Auditory Stimulus-response Modeling with a Match-Mismatch Task

Alain de Cheveigné, Malcolm Slaney, Søren Fuglsang, Jens Hjortkjaer

Auditory Stimulus-response Modeling with a Match-Mismatch Task

Alain de Cheveigné(1, 2, 3), Malcolm Slaney (4), Søren A. Fuglsang (6), Jens Hjortkjaer (5, 6)

AUTHOR AFFILIATIONS:

(1) Laboratoire des Systèmes Perceptifs, UMR 8248, CNRS, France.

(2) Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL, France.

(3) UCL Ear Institute, United Kingdom.

(4) Google Research, Machine Hearing Group. USA.

(5) Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Denmark.

(6) Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark.

CORRESPONDING AUTHOR:

Alain de Cheveigné, Audition, DEC, ENS, 29 rue d'Ulm, 75230, Paris, France, Alain.de.Cheveigne@ens.fr, phone 0033144322672, 00447912504027.

Abstract

The relation between a continuous ongoing stimulus and the brain response that it evokes can be characterized by a stimulus-response model fit to the data. This systems-identification approach offers insight into perceptual processes within the brain, and it is also of potential practical use for devices such as Brain Computer Interfaces (BCI). The quality of the model can be quantified by measuring the fit with a regression problem, or by applying it to a classification task and measuring its performance. Here we focus on a *match-mismatch* task that entails deciding whether a segment of brain signal matches, via a model, the auditory stimulus that evoked it. The match-mismatch task can be used to compare performance of different stimulus-response models. We show that performance in a match-mismatch task and metrics summarizing regression accuracies can provide complementary insights in the relation between stimulus and response. Importantly, the match-mismatch task provides information about discriminatory power, making it directly applicable to BCI applications. Evaluation is performed on a freely available database, and code is available for scripts and functions to allow scrutiny of our results and facilitate comparative evaluation of future developments.

Introduction

Continuous stimuli such as speech or music elicit an ongoing brain response (Ahissar et al., 2001; Aiken and Picton, 2008; Power et al., 2011; Ding and Simon, 2012; Kubanek et al., 2013) that can be detected with electroencephalography (EEG) or magnetoencephalography (MEG). The relation between stimulus and response can be characterized by fitting a model to the data (Lalor et al., 2009; Crosse et al., 2016). Most work has used a *linear* stimulus-response model to relate some feature transform of the stimulus (envelope, spectrogram, etc.) to the brain response. Such models come in three main flavors: a *forward model* that attempts to predict the neural response from the stimulus (Lalor et al., 2009; Ding and Simon, 2012; Crosse et al., 2016), a *backward model* that attempts to infer the stimulus from the response (Mesgarani and Chang, 2012; O'Sullivan et al., 2015; Puvvada and Simon, 2017; Hausfeld et al., 2018; O'Sullivan et al., 2019; Akbari et al., 2019), or a *hybrid forward-backward model* that transforms both stimulus and response to better reveal their relation (Dmochowski et al., 2017; de Cheveigné et al., 2018; Zhuang et al., 2020). The fit of the model is usually quantified by calculating the correlation coefficient between stimulus and response, or their transforms: the observation of a significant correlation suggests that the model captures some aspect of neural processing. Details of the model (e.g. latency or shape of a temporal response function) can then provide insights into the sensory processing mechanisms at work within the brain.

In this paper, we consider a simple classification task (*match-mismatch*, MM), that applies to listening scenarios with only one sound source. This task consists of deciding whether a segment of EEG or MEG is temporally aligned with a segment of audio (i.e. that segment of response was evoked by that segment of stimulus), or not. This can be framed as a classification task, and performance can be quantified by the *sensitivity index*, defined here as the standardized mean of the distribution of the decision metric, or the *error rate*. Together, correlation, sensitivity index, and error rate form a trio of complementary performance metrics.

Auditory attention decoding (AAD), a different task, has played an important role in past studies (Kerlin et al., 2010; Power et al., 2011; Ding and Simon, 2012; Mesgarani and Chang, 2012). A subject is instructed to attend to one of two concurrent streams, usually speech, and the algorithm decides which stream was attended based on the brain activity. Performance is measured in terms of how reliably the algorithm identifies the appropriate speech stream, and can be used to judge the quality of the model, as with the MM task. However, unlike the AAD task, the MM task can be evaluated in listening scenarios where there is only one speaker, and does not depend on whether the listener followed instructions as to which stream to attend.

The AAD task taps a richer phenomenology than MM and is thus more elaborate: data collection requires a two-voice stimulus, specific instructions to subjects, and a well-controlled experimental setup. However, AAD models rely on data labels defined by the experimental task (which voice the subject is attending). Moreover, we cannot rule out that the listener's attentional state differs momentarily from instructions (e.g. attentional capture by the "unattended" stream), and so some proportion of the data may be *mislabeled*. This can be a problem if we wish to evaluate algorithms in the limit of small error rates (which is where we want to be). The simpler MM task, in contrast, is applicable to the evaluation of high performance algorithms with vanishing error rates. Also, avoiding data labels allows models to be trained for this task in a self-supervised way. In this paper, we use the MM task to compare stimulus-response models that relate speech to EEG responses. This allows us to compare models even in the limit of vanishing error rates. We speculate that the MM task, like AAD, might find use in a BCI application, for example to monitor the attentional (or inattentional) state of a user.

Accurate model performance is critical in this case.

Building on prior work, we introduce a set of refinements of stimulus-response models that lead to significant improvements. These refinements allow more detailed models and limit the curse of overfitting. As we will show, error rates averaged over subjects for 5s segments fall from ~30% for the simplest model to ~3% for the best (0% error for a subset of subjects) indicating highly reliable stimulus-response models. Our focus is on understanding which processing steps improve performance, and why.

Recently, intense activity has been devoted to stimulus-response models to gain insight into perceptual processes for speech or music (Di Liberto et al., 2015; Goossens et al., 2018; O'Sullivan et al., 2019; Broderick et al., 2019; Decruy et al., 2020; Bednar and Lalor, 2020; Zuk et al., 2020), and for BCI applications (Jaeger et al., 2020; Jalilpour Monesi et al., 2020). However, progress is slowed by the lack of reliable comparative evaluation due to the diversity of experimental conditions and data, the absence of state-of-the-art algorithms in the "line-up", and the aforementioned issue of segment mislabeling that hobbles evaluation based on the commonly-used AAD task. We use a publicly available database, metrics based on the simpler MM task, and we propose a well-defined benchmark implementation to facilitate evaluation of future advances.

This study offers two main contributions. First, it introduces a simple objective task, match-mismatch, to help in the evaluation of stimulus-response models. Second, it documents a set of techniques that boost performance beyond state of the art.

1 Methods

This section describes the stimulus-response model and provides details of the evaluation methods and experiments. The busy reader is encouraged to read the next Subsection, then skip to Results and come back for more details as needed. We assume that brain responses are recorded by EEG, but the same methods are applicable to MEG or other recording modalities.

1.1 Models and metrics

In this subsection we define the mathematical tools to describe what we wish to accomplish, and the metrics to judge success.

Data Model. The brain response data consist of a time series matrix $\mathbf{X}$ of dimensions T (time) $\times$ J (channels). Each channel of the response is assumed to be a weighted sum of sources, including brain sources of interest that reflect processing of sound as well as undesired noise and artifacts:

$$x_j(t) = \sum_i s_i(t) m_{ij}, \tag{1}$$

where t is time, $[s_i(t)], i = 1 \dots I$ are sources, and the m_{ij} are unknown source-to-sensor mixing weights. In matrix notation **X=SM**. This model matches the physical source-to-sensor mixing process which is, to a good approximation, linear and instantaneous. The stimulus is represented as a matrix or column vector $\mathbf{A}$, usually a transform of the acoustic stimulus designed to mimick known aspects of processing within the auditory system. Typical transforms are the waveform envelope (akin to a measure of "instantaneous loudness") or the spectrogram (akin to an "auditory nerve activity pattern"). $\mathbf{A}$ is of size $T \times K$, where K is the number of channels of the stimulus representation (e.g. number of frequency bands of a spectrogram). In the following, $K = 1$.

Stimulus-Response Model. We assume that some transform f of the stimulus representation is non-trivially related to some transform g of the EEG:

$$f(\mathbf{A}) \approx g(\mathbf{X}) \tag{2}$$

where $\approx$ indicates similarity according to some metric. By non-trivial we mean that Eq. 2 can be used empirically to decide whether or not some segment $\mathbf{X}_s$ of the brain data was recorded in response to a segment $\mathbf{A_s}$ of the stimulus.

