PAPER

Classification of visual comprehension based on EEG data using sparse optimal scoring

Journal of Neural Engineering

PAPER

Classification of visual comprehension based on EEG data using sparse optimal scoring

RECEIVED
22 September 2020

REVISED
21 December 2020

ACCEPTED FOR PUBLICATION
13 January 2021

PUBLISHED
3 March 2021

Linda K Ford[1], **Joshua D Borneman**[2], **Julia Krebs**[3,4], **Evguenia A Malaia**[5] **and Brendan P Ames**[1]

[1] Department of Mathematics, The University of Alabama, Tuscaloosa, AL 35487-0350, United States of America
[2] Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN 47907, United States of America
[3] Centre for Cognitive Neuroscience (CCNS), University of Salzburg, Salzburg, Austria
[4] Research Group Neurobiology of Language, Department of Linguistics, University of Salzburg, Salzburg, Austria
[5] Department of Communicative Disorders, The University of Alabama, Tuscaloosa, AL 35487-0350, United States of America

E-mail: lkford1@crimson.ua.edu

Keywords: discriminant analysis, classification, optimal scoring, EEG, sign language

Abstract

Objective. Understanding and differentiating brain states is an important task in the field of cognitive neuroscience with applications in health diagnostics, such as detecting neurotypical development vs. autism spectrum or coma/vegetative state vs. locked-in state. Electroencephalography (EEG) analysis is a particularly useful tool for this task as EEG data can detect millisecond-level changes in brain activity across a range of frequencies in a non-invasive and relatively inexpensive fashion. The goal of this study is to apply machine learning methods to EEG data in order to classify visual language comprehension across multiple participants. *Approach.* 26-channel EEG was recorded for 24 Deaf participants while they watched videos of sign language sentences played in time-direct and time-reverse formats to simulate interpretable vs. uninterpretable sign language, respectively. Sparse optimal scoring (SOS) was applied to EEG data in order to classify which type of video a participant was watching, time-direct or time-reversed. The use of SOS also served to reduce the dimensionality of the features to improve model interpretability. *Main results.* The analysis of frequency-domain EEG data resulted in an average out-of-sample classification accuracy of 98.89%, which was far superior to the time-domain analysis. This high classification accuracy suggests this model can accurately identify common neural responses to visual linguistic stimuli. *Significance.* The significance of this work is in determining necessary and sufficient neural features for classifying the high-level neural process of visual language comprehension across multiple participants.

1. Introduction

The goal of cognitive neuroscience is to describe in a mechanistic model the human ability for perceiving sensory signal, and manipulating it as input for higher cognition. To understand the interaction of perception and cognition, and predict the behavioral outcome that results from the two, is the ultimate challenge of the field. Within this long-term goal, a more specific task of classifying human brain states based on neural data has been tackled using multiple methods. Electroencephalography (EEG) data has been particularly useful for this purpose, as it provides high temporal resolution, capturing millisecond-level changes in brain activity across a range of frequencies. It is also non-invasive and relatively cheap to collect, as compared to other neuroimaging modalities. However, the challenge of interpreting neural signals—i.e. connecting the dynamics in the data with specific brain states and cognitive processing—remains a complex problem.

Multiple EEG-data-based analyses relied on machine learning approaches to help interpret the functionality of brain states in the data [1]. Achieving robust classification of EEG data across populations or tasks could yield critical information for use in health applications (e.g. early, non-invasive diagnostics), and to understanding functional significance of specific features of brain activity in BCI applications. To situate the present study within the

body of EEG classification research, we first briefly review the approaches toward EEG data classification.

1.1. Classification targets

The goals of classification can differ substantially between studies, and theoretical understanding of differences in brain states is often the grounds for methods selection. One frequently posed goal is that of differentiating between populations on the basis of patient EEG with the purpose of health diagnostics, e.g. for early detection of Autism Spectrum Disorders in young children; Parkinsons's or Alzheimer's in elderly patients; or attempting to differentiate between coma/vegetative state vs. locked-in states using individual brain activity, cf [2–4]. A substantial body of work also attempts to classify brain states as being related to specific tasks in higher cognition, either within a single individual (e.g. for BCI applications for locked-in patients), or, more generally, in neurotypical participants as they engage in a specific task (e.g. as experiencing a specific emotion, viewing/listening to a specific type of stimulus, performing under increased mental workload), or assume non-intentional states, such as seizure onset, or sleep stages [1]. Among the types of tasks that focus on identifying higher-level cognitive processes, detection of language comprehension is of special importance in health research. To give one example, patients in a medically induced coma, or are in a minimally conscious state (MCS) who are incapable of providing overt responses, often demonstrate changes in neural activity in response to verbal prompts. Cruse *et al* [5] detected significant sensorimotor activations in neural activity of 19% of the participants in minimally conscious states, suggesting that the patients could process language. Coleman *et al* [6] demonstrated that a subset of patients fulfilling the behavioural criteria for the vegetative state retained islands of preserved cognitive function, as identified by EEG data. Classifying features of neural activity characteristic of comprehending language would be critical for health applications, as EEG can be easily recorded at bedside and does not depend on explicit cooperation of the patient. Depending on the goal of classification, a variety of EEG-based inputs can be used in the analysis, from raw EEG signal, to calculated features (e.g. ERPs/evoked response potentials—deviations at the specific time in response to specific stimuli), or spectrotemporal parameters of the data [1, 7, 8]. Frequency-domain EEG data especially can characterize differences between brain states that appear similar in the time domain [9], but are associated with objective differences in behavior.

1.2. Visual communication/comprehension as classification target

When people listen to speech, neural activity tracks the entropy fluctuation in the acoustic envelope of the signal [10, 11]. For spoken languages, the ability for this sensory signal-based entrainment has been identified as the basis of speech parsing and comprehension [12]. At the same time, humans rely on the visual system as the primary source for early conceptual feature development, and fundamental cognitive processes, as scene/event segmentation [13]. Machine learning approaches such as support vector machines (SVM) and decision trees have been previously successfully applied for Parkinson's disease detection/classification, based on spectral EEG during visual stimulation [4]. Sign languages, which rely on visual signal in information transfer, also rely on entropy of the visual signal [14, 15]. Specifically, sign language communication has higher Shannon entropy in the signal [16, 17], as compared to non-linguistic human motion. Building on our understanding of signers' sensitivity to entropy of the information-bearing visual signal, we tracked the cortical dynamics of comprehension in the visual modality using optical flow measures in the visual signal. In order to focus on the higher cognitive task of language processing, we controlled for lower-level perceptual (i.e. spectrotemporal) features of the stimuli by using the same videos reversed in time as control condition, and eliciting overt behavioral judgements of their interpretability from participants. We then used peak coherence between the visual entropy dynamics of the stimuli, and individual EEG, to classify EEG segments into 'comprehension' and 'no comprehension'.

1.3. Machine learning approaches for classification

Classification is a classical supervised learning task with many well-established heuristics for the task, including but certainly not limited to nearest neighbors and centroids methods, discriminant analysis, support vector machines, Bayesian methods, perceptron methods, generative additive and tree models, logistic regression, as well as convolutional neural networks and other deep learning approaches; Hastie *et al* [18] provides an excellent summary of classical approaches while the surveys [19, 20] provide an introduction to deep learning approaches for classification.

We use sparse discriminant analysis as the foundation of our classification process due to the increased interpretability of discriminative models over other classical methods. Linear discriminant analysis (LDA) can be thought of as a supervised variant of principal component analysis (PCA). Instead of mapping the data to a lower dimensional space where the coordinate axes align with the directions of maximum variance, as in PCA, we want to find a projection onto a lower dimensional space where the linear separability between different classes in the training data is maximized. We do so by identifying a subspace where the ratio of the spread between projected class means

and total variance within projected classes is maximized. The basis vectors for this subspace, called discriminant vectors, provide one means of interpreting the data: the discriminant vectors indicate the directions in the original feature space in which the classes differ most significantly. By adding a sparsity inducing penalty function to the chosen optimization model for LDA, we can add a second source of interpretability via feature selection. In this case, many of the features of the calculated discriminant vectors are zero; the remaining nonzero entries reveal which features are significant for distinguishing between classes. We discuss the application of sparse linear discriminant analysis, via the SOS criterion, in section 2.5.1; a detailed summary of linear discriminant analysis can be found in [18, section 4.3] and more information about sparse discriminant analysis can be found in [21, section 8.4], as well as the foundational works [22–27].

1.4. Significance

The significance of the present work is in isolating features of neural response that are necessary and sufficient for classification of high-level neural process of language comprehension across multiple participants. The oscillatory response features revealed by successful classification are promising as biomarkers to elucidate the underlying neurophysiology of language comprehension, and, more generally, higher cognition.

2. Methods

This section is organized as follows. Sections 2.1–2.4 outline how the data was collected and processed. The various classification methods used to analyze the data are discussed in sections 2.5.1–2.5.3. Specifically, in section 2.5.1 we discuss the method of sparse optimal scoring in depth, as it was the primary method used for our analysis. These classification methods were applied to two forms of the EEG data, time-domain and frequency-domain, which are discussed in sections 2.6.1 and 2.6.2 respectively. Finally, specific algorithms used for the various classification methods are given in sections 2.7.1–2.7.3. In particular, we introduce a novel algorithm for sparse optimal scoring in section 2.7.2 as our primary algorithm used in our analysis.

