# Automatic identification of received language in MEG

*Emilio Parisotto*[1], *Youness A. Ghassabeh*[2], *Matt J. MacDonald*[3], *Adelina Cozma*[4], *Elizabeth W. Pang*[5], *Frank Rudzicz*[1,2,*]

[1] Department of Computer Science, University of Toronto; [2] Toronto Rehabilitation Institute-UHN; [3] SickKids Research Institute; [4] University of Calgary; [5]Neurology, Hospital for Sick Children; Toronto ON Canada

(*) frank@cs.toronto.edu

## Abstract

We identify the language being received during English and Romanian auditory stimuli in 11 subjects before and after a period of learning 50 words in the latter using only magnetoencephalographic measures. To accomplish this, we extract on the order of 100,000 features (based on wavelets and descriptive statistics over windowed signals), and identify the most salient features. While we achieve very high accuracy in pre-training (up to 90% mean accuracy across 10-fold cross-validation for some subjects), it is significantly more difficult to tell received languages apart after training. We also identify significant effects of semantic word category and the subject's ability to play a musical instrument on classification accuracy.

**Index Terms**: Magnetoencephalography, feature selection, language classification

## 1. Introduction

Automatically classifying magnetoencephalography (MEG) data presents several challenges including high dimensionality, low signal-to-noise ratio, and high inter-channel redundancy. These factors, along with comparatively low trial counts in many MEG experiments, can lead to overfitting. To overcome these challenges, we examine a method of feature selection and dimensionality reduction that reduces an initially very high-dimensional feature space into a more succinct, low-dimensional representation which still maintains discriminative information. Specifically, we classify the language of received heard utterances given only MEG signals recorded during the time the word was spoken.

Previous work on MEG classification includes detecting hand movement [1], identifying schizophrenia [2], and on discriminating between sets of imagined words [3]. To classify between three different hand movements[1], Asano *et al.* [1] used an adaptive spatial filter, principal components analysis (PCA) and a support vector machine (SVM) to achieve 62.6% on held-out test data. In Ince *et al.* [2], a subject performed a working memory functional task while MEG data were recorded; an SVM with recursive feature elimination (SVM-RFE) was then used to both select a concise feature set and to identify schizophrenia. SVM-RFE recursively discarded features that did not significantly contribute to the margin of the SVM classifier to prevent excessive overfitting on the training set, and achieved 83.8% to 91.9% on the test data.

Closer to our work, Guimaraes *et al.* [3] classified sets of 7-9 imagined words in two subtasks. In the first, the subject was simply required to attentively listen to a spoken word, while in the second the subject was shown each word visually and told to recite it silently. Those data were then examined using linear discriminant classification and SVM algorithms to classify each channel, and further analyzed in terms of the effects of spatial PCA, independent components analysis (ICA) and second-order blind identification decomposition. By combining channels, Guimaraes *et al.* achieved 60.1% mean classification rate on nine auditory words and 97.5% maximum mean classification rate on two-word problems.

[1]Corresponding to the signs in the game of 'rock, paper, scissors'.

## 2. Data

The data used in this paper were originally from a neuroimaging study examining how language learning over an extended period affects semantic processing [4]. Each subject learned 50 words in a new language, Romanian, over a two-week period; this length of time avoids possible effects of short-term memory on language discrimination. The MEG data were collected for each subject during a receptive language comprehension task involving one session prior to learning any Romanian words and another after language training. In each case, subjects were presented with an auditory word in either English or Romanian and instructed to choose one of two pictures whose meaning coincided with that word. These sessions used 50 English and 50 Romanian words, all of which were distinct in meaning. Each word was repeated twice for a total of 200 trials per subject, although visual stimuli were all unique images. Each 4.5 second epoch started with the auditory word being presented during a 1 second interval, after which the two images were presented for 3.5 seconds.

The data were continuously acquired by a whole-head 151-channel MEG with a 625 Hz sampling rate. The signals were segmented into the 200 trials and then downsampled to 100 Hz to remove the high frequency noise components and transform the data into a more manageable form.

Fourteen subjects participated (7 females, 7 males; mean age = 28.4 ($\sigma = 4.7$)), 11 of whom are considered here due to missing demographic information and unresolved inconsistencies in the data for two subjects. All subjects were right-handed, as confirmed by the Edinburgh Handedness Inventory and spoke English as a first language except two (French and Swiss German, respectively). All participants completed the Peabody Picture Vocabulary Test (3rd ed.; PPVT) [5] and the Expressive Vocabulary Test (EVT) [6]. Subjects also completed a questionnaire that determined, among other things, their primary language, second language (if applicable), and whether they play a musical instrument. These factors are explored in

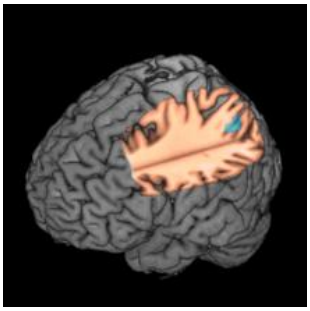

Figure 1: Regions showing maximal changes post-training for Romanian during the receptive task, adapted from [4].

section 4.1.