Equation 2 is quite general, but we focus on three special cases for which the transforms f and g are linear. In the linear *forward model*, $\mathbf{AF} \approx \mathbf{X}$, a transform matrix $\mathbf{F}$ (possibly convolutive) is used to predict the response from the stimulus. In the *backward model*, $\mathbf{A} \approx \mathbf{XG}$, a transform matrix $\mathbf{G}$ (possibly convolutive) is used to infer the stimulus from the response. Forward and backward models are also referred to as "encoding" and "decoding" (Naselaris et al., 2011), or "temporal response function" (TRF) and "stimulus reconstruction" models, respectively. A third "hybrid" model involves linear transforms of both: $\mathbf{AF} \approx \mathbf{XG}$. Tradeoffs between these three approaches are reviewed in the Discussion.

The transform matrices $\mathbf{F}$ and/or $\mathbf{G}$ are found by a data-driven algorithm, regression for the first and second models, or canonical correlation analysis (CCA) for the third. Given datasets $\mathbf{A}$ and $\mathbf{X}$, CCA finds transform matrices $\mathbf{F}$ and $\mathbf{G}$ such that (a) columns of $\mathbf{Y}_A = \mathbf{AF}$ are orthonormal (variance 1 and mutually uncorrelated), (b) columns of $\mathbf{Y}_X = \mathbf{XG}$ are orthonormal, (c) the first pair of columns y_{A1} and y_{X1} have the greatest possible correlation, the second pair of columns has the greatest possible correlation once the first pair has been projected out, and so-on. CCA transform matrices $\mathbf{F}$ and $\mathbf{G}$ are of size $J \times H$ and $K \times H$ respectively, where H is at most equal to the smaller of J and K.

The Match-mismatch Task. To assist evaluation, we define a task as follows. Given a segment of stimulus signal $\mathbf{A_s}$, the segment of EEG signal $\mathbf{X_s}$ that it

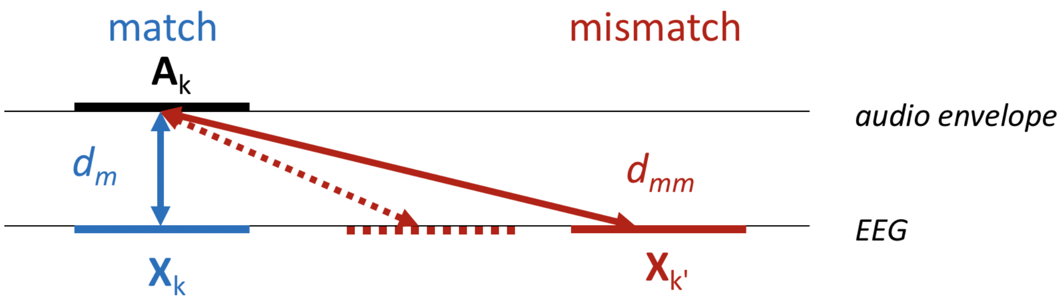

Figure 1: Match-mismatch task. The distance between audio feature and EEG data is quantified over time-aligned (matching) segments and misaligned (mismatch) segments.

evoked, and some unrelated segment of EEG signal $\mathbf{X}_{\mathbf{s}'\neq\mathbf{s}}$, decide which of the two EEG segments matches, via a model, the stimulus (Fig. 1). A practical application might be to determine whether a user is attentive to sound, or whether a particular alarm sound was noticed. Here we use it simply to mesure the quality of the stimulus-response model.

Metrics. The goodness-of-fit of stimulus-response models will here be evaluated using three metrics: correlation, sensitivity index, and classification error rate, the last two contingent on the MM task. The first, correlation, is calculated between transforms $f(\mathbf{A})$ and $g(\mathbf{X})$ over a segment of duration D or over the full duration of the data. When the data are normalized, as they are in this paper, correlation is related to Euclidean distance by the relation $r = 1 - d^2/2$. A perfect match is characterized by $r = 1$, $d = 0$, lack of correlation by (in expectation) $r = 0$, $d = \sqrt{2}$.

The second, sensitivity index, is based on the distribution of the difference $\Delta_s = d_{mm} - d_m$ of Euclidean distances for matched and mismatched segments. For each segment s of stimulus (transformed and z-scored), d_{mm} is calculated as the distance to mismatched segments s' of EEG (transformed and z-scored), averaged over all $s' \neq s$, while d_m is the distance to the matched segment of EEG features. Values of d_{mm} cluster around $\sqrt{2}$ because the data are normalized and mismatched segments are uncorrelated. Matched distances d_m tend to have smaller values, and so the difference Δ_s is (hopefully) positively distributed. The sensitivity index is calculated as the mean of this distribution divided by its standard deviation (standardized mean):

$$z = m/\sigma. \tag{3}$$

This definition is analogous to that of the "standardized mean difference" or "d-prime", but differs in that it reflects the distribution of the difference between d_m and d_{mm}, rather than the distributions of those values themselves.

The third, error rate, counts the proportion of segments classified incorrectly in the MM task (the proportion of segments s for which $\Delta_S < 0$). Both metrics depend on segment duration D, which is varied as a parameter: the shorter the segment, the noisier the correlation or decision calculation, and the harder the task. Error rate (e) is preferred to proportion correct ($1 - e$) because, plotted on a logarithmic scale, it better reveals incremental steps towards better performance.

Each metric has its virtues, as elaborated in the Discussion. Error rate is relevant for applications, but it hinges on a few samples near the classifier decision boundary, and thus may be noisy and insensitive to small (but real) increments in model quality. The sensitivity index is more stable and sensitive to subtle differences between distributions, and it is predictive of error rate for a balanced two-class classification problems if samples are normally distributed. Both metrics require a task. In contrast, correlation (or Euclidean distance) requires no task, but is sensitive to trivial manipulations such as low-pass filtering (see below). Thus the metrics are complementary and we report all three.

Cross-validation. Regression and CCA are data-driven and thus prone to overfitting. To avoid inflating results, a model can be trained and tested on different sets of data. Supposing the data are divided into Q "trials", the model is fit on $Q-1$ trials and correlations are evaluated on the Qth (left-one-out). Correlations then are averaged over all Q choices of the left-out trial. In the absence of a true effect, cross-validated correlation is approximately distributed around zero, but any individual estimate might be non-zero. To test whether an empirically-observed value is *significantly* non-zero, cross-validated correlation may be calculated repeatedly using samples of surrogate data of similar characteristics, but for which no relation is expected. A well-known technique is to apply a Fourier Transform to each column of the real data, replace the phases by random values, and apply the inverse Fourier Transform to obtain surrogate data with the same power spectrum (and hence same autocorrelation) as the original, but with random phases and hence zero expected correlation.

1.2 Extending and reducing the model

At least three factors degrade the model fit: *latency* and *spectral mismatch* between the stimulus representation and the brain response, and *additive noise* in the response. These can be alleviated by augmenting the data with a set of time lags (or a filter bank). The resulting increase in free parameters may be compensated for by dimensionality reduction techniques to reduce the tendency to overfitting.

Lags and Time Shift. It may be useful to augment the stimulus and/or brain signals with *time lags*. Applying a set of lags $0 \dots L_A - 1$ to $\mathbf{A}$ and concatenating the time-lagged channels side by side yields a matrix of size $T \times KL_A$. Similarly, applying L_X lags to $\mathbf{X}$ yields a time-lagged matrix of size $T \times JL_X$. The motivation for applying lags is that it allows the algorithm (univariate regression or

CCA) to automatically synthesize a *finite impulse response filter* (FIR) or, in the case of multichannel data, a multichannel FIR. This allows the model to minimize spectral mismatch (amplitude and phase) between $\mathbf{A}$ and $\mathbf{X}$, greatly enhancing its flexibility. The number of lags L determines the order of the synthesized FIR filter. A larger L confers the ability to select or reject temporal patterns on a longer time scale (lower frequencies), at the cost of greater computational cost and greater risk of overfitting.

In addition to these lags, we introduce an overall *time shift* S between stimulus and response. This parameter, distinct from the lags, is intended to absorb any gross temporal mismatch due to instrumental or sensory latencies. This frees the lag parameters to fit finer spectro-temporal characteristics. Without it, a larger value of L might be needed, with greater computational cost and risk of overfitting. S is treated as a hyperparameter: the fit is repeated for several values and the one that yields the highest correlation value is retained.