2.1. Participants

24 participants (13 males) aged between 28 and 68 years (M =42.04, SD = 12.27) took part in the experiment. All participants were assessed by a certified sign language interpreter as proficient users of Austrian Sign Language (ÖGS). All reported normal or corrected-to-normal vision, and no history of neurological disorders. All procedures were undertaken in accordance with the Declaration of Helsinki; all participants provided informed consent prior to the study, and were compensated for their time. At the time of data collection, there was no Institutional Review Board at the University of Salzburg. Since then, a University of Salzburg Institutional Review Board has been established, and the methods described in the study are covered by an approved protocol *Neurophysiologische Untersuchung von Sprachverarbeitungs- und Sprachlernprozessen*, Protocol number: EK-GZ: 07/2018. Before the start of the study, each participant was shown a block of videos to allow for a short practice in carrying out the task, and clarification in case of additional questions. Participants were instructed to avoid excessive motion during the presentation of the video material, and were allowed breaks between sets as needed.

2.2. Stimuli and procedures

The stimulus set consisted of 40 videos of full sentences in Austrian Sign Language (ÖGS), which were produced by a Deaf fluent signer. An illustration of a stimulus video with still frames is given in figure 1. Each video was, additionally, time-reversed, resulting in a video that was uninterpretable in sign language, but identical to the original video in spectrotemporal parameters (i.e. luminance, color, vertical and horizontal frequencies, speed of motion). Thus, a total of 80 videos were presented to the participants: 40 were videos of sign language sentences in time-direct condition, with comprehensible sign language; 40 were time-reversed videos, with uninterpretable contents. As distractors, 200 additional videos of meaningful ÖGS sentences were interspersed with the 80 trial videos. The duration of videos ranged from 5 to 8 s. Participants performed a Likert scale judgement task by rating each video's acceptability as interpretable sign language sentence, with 1 meaning 'this video is not (interpretable as) sign language', and 7 meaning 'very good sign language'.

Before each video presentation, an attentional fixation crosshatch was presented on the screen for 2000 ms; this was followed by an empty black screen for 200 ms. Stimulus video was then presented in the center of the screen. After the video, a green question mark appeared for 3000 ms, indicating the time during which participants were asked to rate the videos on a Likert scale between 1 and 7. Participants provided the ratings by a button-press on the keyboard.

2.3. Data acquisition and pre-processing

26-channel EEG with two additional mastoid channels was recorded using active electrodes; AFz electrode functioned as the ground. Raw EEG signal was recorded with a sampling rate of 500 Hz.

For amplifying the EEG signal we used a Brain Products amplifier (high pass: 0.01 Hz). In addition, a notch filter of 50 Hz was used, which is the main frequency source in Austria/EU and a potential source of noise.

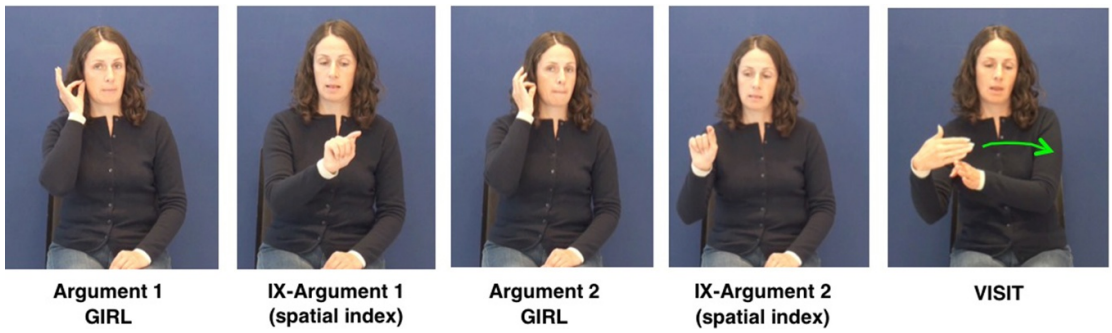

Figure 1. Illustration of sign language video with still frames.

All electrodes were referenced against the electrode on the left mastoid bone. Active electrodes in the elastic cap (Easy Cap, Herrsching-Breitbrunn, Germany) were arranged according to the standards of the 10/10 system (Fz, Cz, Pz, Oz, F3/4, F7/8, FC1/2, FC5/6, T7/8, C3/4, CP1/2, CP5/6, P3/7, P4/8, O1/2). Horizontal and vertical eye movements (HEOG, VEOG) were recorded by additional electrodes at the lateral ocular muscles and above and below the left eye, respectively; this data was used for artifact rejection procedure at a later step. Electrode impedances were kept below 5 kΩ throughout the recording. Time stamps for the start of each trial were sent by the stimulus presentation computer and recorded as part of the EEG file.

Following each recording, electrodes were re-referenced against the mastoid averages in Brain Analyzer software. The signal was filtered with a bandpass filter (Butterworth Zero Phase Filters; high pass: 0.1 Hz, 48 dB Oct^{-1}; low pass: 20 Hz, 48 dB Oct^{-1}). High-pass 0.1 Hz filtering is a standard step in EEG studies of language and cognition due to the need to eliminate the slow signal drift [28]. 20 Hz low-pass filter was chosen in view of the plan to perform correlation analysis between EEG data and optical flow in the video. As the frequencies in the video are limited by Nyquist frequency, i.e. 30 fps/2 [16], 20 Hz was selected as a reasonable compromise for low-pass threshold for EEG filtering. The signal was further corrected for ocular artifacts by the Gratton and Coles method [29], and screened for artifacts using minimal/maximal amplitude thresholds at $-75/+75$ μV. EEG was segmented into epochs from the onsets of video to 5 s following the onset, such that only neural responses to ongoing video stimuli were used. This resulted in time-series data consisting of 5000 time points per channel for each stimulus video.

2.4. Calculation of optical flow and EEG coherence to optical flow

We quantified the signal in each stimulus video, which are the time-direct and time-reversed videos fully described in section 2.2, using changes of optical flow across multiple visual frequencies in time. Optical flow of a video frame is the distribution of apparent velocities of objects in an image. To compute optical flow, a velocity vector (in pixels/frame) is found for each pixel, based on how fast and in which direction the feature shown in that pixel has moved from the frame before. Optical flow for each video was determined using the MathWorks' MATLAB vision toolbox optical flow function. This function was utilized to compare each video frame with the prior frame using Horn–Schunck method [30]. This resulted in an output matrix of size equal to the input video frame. Each element of the matrix identified the magnitude of optical flow velocity (pixels per frame) between the two frames for each corresponding pixel in the video. An optical flow histogram (i.e. a velocity spectrum) was then created for each video frame. The amplitudes across all velocity bins for each frame were added to calculate an integrated magnitude of optical flow for each frame, thereby generating an optical flow time-series [16]. The power spectral density (PSD) of optical flow per frame was calculated using MATLAB's 'pwelch' PSD estimate. This produced a single measure of power spectral density dynamics for each frame of the video (30 fps).

This measure was then linearly regressed against EEG signal of each participant, and peak cross-correlation frequency between visual stimuli and EEG was extracted for each video and each participant's every channel in the EEG data. To do this, coherence was calculated between the optical flow time-series of each stimulus video and the EEG in every electrode for each participant. To compute coherence at a given frequency (with the frequencies ranging from 0.04 Hz to 12.4 Hz in 0.2 Hz bins, as limited by the 30 fps recording rate of the video), both time-series were first filtered at that frequency using a second-order IIR bandpass filter. The cross-correlation was then calculated using canonical component analysis of MATLAB NoiseTools toolbox (cf [31]). Peak correlation was then extracted for each frequency for each electrode, participant, and video. The coherence was calculated separately for each condition, averaging the response data across videos grouped by comprehended sign language vs. non-comprehended reversed videos, for each of the four regions: frontal (comprising data from electrodes in positions F3, F4,

Fz, FC1, and FC2), left (electrodes C3, FC5, T7, CP1, CP5), right (C4, FC6, T8, CP2, CP6), posterior (P7, P8, P3, P4, Pz). This resulted in 62 entrainment measures, one for each 0.2 Hz frequency bin, for each of the four regions.

2.5. Classification methods

2.5.1. Sparse optimal scoring (SOS)

We used Sparse Optimal Scoring (SOS) to train a classifier to predict which type of video each participant was watching, either time-direct sign language or time-reversed sign language, based on time-domain and frequency-domain data. SOS is a form of Linear Discriminant Analysis (LDA), which is a classical method for performing supervised classification and dimension reduction. Specifically, LDA is used to project high-dimensional data to a lower dimensional space so that spread between data in different classes is maximized while variability within classes is minimized in the low-dimensional space; the basis vectors for this low-dimensional space generated by LDA are called discriminant vectors. We then perform nearest centroid classification in the low-dimensional space. Specifically, we first calculate the projection of each class mean or centroid onto the span of the discriminant vectors and then assign any unlabeled data to the class of the closest centroid following projection onto the span of the discriminant vectors. When the number of training observations (n) exceeds the number of predictor variables (p), we can obtain the discriminant vectors by solving a generalized eigenproblem. Unfortunately, the change of variables needed to solve this eigenproblem is undefined in the high-dimensional setting where $p > n$.

