [14] with the number of trees $n_t = 100$ and the maximum depth of the tree $max_d = 11$
- Random forests (RF) [3] with $n_t = 100$ and $max_d = 11$
- Decision tree (DT) $max_d = n$
- Randomized decision trees (RDT) [16]
- Nearest-centroid classifier using Euclidean distance (NCC) [26]

When no information about the parameters is provided above, the classifiers were applied with their defaults parameters in scikit-learn[1].

The classifiers investigated cover the methods most commonly applied to BCI implementations [19]. Some of these classifiers consider interactions between the features, some others incorporate regularization techniques, or take into account similarity metrics between the data.

4 Results

The goal of our experiments were: 1) Evaluate the performance of the classifiers across subjects and tasks when all the information is used. 2) Determine how the choice of the CSP component impacts the classification accuracy.

[1] See `http://scikit-learn.org/stable/index.html` for more details on the code.

4.1 Comparison between the classifiers

We applied the ten classification methods to the set of 512 features. Classifiers were learned using the training data from which the CSPs had been extracted and evaluated on test data. 30 repetitions of the learning process were run. In each repetition, one classifier was learned using 29 of the 30 epochs, for each of the two tasks involved in the binary classification process (e.g., vowel /a/ versus /u/). This is a framework similar to leave-one-out cross-validation but instead of evaluating the classifiers in the fold left out, they were all evaluated in the test data that comprises 20 epochs. Notice that the application of the standard leave-one-out cross-validation method would have implied learning different CSPs in each of the repetitions. By dividing, the cases in two groups, train and test, we compute the CSPs using the complete set of training data only once. We can still compute estimates on the accuracies because each classifier is learned with a different subset of the training data.

Mean and standard deviation of the accuracy of the classifiers on the test data were computed and are shown in Table 1. This table also includes results using a non-linear SVM [29] as presented in [9]. For SVM, only 20 repetitions in two groups were applied. Therefore, these results are included here just as a reference.

In Table 1, the best accuracy for each combination of pairs of tasks and subjects is highlighted in bold. It can be seen from the table that there is a clear split in the behavior of the algorithms among the subjects. GNB is the best algorithm for subject $S1$, reaching accuracies over 89% for all tasks. For subject $S2$, RDT clearly achieves the best accuracies, and for subject $S3$, RDT and RF exhibit the same behavior for task F1, while GNB obtains the best accuracies for tasks F2 and F3. SVM results are clearly improved for subjects S1 and S2 and slightly outperformed for subject S3.

A multiple comparison test using the Tukey's honestly significant difference criterion was applied to the classification results to look for significant differences between algorithms. The output of 30 classifiers was used to assess for these differences. Results are summarized in Table 2 where cell (r, c) indicates the number of times algorithm c was significantly better than algorithm r in the 9 possible combinations. For instance, cell $(1, 4)$ indicates that Ll1 achieved significant better results than KNN in 2 scenarios. Cell $(4, 1)$ indicates that KNN was better than Ll1 in 4 of 9 scenarios. In the remaining 3 scenarios there were not significant differences in the behavior of these two algorithms.

The last row in Table 2 shows the number of times each algorithm was outperformed by the others (o^-). The last column shows the number of times each algorithm outperforms the rest (o^+). The algorithms that showed a clear difference between o^+ and o^- were: RF (22), GNB (20), and RDT (12).

4.2 Influence of the CSP components

In the second part of the experiments, the classifiers were applied to the 128 features associated to each of the 4 CSPs. The goal of the experiment was two-fold: to evaluate the ability of the classifiers to use partial information about the brain signals, and to determine if the classifiers' performance, and consequently the ranking between the algorithms, held for the four groups of variables.

Table 3 shows the best absolute classification accuracies obtained by the algorithms across the four groups. Cells in bold indicate situations where the classification accuracies