Dyadic filter basis. Lags $0 \dots L-1$ form a basis of the space of FIR filters of order L, but one can choose a different basis, for example outputs of a L-channel *filter bank* of FIRs of order L. To reduce dimensionality, one can then choose a subset $L' < L$ of that basis, defining a L'-dimensional *subspace* of the space of FIRs of order L. With a judicious choice of filter bank, performance with $L' < L$ channels may be superior to merely choosing $L' < L$ lags, in part due to a lower risk of overfitting. For example, a logarithmic filter bank (e.g. wavelet, or dyadic) can capture patterns of both short and long time scale with a limited number of channels, whereas capturing the same long time scale with a basis of lags would entail a much larger dimensionality. Here, we use a dyadic filter basis.

Dimensionality reduction. The models we describe here can be large, including a large number of parameters, yet we might not have enough training data so the fitting process may be prone to overfitting. Overfitting can be made less severe by *reducing the dimensionality* of the data before fitting the model, or by applying *regularization* within the fitting algorithm (Wong et al., 2018). The two approaches are closely related (Tibshirani et al., 2017, Sect 3.4.1). Here, we use dimensionality reduction because it can be applied in stages and separately for stimulus and EEG representations. Typically, data are submitted to Principal Component Analysis (PCA) and principal component (PCs) beyond a certain rank N are discarded, thus ignoring directions of low variance within the data. This enforces the reasonable assumption that low-variance directions are dominated by a noise floor (for example due to sensor noise). Since brain activity along those dimensions, if any, would be swamped by the noise, little is lost by removing them. Ridge regularization has a similar effect (Tibshirani et al., 2017). As an alternative to PCA, we consider also *shared component analysis* (SCA) (de Cheveigné, 2020). Whereas PCA favors directions with large variance, SCA favors directions shared across multiple channels.

1.3 Evaluation

Given the task described above, there are several ways we can measure success. This subsection describes metrics, using cross-validation to limit overly optimistic measures of success.

Data The data we use here are from a study that aimed to characterize cortical responses to speech for both normal-hearing and hearing-impaired listeners (Fuglsang et al., 2020). Experimental details are provided in that paper, and the data themselves are available from `http://doi.org/10.5281/zenodo.3618205`. In brief, 64-channel EEG responses to acoustic stimuli were recorded at a sampling rate of 512 Hz from 44 subjects, including both normal-and hearing-impaired. Stimuli for the latter were equalized to compensate for the impairment, and we pool data from both. Stimuli presented to each subject included 16 segments of single-talker speech with a male or female talker speaking in quiet, each of 50 s duration, that we consider in this study. Other stimuli presented in the same recording session (concurrent speech, tones) are not used. The publicly available dataset includes the temporal envelope of the speech stimulus, sampled at the same rate as the EEG, calculated by a model of instantaneous loudness that has been shown to be a predictor of cortical responses (Lalor et al., 2009; Ding and Simon, 2012; Di Liberto et al., 2015; Crosse et al., 2016).

Preprocessing The EEG data were smoothed by convolution with a square window of duration 1/50 Hz (implemented with interpolation) to suppress the line artifact (50 Hz and harmonics) and downsampled by smoothing with a 4-sample square window and decimation by a factor of 4 to 128 Hz. The data were detrended by applying a robust detrending algorithm (de Cheveigné and Arzounian, 2018) that robustly fit a 2nd order polynomial to overlapping intervals of size 15 s, subtracted the fit, and "stitched" detrended intervals together with a standard overlap-add procedure. The data were then high-pass filtered at 0.5 Hz using an order-2 Butterworth filter, then low-pass filtered at 30 Hz also with an order-2 Butterworth filter, and cut into 16 trials of 50 s duration. To remove eyeblink artifacts, a temporal mask was derived from the absolute power on a combination of two EOG channels and three frontal channels (F1, F2, Fz). Using this mask as a bias, the DSS algorithm was applied to find a transform maximizing eyblink activity (de Cheveigné and Parra, 2014) and the first two components (representing eyeblink artifact) were projected out of the EEG data.

To avoid aggravating the mismatch between stimulus and brain response, the stimulus envelope was filtered using the same high pass and low pass filters as for the EEG. All filters were "single pass" (causal).

Basic Models. To ease comparison with other studies, we define six models (Fig. 2) that illustrate basic processing choices, some of which have been made in prior studies and all of which are useful to understand in detail. For each, an overall time shift S is applied to the stimulus relative to the EEG.

- *Model* **A** compares one EEG channel with the stimulus envelope, with no spatial or temporal filtering ($L_A = 1$, $L_X = 1$) other than the time shift S.
- *Model* **B** compares one EEG channel with a linear combination of time-lagged envelope signals (L_A=11, L_X=1) obtained by regression. This corresponds to a standard forward model as reported in the literature.
- *Model* **C** compares the envelope to a linear combination of EEG channels (without lags; $L_A = 1$, $L_X = 1$) obtained by regression. This is analogous to the basic backward model considered in de Cheveigné et al. (2018), or the single-delay model of Hausfeld et al. (2018)
- *Model* **D** compares linear combinations of time-lagged envelope signals with linear combinations of EEG channels ($L_A = 11$, $L_X = 1$), obtained by CCA. This is analogous to CCA model 1 of de Cheveigné et al. (2018).
- *Model* **E** compares the envelope with a linear combination of time-lagged EEG channels ($L_A = 1$, $L_X = 11$) obtained by regression. This is analogous to the backward model of e.g. Fuglsang et al. (2017), or the multiple-delay model of Hausfeld et al. (2018).
- *Model* **F** compares linear combinations of time-lagged envelope signals with linear combinations of time-lagged EEG channels ($L_A = 11$, $L_X = 11$),

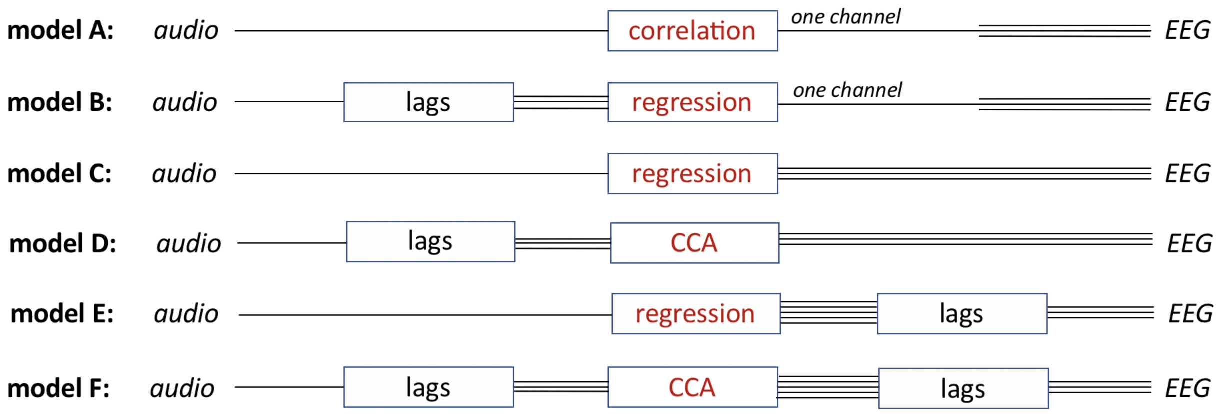

Figure 2: Baseline models. Models **A** and **B** compare the stimulus time series to a single EEG channel time series, models **C** to **F** compare the stimulus to the ensemble of EEG channels. Depending on the model, the stimulus may or not be augmented by applying time shifts of 0 to $L_A - 1$ samples, and the EEG may or may not be augmenting by applying time shifts of 0 to $L_X - 1$ samples. The fit between stimulus and EEG response is quantified by normalized cross-correlation, preceded by regression for models **B**, **C** and **E**, and by CCA for models **D** and **F**. All models include an additional time shift S of the stimulus relative to the EEG (not shown).

obtained by CCA. This is analogous to "CCA model 2" in de Cheveigné et al. (2018).

To summarize the similarities and differences: models **A** and **B** relate the stimulus to just one of the J EEG channels. In contrast, all other models relate the stimulus to the ensemble of EEG channels. For models **A**, **B**, **C** and **E** the fit is based on univariate regression, and for models **D** and **F** on a multivariate CCA model. For univariate regression models, the fit is quantified by a single correlation coefficient, and for CCA by as many coefficients as CC pairs (Fig. 2).

Not counting S, the number of parameters in the fit is 1 for model **A**, $L_A = 11$ for model **B**, $J = 64$ for model **C**, $L_A + J = 55$ for model **D**, $JL_X = 704$ for model **E**, and $L_A + JL_X = 715$ for model **F**.