Classical optimal scoring recasts LDA as a regression problem using a sequence of scores to transform the categorical labels into quantitative variables. Specifically, optimal scoring calculates each discriminant vector using linear regression, with class scores used as response variables and scores chosen so that fitting error in the resulting linear model is minimized. In order to perform LDA in the high-dimensional setting where $p > n$, Clemmensen *et al* [23] impose an elastic-net penalty on the optimal scoring problem. When many predictor variables are highly correlated, similar classification performance can be obtained using relatively few variables, rather than all predictor variables; the correlation ensures that information from all variables is represented by only a few variables. It is well-known that the elastic-net penalty induces sparse solutions when applied to linear regression problems (see [32]); in this case, the use of the elastic-net penalty ensures that the discriminant vectors generated by SOS are sparse. This simultaneously solves the undersampling problem when $p > n$, since the number of nonzero predictor models in the discriminant model is typically much fewer than n, and yields a more interpretable or explainable model since only a relatively small number of variables are used to make classifications.

We calculate the set of discriminant vector and optimal scoring vector pairs in sequence as follows. Suppose that the first j discriminant vector pairs $(\boldsymbol{\beta}_1, \boldsymbol{\theta}_1)$, $(\boldsymbol{\beta}_2, \boldsymbol{\theta}_2), \ldots, (\boldsymbol{\beta}_j, \boldsymbol{\theta}_j)$ have been calculated. We calculate the next pair by solving the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \ \|\boldsymbol{Y}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \gamma \boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_1 \\ &\text{s.t.} \ \frac{1}{n}\boldsymbol{\theta}^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta} = 1 \\ &\quad \boldsymbol{\theta}^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_\ell = 0, \ \ \ell = 1, 2, \ldots, j, \end{aligned} \tag{1}$$

where $\boldsymbol{X}$ is an $n \times p$ data matrix consisting of n observations and p predictors, K is the number of classes, $\boldsymbol{Y}$ is an $n \times K$ indicator matrix for the K classes, $\boldsymbol{\theta}$ is a K-vector of class scores, $\boldsymbol{\beta}$ is a p-vector of variable coefficients, $\boldsymbol{\Omega}$ is a $p \times p$ positive definite regularization matrix, and λ and γ are non-negative tuning parameters. The data matrix $\boldsymbol{X}$ is assumed to be centered and standardized so that $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = 1 \ \ \forall j = 1, \ldots, p$. It is easy to see that the problem when $j = 0$, i.e. to calculate the first pair, has a trivial solution given by $(\boldsymbol{\beta}_1, \boldsymbol{\theta}_1) = (\boldsymbol{0}, \boldsymbol{e})$, where $\boldsymbol{e}$ is the all-ones vector in $\mathbb{R}^K$. This process is continued until K discriminant and scoring vector pairs are obtained (including the trivial solution). The constraints imposed on $\boldsymbol{\theta}$ ensure that the set of optimal scoring vectors forms an orthonormal basis for $\mathbb{R}^K$ with respect to the inner product defined by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{n}\boldsymbol{x}^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{y}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K$. For more information on the sparse optimal scoring formulation of LDA, see [23].

In this study, we have $K = 2$ classes corresponding to participants viewing time-direct or time-reversed videos. The SOS problem simplifies significantly in this setting. In particular, we need only calculate one nontrivial solution: $(\boldsymbol{\beta}_2, \boldsymbol{\theta}_2)$. Moreover, there are exactly two unit vectors orthogonal to $\boldsymbol{e}$, both of which span the same line in $\mathbb{R}^2$; when the two classes contain the same number of training observations, these are $(1, -1)/\sqrt{2}$ and $(-1, 1)/\sqrt{2}$. Thus, we need only calculate the discriminant vector $\boldsymbol{\beta}$ with $\boldsymbol{\theta}$ chosen to be one of these optimal scoring vectors; the SOS problem reduces to the following elastic net regularized linear regression problem in this case:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \gamma \boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_1. \tag{2}$$

We delay discussion of numerical methods for solving this problem until sections 2.7.1 and 2.7.2.

To perform nearest centroid classification in the two-class setting, we calculate the dot product between the nontrivial discriminant vector $\boldsymbol{\beta}_2$ and the sample mean for each class, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, to compute the projection of the centroids $\boldsymbol{\beta}_2^T \boldsymbol{\mu}_1$ and $\boldsymbol{\beta}_2^T \boldsymbol{\mu}_2$ onto the line spanned by $\boldsymbol{\beta}_2$. For each test observation $\boldsymbol{x}$, we obtain a class label by projecting onto

the span of $\boldsymbol{\beta}_2$ by $\boldsymbol{\beta}_2^T\boldsymbol{x}$; we then assign $\boldsymbol{x}$ to the class of the nearer of $\boldsymbol{\beta}_2^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta}_2^T\boldsymbol{\mu}_2$. In the special case that the two classes have the same number of training observations, then the assumption that the data matrix is centered ensures that $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ and $\boldsymbol{\beta}_2^T\boldsymbol{\mu}_1 = -\boldsymbol{\beta}_2^T\boldsymbol{\mu}_2$. In this case, we assign $\boldsymbol{x}$ based on the sign of the projection $\boldsymbol{\beta}_2^T\boldsymbol{x}$. This leads to a natural interpretation of the contribution of the individual entries of $\boldsymbol{\beta}_2$ to the classification process. Recall that the dot product is defined by

$$\boldsymbol{\beta}_2^T\boldsymbol{x} = \sum_{i=1}^{p} [\boldsymbol{\beta}_2]_i x_i \tag{3}$$

i.e. the sum of the products of corresponding entries of $\boldsymbol{\beta}_2$ and $\boldsymbol{x}$. Thus, a relatively large positive product of the form $[\boldsymbol{\beta}_2]_i x_i$ would contribute heavily to (potentially) classifying $\boldsymbol{x}$ as belonging to one class, while a large negative value of $[\boldsymbol{\beta}_2]_i x_i$ would contribute more significantly towards classification in the other class; it is important to note that the class score depends on all summands, and the presence of a single large positive or negative term is not an indicator of class membership on its own. The entries of the Hadamard or entry-wise product $\boldsymbol{\beta}_2 \circ \boldsymbol{x}$ are defined by

$$[\boldsymbol{\beta}_2 \circ \boldsymbol{x}]_i = [\boldsymbol{\beta}_2]_i x_i. \tag{4}$$

Following the previous argument, the value of each entry of the Hadamard product can be used to interpret the behaviour of the classifier function defined by the discriminant vector $\boldsymbol{\beta}_2$. We will discuss this in more detail later with specialization to our frequency-domain analysis.

2.5.2. Support vector machine (SVM)

To test if a classifier which produces a non-linear decision boundary would perform better on the time-domain data than SOS, which produces a linear decision boundary, we trained classifiers using support vector machines with a radial kernel and with a polynomial kernel. The support vector machine (SVM) is an extension of the support vector classifier which can accommodate non-linear class boundaries. For the two-class case, the support vector classifier generates a separating hyperplane to classify the observations based on which side of the hyperplane the observation lies. The hyperplane is chosen to maximize the distance between the observations and the hyperplane while minimizing the number of observations that are misclassified. It can be shown that the support vector classifier can be computed using only the inner products of the observations, where the inner product of two observation $\boldsymbol{x}_1, \boldsymbol{x}_2$ is given by $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \sum_{j=1}^{p} [\boldsymbol{x}_1]_j [\boldsymbol{x}_2]_j$ where p is the number of predictors [33]. The support vector machine extends this approach to non-linear class boundaries by enlarging the feature space using a generalization of the inner product called a kernel. The polynomial kernel is given by

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left(1 + \sum_{j=1}^{p} [\boldsymbol{x}_1]_j [\boldsymbol{x}_2]_j\right)^d \tag{5}$$

where d is a positive integer called the degree [33, chapter 9]. The radial kernel given by

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\gamma \sum_{j=1}^{p} ([\boldsymbol{x}_1]_j - [\boldsymbol{x}_2]_j)^2\right) \tag{6}$$

where γ is a positive constant. For more details on support vector machines, see chapter 9 of [33] and chapter 12 of [18].

2.5.3. Elastic-net regression

To test if the behavioral responses to the stimuli could be used to classify the data, we fit a naive model using elastic-net regression to predict the rating given to each stimulus video by each participant. Elastic-net regression uses standard linear regression to predict a quantitative outcome with an elastic-net penalty applied to improve the fit and reduce the number of non-zero predictors in the model. The elastic-net penalty is given by

$$\gamma \sum_{j=1}^{p} \boldsymbol{\beta}_j^2 + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \tag{7}$$

where p is the number of predictors, $\boldsymbol{\beta}_j$ is the coefficient for the jth predictor, and γ and λ are non-negative tuning parameters. For more details about elastic-net regression, see [18, section 3.4] and [18, section 18.4].