Source localization analyses in [4] showed pre-dominant and significant changes pre-/post-training for Romanian in the receptive task in the left superior parietal lobule (Fig. 1).

# 3. Methods

To classify the MEG data, we first transform all raw data from their original sensor space into 'source' space using independent components analysis (ICA). Many potentially redundant features are then extracted from the transformed data and concatenated together for each trial, as described in section 3.2. To offset the resulting high dimensionality, we score features heuristically according to their discriminability (according to Welch's $t$-test) between the two language classes. This precedes a method of coarse dimensionality reduction that maximizes the retained information. Finally, the transformed data are sent through a non-linear support vector machine, which classifies each MEG trial as resulting from either a Romanian or English spoken word.

To test the generalizability of our system, we use 10-fold cross-validation. In each fold, for each subject independently, 20 trials (10 from each language) are held out as the test set and all ICA weights, PCA components, and classifiers are trained on the remaining 180 trials. This was repeated using 10 distinct subsets of 20 trials and the final averaged classification performance is reported.

## 3.1. Blind source separation

Blind source separation (BSS) tries to recover the original source signals from their mixture without *a priori* knowledge of the source signal [7]. This can be reduced to finding a linear representation with (maximally) statistically independent components. Given a vector of $n$ observations at time $t$, $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]^\intercal$, we apply ICA, which models each $x_i$ as a linear mixture of independent components (sources), $s_i, i = 1, \ldots, n$. We then have $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^\intercal$, and an unknown mixture matrix $\mathbf{A}$[7]. ICA looks for an 'un-mixing' matrix $\mathbf{W} = \mathbf{A}^{-1}$ such that $\mathbf{s} \approx \mathbf{W}\mathbf{x}$. There are two main families of ICA algorithms [8, 9] which generate slightly different $\mathbf{W}$. Some implementations try to minimize the mutual information and use measures such as Kullback-Leibler divergence or maximum entropy [10]. Others are based on the maximization of non-Gaussianity, as measured by kurtosis or negentropy [7]. Before applying ICA on our data, we first zero-mean all observations and 'whiten' the observed variables; that is, we linearly transform observation vector $\mathbf{x}$ such that the components of the new vector $\tilde{\mathbf{x}}$ are uncorrelated and $E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal) = \mathbf{I}$. For our ICA decomposition, we use the logistic infomax algorithm [11]. This is a gradient-based neural network algorithm that uses higher-order statistics for the information maximization [12]. Specifically [13, 14],

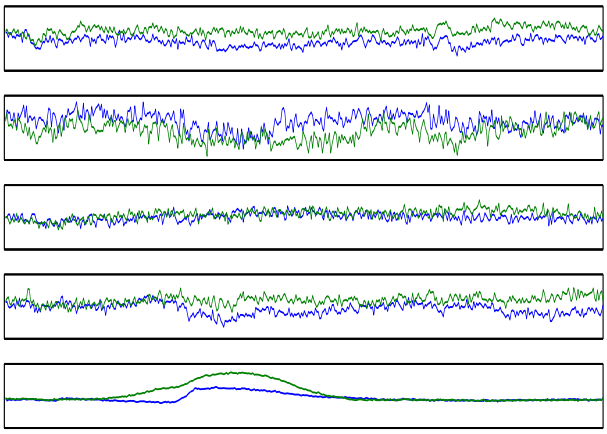

Figure 2: ICA channels with top 5 greatest mean projected variance, averaged over all receptive trials. Blue curves are given Romanian stimuli, and green curves are given English.

1. We choose an initial un-mixing matrix $\mathbf{W}_0$,
2. We update $\mathbf{W}_{k+1} = \mathbf{W}_k + \eta_k(\mathbf{I} - g(\mathbf{y})\mathbf{y}^\intercal)\mathbf{W}_k$,
3. Normalize $\mathbf{W}_{k+1} = \mathbf{W}_{k+1}/\|\mathbf{W}_{k+1}\|$, and
4. If not converged, go back to 2.