Model G. In addition to basic models **A**-**F**, we define a reference or "gold standard" model **G**, variant of model **F**, with a performance close to the best we found, and with a relatively straightforward and precisely defined implementation that can help future studies to document further improvements in performance. Details of this model are given in the Results section.

Display of results, statistics, implementation. Results are evaluated using the three metrics described above, and plotted as a function of selected parameters chosen to offer insight. Effects are tested for statistical significance using a non-parametric Wilcoxon signed rank test over subjects. Processing scripts in Matlab make use of the NoiseTools toolbox (`http://audition.ens.fr/adc/NoiseTools/`). Scripts are available at `http://audition.ens.fr/adc/NoiseTools/src/NoiseTools/EXAMPLES/match_mismatch/`.

2 Results

In the following, we evaluate and compare the models, focusing on the factors that affect performance. Section 2.1 compares performance of the six basic models (**A**, **B**, **C**, **D**, **E**, **F**; Fig 3), using the correlation metric for simplicity and to allow comparison with prior studies. Section 2.2 then introduces the MM classification task, and explores how sensitivity and error metrics depend on segment duration. Section 2.3 explores the dependency of all three metrics on the number of spatial dimensions (number of channels or principal components) and temporal dimensions (lags or filter channels). Based on this, Section 2.4 proposes a new model, **G**, for use as a comparison point in future studies. Section 2.5 investigates factors that cause the classifier to fail, and Sect. 2.6 summarizes performance across models.

2.1 Correlation metric for basic Models

Figure 3 summarizes results obtained with the basic models. The first and second rows display correlation (calculated over the duration of each trial, $\sim$50 s) for models **A** to **F** for one subject (subject 4), plotted as a function of overall time shift S between stimulus and response. Thick black lines are cross-validated correlation, thin black lines are correlation without crossvalidation. Colored lines, where present, are cross-validated correlation for CCs beyond the first. Figure 3 (bottom right) summarizes these results by plotting, for each model, the peak cross-validated correlation averaged over subjects (red) and for individual subjects (gray, black for subject 4). In general, note that peak correlation increases from model **A** to **F**, and that this peak occurs for an overall shift value of $\sim$150 ms for these data.

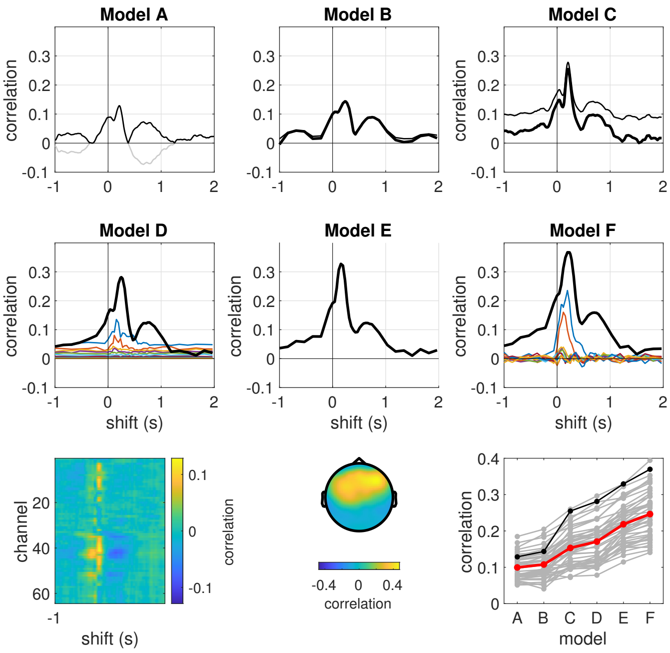

Figure 3: Baseline models **A** to **F**. Top two rows: correlation as a function of overall time shift S for each model, for subject 4. Model **A**: cross-correlation function (gray), or absolute value of same (black) between the stimulus envelope and channel 10 of EEG (FC3). Model **B**: correlation between channel 10 of EEG and the projection of channel 10 on the time-lagged stimulus, with crossvalidation (thick) or without (thin). Model **C**: same, between the stimulus and its projection on all EEG channels. From now on in this paper we consider only crossvalidated correlation. Model **D**: cross-validated correlation between CC pairs for first (black) and subsequent (color) CCs, for CCA between time-lagged stimulus and EEG channels. Model **E**, cross-validated correlation between stimulus and projection on time-lagged EEG channels. Model **F**: same as **D** for CCA between time-lagged stimulus and time-lagged EEG channels. Bottom left: cross-correlation functions between stimulus and EEG for all electrodes (Model **A**). Bottom center: topography of correlation coefficients beween EEG-derived component of first CC pair of model **D** and individual EEG channels. Bottom right: peak correlation for each model, for all subjects (gray) and average over subjects (red). The black line is subject 4.

Model A. This is the simplest incarnation of a stimulus-response model (Eq. 2), with f and g both identity functions. Figure 3 (top left) shows correlation (gray) and absolute correlation (black) between stimulus and EEG as a function of shift S for the best EEG channel (FC3). Correlation vs shift is identical to the cross-correlation function between stimulus and response. Peak absolute correlation is 0.13 for this subject; peak values for other subjects can be read off Fig. 3 (bottom right). Translating correlation to variance explained, for the best subject, about 4% of the variance of the best channel is explained by the stimulus. This might seem small, but still it is remarkable given the many other processes active in the brain, as well as non-neural artifacts. The shape of the cross-correlation differs slightly between electrodes (Fig. 3 bottom left), implying that response properties are not uniform across the brain.

Model B. In this model, the same EEG channel is projected onto the subspace spanned by the L_A=11 time-lagged stimulus signals, thus yielding weights that define an *FIR filter* applied to the stimulus. Figure 3 (top center) shows correlation (thin) and cross-validated correlation (thick) as a function of shift S for the best channel (FC3). Cross-validated correlation differs only slightly from raw correlation (thick vs thin) suggesting minimal overfitting in this simple model. Peak correlation is greater than for model **A**, suggesting that the FIR filter has improved the fit. This improvement is robust across subjects (Fig. 3, bottom left), as confirmed by a Wilcoxon signed rank test (p$<10^{-8}$).

Model C. In this model, the stimulus is projected onto the subspace spanned by the J=64 EEG channels, resulting in a *spatial filter*. Figure 3 (top right) shows correlation (thin) and cross-validated correlation (thick) between the stimulus signal and its projection (spatially-filtered EEG) as a function of shift S. The peak correlation is higher than for the previous two models (p$<10^{-8}$). The topography associated with the projection (correlation with individual EEG channels) shows a pattern typical of auditory responses (Fig. 3, bottom center).

Model D. In this model, time lags are applied to the stimulus but not the EEG. Time-lagged stimulus and multichannel EEG being both multivariate, the appropriate tool is CCA, which results in multiple CC pairs, each associated with a correlation value. Figure 3 (middle left) shows cross-validated correlation for the first CC (thick black) and subsequent CCs (color). Peak cross-validated correlation is larger compared to previous models ($p<10^{-11}$). Additional CCs appear to show elevated correlation: each is associated with a distinct FIR filter applied to the stimulus, and a distinct spatial filter applied to EEG. The existence of multiple correlated CCs suggests that the stimulus-response model captures multiple brain processes each sensitive to different frequency bands within the stimulus.

Model E. In this model, time lags are applied to all EEG channels, resulting in a backward model with both spatial and temporal filtering, analogous to the backward model of e.g. Fuglsang et al. (2017). Peak cross-validated correlation is higher than all previous models (Fig. 3, middle center, $p<10^{-6}$).

Model F. Finally, a logical step is to apply lags to both stimulus and EEG. Each CC then associates a distinct FIR filter applied to the stimulus with a distinct *multichannel FIR filter* applied to the EEG. Peak cross-validated correlation is again higher than all previous models (Fig. 3 middle right), $p<10^{-12}$.

Figure 3 (bottom right) shows that this progression across models is observed in most subjects (gray lines), as summarized by their mean (red line). Three features seem to contribute to a better fit: *spatial filtering* made possible thanks to the multichannel nature of EEG (models **C-F**), *temporal filtering* allowed by augmenting the data with time shifts (models **B**-**F**), and *CCA* which allows multivariate representations of both stimulus and response to be optimally related (models **D** and **F**). It is worth noting that these models differ also in their number of *free parameters*, from 1 for model **A** (not counting shift) to 735 for model **F** (see Methods). One might speculate that the increasing number of free parameters, rather than any particular feature, is what explains the progression in correlation scores. However, these results were obtained for cross-validated correlation for which overfitting is expected to be detrimental. Instead, it seems that the more complex models genuinely provide a better fit for this dataset, as confirmed with other metrics, below.

