EEG DIMENSIONALITY REDUCTION IN AUTOMATIC IDENTIFICATION OF SYNONYMY

Emilio Parisotto[1]*, Youness Aliyari Ghassabeh*[2]*, Siamak Freydoonnejad*[1]*, Frank Rudzicz*[1,2]

[1] Department of Computer Science, University of Toronto; [2] Toronto Rehabilitation Institute-UHN; Toronto ON Canada

ABSTRACT

Recent work has demonstrated the feasibility of extracting semantic categories directly from cortical measures (e.g., electroencephalography, EEG) during receptive tasks. Here, we automatically classify speech stimuli as either synonymous or non-synonymous with a prior prime in a speech-receptive task given only EEG data with up to 86.84% accuracy. An analysis of variance reveals no significant difference among support vector machine and k-nearest neighbours classifiers, but a significant effect of the individual subject on accuracy. To perform classification, we reduce the highly-parameterized space by three successive techniques: a ranking based on t-test similarity, another based on principal components analysis (PCA), and a third on linear discriminant analysis.

***Index Terms*—** Electroencephalography, feature selection, semantic classification

1. INTRODUCTION

Feature extraction and selection are especially pertinent to electroencephalography (EEG) signal classification, given the high degree of noise, low spatial resolution, and relatively large channel redundancy of EEG data. Despite these inherent challenges, EEG signal classification is increasingly popular across applications including silent-speech interfaces [1, 2], biometric authentication [3], epilepsy prediction [4], and brain-computer interfaces (BCI) for spelling [5].

This paper explores a variety of EEG-based features in a semantic binary classification task to distinguish between speech stimuli that are either synonymous or non-synonymous with an initial prime word. To find relevant features, we rank a large pool of stochastic features with t-tests of significant difference between classes. Given this ranking, we further reduce the dimensionality using principal components analysis (PCA) and linear discriminant analysis (LDA). The final reduced feature set is then sent through two binary classifiers: a support vector machine (SVM) and k-nearest neighbours (KNN).

1.1. Background

Previous work in classification of EEG signals has used a variety of features including auto-regressive models [2, 3] and common spatial pattern (CSP) filters, the latter of which has been effective within silent-speech interfaces [1]. The participants in that study either imagined speaking one of two vowels (/*u*/ or /*a*/) or, as a control, remained alert without any conscious effort [1]. The CSP method was used to maximize the discriminative variance between each of the 3 pairwise classification subtasks (/u/:/a/, /u/:control, /a/:control) and a binary non-linear SVM classifier was run on each subtask. Classification accuracies ranged from 67-79% for /a/:control, 72-82% for /u/:control, and 56-72% for /u/:/a/.

A more typical approach is to preprocesses EEG signals using ICA and to learn the coefficients of a univariate autoregressive (AR) model as a feature set. For example, Brigham and Kumar [2] used this approach along with artefact-rejection and source-selection based on the Hurst exponent to distinguish syllables /*ba*/ and /*ku*/. While accuracy was not significantly higher than chance across all their subjects, the same method without the Hurst exponent selection process was shown to be highly effective at identifying the thinker's identity given their EEG data [3]. This AR approach was extended to a much larger data set of 120 subjects where subjects were presented with visual stimuli intended to induce a visual evoked potential, providing up to 98.96% accuracy in subject identification [3].

Dal Seno *et al.* used a genetic algorithm to choose a discriminative set of features for use in a logistic classifier to detect a P300 event-related potential [5], which further drove an online BCI speller. Similarly, D'Alessandro *et al.* used genetic algorithms to select a subset of a wide range of signal features, with the goal of predicting epilepsy in a patient using a probabilistic neural network classifier [4].

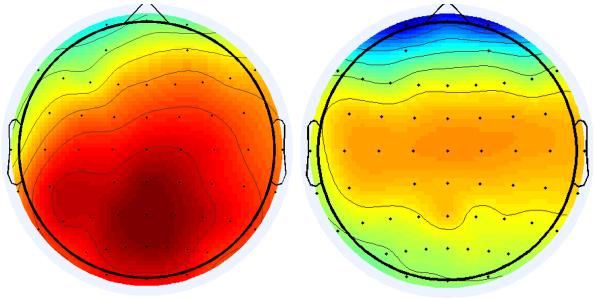
2. DATA

Our EEG data are adapted from Hohlfeld *et al.* [6]. Here, subjects were first presented with an auditory 'prime' word, followed immediately by a second auditory 'target' word. Subjects were instructed to determine whether the target was synonymous or non-synonymous with the prime as quickly as possible, and press one of two buttons to express that decision. Each subject was presented with 120 distinct synonymous noun pairs and 120 distinct non-synonymous noun pairs. Each pair was used twice for a total of 480 trials pre-

sented in different random orders for each subject. Here, we use all data from all 5 subjects that were available to us. Averages of event-related potentials across all stimuli for one subject, for both synonym and non-synonym targets, are shown in Figure 1