Here, $\|\mathbf{W}\|$ is the matrix norm, $\mathbf{y} = \mathbf{W}\mathbf{x}$, $\mathbf{I}$ is the identity matrix, and $g(\mathbf{y}) = \mathbf{y} - \tanh(\mathbf{y})$. Figure 2 shows the ICA channels having the maximal variance, averaged over all receptive trials. The sudden inflection in the fifth ICA channel appears to correspond strongly to the onset of the image one second after the auditory stimulus.

## 3.2. Feature extraction

Features are extracted from the ICA-transformed source space. Specifically, we apply the discrete wavelet transform (DWT), often used in electroencephalography (EEG) classification tasks [15, 16], and also extract some descriptive statistics within sliding windows over each trial.

In contrast to the short-time Fourier transform (STFT), continuous wavelet transforms (CWTs) can decompose a signal to have both high temporal resolution (for short, high-frequency events) and high spectral resolution (for long, low-frequency events). They accomplish this by shifting and scaling a base function (i.e., the 'mother wavelet') and convolving it with the input signal [17]:

$$X(\tau, \alpha) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} x(t) h^* \left( \frac{t - \tau}{\alpha} \right) dt, \qquad (1)$$

where $^*$ is the complex conjugation operator, $\alpha$ is the scale parameter, $\tau$ is the shift parameter, and $h$ is the mother wavelet. The DWT is a discretized version of the CWT where the scale and shift factors are sampled at discrete points [17].

To reconstruct the original input signal from the wavelet coefficients, certain conditions must hold with regards to the discretized mother wavelet. In this paper, we choose the scale and

shift factors used for computing the DWT using *dyadic sampling*, where $\alpha_m = 2^m$ and $\tau_n = n\alpha_m$ with $n, m \in \mathbb{N}$. For discrete time signals, the DWT coefficients can be calculated as [17]:

$$c_{n,m} = \sum_{p=-\infty}^{\infty} x[p]h^*_{n,m}[p]. \quad (2)$$

The DWT has an efficient implementation wherein both a high-pass and low-pass decomposition are applied to the input signal, with each filtered signal then downsampled by a factor of 2. This is repeated recursively on the low-passed signal until a stopping criterion is reached. The values of the low- and high-passed signals on each iteration are called the 'approximation' and 'detail' coefficients, respectively. The low- and high-pass filters ($q$ and $r$, respectively) are calculated from the mother wavelet as follows [17]:

$$q[n] = (-1)^n h[-n+1], \quad (3)$$

$$r[n] = h[n]. \quad (4)$$

Here, the mother wavelet is the Daubechies DB4 wavelet with 5 levels of decomposition. The Daubechies family of orthogonal wavelets are commonly used in the context of EEG epilepsy prediction [15, 16]. We additionally use the wavelet band energy as a feature, which is calculated by taking the energy of each group of detail or approximation coefficients. The wavelet bands are 50-25 Hz, 25-12.5 Hz, 12.5-6.25 Hz, 6.25-3.125 Hz, 3.125-0 Hz. Once the wavelet coefficients and their energies have been calculated, estimates of their velocities and accelerations are also calculated.

We additionally derive descriptive statistics on overlapping windows of each MEG channel. These windows are empirically chosen to be approximately 10% of the total epoch length, with 50% overlap between adjacent windows. From each of these windows, we calculate: minimum, maximum, mean, maximum $+/-$ minimum, standard deviation, variance, skewness, kurtosis, sum, median, energy, and an estimate of the integral (by trapezoidal numerical integration).

### 3.3. Feature selection and dimensionality reduction

The wavelets, descriptive statistics for each window, and their velocities and accelerations, for each channel are all concatenated into a single (long) vector trial on the order of 100,000 features. To prevent overfitting, we perform Welch's $t$-test separately on each feature and sort the results by the resulting $p$-values in decreasing order. This approximates the discriminability of the two language classes for each feature. Empirically, the 150 features with the lowest $p$-values gave the lowest classification error, among all alternatives tested. Welch's $t$-test depends on the data under comparison to be normally distributed. The Lilliefors test [18] on a subset of the data reveals that only 16.4% and 16.8% of features are not normally distributed in the English and Romanian stimuli, respectively, at $\alpha = 0.05$.

Before classification, these 150-dimensional feature vectors (for each trial) are further processed using principal components analysis (PCA)[19]. The number of components, $N$, was selected for each subject individually in order to capture 97% of the variance (typically, $10 \leq N \leq 20$).