2.6. Data setup

Throughout the remainder of the paper, we use MATLAB to perform the necessary calculations used in the classification methods, unless otherwise stated.

2.6.1. Time-domain data

To analyze the time-series data using SOS, we used time and channels as predictors and each video watched by each participant as observations. In this way, each observation belonged to one of two classes, time-direct or time-reversed video. Each of the 24 participants watched 40 time-direct videos and 40 time-reversed videos resulting in $24 \cdot 80 = 1920$ total observations. The time-series data for each video included 5000 time points per video for each of the 26 channels. Using time and all channels as predictors resulted in 130 000 predictors. In this way, the resulting data matrix was $1920 \times 130\,000$.

When finding the optimal solution to the SOS problem, we must calculate a $p \times p$ matrix. In this case, we would require a $130\,000 \times 130\,000$ matrix which is too large for MATLAB to handle on most computers without the use of high performance computing or various parallel computing tools. To reduce the number of predictors, we averaged every

10 time points into 1 measurement. This reduced the number of time points per stimulus video to 500 and the total number of predictors to 13 000. Recall that the original EEG sampling rate was 500 Hz, as noted in section 2.3; thus, this averaging resulted in an equivalent sampling rate of 50 Hz. Note that this lower sampling rate is still higher than the Nyquist frequency for optical flow analysis (15 Hz) and the range of coherence in the frequency domain analysis which focused on 0.4–12.4 Hz. This smoothed time series also served to account for minor drift between each video since each word in each sentence may not occur at exactly the same time point in every video.

To analyze the response of different areas of the brain to each type of condition, we used subsets of the channels as predictors, in addition to using all channels as predictors. The four areas studied were anterior (electrodes F3, F7, FC5, FC1, F4, F8, FC6, FC2, and Fz), posterior (electrodes O1, Oz, O2, P3, P7, P4, P8, Pz), left hemisphere (electrodes F3, F7, FC5, FC1, T7, C3, CP5, CP1, P7, P3, O1), and right hemisphere (electrodes F4, F8, FC6, FC2, T8, C4, CP6, CP2, P8, P4, O2). Let c_r be the number of channels used for region r. To construct the data matrix, we took the 500 time points for a particular video for each channel used and concatenated them into a $1 \times 500c_r$ row vector where components $1, \ldots, 500$ correspond to measurements from the first channel, components $501, \ldots, 1000$ correspond to measurements from the second channel, and so on. This was done for each video and each participant resulting in 1920 row vectors, each consisting of $500c_r$ predictors.

To ensure all EEG measurements for each participant were on the same scale, we first normalized the measurements for each channel across all 80 videos so that each channel had mean 0 and variance 1. We then averaged every 10 time points and arranged the data so that we had 1920 row vectors each with $500c_r$ predictors, as previously mentioned. Finally, the observations used for each training set were extracted and the predictors were normalized so that each had mean 0 and variance 1, in order to meet the assumption for the SOS problem. The remaining observations were used for testing, and the testing data set was transformed to be in the same space as the normalized training data.

2.6.2. Frequency-domain data

To analyze the frequency-domain data using SOS, we used the coherence values from the 62 frequency ranges over each of the four regions (frontal, posterior, left, and right) and concatenated them into a 1×248 row vector where components $1, \ldots, 62$ correspond to the frontal region, components $63, \ldots, 124$ correspond to the posterior region, components $125, \ldots, 186$ correspond to the left hemisphere, and components $187, \ldots, 248$ correspond to the right hemisphere. In this way, we analyzed the frequency-domain data from all regions at once. Recall from section 2.4 that coherence was calculated separately for each condition, time-direct and time-reversed. Therefore, each 1×248 row vector represents coherence values for a particular condition corresponding to an individual participant. Hence, there are two vectors of coherence values for each participant resulting in 48 observations, each consisting of 248 predictors. The observations used for each training set were extracted and the predictors were normalized so that each had mean 0 and variance 1, in order to meet the assumption for the SOS problem. The remaining observations were used for testing, and the testing data set was transformed to be in the same space as the normalized training data.

2.7. Algorithms

2.7.1. Existing algorithms for SOS

Clemmensen *et al* [23] proposed a block coordinate scheme for alternately optimizing the two vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by first fixing $\boldsymbol{\beta}$ and solving for $\boldsymbol{\theta}$ and then fixing $\boldsymbol{\theta}$ and solving for $\boldsymbol{\beta}$; this process is repeated until a maximum number of iterations or stopping criterion is met. For fixed $\boldsymbol{\beta}$, the optimization problem admits a closed form solution for $\boldsymbol{\theta}$ given by [23, equation (11)]. For fixed $\boldsymbol{\theta}$, the optimal $\boldsymbol{\beta}$ is found by solving the generalized elastic-net problem given by (2). This problem can be solved using the least angle regression (LARS-EN) algorithm proposed in [32] or by various proximal methods (ASDA) proposed in [34]. In practice, the accelerated proximal gradient method proposed in [34] seems to perform best for medium and large-scale data (p larger than a few hundred), in terms of both computational efficiency, while LARS-EN performs best for small-scale data.

2.7.2. The CDlr algorithm for SOS

For this analysis, we used the novel CDlr algorithm proposed in [35]. It applies a coordinate descent method for solving (2). This provides several clear improvements upon the LARS-EN and ASDA methods when applied to large-scale data. First, CDlr is more easily parallelizable than LARS-EN and ASDA; each iteration of CDlr updates a single entry of the discriminant vector iterate and all updates can be calculated simultaneously, while the arithmetic operations performed in each iteration of ASDA and LARS-EN must be performed in series. A parallel version of CDlr could then be used for future analysis of the time-domain data in order to analyze all the data at once without the need for data reduction methods such as averaging groups of predictors. Second, the sequences of iterates generated by each method converge to the same optimal $\boldsymbol{\beta}$ since this subproblem has a unique optimal solution for each fixed $\boldsymbol{\theta}$. In practice, the sequence of iterates generated by CDlr typically have fewer nonzero entries than the corresponding iterates generated by ASDA; in this

Algorithm 1. CDlr Algorithm: coordinate descent method for solving (2)

Data: Initial iterate $\boldsymbol{\beta}_0$
Result: Optimal solution $\boldsymbol{\beta}^*$
Precompute: $\boldsymbol{V} = \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\theta}$, $\boldsymbol{W} = \boldsymbol{X}^T\boldsymbol{X}$
while not converged **do**
  Choose index $j \in \{1, 2, \ldots, p\}$
    $Z = 2V_j - 2(\boldsymbol{W}_j^T\boldsymbol{\beta} - \beta_j)$
    $\beta_j = \frac{\text{sign}(Z)\max\{|Z| - \lambda, 0\}}{2(\gamma+1)}$
end while
where V_j and β_j are the jth components of $\boldsymbol{V}$ and $\boldsymbol{\beta}$ and $\boldsymbol{W}_j^T$ is the jth row of $\boldsymbol{W}$

sense, CDlr converges more quickly to a sparse solution than ASDA and LARS-EN. It is important to note that we used CDlr over ASDA or LARS-EN due to the large size of the time-series data in preparation for applying a parallel version of CDlr in future analyses. When performing the frequency-domain analysis, we used CDlr for consistency, although the size of the frequency-domain data makes it better suited for use with LARS-EN.

2.7.3. Algorithms used for SVM and elastic-net regression

We used R version 3.6.2 [36], the e1071 [37], and the glmnet [38] packages to generate SVMs and elastic-net models. The function `svm()` in the e1071 library was used to generate the SVM models with polynomial and radial kernels, and the `tune()` function was used to perform ten-fold cross-validation to select the best degree and γ for the polynomial and radial kernels, respectively. We used the `glmnet()` function in the glmnet package to generate the elastic-net regression model and the `cv.glmnet()` function to perform ten-fold cross-validation in order to select the best tuning parameters.

3. Results

3.1. Time-domain

We wish to predict 2 conditions, time-direct sign language and time-reversed sign language. In this case, $K = 2$ and SOS will generate one nontrivial optimal $(\boldsymbol{\theta}, \boldsymbol{\beta})$ pair. Each observation in the test set was projected onto the line spanned by the nontrivial discriminant vector. This yielded a value indicating which class the test observation should be assigned. If the value is less than the midpoint between the projected values of the two class means then we assign the observation to the set of time-direct stimuli; otherwise, we assign the stimulus to the set of time-reversed stimuli. In the special case that the two classes contain the same number of training samples, this midpoint is zero and we assign an observation to the set of time-direct stimuli if its projection onto span $\boldsymbol{\beta}$ is negative and to the set of time-reversed stimuli otherwise. We then compared the calculated labels with the actual condition of the stimuli to determine the accuracy of our predictions.

Table 1. Results from analyzing time-domain data from four regions of the brain as well as from all regions using SOS.