S/F	Logistic l1			Gaussian Naive Bayes			Randomized decision tree		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
$S1$	72 ± 3	83 ± 2	68 ± 3	**90** ± 1	92 ± 1	**92** ± 0	87 ± 2	95 ± 2	84 ± 3
$S2$	62 ± 3	67 ± 2	57 ± 3	78 ± 2	64 ± 2	59 ± 1	**79** ± 4	**78** ± 4	**71** ± 4
$S3$	57 ± 3	65 ± 2	55 ± 4	62 ± 2	**80** ± 1	**62** ± 2	**68** ± 4	75 ± 3	57 ± 4
	Logistic l2			Gradient boosting			Nearest centroid classifier		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
$S1$	68 ± 3	83 ± 2	61 ± 3	66 ± 6	82 ± 6	64 ± 6	82 ± 1	88 ± 2	64 ± 2
$S2$	74 ± 2	67 ± 2	57 ± 3	64 ± 3	68 ± 5	70 ± 5	66 ± 2	63 ± 4	59 ± 2
$S3$	57 ± 3	64 ± 1	57 ± 3	61 ± 6	69 ± 5	50 ± 4	59 ± 1	66 ± 1	52 ± 3
	LDA			Random forest			SVM		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
$S1$	72 ± 3	79 ± 4	66 ± 3	87 ± 3	**96** ± 2	81 ± 4	79 ± 3	82 ± 4	72 ± 3
$S2$	65 ± 4	65 ± 4	60 ± 4	75 ± 3	74 ± 5	70 ± 5	71 ± 5	72 ± 4	60 ± 5
$S3$	57 ± 3	65 ± 3	47 ± 3	**68** ± 4	77 ± 3	59 ± 6	67 ± 4	80 ± 3	56 ± 4
	KNN			Decision tree					
	F1	F2	F3	F1	F2	F3			
$S1$	77 ± 2	87 ± 1	77 ± 2	67 ± 4	84 ± 5	62 ± 6			
$S2$	61 ± 2	66 ± 1	64 ± 1	63 ± 4	66 ± 6	69 ± 5			
$S3$	67 ± 2	71 ± 1	58 ± 2	61 ± 6	68 ± 4	50 ± 4			

Table 1: Mean and standard deviation of the classification accuracies obtained by the classifiers on the test data. Classifiers use the 512 features corresponding to the four CSP.

achieved by one of the sets of features associated to any of the four CSPs improved the accuracy obtained by the same classifier for the complete set of features (corresponding cell in Table 1). For all classifiers, there are situations where a subset of the variables improves the accuracies achieved using all the features. Furthermore, in 2 out the 9 cases, corresponding to pairs (S3,F3) and (S3,F1), the best accuracy results were obtained using only a subset of the features.

Figure 1 focuses on the 3 classifiers that produce the highest accuracies: GNB, RF, and RDT which are respectively represented in the figure by their index in Table 2 (i.e., 5, 7, and 9). For these classifiers, the figure shows the accuracies obtained using the features associated to each of the four CSPs. It can be appreciated in the figure how accuracies are highly influenced by the CSPs. For instance, best accuracies for the pair (F1,S1) are obtained by all classifiers by the third CSP. However, for the pair (F2,S1), the best accuracies are achieved using the fourth CSP.

Classifiers	Ll1	Ll2	LDA	KNN	GNB	GB	RF	DT	RDT	NCC	Tot.
Logistic l1	0	5	5	2	1	3	0	3	0	3	22
Logistic l2	3	0	4	2	1	3	0	4	1	3	21
LDA	4	5	0	1	2	3	0	3	0	3	21
KNN	4	2	2	0	3	6	0	6	2	6	31
Gaussian Naive Bayes	1	2	0	2	0	7	5	7	4	7	35
Gradient boosting	4	3	5	3	0	0	1	7	0	5	28
Random forest	1	2	0	5	2	1	0	8	5	9	33
Decision tree	4	2	5	3	1	2	1	0	0	4	22
Randomized decision tree	1	1	0	2	3	2	4	2	0	9	24
Nearest centroid classifier	4	5	4	2	2	3	0	4	0	0	24
Total	26	27	25	22	15	30	11	44	12	49	0

Table 2: Results of the statistical test on the difference between the performance of the classifiers. Cell (r, c) indicates the number of times algorithm c was significantly better than algorithm r in the 9 possible combinations of tasks and subjects.