## 4. Experiments

The processed MEG data are classified using a nonlinear support vector machine (SVM) [20]. Given a set of training pairs, $(\mathbf{x}_i, y_i), i = 1..n$, where $\mathbf{x}_i \in \mathbb{R}^D$ is the observed vector and $y_i \in \{-1, 1\}$ is the corresponding label, we find the optimal hyperplane by optimizing its normal vector $\mathbf{w}$, intercept $b$, and slack variables $\xi_i, i = 1..n$ [21] by:

$$\begin{aligned} &\arg\min_{\mathbf{w},b,\xi} 1/2\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\ &\text{s.t. } y_i(\mathbf{w}^\intercal x_i + b) \geq 1 - \xi_i, \text{ and } \xi_i \geq 0, \text{ for } i = 1..n \end{aligned} \quad (5)$$

where the constant $C > 0$ is the penalty parameter of the error term. We tried the polynomial ($K_{poly}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\intercal \mathbf{x}_j + 1)^d$, $d$ varied empirically) and radial-basis ($K_{rbf}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma > 0$) kernels [20]. The $K_{poly}$ kernel with $d = 2$ gave the best results and is hereafter reported.

Figure 3: Mean accuracy for each subject on the pre- and post-learning trials with 95% confidence intervals.

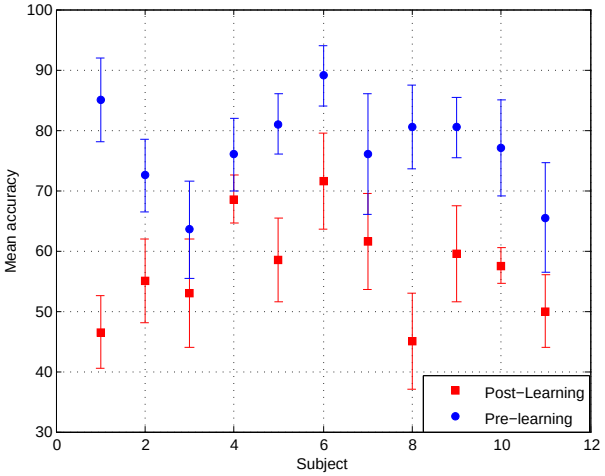


Figure 3 shows the mean accuracy for each subject on the pre- and post-learning trials. A full 2-way ANOVA reveals significant effects of subject ($F_9 = 8.18, p < 0.001$) and training ($F_1 = 243.87, p < 0.001$) on classification accuracy, and significant interaction between subject and training ($F_9 = 4.94, p < 0.001$). Intuitively, integration of Romanian into the left superior parietal lobule may partially explain why identifying the received language from MEG becomes closer to chance after training, and this is being investigated.

### 4.1. Effects of behavioural demographics

Table 1 shows a linear 5-way ANOVA for behavioural demographics, namely first and second language, PPVT and EVT fluency scores, and whether the subject plays a musical instrument. Interestingly, one's first language has a mild effect on accuracy, but bilingualism does not. If subjects play a musical instrument, the received language becomes significantly harder to discriminate, on average.

### 4.2. Effects of stimuli

Across both languages, stimuli words were designed to evenly distribute across five semantic categories: *Animal* (e.g., 'zebra'), *Clothing and accessories* (e.g., 'necklace'), *Food products* (e.g., 'pumpkin'), *Found in the home* (e.g., 'fireplace'), and *Found in the workplace* (e.g., 'stapler'). The average accuracies across 11 subjects are shown in table 2, given the period of recording and the stimulus language.

Table 1: ANOVA given behavioural demographics.

| Source | Sum Sq. | F | $p$ |
|---|---|---|---|
| $1^{st}$ Lang. | 1466.8 | $F_2 = 3.25$ | 0.04 |
| $2^{nd}$ Lang. | 831.2 | $F_2 = 1.84$ | 0.16 |
| Music | 1544.16 | $F_1 = 6.84$ | <0.01 |
| PPVT | 11.8 | $F_1 = 0.05$ | 0.82 |
| EVT | 222.2 | $F_1 = 0.98$ | 0.32 |

A 3-way linear ANOVA reveals significant effects of the subject ($F_{10} = 5.58, p < 0.001$), received language ($F_1 = 31.65, p < 0.001$), and category of word ($F_4 = 29.48, p < 0.001$) on the accuracy of language classification. Here we use all 11 subjects – in section 4.1 we had to exclude one person with incomplete demographics. Across stimuli language and period of testing, the *Food* was most discernible (69.66%) followed by *Animals* (67.73%), *Found in the home* (67.71%), *Found at work* (65.11%), and *Clothing* (64.66%). Whether this order corresponds to the increasing age-of-acquisition of their component words is yet to be determined.