2.2 Task-based metrics

Here, we take the best model so far in terms of correlation (**F**), and rate its performance of in terms of sensitivity and error in the MM task. As explained in the Methods, the task is to decide whether a segment of the audio stimulus of duration D matches, via a model, the segment of EEG data that it evoked better than unrelated segments. For every segment s of audio we calculate the Euclidean distance d_m with the corresponding EEG segment and compare it with the average Euclidean distance to unrelated segments d_{mm}. A successful match is declared if $\Delta_s = d_{mm} - d_m > 0$. The chance error rate is 50%.

Distances are noisier for shorter segments, so we expect values of Δ_s to be more scattered, and errors more common, for smaller values of D. Figure 4 (left) shows the distribution of Δ_s for segment durations D=10s (red) and D=1.25s (blue). For longer segments, the distribution includes mostly positive values (correct classification), for shorter it includes a greater proportion of negative values (incorrect). The degree to which the distribution extends to positive values, minimizing error, is captured by the sensitivity index defined as the standardized mean

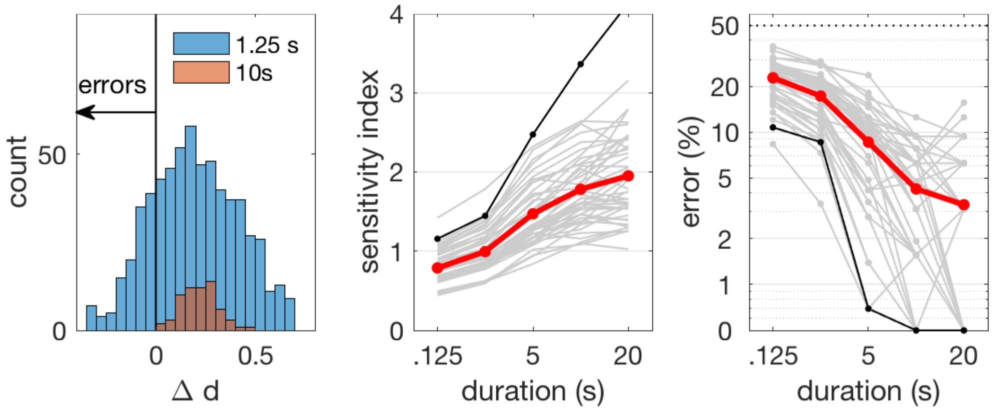

Figure 4: Match-mismatch task. Left: histogram of Δ_s for segment durations of 1.25 s (blue) or 10 s (red), for subject 4. For shorter segments the histogram is wider and there are more errors. Center: sensitivity index μ/σ as a function of segment duration averaged over subjects (red) and for each individual subject (gray, subject 4 is black), for model **F**. Right: error rate. A lower error rate here indicates that the single-talker stimulus-response model more faithfully discriminates between match-mismatch segments.

of Δ_s. Larger is better.

Figure 4 (center) shows the sensitivity index as a function of segment duration averaged over subjects (red) and for individual subjects (gray, subject 4 is black). Figure 4 (right) likewise shows error rate as a function of duration. We expect the sensitivity index to be greater, and the error smaller, for a longer segment duration because the task is easier, and indeed this is the case. In the following we focus on D=5s, for which the error rate averaged over subjects is $\sim$9% for this model (model **F**, $L_A = L_X = 11$). The variability over subjects is remarkable: at 5s duration the error rate ranges from close to 0 (perfect classification) to more than 20%.

2.3 Spatial and temporal dimensionality.

This section explores ways to boost performance beyond that of model **F**. Comparing basic models (Fig. 3) it appears that performance can benefit from both spatial filtering and lags. However, a recurring issue for stimulus-response models is overfitting, which depends on the complexity of the model, function here of both the number of spatial dimensions (channels or PCs), and the number of lags. Both factors are explored here.

Number of spatial dimensions. Using model **C** (no lags) as a reference point, Fig. 5 (red) shows the effect of applying PCA to the EEG data and discarding PCs beyond a certain rank N. The sensitivity index peaks, and the error rate is minimal, for $N \approx 32$, suggesting that overfitting may be occurring due to excess dimensionality and that reducing dimensionality can mitigate its effects.

Truncating the series of PCs is markedly better than the simple expedient of discarding channels (dotted line; channels were sorted by decreasing correlation with the stimulus and the least correlated were discarded). This result is interesting in relation to claims that reducing the number of electrodes can yield equivalent performance to the full set, or even better performance due to less overfitting (Montoya-Martínez et al., 2019). Such is not the case here: the sensitivity index (Fig. 5 center, dotted line) rises monotonically, implying that a reduced set of electrodes is inferior to the full set. At no point does performance reach the level that can be attained by selecting PCs from a PCA applied to the full set of electrodes. The conclusion is simple: more electrodes is better.

The benefit is slightly greater if PCA is replaced by a different transform, SCA (Shared Component Analysis, de Cheveigné, 2020, Fig. 5, blue) that favors components that are shared across electrodes ($p<10^{-3}$, Wilcoxon rank sum test).

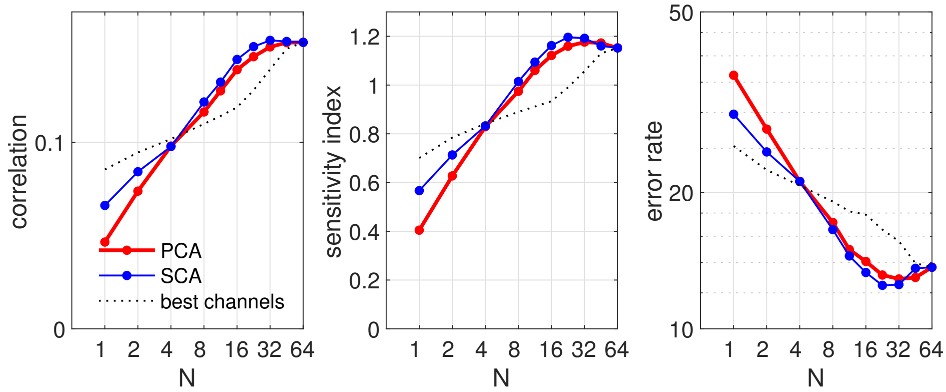

Figure 5: Performance as a function of the number of spatial dimensions. Left, red: cross-validated correlation averaged over subjects as a function of the number of PCs retained after PCA of the 64-channel EEG data. Blue: same, using SCA instead of PCA (see text). The dotted line represents subect-averaged correlation for subsets of EEG channels chosen for maximal correlation with the stimulus. Center: sensitivity index. Right: error rate. The model here includes no lags (similar to model **C**). Segment size is 5s.

Number of lags ($L = L_A = L_X$). Figure 6 shows metrics of correlation, sensitivity index, and error rate as a function of L averaged over subjects (red) and for individual subjects (gray, subject 4 is black). As the number of lags is increased, correlation and sensitivity increase until L=32 (250 ms), then decrease beyond. This peak is mirrored by a dip in error rate at $L = 32$. The best error rate for lags is 2.8% on average over subjects.

Referring to Eq. 3, the downturn at $L = 32$ might reflect an increase in d_{mm} relative to d_m, thus reducing the numerator m, or an increase in their variablity, thus increasing the denominator σ.

This non-monotonic pattern is suggestive of overfitting because a larger number of lags implies also a larger number of parameters. However, it might also be that large lags are deleterious for some other reason, for example because they capture slow patterns that do not generalize well. The blue lines in Fig. 6 represent the same metrics for a model in which lags $1 \dots L$ have been replaced by channels $1 \dots L'$ of a dyadic filter bank with FIR filters of order L. The number L' of channels is smaller than the order L of the filters ($L' = 10$ for $L = 32$; $L' = 12$ for $L = 64$, etc.). Since fewer CCA parameters are required for a dyadic filterbank of order L than for L lags, we would expect less overfitting. Contrary to that expectation, sensitivity and error metrics for lags and dyadic filter show a knee at the same value of L (32), compare red and blue in Fig. 6, suggesting that overfitting is *not* a critical factor in this pattern. This conclusion is reinforced by the fact that replacing the 64 EEG channels by N=32 PCs (or SCs) before applying the dyadic filter bank also results in a knee at $L = 32$ (not shown).