Region	Accuracy	Region	Accuracy
Posterior ($p = 4000$)	69.01%	Left ($p = 5500$)	67.60%
Anterior ($p = 4500$)	63.18%	Right ($p = 5500$)	67.60%
All regions ($p = 13\,000$)	70.05%		

To evaluate the performance of the model, we trained the model on observations from 23 of the 24 participants. That is, 1840 training observations were used. The remaining 80 observations corresponding to the remaining participant were used for testing. This was repeated for all 24 participants, and the results were averaged across all 24 trials. In this way, we performed a form of leave-one-out cross validation with the testing fold consisting of one participant. For each trial, we set $\boldsymbol{\Omega}$ to be the $p \times p$ identity matrix, $\lambda = 0.06$, and $\gamma = 0.24$. Recall that the number of features p is equal to 500 times the number of channels associated with the used regions; e.g. for posterior channels we have $p = 4000$, for anterior $p = 4500$, and $p = 13\,000$ if we use all channels. We terminated the algorithm once a 10^{-3} suboptimal solution was found. The results are summarized in table 1.

To compare the results obtained from applying SOS to the time-domain data with a classifier which produces a non-linear decision boundary, we fit a model using SVM with radial and polynomials kernels. The performance of each was evaluated using the same form of leave-one-out cross validation with each testing fold consisting of one participant as was used with SOS. The cost, γ, and degree parameters were determined using cross validation. The results are summarized in table 2. On average, SVM with radial kernel achieved 50% classification accuracy and SVM with polynomial kernel achieved 65.3% classification accuracy. It is important to note that SVM with radial kernel classified all testing observations into one class. Therefore, the 50% classification accuracy is only a result of having two balanced classes in the testing set.

3.2. Behavioral analysis

Analysis of the behavioral data indicated that only sentences in the direct video condition were rated as linguistically acceptable (signing videos acceptability $M = 5.8$, $SD = 1.05$; reversed videos $M = 1.72$, $SD = 0.76$). To evaluate the significance of the behavioral data to classification performance, we fit a model using elastic-net regression to predict the behavioral response rating based on the time-domain EEG data. We then classified each video as time-direct if the predicted rating was greater than 4

Table 2. Results from analyzing time-domain data using SVM with radial and polynomial kernels and elastic-net regression.

	SVM		Regression
	radial	polynomial	elastic-net
Average Accuracy	50%	65.3%	69.67%

Table 3. Results from performing 10-fold cross validation on frequency-domain data using SOS.

Trial	Accuracy	Trial	Accuracy
1	100%	6	94.44%
2	97.22%	7	100%
3	100%	8	100%
4	100%	9	100%
5	100%	10	97.22%
Average	**98.89%**		

(i.e. acceptable sign language) or time-reverse otherwise. The performance was evaluated using the same form of leave-one-out cross validation with each testing fold consisting of one participant, as was discussed in section 3.1. For each trial, we set $\alpha = 0.2$ (the balancing parameter between lasso and ridge penalty) and used cross validation to determine the best tuning parameter λ. The results are summarized in table 2. On average, this elastic-net model resulted in 69.67% classification accuracy.

3.3. Frequency-domain

The performance of the model was evaluated using 10 randomized training/testing splits. In this way, 25% of the 48 observations were randomly selected as the training set and the remaining 75% were selected as the test set. For each trial we set $\boldsymbol{\Omega}$ to be the 248×248 identity matrix, $\lambda = 0.06$, and $\gamma = 0.24$. We terminated the algorithm once a 10^{-3} suboptimal solution was found. The results are summarized in table 3. On average, the algorithm achieved 98.89% out-of-sample prediction accuracy, which is less than 1 observation misclassified per trial. Trial 6 resulted in the lowest accuracy with 2 misclassified observations.

To illustrate the benefit of CDlr over ASDA, we ran the same trials using the same $\boldsymbol{\Omega}$, λ, γ, and tolerance parameters as above on the frequency domain data with ADMM and APG and compared the results. For ADMM we set the augmented Lagrangian parameter $\mu = 5$. On average, CDlr and ADMM achieved the same level of prediction accuracy, while APG achieved slightly poorer accuracy (although still fewer than 2 misclassified observations per trial on average). The average number of nonzero features in the CDlr discriminant vectors was 9.93% less than for ADMM and 78.61% less than for APG. The results are summarized in table 4. It is important to note that the initial solutions used for APG were already 10^{-3} suboptimal when given the same parameters as CDlr; in this case, no subiterations performed. In order for APG to perform any further updates to the discriminant vector, a more strict tolerance or different tuning parameters are needed.

Table 4. Comparison of APG, ADMM, and CDlr algorithms on frequency-domain data.

Algorithm	APG	ADMM	CDlr
Average accuracy	96.67%	98.89%	98.89%
Average number of nonzero features	229.1	54.4	49

To visualize how SOS classifies the data, the test observations are projected onto the optimal discriminant vector with the class boundary indicating how SOS performs nearest centroid classification. In this lower dimensional space, the classes should be well separated. As shown in figure 2, when projected onto the optimal discriminant vector, $\boldsymbol{\beta}$, the 2 classes are well separated which results in very high prediction accuracy.

The high prediction accuracy of SOS in classifying the time-direct vs time-reversed conditions suggests this model can accurately identify common neural responses to visual linguistic stimuli. To have a better understanding of the model over all trials, we take the average of each $\boldsymbol{\beta}$ value corresponding to the various frequencies over all 10 discriminant vectors. To ensure that no particular trial has more or less impact on the averages, we normalized each discriminant vector before computing any averages. Frequencies that are less informative are represented as zero values in the original discriminant vector. Therefore, a smaller value (in terms of absolute value) in the average vector indicates less significance of that frequency over all the trials. We can then plot the average vector over the range of frequencies, for each region (see appendix figure A1). We can also view the impact of each frequency on the nearest centroid classification by considering the Hadamard product of the average coherence value from each condition and each region with the average discriminant vector for each region. In terms of magnitude, a larger positive value would contribute more toward classifying an observation to the time-reversed class while a larger negative value would contribute more toward classifying an observation to the time-direct class. The average coherence values for each condition and the Hadamard product of coherence with the average discriminant vector from each region can be found in appendix figures A2 and A3. In order to view the contribution of each frequency to the classification, we can plot the difference in this product between the two conditions as shown in figure 3.

4. Discussion

The rich dynamics of visual environment in sign language delivers sensory input that has entropy-rich temporal structure [14, 39]. In another linguistic

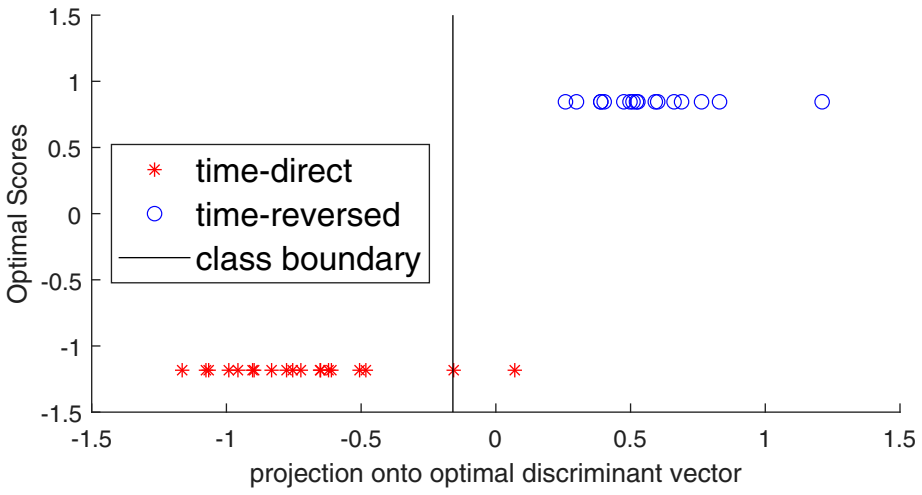

Figure 2. Plot of test observations in trial 6 projected onto the SOS optimal discriminant vector.

modality—auditory perceived speech—brain activity dynamically tracks speech streams using both low-frequency phase and high-frequency amplitude fluctuations [40]. The difference is between sensory and cognitive processing: near auditory cortices (low-level sensory processing), attention enhances cortical tracking of attended streams; while in higher-order regions, the representation appears to become more 'selective'. In our data, the equivalent would be attended direct sign language, and unattended (uninterpretable) reverse videos.