The EEG data were recorded using 60 channels sampled at 200 Hz, with electro-oculogram (EOG) signals simultaneously recorded to facilitate artefact removal. The EEG data was re-referenced to the average of the A1 and A2 electrodes and segmented into fixed-length trials. Certain single trials were rejected if at any point the signal amplitude went above a $\pm$ 75 μV threshold [6]. In this paper, we performed a preprocessing step to further remove ocular artefacts using EEGlab's eye movement correction procedure (EMCP) [7, 8]. Each trial was cropped to a period beginning at the target stimulus onset and ending one second later.

Fig. 1: The neural activity of subject 2, averaged over a) synonym, and b) non-synonym sets.



(a) (b)

3. FEATURE EXTRACTION

For each trial and for each channel, the signal is first segmented into overlapping windows, empirically determined to be about 10% of the epoch length, with a window overlap of 50%. We then capture a number of features from each window, namely the minimum and maximum values, the mean, standard deviation, and variance over the epoch, maximum - minimum, maximum + minimum, as well as skewness and kurtosis where

$$\textit{skewness} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \Big/ \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3,$$

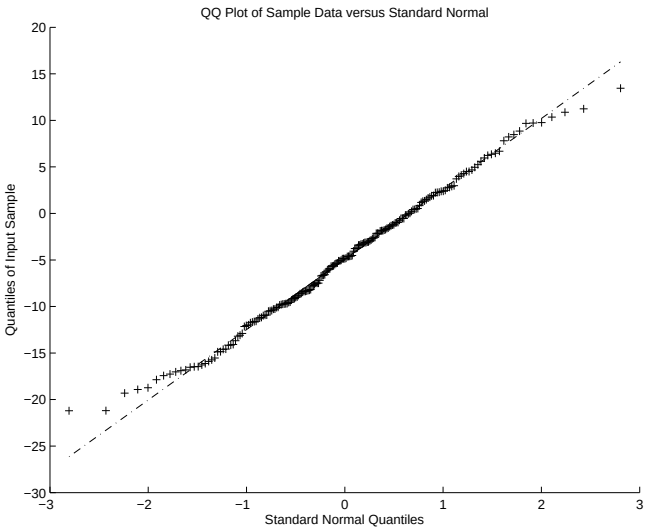
and

$$\textit{kurtosis} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \Big/ \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2,$$

where x is the windowed signal, n is the number of samples in the window and $\bar{x}$ is the window mean. Additionally, we take the mean and minimum value over the absolute values of the window. These features are concatenated into a vector for each window, along with their velocities and accelerations (1^{st} and 2^{nd} frame differences) .

The matrix of features, their velocities, and their accelerations over all windows in an epoch yields a matrix on the order of 10,000 dimensions, which is unsuitably large given the amount of training data. To overcome overfitting, we score each dimension according to the t-test, and take a subset with empirically low associated p-values. These p-values are computed using Welch's t-test between the synonymous and the non-synonymous groups. Theoretically, lower p-values indicate that the groups are more clearly differentiable along the given dimensions, which can in turn be ranked. Empirically, choosing the 100 dimensions with the lowest p-values was found to be most effective. Since we are not performing hypothesis testing, handling multiple comparisons does not apply (applying Bonferroni correction would not affect the ranking, in any case). Although the p-values calculated here do not assume equal variances between the distributions of the synonyms and non-synonyms, it does assume that the feature dimension in each case is normally distributed. The Lilliefors test [9] on each dimension reveals that 59.4% and 59.9% of all features are normally distributed in the synonym and non-synonym data, respectively, at $\alpha = 0.05$. An example is shown in Figure 2.

Fig. 2: A quantile-quantile plot showing an example feature dimension in the synonym case, revealing the dimension is approximately normal.