Once again, the variability of these metrics over subjects is remarkable. For L=32, the error rate for 5-second segments ranges from 0% for the best 10 subjects to $\approx$9 % for the worst. Incidentally, the error rate averaged over hearing-impaired subjects (1.8%) is smaller than for normal hearing subjects (2.8%), p<0.005, t-test. Several studies have reported stronger cortical responses for hearing impaired than normal hearing subjects (Goossens et al., 2018; Decruy et al., 2020; Fuglsang et al., 2020).

2.4 Model G ("gold standard")

Given the techniques described above and their results we can define a new gold-standard model applicable to this dataset. This model embodies one choice of processing and parameters among those explored so far. It is intended as a precisely-defined and easy-to-implement reference with which to evaluate new algorithms.

We define model **G** with the following steps: (a) a time shift of 200 ms is applied to the EEG relative to the audio envelope to compensate for the latency of the neural response, (b) data are preprocessed as described in Methods, (c) PCA is applied to EEG and the first 32 PCs are selected, (d) audio envelope and EEG

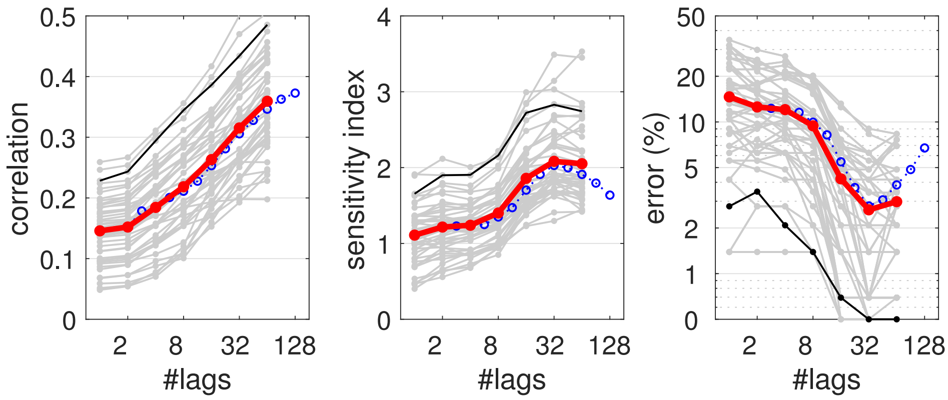

Figure 6: Performance as a function of the number of lags applied to the stimulus and to the EEG. Left: crossvalidated correlation as a function of number of lags L_A=L_X averaged over subjects (red) and for all subjects (gray, black is subject 4). Blue are for a dyadic filter bank instead of lags (see text). Center: sensitivity index. Right: error rate. Segment size is 5s ().

are augmented by applying lags 0 . . . 31, (e) the augmented data are fit by a linear stimulus/response model based on CCA, (f) sensitivity and error rate metrics are derived from the MM task applied to consecutive segments of duration 5s.

To be precise: the CCA solution is trained on subsets of 15 trials and tested on the 16th (left out), and this is repeated for all 16 choices of left-out segments (folds). For each fold, all consecutive 5s segments of audio within the left out trial are considered. For each segment, the Euclidean distance d_m between that segment of audio and the corresponding segment of EEG is calculated (matched distance), and the Euclidean distance d_{mm} between that segment of audio and all consecutive segments of EEG of all 15 *other* trials is calculated and averaged over those trials and segments (mismatched distance). Those two numbers are subtracted to yield a difference score (one per segment), and these scores are aggregated over all 16 folds, forming a distribution of difference scores. The ratio between the mean of this distribution and its standard deviation yields the *sensitivity* metric, and the proportion of samples for which the difference $d_{mm} - d_m$ falls below 0 yields the *error rate* metric. All distance calculations take into account the first 5 CCs of the CCA solution. A Matlab implementation of these steps is available at `http://audition.ens.fr/adc/NoiseTools/src/NoiseTools/EXAMPLES/match-mismatch/`.

To evaluate a new method, the recommended procedure is (1) implement model **G** on the system used to implement the new algorithm, (2) test it using the same publicly available database as we use to verify that the metrics yield scores consistent with what we report, and (3) apply the new method to the same database and compare scores with (2). The reason for step (2) is to control for implementation-specific differences (e.g. single vs double precision, etc.).

Alternatively, if a different database is to be used, do (1) as above then (2') test model **G** using that database, and (3') test the new method on that database and compare scores with (2'). In any event, it is not recommended to compare a new method with prior methods on a different database, or with different metrics, or with a different task. For example, there would be little merit in comparing the scores we report here to those reported in the literature for AAD.

2.5 Anatomy of an error

One of our goals is to gain a better understanding of factors that determine model quality. The difference $\Delta_s = d_{mm} - d_m$ might fall below zero as a result of a relatively small value of d_{mm} or a relatively large value of d_m. It is clear from Fig. 7 that for subject 3 (relatively poor model performance) the latter is the main factor. The top panel shows d_m (dots) and d_{mm} (crosses) for all segments of all trials. The mismatched distances are distributed tightly around $d_{mm} \approx 1.4$ as expected (Sect. 1.1, Metrics) whereas matched distances d_m mostly fall well outside this distribution. This is clear also from the scatterplot of d_m (dots) vs d_{mm}(bottom left). The diagonal line represents the classification boundary $\Delta_s =$

0: all points to the right and below (red) are misclassified. Another plausible boundary, $\bar{d}_{mm} - d_m$, is shown as a vertical dotted line.

The matched distance d_m is a good predictor of classification reliablity: for $d_m < 1.3$ the classification statistic Δ_s is distributed far from the decision boundary (Fig. 7, bottom right, brown), so the classification is highly reliable. For larger values of d_m the classification is less reliable. This implies an asymmetry in the conclusions that can be drawn from the classifier. For example a hypothetical "attention-monitoring" device might rapidly and reliably detect that a stimulus *has* registered within a subject's brain, but the opposite conclusion that it *has not* registered would take longer and/or be less reliable.

What factors might inflate d_m? Regressing d_m on RMS EEG shows a significant but weak correlation (r=0.12, p<10^{-7}), suggesting that high-amplitude glitches in the EEG might be a contributory factor. Likewise, a significant but weak negative correlation with RMS stimulus (r=-0.07, p<10^{-20}) suggests a possible small contribution of lulls in the stimulus. However, the small correlation values suggest that other factors, unknown, dominate.

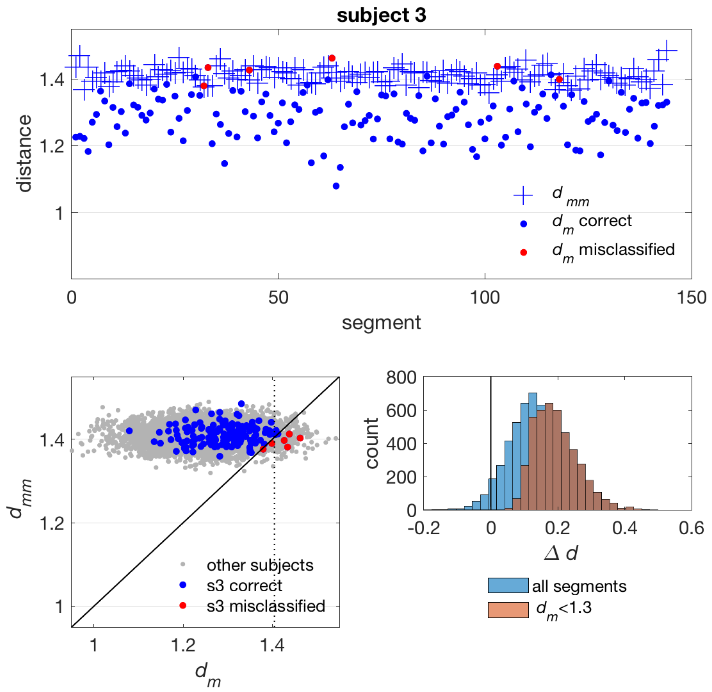

Figure 7: Top: Euclidean distance between matched (dots) and mismatched (+) segments of duration 5 s for all trials of one subject (subject 3, chosen for a relatively high error rate). Red dots indicate classification errors. Bottom left: scatterplot of mismatched vs matched distances for subject 3 (blue/red) and all other subjects (gray). The diagonal represents the classification boundary $\Delta_s = 0$. Points below that line (red) are misclassified. Bottom right: histograms of values of Δ_s for all segments (blue), and for segments for which the matched distance d_m is less than 1.3 (brown).