Prior work has indicated that electrophysiological oscillations in human cerebral cortex become entrained to quasi-periodic fluctuations of visual movement in sign language [41], with cortical entrainment peaking at 1 Hz, most robustly over occipital and central EEG channels. While the results suggested that signers' brain entrained to low-frequency variability in language, lack of a control condition (i.e a dynamic visual stimuli with spectrotemporal properties of signing, such as reversed signing) did not allow for interpretation that this entrainment was specific to higher cognition (i.e. language comprehension), as opposed to sensory entrainment to the stimuli with spectrotemporal parameters characteristic of the environment [15]. Investigations of brain rhythms underlying visuospatial selective attention [42] indicated that attention filters behaviorally relevant stimuli from the stream of sensory information, while acting as a signal gain, i.e. biasing processing toward attended stimuli. Senoussi *et al* [43] further demonstrated that visuospatial attention reorients periodically at 4 Hz (theta range) between stimulus locations (attentional exploration), while sampling each location periodically at 11 Hz (alpha) (the latter being an ongoing sensory sampling rhythm). Mai *et al* [44] observed significant effects of EEG power and EEG-acoustic entrainment at δ and θ bands during sensory-level phonological processing in speech [44]. Thus, the low-frequency response seen in present data and other frequency-domain EEG investigations of sign language [41] may reflect an equivalent of attentional orientation to spatiotemporally-familiar stimuli in vision. The finding of EEG locking to stimulation in the delta band (1–4 Hz) also suggests a special role for narrow-band low frequency periodic brain responses. Keitel *et al* [42] previously demonstrated that neural activity continuously reflects stimulus temporal structure, by assessing the neural response of visual cortex to quasi-rhythmic stimulation, with frequencies continuously varying within ranges of classical theta (4–7 Hz), alpha (8–13 Hz) and beta bands (14–20 Hz); the results indicated phase-locking to stimulation in all three frequency ranges. This leads to a conjecture that EEG-stimulus locking is a continuous neural signature of processing dynamic sensory input in early visual cortices, which serves to trace the temporal evolution of visual input (whether rhythmic or quasi–rhythmic) and is subject to attentional bias.

The present work focused on a more fundamental question than that of sensory entrainment to stimuli: we asked whether neural signatures of higher cognition could be identified based on cortical entrainment to visual stimuli of equivalent spectrotemporal parameters, with and without comprehensible sign language. We found that neural entrainment to dynamic changes in visual stimuli occurs in both comprehension and non-comprehension conditions. EEG brain responses continuously reflected quasi-rhythmic dynamics in visual stimulation across different time scales for both linguistic (interpretable/comprehensible) and reversed-video (incomprehensible) stimuli. However, entrainment

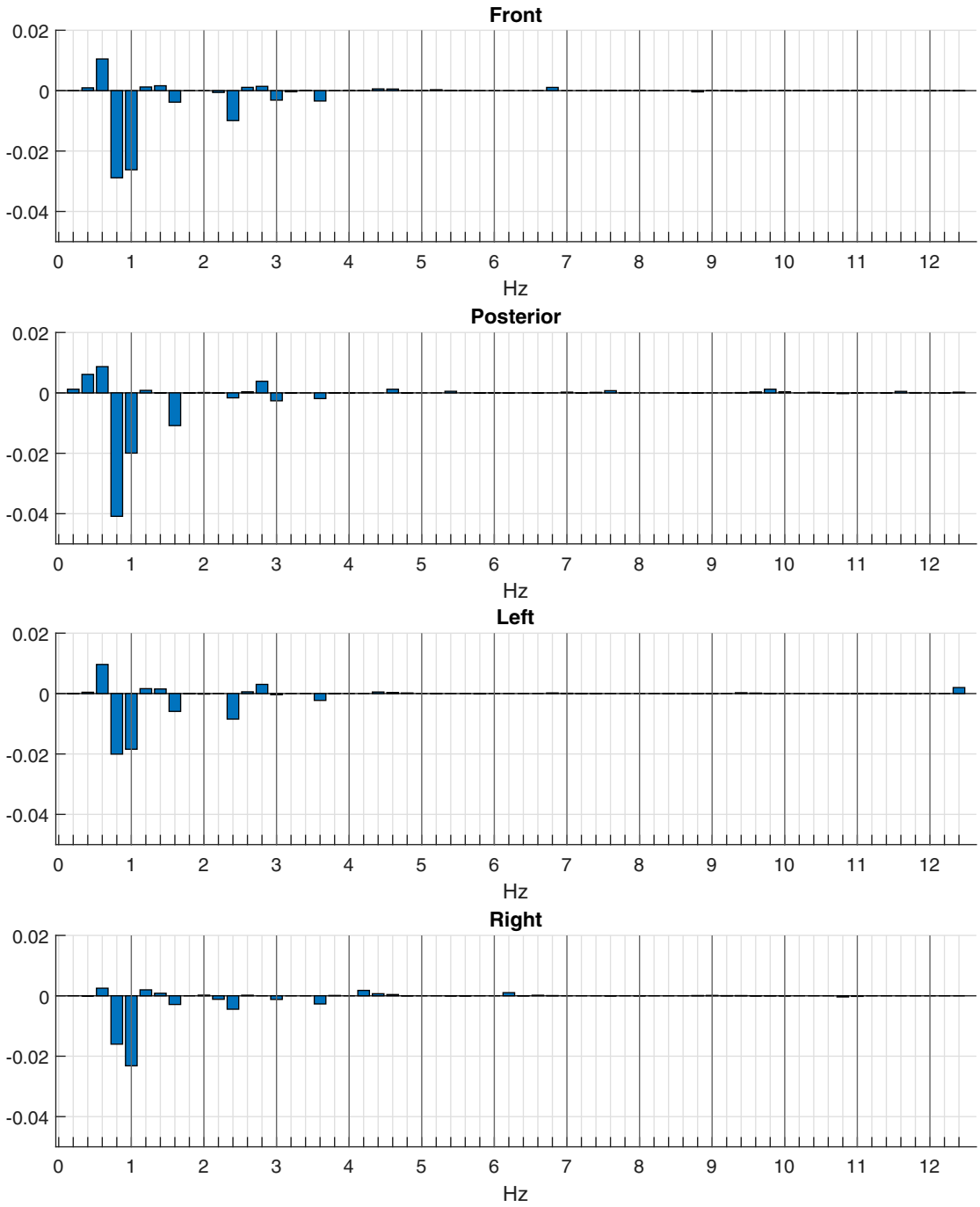

Figure 3. Difference in the value of the Hadamard product (calculated as the element-wise product of coherence with the average SOS discriminant vector) from each condition plotted at each frequency.

frequencies and their topographic distribution on the scull differ substantially between comprehension and non-comprehension conditions, which allowed for detection of sign language comprehension with high fidelity in few participants. The measures of brain-stimulus coupling increased in 0.8–1 Hz frequency ranges when the stimulus was sign language—i.e. was conveying information, e.g. was behaviourally relevant.

In sign language, tracing the dynamics of visual stimuli on different time scales may subserve integration of signal on different temporal scales, such as handshape changes at the semantic level, non-manual markers at the syntactic level (e.g. brow furrows that scope over interrogative sentences), and pragmatic parameters, such as head and body leans, which are interpreted at discourse level.

5. Conclusion

In this work, we have identified that although coherence to visual stimuli with spectrotemporal parameters of sign language is driven bottom-up, sensory stimulation in signers, the differences in coherence at 0.8–1 Hz across neural regions appear to reflect top-down processing based on linguistic knowledge in signers. Together with prior work, our results suggest that, although sensory cortical entrainment to visual stimuli occurs in all conditions, higher cognitive processes in signers elicit a distinct temporal signature that allows for near-perfect classification. Our findings suggest that while the brain entrains to sensory visual information in sign language signal, it is possible to detect the contribution of higher cognition in EEG signal robustly and reliably. Analysis of coherence between optical flow dynamics in the visual signal and EEG data resulted in an average out-of-sample classification accuracy of 98.89%, which was far superior to the time-domain analysis of EEG data alone. This high classification accuracy suggests that the models based on neural response to stimuli, rather than neural data alone, can more accurately identify the underlying features of relevant brain states, such as instances of successful higher level cognition, such as language comprehension. The work thus demonstrates the importance of using the relationship between the external signal and frequency-domain neural response in identifying the parameters of neural response necessary and sufficient for robust classification of higher-level cognition states. It also suggests that SOS is robust to a high level of variability across participants, and effective for classification with appropriate data reduction.

Acknowledgments

We want to thank all Deaf informants taking part in the present study, Waltraud Unterasinger for signing the stimulus material, and Ronnie Wilbur and Dietmar Roehm for valuable discussions. Preparation of this manuscript was partially funded by the Grants #1734938 and #1932547 from the U.S. National Science Foundation to E M, as well as University of Alabama Cyberseed Grant SP14572, University of Alabama Research Grant RG14838, and Grant #2012554 from the U.S. National Science Foundation to B A.