Table 2: Average (and $\sigma$) classification accuracies (%) across word categories, language, and pre/post-learning.

| | English | | Romanian | |
|---|---|---|---|---|
| | PRE | POST | PRE | POST |
| Animal | 82.73 (10.8) | 56.82 (19.0) | 75.91 (14.1) | 55.45 (10.8) |
| Clothing | 71.82 (12.1) | 53.64 (17.6) | 77.27 (13.3) | 55.91 (14.3) |
| Food | 82.27 (10.8) | 57.73 (9.8) | 80.91 (8.9) | 57.73 (10.3) |
| In home | 78.64 (15.5) | 54.09 (13.0) | 76.82 (10.1) | 60.91 (12.4) |
| At work | 69.55 (11.3) | 57.73 (12.1) | 73.64 (11.6) | 59.55 (8.5) |

## 5. Discussion

This paper is the first, to our knowledge, that identifies received language in MEG (or EEG, for that matter). After reducing a very large feature space of wavelets and descriptive statistics across the 151 available channels using a combination of Welch's $t$-test and PCA, we achieve high accuracy pre-training (up to 90%), and note that it becomes significantly more difficult to identify received languages after training. This could suggest that the newly-learned Romanian words are being integrated into the subject's existing word knowledge base. We also find that both the semantic category of stimuli words and the subject's ability to play a musical instrument significantly affects classification accuracy.

Future work involves extending these methods to an expressive task in MEG where subjects were shown an image and asked to name the object in either English or Romanian [4]. We are currently looking into which parts of the brain contain the most discriminative features for language- and category-identification, and how they relate across receptive and expressive tasks. The joint tasks will allow us to correlate discriminative features across both receptive and expressive language, which will hopefully illuminate possible shared semantic processes underlying both tasks, within theoretical models of joint speech production and perception [22].

## 6. Acknowledgements

The infomax function was implemented in EEGlab [23]. This research is funded by the Toronto Rehabilitation Institute - UHN, the Natural Sciences and Engineering Research Council of Canada (RGPIN 435874), and a grant from the Nuance Foundation. Data acquisition was funded by CIHR MOP-89961.

## 7. References

[1] F. Asano, M. Kimura, T. Sekiguchi, and Y. Kamitani, "Classification of movement-related single-trial meg data using adaptive spatial filter," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 357–360.

[2] N. Ince, F. Goksu, G. Pellizzer, A. Tewfik, and M. Stephane, "Selection of spectro-temporal patterns in multichannel MEG with support vector machines for schizophrenia classification," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, Aug 2008, pp. 3554–3557.

[3] M. Guimaraes, D. Wong, E. Uy, L. Grosenick, and P. Suppes, "Single-trial classification of MEG recordings," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 436–443, March 2007.

[4] M. MacDonald, A. Cozma, A. Oh, and E. Pang, "Object naming in a foreign language: a pre/post MEG study of language learning," in *Proceedings of the 18th International Conference on Biomagnetism*, Paris France, August 2012.

[5] L. Dunn and L. Dunn, *Peabody picture vocabulary test*, 3rd ed. Circle Pines, MN: American Guidance Service, 1997.

[6] K. Williams, *Expressive vocabulary test*. Circle Pines, MN: American Guidance Service, 1997.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, 2001.

[8] S. Haykin, *Neural Networks and Learning Machines*. New Jersey: Pearson Education, Inc., 2009.

[9] D. Langlois, S. Chartier, and D. Gosselin, "An introduction to independent component analysis: InfoMax and FastICA algorithms," *Tutorials in Quantitative Methods for Psychology*, vol. 6, no. 1, pp. 31–38, 2010.

[10] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[11] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computing*, vol. 7, no. 6, pp. 1129–1159, 1995.

[12] G. R. Naik and D. K. Kumar, "An overview of independent component analysis and its applications," *Informatica*, vol. 35, no. 1, pp. 63–81, March 2011.

[13] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Adv. Neural Inf. Process. Syst., MIT Press, Cambridge, MA*, 1998, pp. 757–763.

[14] A. Delorme, J. Palmer, J. Onton, R. Oostenveld, and S. Makeig, "Independent EEG Sources Are Dipolar," *PLoS ONE*, vol. 7, no. 2, 2012.

[15] M. Saab and J. Gotman, "A system to detect the onset of epileptic seizures in scalp EEG," *Clinical Neurophysiology*, vol. 116, pp. 427–442, 2005.

[16] A. Zandi, M. Javidan, G. Dumont, and R. Tafreshi, "Automated real-time epileptic seizure detection in scalp EEG recordings using an algorithm based on wavelet packet transform," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1639–1651, July 2010.

[17] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2001.

[18] H. 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.

[19] 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.

[20] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*. The MIT Press, 1998.

[21] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Tech. Rep., 2003.

[22] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[23] 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.