2.6 Summary of methods

Figure 8 summarizes error rates obtained with each of the models **A**-**G**, averaged over subjects. Models **A** and **B** are classic forward models that attempt to predict one channel of EEG from the stimulus representation. Models **C** and **E** are classic backward models that attempt to infer the stimulus representation from the EEG. Models **D**, **F** and **G** are hybrid models based on CCA. The best model (**G**) makes an order of magnitude fewer mistakes than the worst (**A**). For a 5s window the error rate for model **G** is less than 3% on average over subjects (0% for 10 subjects). Extrapolating from progress so far, we think that further progress is possible. Associated with the publicly available dataset that we used, model **G** might serve as a "gold standard" for comparative evaluation of such future progress.

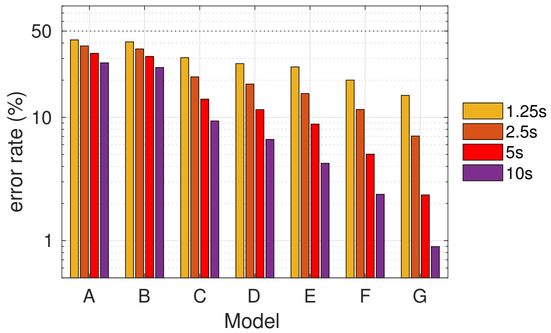

Figure 8: Summary of error rates for models **A**-**G**, averaged over subjects, for several values of duration D. The dotted line represents chance (50%)

3 Discussion

This study offers two main contributions. First, it introduces a simple objective task to help in the evaluation of stimulus-response models. Second, it documents a set of techniques that boost performance beyond state of the art.

The need for an objective task. A straightforward quality metric for a stimulus-response model is *correlation*, which can be calculated between response and predicted response in a forward model, between stimulus and inferred stimulus in a backward model, or between transforms of both in a hybrid model. That metric is simple and informative: better models tend to yield higher scores. However an elevated score can also result from chance correlations, which tend to be more widely distributed for data dominated by low frequencies. This could mislead a researcher to conclude that lowpass filtering improved the model, despite losing potentially relevant information carried by higher frequencies (Kriegeskorte and Douglas, 2019). The performance metrics of an objective task alleviate this problem, because loss of relevant information must impair task performance. Another argument in favor of an objective task is that success is a good measure of the model's "real world" value.

Why three metrics? Firstly, they are not equivalent: referring to Fig. 6, sensitivity and error rate (center and right) show a reversal at $L = 32$ indicative of overfitting that is not visible with the correlation metric (left). The appeal of *error rate* is that it is directly relevant for applications, the downside is that it is somewhat coarse and brittle (it depends on a few samples near the classification boundary). The appeal of *sensitivity* is that it depends on all samples by summarizing them based on their mean and standard deviation, but like error rate it requires a task. The appeal of *correlation* is that it is task-agnostic. Thus, the three metrics are complementary.

Selective versus sustained attention. Auditory attention is often investigated in a situation where multiple stimuli compete for attention, for example two concurrent pulse trains (Hillyard et al., 1973), or competing voices (Kerlin et al., 2010), or competing instruments (Treder et al., 2014). Attention may also be characterized as the difference in response to a stimulus in the presence, or absence, of concurrent visual stimulation (Molloy et al., 2015), or of a behavioral task (Scheer et al., 2018). In each case, a comparison is made between brain responses to the same stimulus recorded in two situations. In contrast, the MM task requires only a single recording, and, more importantly, assumes no competition for attentive resources. As such it might be of use to monitor the general attentive (vs inattentive) state of a subject, for example to determine whether an alert has been perceived, or a message is likely to have registered, or to detect drowsiness.

The AAD task is attractive because it is directly relevant to a BCI application such as cognitive control of acoustic processing in a hearing aid. Improvements in performance on that task are critical for the usability for the device. However, even in that case, we believe it may still be fruitful to optimize the stimulus-response models using the MM task. Improvements obtained for the simpler task should transfer to the harder task.

A drawback of AAD is that it relies on specific experimental setups with competing voices, attention task instructions, and greater demands for listening effort. The MM task does not rely on data labels defined by the experimental setup but derives the labels (match vs mismatch) from manipulations of the input data. It can therefore be used with any type of speech listening data. An analogous task has been used successfully for self-supervised learning, for instance, by training neural networks to predict whether video and audio segments are temporally aligned (Owens and Efros, 2018; Arandjelović and Zisserman, 2018). Here, we focus on linear models, but the task and metrics can be readily extended for self-supervised training of large-scale neural networks that require extensive data. Being free of reliance on particular 'attention labels', the MM-approach is better suited to evaluate and compare and evaluate models across datasets with different experimental setups.

Another downside to the AAD task is potential mislabeling due to attentional capture by the wrong stream. Administering a questionnaire about the attended offers some degree of control, but it we cannot be sure that a subject consistently followed the instructions throughout. Thus, a certain proportion of the database might be *mislabeled*, an important concern when evaluating well-performing models for which the error rate might drop to a level comparable to the proportion of mislabeled data. The MM task is better in these respects.

Encoding, decoding, and hybrid models. A forward (encoding) model is judged by the proportion of brain signal variance that it can account for (Naselaris et al., 2011; Kriegeskorte and Douglas, 2019). However, much of the activity recorded on any single EEG or MEG channel is not stimulus-related, so that number is necessarily small, even for a model that perfectly predicted all stimulus-related brain activity. Unrelated variance can be reduced by selecting the best channels (e.g. located over sensory cortex), or by applying a spatial filter (e.g. Models **C** or **D**) or a spatiotemporal filter (e.g. Models **E** or **F**) to the brain response. This, in essence, is what is accomplished by a hybrid model such as CCA (Dmochowski et al., 2017; de Cheveigné et al., 2018; Zhuang et al., 2020). CCA is effective because it allows response variance unrelated to stimulation to be stripped away, leaving a remainder that can be more meanifully related to the stimulus.

The model then is predictive of a *transform* of the measured brain response, rather than of the response itself, which makes it harder to interpret than a forward model. For example, Model **F** defines a set of linear transforms of the time-lagged EEG signals (multichannel FIR), which are then each predicted from the stimulus envelope via an FIR filter. This is harder to interpret than Model **B** that defines the impulse response (or TRF) of a filter that directly predicts the response of one EEG channel from the stimulus envelope, or even Model **D** that defines a filter that predicts a linear combination of EEG channels (spatially filtered EEG).

Analogous comments can be made with respect to backward models (stimulus reconstruction): a hybrid model reconstructs only a select transform of the stimulus representation rather than its entirety. This difficulty of interpretation is a downside of hybrid models, an upside is that the transformed response $g(X)$ (right hand side of Eq. 2) is more reliably predicted by the stimulus than X, and thus arguably offers a closer (less noisy) view of sensory-dependent parts of brain activity, and of the information that they encode (Kriegeskorte and Douglas, 2019).

As an aside, it is worth noting that our classification task differs from typical decoding tasks (Kriegeskorte and Douglas, 2019) in that it operates on the [stimulus, response] pair, rather than only the brain response.

Equation 2 allows arbitrary transforms $f(A)$ and $g(X)$ that are more general than the linear transforms that we actually use. The aim for this more general framework is to leave room for more complex models, for example relating the stimulus to gamma power, etc. It could be further extended by allowing g(.) to depend on X (e.g. allowing for sensory processing to depend on brain state).

Improving the model. What do we expect of a stimulus-response model? Activity within the brain is largely unrelated to auditory stimulation, and conversely, some features of the stimulus might not affect the response (i.e. different stimuli might evoke the same response). This necessarily drives down the correlation for matched segments. Worse, spurious correlations may favor mismatched segments by chance, thus driving up the error rate. The role of the model is to factor out such dimensions of mismatch from both stimulus and EEG.

Linear models achieve this by linearly separating relevant and irrelevant patterns, projecting them into different subspaces. This can occur in at least three domains: *spatial* (exploiting cross-channel correlation structure), *spectral* (exploiting difference in spectral properties between relevant and irrelevant sources), or *temporal* (exploiting temporal sparsity of relevant and irrelevant sources). The transforms involved are linear, but discovered by data-driven algorithms that are not. There are many such algorithms, some that we explored, others that remain to be explored.

Prior studies have considered mainly either a forward model (similar to model **B**) or a backward model (similar to **C** or **E**). Reported correlation values are typically "above chance" but still rather low. For example, a score of r=0.1 to 0.2

means only 1 to 4% variance explained, and a correct-classification score of 90% for a segment of 60 s duration (as reported for a typical subject of O'Sullivan et al., 2015), implies a one-minute wait for a decision that might be wrong on one trial out of every ten. For applications, it is crucial to achieve better reliability and smaller latency, and from the scientific perspective it is desirable to find models that offer a better fit to the data.