Appendix

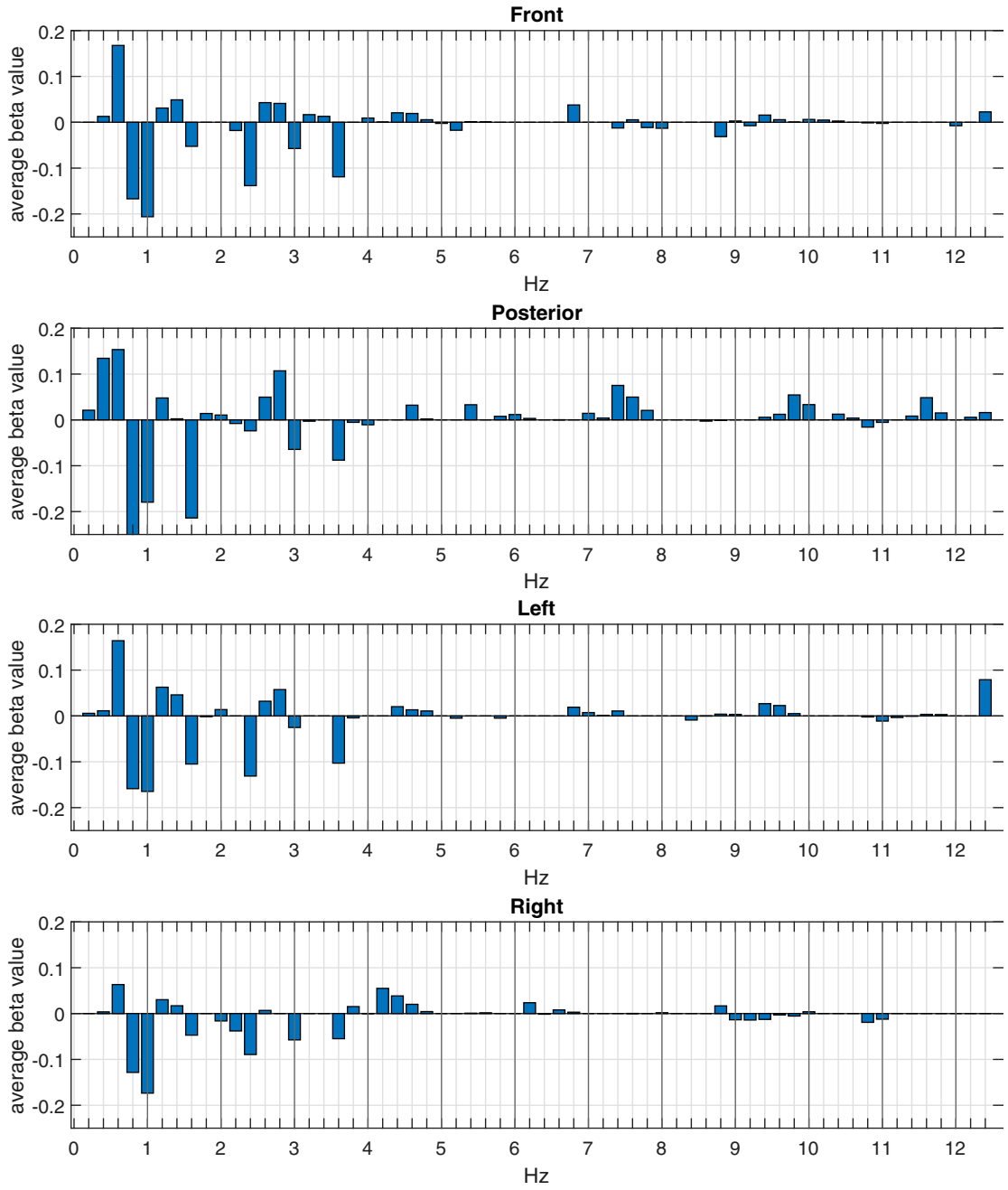

Figure A1. Average SOS discriminant vector values at each frequency for each region.

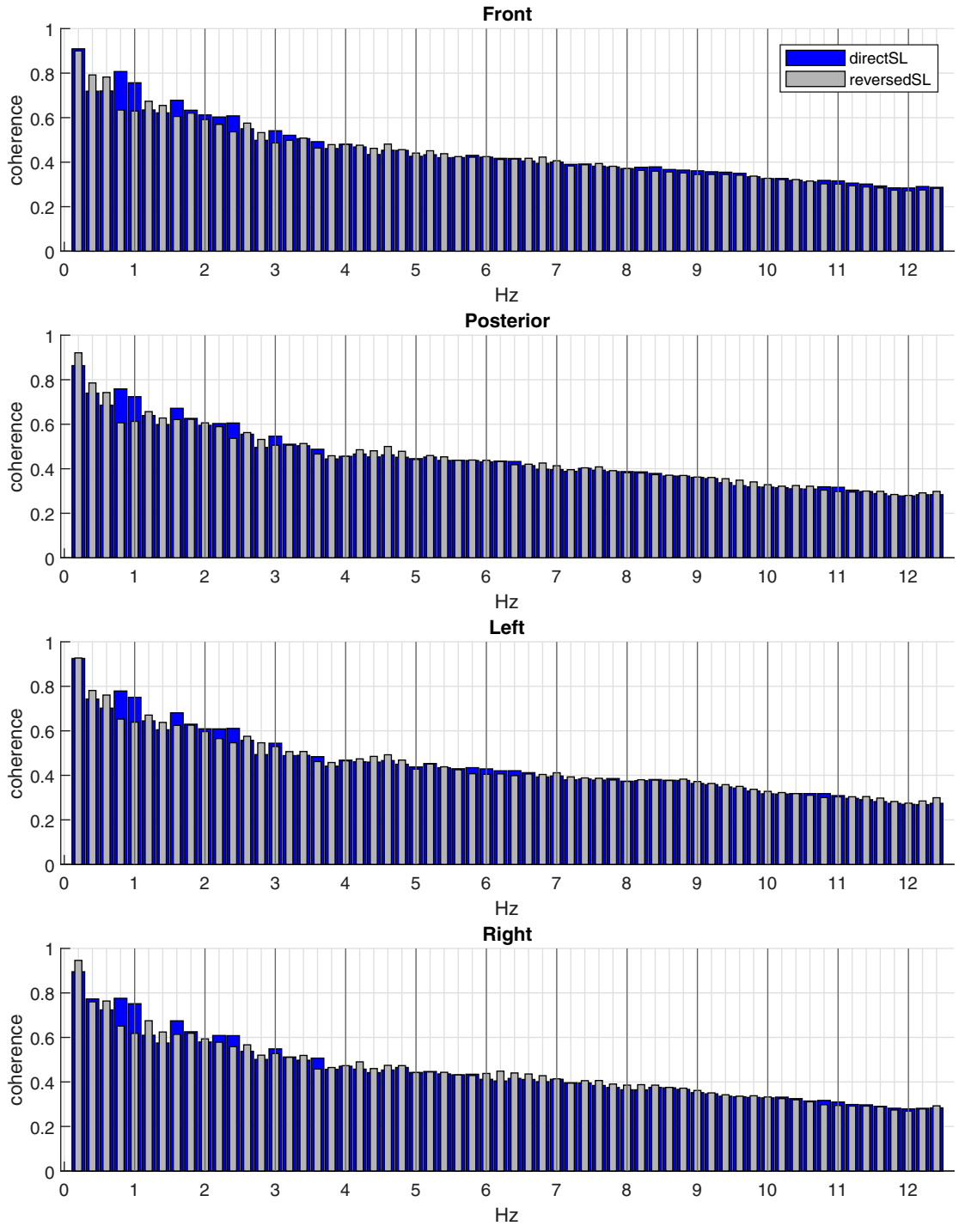

Figure A2. Average coherence values for each condition plotted at each frequency for each region.