4. DIMENSIONALITY REDUCTION

PCA linearly transforms a set of variables into new set of orthogonal variables called principal components in which each

successive dimension contains a maximum amount of the iteratively remaining data. Here, $\boldsymbol{x} \in \mathbb{R}^D$ is a vector consisting of D correlated zero mean random variables. We linearly transform the input data to a d dimensional ($d << D$) linear subspace that captures most of the variability in $\boldsymbol{x}$. To find the first principal component, we find a unit vector $\boldsymbol{b}_1 \in \mathbb{R}^D$ such that the variance of the projected data $\boldsymbol{b}_1^\intercal \boldsymbol{x}$ is maximized. If $\boldsymbol{y} = \boldsymbol{b}_1^\intercal \boldsymbol{x}$ is the projected data, then the variance is given by $E(\boldsymbol{y}^2) = \boldsymbol{b}_1^\intercal \boldsymbol{\Sigma} \boldsymbol{b}_1$, which $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{x}$, i.e., $\boldsymbol{\Sigma} = E(\boldsymbol{x}\boldsymbol{x}^\intercal) - E(\boldsymbol{x})E(\boldsymbol{x})^\intercal$. We find the unknown vector $\boldsymbol{b}_1$ by solving a constrained optimization problem using Lagrange multipliers as follows [10]

$$\arg\max_{\lambda, \boldsymbol{b}_1} \boldsymbol{b}_1^\intercal \boldsymbol{\Sigma} \boldsymbol{b}_1 - \lambda_1(\boldsymbol{b}_1^\intercal \boldsymbol{b}_1 - 1), \tag{1}$$

where λ_1 is the Lagrange multiplier. It is straightforward to show that λ_1 must be the largest eigenvalue of the covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{b}_1$ is the eigenvector corresponding to the largest eigenvalue. In general it can be shown that for the kth principal component of $\boldsymbol{x}$, $\boldsymbol{b}_k^\intercal \boldsymbol{x}$, we have $E(\boldsymbol{b}_k^\intercal \boldsymbol{\Sigma} \boldsymbol{b}_k) = \lambda_k$, where λ_k is the kth largest eigenvalue of the covariance matrix and $\boldsymbol{b}_k$ is the corresponding eigenvector.

LDA searches for directions providing the maximum linear discrimination of classes while reducing overall dimensionality [11, 12]. To achieve this, within- and between-class scatter matrices are defined. The within-class scatter matrix, $\boldsymbol{\Sigma}_W$, represents the scatter of samples around their class means, and the between-class scatter matrix, $\boldsymbol{\Sigma}_B$, represents the scatter of class means around the total mean. LDA looks for the direction in which maximum class separability is achieved by projecting the data into those directions. After this projection, all the samples belonging to the same class stay close together and well-separated from those of the other classes. The LDA transformation matrix, $\boldsymbol{\Phi}_{LDA,\Delta}$, into a Δ-dimensional ($\Delta < D$) space is given by Δ leading eigenvectors of $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$ [12]. If K denotes the number of classes and $Rank(\boldsymbol{\Sigma}_B) \leq K - 1$, then the reduced dimension by the LDA technique is at most $K - 1$, i.e., $\Delta \leq K - 1$. Therefore, instead of finding leading eigenvectors of $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$, one can solve the generalized eigenvalue problem $\boldsymbol{\Sigma}\boldsymbol{\Phi}_{LDA} = \boldsymbol{\Sigma}_W \boldsymbol{\Phi}_{LDA} \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the diagonal eigenvalue matrix and the desired Δ LDA features are given by p columns of $\boldsymbol{\Phi}_{LDA}$ corresponding to the largest eigenvalues of $\boldsymbol{\Lambda}$.

5. EXPERIMENTS

In this paper, we use the proportion of total variance [10] to estimate the optimal number of principal components. In other words, the number of principal components, l, is the smallest integer that satisfies the inequality

$$\frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \geq 0.98. \tag{2}$$

where D is the total number of principal components, $\lambda_i, i = 1, \ldots, D$ is the ith eigenvalue of the covariance matrix of the observed data, and l is the number of components that compromise 98% of the total variance. The value of l depends on the statistics of the data and is not fixed, for example for the first and the third subject we found $l = 25$ and for the second subject $l = 30$.

As mentioned, the total number of LDA features, Δ, is always less than or equal to the number of classes minus one. Since we have just two classes in these experiments (synonym or non-synonym), we must set $\Delta = 1$, meaning the output of the PCA step will project into a one-dimensional space.

We compare two classifiers on the reduced dimensions. The linear support vector machine (SVM) classifier [13], which tries to find a decision boundary (a hyper-plane for linear SVM) to separate two classes with the largest possible margin [14]. The k-nearest neighbours (KNN) classifier is a simple supervised non-parametric algorithm that classifies observations based on a similarity measure (i.e., Euclidian distance here) to previously seen examples. For the KNN classifier, we set $k = 1$ (see section 6). Figure 3 shows the average accuracy of the SVM classifier over varying number of principal components, with standard error.