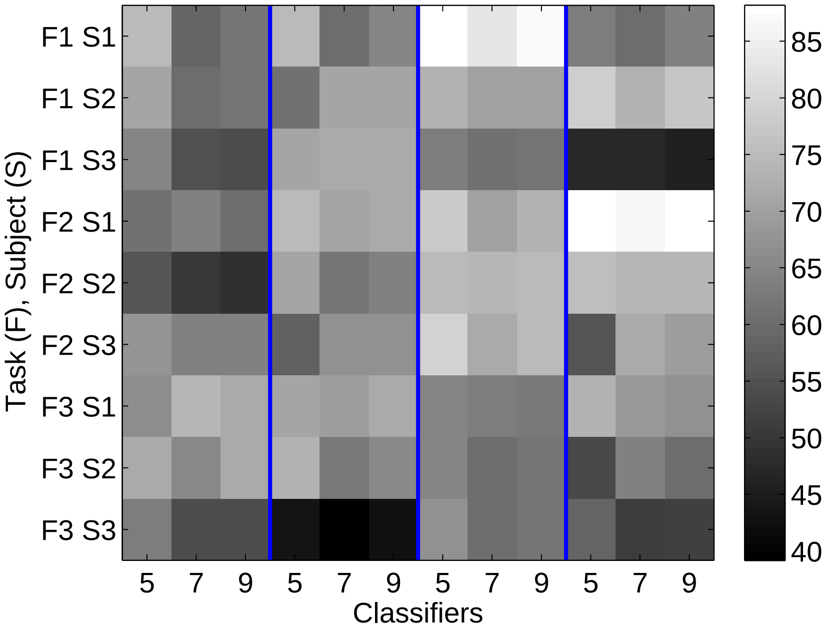

Figure 1: Classification accuracies obtained by GNB, RF, and RDT using features associated to each of the four CSP for all pairs of tasks and subjects.

4.3 Discussion

The analysis of the experiments reveals that tasks can be optimally solved for different subjects using different classifiers. This is particularly evident when observing the behavior of GNB for subject $S1$, and of RDT for subject $S2$. The relevant point here is that classification algorithms may be able to capture different mechanisms related to the vowel imagery decoding. However, it is clear that there is only a small number of classifiers that achieve good results as the statistical tests show. Therefore, it makes sense to split the search for classifiers into two phases. A screening phase in which classifiers are evaluated in the training data and a second phase where the best classifiers found for the combinations of tasks and subjects are applied to test data.

The fact that some problems are better solved using only a subset of variables associated to a single CSP indicates that also for imagery classification problems feature subset selection is critical. Our analysis suggests that separating the features according to the CSP and evaluating the accuracy of each subset of features separately can provide a natural way to diminish the cost of the feature subset selection process. It is also clear that using all the features does not always produces the best accuracies. However, regularized classifiers were not among the best contenders for any of the classification tasks.

5 Related work

A number of papers have proposed different variants for decoding imagined speech from EEG data. Usually, only a subset of vowels and consonants is used for the experiments or the word alphabet is very reduced. Similarly, one or few classifiers are usually used to analyze the data without an exhaustive investigation of the role played by the classifier. In this section, we review some of the previous approaches.

In [25], EEG data was collected from 5 subjects that were instructed to imagine 7 words (internal speech). Prototype waves of each word were learned from the filtered EEG signals from each sensor Test samples were correctly recognized with accuracies between 34% and 97%. One aspect that might help to explain the good classification results achieved in this paper is that the training and test cases were constructed from multiple (10) trials. Another issue, relevant to the decoding results is that no traditional classification algorithm was used since classification was based on the best fit between prototype and test samples.

In [10], EEG signals were analyzed to decode the rhythm in which imagined syllables were produced. 7 subjects performed 120 experiments for each combination of 2 syllables and 3 rhythms. Joint time-frequency analysis was conducted using the Hilbert spectrum (HS) [17]. Features were extracted from the normalized HS, and a Bayesian classifier based on multi-class LDA was applied. Accuracies between 48.33% and 72.67% were obtained for the different subjects. Notably, the relevant features found by the method helped to classify the imagined speech rhythm but failed to classify both rhythm and syllable.

Few papers have addressed the comparison of different classifiers in word imagery decoding problems. One of the few exception is the work presented [27, 28]. This work investigated the ability to decode the imagined words from a reduced five-word vocabulary in EEG signals taken from 21 subjects. Information about four EEG-channels was used for classification. They used discrete wavelet transform as features, and applied three classifiers: naive Bayes, RF, and SVM. RF obtained the best average results considering the 21 subjects. However, as also occurred in our experiments, the ranking between the classifiers varied depending on the subject.

Chi et al. [7] achieved classification accuracies above 70% on pairwise comparisons between five imagined phonemes using LDA. Lower classification accuracies were achieved in the same work using the Naive Bayes classifier. D'Zmura et al [12] asserted the importance of using spectral features in the problem of classifying two different syllables with three different rythms. Hilbert envelopes of each electrode waveform were computed and the average signal envelope across each electrode was used to form a template for each class. Finally matched filters were used for classification. A classification accuracy of 87% was obtained for one of the four subjects included in the experiment. The same dataset was used by Brigham and Kumar [4] that applied autoregressive methods and a 3-Nearest Neighbor classifier. Classification for each subject did not perform better than chance. This is an example of word imagery decoding problems where two classification approaches have produced drastic differences in the obtained accuracies.