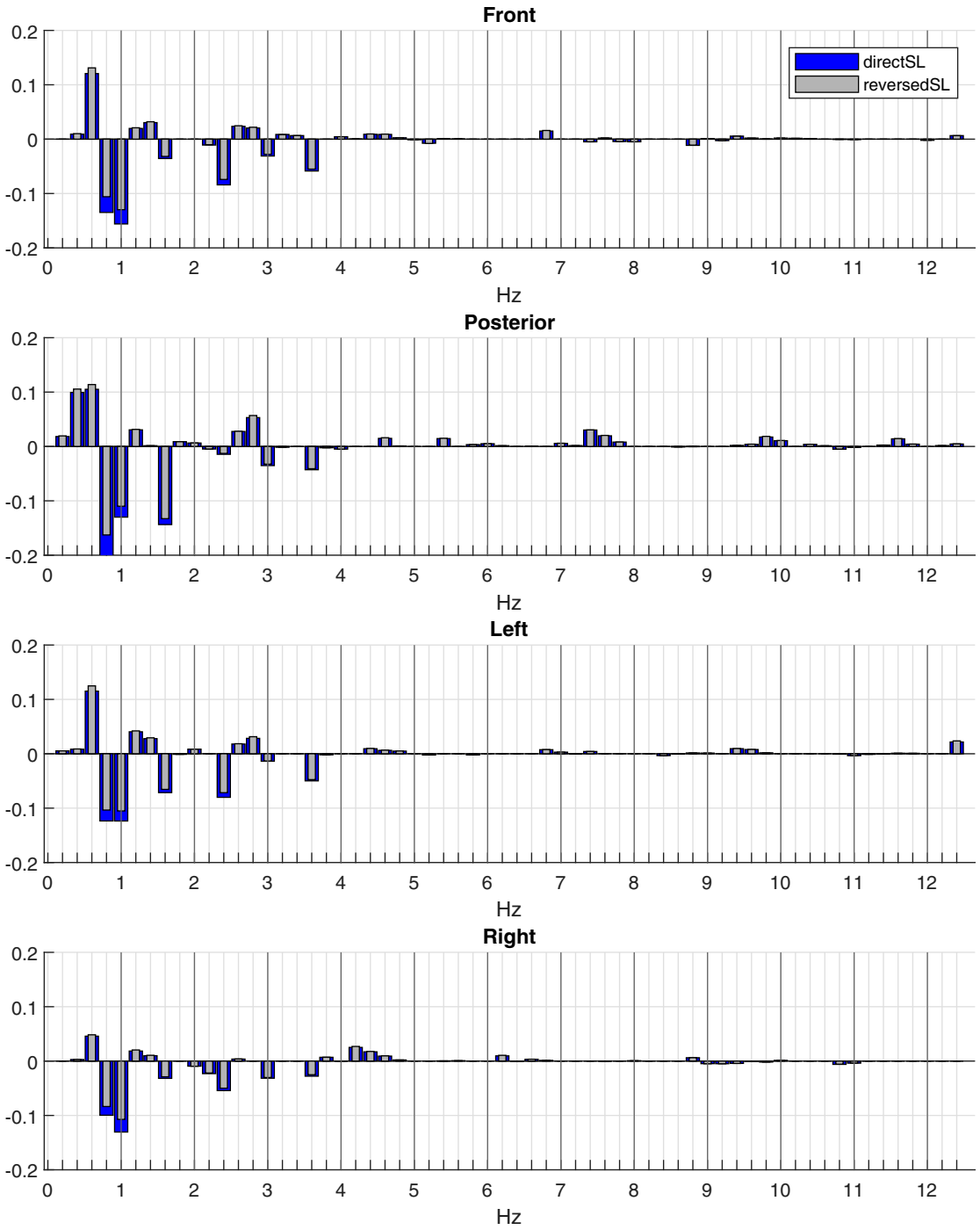

Figure A3. Hadamard product of coherence with the average SOS discriminant vector from each region plotted at each frequency.

ORCID iD

Linda K Ford https://orcid.org/0000-0002-9230-2043

References

[1] Craik A, He Y and Contreras-Vidal J L 2019 Deep learning for electroencephalogram (EEG) classification tasks: a review *J. Neural Eng.* **16** 031001

[2] Rubchinsky L L, Park C and Worth R M 2012 Intermittent neural synchronization in Parkinson's disease *Nonlinear Dyn.* **68** 329–46

[3] Malaia E, Bates E, Seitzman B and Coppess K 2016 Altered brain network dynamics in youths with autism spectrum disorder *Exp. Brain Res.* **234** 3425–31

[4] Isabel Vanegas M, Felice Ghilardi M, Kelly S P and Blangero A 2018 Machine learning for EEG-based biomarkers in Parkinson's disease *2018 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)* (IEEE) pp 2661–5

[5] Cruse D, Chennu S, Chatelle C, Bekinschtein T A, Fernández-Espejo D, Pickard J D, Laureys S and Owen A M 2011 Bedside detection of awareness in the vegetative state: a cohort study *Lancet* **378** 2088–94

[6] Coleman M R, Rodd J M, Davis M H, Johnsrude I S, Menon D K, Pickard J D and Owen A M 2007 Do vegetative patients retain aspects of language comprehension? Evidence from fMRI *Brain* **130** 2494–507

[7] Liu Y, Liu Y, Tang J, Yin E, Hu D and Zhou Z 2020 A self-paced BCI prototype system based on the incorporation of an intelligent environment-understanding approach for rehabilitation hospital environmental control *Comput. Biol. Med.* **118** 103618

[8] Jiang J, Wang C, Wu J, Qin W, Xu M and Yin E 2020 Temporal combination pattern optimization based on feature selection method for motor imagery BCIs *Front. Hum. Neurosci.* **14**

[9] Roehm D, Bornkessel-Schlesewsky I and Schlesewsky M 2007 The internal structure of the N400: frequency characteristics of a language related ERP component PhD Thesis

[10] Peelle J E, Gross J and Davis M H 2012 Phase-locked responses to speech in human auditory cortex are enhanced during comprehension *Cereb. Cortex* **23** 1378–87

[11] Stilp C E and Kluender K R 2016 Stimulus statistics change sounds from near-indiscriminable to hyperdiscriminable *PLoS One* **11** e0161001

[12] Ding N, Melloni L, Zhang H, Tian X and Poeppel D 2016 Cortical tracking of hierarchical linguistic structures in connected speech *Nat. Neurosci.* **19** 158–64 (PMID: 26642090 PMCID: PMC4809195)

[13] Richmond L L and Zacks J M 2017 Constructing experience: event models from perception to action *Trends Cogn. Sci.* (https://doi.org/10.1016/j.tics.2017.08.005)

[14] Malaia E, Borneman J D and Wilbur R B 2016 Assessment of information content in visual signal: analysis of optical flow fractal complexity *Vis. Cogn.* **24** 246–51

[15] Bosworth R G, Wright C E and Dobkins K R 2019 Analysis of the visual spatiotemporal properties of American sign language *Vis. Res.* **164** 34–43

[16] Borneman J D, Malaia E A and Wilbur R B 2018 Motion characterization using optical flow and fractal complexity *J. Electron. Imaging* **27** 1

[17] Gurbuz S Z *et al* 2020 A linguistic perspective on radar micro-doppler analysis of American sign language *2020 IEEE Int. Conf. (RADAR)* (IEEE) pp 232–7

[18] Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (*Springer Series in Statistics*) 2nd edn (New York: Springer)

[19] Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes M P, Shyu M-L, Chen S-C and Iyengar S S 2018 A survey on deep learning: algorithms, techniques and applications *ACM Comput. Surv.* **51** 1–36

[20] Alom Md Z *et al* 2019 A state-of-the-art survey on deep learning theory and architectures *Electronics* **8** 292

[21] Hastie T, Tibshirani R and Wainwright M 2015 *Statistical Learning With Sparsity: The Lasso and Generalizations* (Boca Raton, FL: CRC Press)

[22] Witten D M and Tibshirani R 2011 Penalized classification using fisher's linear discriminant *J. R. Stat. Soc.* B **73** 753–72

[23] Clemmensen L, Hastie T, Witten D and Ersbøll B 2011 Sparse discriminant analysis *Technometrics* **53** 406–13

[24] Leng C 2008 Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data *Comput. Biol. Chem.* **32** 417–25

[25] Cai T and Liu W 2011 A direct estimation approach to sparse linear discriminant analysis *J. Am. Stat. Assoc.* **106** 1566–77

[26] Ames B P W and Hong M 2016 Alternating direction method of multipliers for penalized zero-variance discriminant analysis *Comput. Optim. Appl.* **64** 725–54

[27] Shao J *et al* 2011 Sparse linear discriminant analysis by thresholding for high dimensional data *Ann. Stat.* **39** 1241–65

[28] Tanner D, Morgan-Short K and Luck S J 2015 How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition *Psychophysiology* **52** 997–1009

[29] Gratton G, Coles M G H and Donchin E 1983 A new method for off-line removal of ocular artifact *Electroencephalogr. Clin. Neurophysiol.* **55** 468–84

[30] Horn B K P and Schunck B G 8 1981 Determining optical flow *Artif. Intell.* **17** 185–203

[31] de Cheveigné A, Wong D D E, Di Liberto G M, Hjortkjær J, Slaney M and Lalor E 2018 Decoding the auditory brain with canonical component analysis *Neuroimage* **172** 206–16

[32] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301–20

[33] James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning* 1st edn (New York: Springer)

[34] Atkins S, Einarsson G, Ames B and Clemmensen L Proximal methods for sparse optimal scoring and discriminant analysis (arXiv:1705.07194v3)

[35] Ford L K and Ames B P Coordinate descent methods for sparse optimal scoring in preparation

[36] R Core Team 2019 *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing)

[37] Meyer D, Dimitriadou E, Hornik K, Weingessel A and Leisch F 2019 e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien *R Package Version 1.7-3*

[38] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent *J. Stat. Softw.* **33** 1–22

[39] Bosworth R G, Bartlett M S and Dobkins K R 2006 Image statistics of American sign language: comparison with faces and natural scenes *J. Opt. Soc. Am.* A **23** 2085–96

[40] Lakatos P, Karmos G, Mehta A D, Ulbert I and Schroeder C E 2008 Entrainment of neuronal oscillations as a mechanism of attentional selection *Science* **320** 110–3

[41] Brookshire G, Lu J, Nusbaum H C, Goldin-Meadow S and Casasanto D 2017 Visual cortex entrains to sign language *Proc. Natl Acad. Sci.* **114** 6352–7 (PMID: 28559320)

[42] Keitel C, Andersen S K, Quigley C and Müller M M 2013 Independent effects of attentional gain control and competitive interactions on visual stimulus processing *Cereb. Cortex* **23** 940–6

[43] Senoussi M, Moreland J C, Busch N A and Dugué L 2019 Attention explores space periodically at the theta frequency *J. Vis.* **19** 22

[44] Mai G, Minett J W and Wang W S Y 2016 Delta, theta, beta and gamma brain oscillations index levels of auditory sentence processing *Neuroimage* **133** 516–28