For each subject, we randomly choose 90% of available samples as the training set and the remaining 10% as the test set, constituting 5 sets of thinker-dependent models. Figure 4 shows the accuracies of applying SVM and KNN on the one-dimensional LDA features, as well as the accuracy from applying SVM on the raw 100-dimensional features. Given a two-way analysis of variance, there is no significant difference among the classifiers ($F_2 = 0.21, p = 0.82$, but there are significant differences due to subjects ($F_4 = 11.91, p < 0.005$), with mean accuracies ranging from 65.22% to 84.21%.

6. DISCUSSION

Recent work has demonstrated the feasibility of extracting semantic categories during receptive language, given fMRI [15] and EEG [16] data. In this paper, we automatically classify speech stimuli as either synonymous or non-synonymous with a prior prime in a receptive task given only EEG data with up to 86.84% accuracy, although this is highly dependent on the individual subject. To do so, we reduce the highly parameterized space by three successive techniques: a ranking based on t-test similarity, PCA, and LDA.

Several approaches have been proposed to determine the appropriate number of principal components (e.g., see [17, 18]). Although the method of proportion of total variance (used in this paper) is generally advocated by statisticians [10], there has been some evidence against its reliability [19]. A comparison of optimization methods, given these data, is the subject of future work. Furthermore, we are considering alternatives to LDA, including locally linear embedding

Fig. 3: Mean classification accuracy (error bars are $\sigma/\sqrt{n}$) using the SVM classifier as a function of the number of principal components.

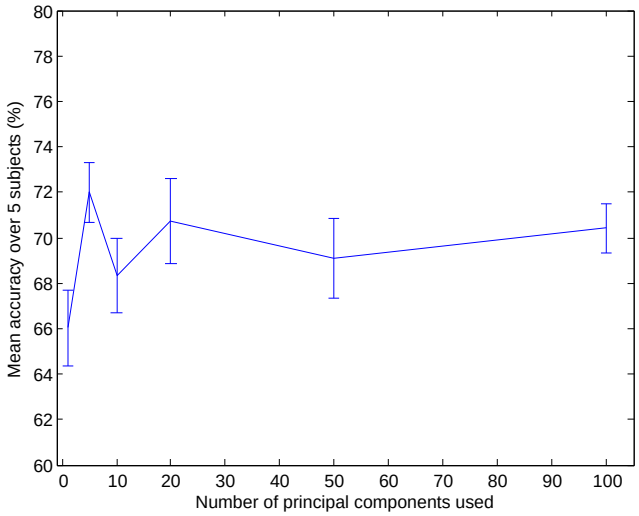
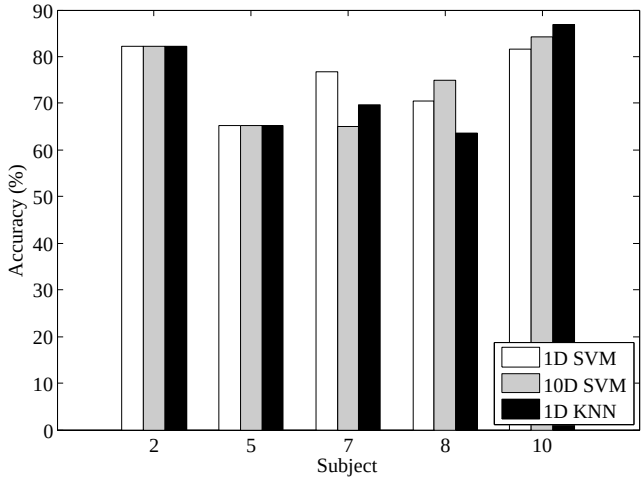


Fig. 4: Classification accuracy for each subject and classifier. Note that not all subjects in the original work by Hohlfeld *et al.* [6] were available to us, but the original subject ID numbers are retained.



(LLE) [20], ISOMAP [21], multidimensional scaling (MDS) [22], and principal curves [23]. For example, we repeated the experiment for the fourth subject by replacing LDA with LLE, which is a well-known nonlinear dimensionality reduction technique that maps the high dimensional data into a lower-dimensional space such that the local structure in the data is preserved during the mapping[1] [20]. However, the resultant accuracy of 70.45% did not provide any gains.