The potential use of covert speech for BCIs have been also investigated with other techniques like functional magnetic resonance imaging (fMRI) [20], electrocorticography (ECoG) [22], magnetoencelography (MEG) [25], and micro-electrode recordings [5, 6].

S/F	Logistic l1			Gaussian Naive Bayes			Randomized decision tree		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
$S1$	69 ± 3	82 ± 2	66 ± 3	87 ± 0	87 ± 0	73 ± 2	$\mathbf{87} \pm 2$	88 ± 2	72 ± 3
$S2$	$\mathbf{62} \pm 2$	65 ± 3	56 ± 3	$\mathbf{78} \pm 1$	$\mathbf{75} \pm 1$	$\mathbf{73} \pm 2$	76 ± 3	74 ± 4	$\mathbf{71} \pm 4$
$S3$	50 ± 4	$\mathbf{72} \pm 2$	$\mathbf{56} \pm 3$	$\mathbf{70} \pm 1$	79 ± 2	$\mathbf{66} \pm 1$	$\mathbf{71} \pm 4$	$\mathbf{75} \pm 3$	$\mathbf{61} \pm 3$
	Logistic l2			Gradient boosting			Nearest centroid classifier		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
$S1$	$\mathbf{76} \pm 1$	77 ± 1	$\mathbf{64} \pm 2$	$\mathbf{75} \pm 5$	76 ± 3	63 ± 3	$\mathbf{85} \pm 1$	81 ± 1	61 ± 3
$S2$	60 ± 3	63 ± 2	55 ± 2	$\mathbf{70} \pm 5$	$\mathbf{70} \pm 4$	59 ± 6	$\mathbf{68} \pm 1$	$\mathbf{64} \pm 2$	58 ± 2
$S3$	$\mathbf{61} \pm 3$	$\mathbf{68} \pm 2$	$\mathbf{59} \pm 3$	60 ± 8	$\mathbf{71} \pm 6$	$\mathbf{61} \pm 3$	58 ± 2	$\mathbf{72} \pm 1$	$\mathbf{65} \pm 3$
	LDA			Random forest					
	F1	F2	F3	F1	F2	F3			
$S1$	59 ± 7	61 ± 8	53 ± 9	83 ± 3	86 ± 2	73 ± 3			
$S2$	54 ± 4	60 ± 5	59 ± 4	73 ± 3	$\mathbf{74} \pm 2$	65 ± 4			
$S3$	$\mathbf{60} \pm 7$	56 ± 8	$\mathbf{50} \pm 6$	$\mathbf{71} \pm 5$	71 ± 3	$\mathbf{60} \pm 3$			
	KNN			Decision tree					
	F1	F2	F3	F1	F2	F3			
$S1$	$\mathbf{77} \pm 0$	82 ± 1	65 ± 0	$\mathbf{73} \pm 6$	72 ± 4	61 ± 3			
$S2$	$\mathbf{61} \pm 2$	69 ± 2	$\mathbf{71} \pm 1$	$\mathbf{69} \pm 5$	$\mathbf{66} \pm 5$	59 ± 7			
$S3$	$\mathbf{69} \pm 2$	70 ± 1	$\mathbf{64} \pm 1$	60 ± 5	$\mathbf{69} \pm 5$	$\mathbf{60} \pm 3$			

Table 3: Mean and standard deviation of the classification accuracies obtained by the classifiers on the test data. Each cell shows the best accuracy among the four classifiers corresponding to the four CSP. Accuracies in bold are equal or better than those achieved by the same type of classifiers using the 512 features.

6 Conclusions

In this paper we have conducted an exhaustive investigation of the performance of classification algorithms for a speech imagery decoding problem. We have shown that previously obtained results [9] can be improved for all tasks and subjects. Our empirical results reveal that simpler classifiers like Gaussian naive Bayes that do not consider dependencies between the features can outperform the results obtained with SVM and with other more complex classifiers. However, for one of the subjects, complex classifiers, able to represent dependencies, outperformed all other methods. The fact that classifiers can critically vary their performance accross subjects and tasks involved in vowel imagery decoding seems to indicate that the classifiers can exploit different mental mechanisms. One possible lesson from this is that the behavior of the classifiers could be used to group subjects with similar underlying mechanism and that, when possible, different classifiers should be tried for the tasks and subjects involved. This will not only help to improve the classification results but to better understand how the variability of the speech imagery process is manifested among the subjects.