Although not mentioned in section 5, we also run a C4.5 decision tree and a naïve Bayes classifier on our data, but accuracies were in all cases lower than the reported SVM results. We have also swept over other values of k in the KNN classifier without significantly improving performance, although further optimization is the subject of future work.

Ongoing work involves exploring additional features and augmenting the simple window functions used here. In preliminary experiments, features based on the Fourier transform of short-time windows have not performed particularly well. This may be due to increased non-normality of those features, or to possible decreases in temporal resolution. We are also exploring AR coefficients [2, 3] and other features [4].

Finally, the recorded EEG signals may be corrupted with artefacts due to scalp impendence or subject movement. Therefore, before the feature selection step, we need to perform noise reduction to estimate the clean EEG signals. The main challenge here is that the statistics of the noise are not known *a priori* and there is no training data available to estimate the clean signal by observing noisy versions. Ongoing work involves applying unsupervised [24] estimation of clean versions of EEG signals as a pre-processing step.

7. ACKNOWLEDGEMENTS

This research is funded by a startup grant from the Toronto Rehabilitation Institute - University Health Network, a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN 435874), and a grant from the Nuance Foundation. We use Matthias Ihrke's EEGlab ocular correction plugin during preprocessing.

8. REFERENCES

[1] C. S. DaSalla, H. Kambara, Y. Koike, and M. Sato, "Spatial filtering and single-trial classification of EEG during vowel speech imagery," in *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology, Singapore*. 2009, i-CREATe 09, pp. 27:1–27:4, ACM.

[2] K. Brigham and B. V. K. V. Kumar, "Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepa-

[1] I.e., nearby points in the high dimensional space remain near to each other in the low dimensional space.

thy," in *Fourth International Conference on Bioinformatics and Biomedical Engineering (ICBBE), Chengdu, China*, June 2010, pp. 1–4.

[3] K. Brigham and B. V. K. V. Kumar, "Subject identification from electroencephalogram EEG signals during imagined speech," in *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS) , Washington, USA*, Sep. 2010, pp. 1–8.

[4] M. D'Alessandro, R. Esteller, G. Vachtsevanos, A. Hinson, J. Echauz, and B. Litt, "Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: A report of four patients," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 5, pp. 603–615, 2003.

[5] B. Dal Seno, M. Matteucci, and L. Mainardi, "Online Detection of P300 and Error Potentials in a BCI Speller," *Computational Intelligence and Neuroscience*, vol. 2010, no. 11, 2010.

[6] A. Hohlfeld, P. Ullsperger, and W. Sommer, "How does the incrementality of auditory word perception interplay with episodic and semantic memory?," *Journal of Neurolinguistics*, vol. 21, no. 4, pp. 279–293, 2008.

[7] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, 2004.

[8] G. Gratton, M. G.H Coles, and E. Donchin, "A new method for off-line removal of ocular artifact," *Electroencephalography and Clinical Neurophysiology*, vol. 55, no. 4, pp. 468–484, 1983.

[9] H.W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, pp. 399–402, 1967.

[10] I. T. Jolliffe, *Principal Component Analysis*, Springer, 2002.

[11] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley-Interscience, 2004.

[12] Y. A. Ghassabeh and H. A. Moghaddam, "Adaptive linear discriminant analysis for online feature extraction," *Machine Vision and Applications*, vol. 24, no. 4, pp. 777–794, 2013.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2000.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[15] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, pp. 1191–1195, 2008.

[16] B. Murphy, M. Baroni, and M. Poesio, "EEG responds to conceptual stimuli and corpus semantics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pp. 619–627.

[17] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, "How many principal components? Stopping rules for determining the number of non-trivial axes revisited," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005.

[18] D. A. Jackson, "Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches," *Computational Statistics and Data Analysis*, vol. 78, no. 8, pp. 2204–2214, 1993.

[19] J. E. Jackson, *A User's Guide to Principal Components*, Wiley-Interscience, 2003.

[20] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[21] J. B. Tenenbaum, V. de Silva, and J. C. Langfordn, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[22] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling,*, Chapman and Hall, 2000.

[23] Y. Aliyari Ghassabeh, T. Linder, and G. Takahara, "On some convergence properties of the subspace constrained mean shift," *Pattern Recognition*, vol. 46, no. 11, pp. 3140–3147, 2013.

[24] Y. A. Ghassabeh, T. Linder, and G. Takahara, "On noisy source vector quantization via a subspace constrained mean shift algorithm," in *26th Biennial Symposium on Communications, Kingston, Canada*, May 2013.