The other question investigated in this paper is to what extent the choice of the CSP influences the accuracies of the classifiers for the different tasks included in the study. We have shown that the combination of features derived for the four most important CSP can decrease, in certain cases the accuracies achieved by using all the features. Furthermore, we have shown that any of the four CSPs (corresponding to the two first and two last eigenvalues) does not always produce the most discriminative sets of features for all combinations of tasks and subjects.

There are a number of ways the results presented in this paper could be extended. One necessary step is to evaluate the behavior of multi-class classifiers to solve the more challenging 3-state classifcation problem. However, a potential obstacle is that the traditional CSP algorithm is only of application to the binary classification problem. Some CSP extensions to multiclass problems have been proposed [11], however some of these extensions are more costly and it is not clear how the integration with the classification algorithms should be conducted. As another possible development, results presented in this paper could be also applied to design ensembles of classifiers that combine the output of the binary classifiers. Finally, a contrastive analysis of the relevant features found by the different classifiers for each task could help to unveil the potential mechanisms involved in the mental imagery tasks. For instance, one of the questions that the classifiers could help to answer is in which situations classification of EEG signals can be be accurately decoded because of the imagined speech muscle movements or the imagined speech itself [4]

Acknowledgments

This work has been partially supported by Saiotek and Research Groups 2007-2012 (IT-242-07) programs (Basque Government), TIN2010-14931, and COMBIOMED network in computational biomedicine (Carlos III Health Institute).

References

[1] D. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[2] B. Z. Allison, E. W. Wolpaw, and J. R. Wolpaw. Brain–computer interface systems: progress and prospects. *Expert review of medical devices*, 4(4):463–474, 2007.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

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

[5] J. S. Brumberg, F. H. Guenther, and P. R. Kennedy. An auditory output brain–computer interface for speech communication. In *Brain-Computer Interface Research*, pages 7–14. Springer, 2013.

[6] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Frontiers in neuroscience*, 5:65, 2011.

[7] X. Chi, J. B. H. D. Schoonover, and M. D'Zmura. EEG-based discrimination of imagined speech phonemes. *International Journal of Bioelectromagnetism*, 13(4):201–206, 2011.

[8] I. Daly, S. J. Nasuto, and K. Warwick. Towards natural human computer interaction in BCI. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 26, 2008.

[9] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Networks*, 22(9):1334–1339, 2009.

[10] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura. EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of neural engineering*, 7(4):046006, 2010.

[11] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Increase information transfer rates in BCI by CSP extension to multi-class. In *Advances in Neural Information Processing Systems*, pages 733–740, 2004.

[12] M. DZmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan. Toward EEG sensing of imagined speech. In *Human-Computer Interaction. New Trends*, pages 40–48. Springer, 2009.

[13] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[14] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[15] D. Garrett, D. Peterson, C. Anderson, and M. Thaut. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):141–144, 2003.

[16] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[17] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[18] M. Lebedev and M. Nicolelis. Brain-machine interfaces: Past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.

[19] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, and B. Arnaldi. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 4:R1–R13, 2007.

[20] D. M. McCorry. *Using Statistical Classification Algorithms to Decode Covert Speech States with Functional Magnetic Resonance Imaging.* PhD thesis, George Mason University, 2010.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4):046028, 2011.

[23] A. Porbadnigk, M. Wester, J.-P. Calliess, and T. Schultz. EEG-based speech recognition impact of temporal effects. In *Proceedings of the 2nd International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2009)*, pages 376–381, Porto, Portugal, 2009. INSTICC Press.

[24] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *Rehabilitation Engineering, IEEE Transactions on*, 8(4):441–446, 2000.

[25] P. Suppes, Z.-L. Lu, and B. Han. Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, 94(26):14965–14969, 1997.

[26] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[27] A. A. Torres-García. Clasificación de palabras no pronunciadas presentes en electroencefalogramas (EEG). Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México, 2011. In Spanish.

[28] A. A. Torres-García, A. C. Reyes-García, and L. Villaseñor-Pineda. Toward a silent speech interface based on unspoken speech. In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2012)*, pages 370–373. SciTePress, 2012.

[29] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York Inc, 2000.

[30] I. C. Wagner, I. Daly, and A. Väljamäe. Non-visual and multisensory BCI systems: Present and future. In *Towards Practical Brain-Computer Interfaces*, pages 375–393. Springer, 2013.

[31] M. Wester and T. Schultz. Unspoken speech-speech recognition based on electroencephalography. Master's thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2006.

[32] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[33] H.-F. Yu, F.-L. Huang, and C.-J. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011.