Forward and backward models transform the stimulus or the response, respectively, but not both, while CCA models transform both. CCA thus allows *both* data streams to be stripped of irrelevant variance, resulting in a better fit as reflected by higher values of the correlation metric (compare models **C** vs **D**, or **E** vs **F**). CCA also produces multiple correlation coefficients that yield a multivariate feature space for classification, with a further boost to task-based metrics.

An important ingredient in the more successful models is lags, that allow the algorithms to synthesize FIR or multichannel FIR filters. FIR filters allow the algorithm to compensate for any convolutional mismatch between the stimulus and EEG signals (e.g. due to latency or smoothing), resulting in better performance (compare models **A** vs **B**, **C** vs **D**, or **E** vs **F**). Adding lags effectively increases the dimensionality of the data space, which is beneficial *as long as the optimal transforms can be found*. Unfortunately, data-driven algorithms to find those transforms may be less effective in a larger space due to overfitting.

Model overfitting was addressed here using dimensionality reduction. This is achieved trivially by discarding sensor channels (with limited success, c.f. dotted line in Fig. 5), or limiting the number of lags (with greater success, Fig. 6 center and right). Replacing the set of lags by a smaller number of channels of a dyadic filter bank also reduces dimensionality ($J \times L' < J \times L$ for time-lagged EEG), with a considerable reduction in computation cost but little difference in performance (compare red and blue lines in Fig. 6). Applying PCA or SCA to the space of sensors and selecting a subset of components also reduces dimensionality ($J' \times L < J \times L$), with a slight boost in performance (Fig. 5). An additional benefit of dimensionality reduction is to reduce computational cost, which can otherwise become prohibitive if many lags are introduced (PCA and CCA require eigendecomposition which costs $o(N^3)$).

The reduction in performance beyond $L = 32$ (∼250 ms) for this dataset (Fig.6) suggests that the benefit of larger L is eventually overcome by overfitting. This could be merely the result of a larger number of free parameters, or more specifically because higher-order FIR filters can enhance slow patterns (low frequencies) that don't generalize from training data to test. The latter seems more likely: replacing L lags by a smaller number$L' < L$ of dyadic filters of order L had little impact on performance (compare blue to red in Fig. 6). The knee occurs at the same value of L (32), suggesting that filter order (or lag span), rather than dimensionality, is the critical factor.

Considering both the shift applied (∼200 ms), and the maximum lag (∼250 ms), the model associates stimulus samples with response samples that occur up to ∼450 ms later. However, we cannot on this basis make a strong statement concerning brain processing latencies, because of the potential smearing effect of the filters applied in preprocessing (Sect. 1.3) (de Cheveigné and Nelken, 2019).

Whither now? Further boosts in performance are needed to enhance the feasibility of potential applications. Based on what we know so far, there are several directions worth pursuing.

One is to improve the stimulus representation. Here, we used the stimulus envelope, a rather crude representation. Richer representations have been explored, such as auditory filterbank (Biesmans et al., 2017), higher-order linguistic structure (Di Liberto et al., 2015), onsets (Oganian and Chang, 2019), or voice pitch

(Forte et al., 2017; Teoh et al., 2019), etc., but they remain to be developed further and integrated. Multi-set CCA (MCCA), which allows merging EEG across subjects, may ease development of such stimulus representations (de Cheveigné et al., 2019).

A second direction is to improve EEG analysis. Standard models (including those reported here) exploit low-frequency components within the EEG, but useful information may also be carried by high-frequency power (Synigal et al., 2020; Forte et al., 2017; Teoh et al., 2019). If the relevant sources have low SNR, they may not be exploitable without appropriate spatial filtering, but standard linear techniques to find the filters (such as CCA) are not directly applicable. One promising approach is to use quadratic component analysis (QCA) to allow power sources to be isolated using standard linear methods (de Cheveigné, 2012). This entails forming cross-products between channels and/or lags, leading to very high-dimensional data, and thus requires an appropriate dimensionality-reduction strategy.

A third direction is better management of the time axis. As Fig. 7 (top) shows, errors occur only for segments for which the mismatch d_m is large, and these occupy only a small fraction of the time axis. A better understanding of what triggers large-mismatch events might allow them to be mitigated. Alternatively, since they are flagged by a high value of d_m, the application may be able to interpolate over them based on the high-reliability (low d_m) context.

A fourth direction is more prosaic: better preprocessing, filtering, artifact rejection, etc. We noted that performance metrics are sensitive to preprocessing parameters, but no attempt was made to tune them in this study.

Finally, a fifth direction is to resort to more recent machine-learning methods in lieu of expertise-based approaches, in the faith that they will discover the same regularities and structure as embodied by hand-crafted methods, and more. Results so far are modest (Ciccarelli et al., 2019; Jalilpour Monesi et al., 2020; Tian and Ma, 2020; Das et al., 2020), but success in many other fields suggests that machine-learning approaches are well worth pursuing.

Acknowledgements

This work was supported by grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL, and ANR-17-EURE-0017. Jens Hjortkjaer and Søren A. Fuglsang were supported by the Novo Nordisk Foundation synergy Grant NNF17OC0027872 (UHeal). We appreciate many helpful discussions with Jonathan Berent and his team.

References

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences* 98:13367–13372.

Aiken SJ, Picton TW (2008) Human Cortical Responses to the Speech Envelope:. *Ear and Hearing* 29:139–157.

Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2019) Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports* 9:874.

Arandjelović R, Zisserman A (2018) Objects that Sound In Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors, *Computer Vision – ECCV 2018*, Vol. 11205, pp. 451–466. Springer International Publishing, Cham Series Title: Lecture Notes in Computer Science.

Bednar A, Lalor EC (2020) Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. *NeuroImage* 205:116283.

Biesmans W, Das N, Francart T, Bertrand A (2017) Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25:402–412.

Broderick MP, Anderson AJ, Lalor EC (2019) Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *The Journal of Neuroscience* 39:7564–7575.

Ciccarelli G, Nolan M, Perricone J, Calamia PT, Haro S, O'Sullivan J, Mesgarani N, Quatieri TF, Smalt CJ (2019) Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods. *Scientific Reports* 9:11538.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience* 10:604.

Das N, Zegers J, Van hamme H, Francart T, Bertrand A (2020) Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding. *Journal of Neural Engineering* 17:046039.

de Cheveigné A (2012) Quadratic component analysis. *Neuroimage* 59:3838–3844.

de Cheveigné A, Arzounian D (2018) Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *NeuroImage* 172:903–912.

de Cheveigné A, Nelken I (2019) Filters: when, why, and how (not) to use them. *Neuron* 102:280–293.

de Cheveigné A, Wong D, Di Liberto G, J. H, M. S, Lalor E (2018) Decoding the auditory brain with canonical component analysis. *NeuroImage* 172:206 – 216.

de Cheveigné A (2020) Shared component analysis. *bioRxiv* .

de Cheveigné A, di Liberto GM, Arzounian D, Wong D, Hjortkjaer J, Asp Fuglsang S, Parra LC (2019) Multiway canonical correlation analysis of brain signals. *Neuroimage* 186:728–740.

de Cheveigné A, Parra LC (2014) Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage* 98:487–505.

Decruy L, Vanthornhout J, Francart T (2020) Hearing impairment is associated with enhanced neural tracking of the speech envelope. *Hearing Research* 393:107961.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current biology : CB* 25:2457–2465.

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology* 107:78–89.

Dmochowski J, Ki J, DeGuzman P, Sajda P, Parra L (2017) Extracting mutlidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. *Neuroimage* 180:134–146.

Forte AE, Etard O, Reichenbach T (2017) The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife* p. 12.

Fuglsang SA, Märcher-Rørsted J, Dau T, Hjortkjær J (2020) Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *Journal of Neuroscience* 40:2562–2572.

Fuglsang SA, Dau T, Hjortkjær J (2017) Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage* 156:435–444.

Goossens T, Vercammen C, Wouters J, van Wieringen A (2018) Neural envelope encoding predicts speech perception performance for normal-hearing and hearing-impaired adults. *Hearing Research* 370:189–200.

Hausfeld L, Riecke L, Valente G, Formisano E (2018) Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage* 181:617–626.

Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical Signs of Selective Attention in the Human Brain. *Science* 182:177–180.

Jaeger M, Mirkovic B, Bleichner MG, Debener S (2020) Decoding the Attended Speaker From EEG Using Adaptive Evaluation Intervals Captures Fluctuations in Attentional Listening. *Frontiers in Neuroscience* 14:603.

Jalilpour Monesi M, Accou B, Montoya-Martinez J, Francart T, Van hamme H (2020) An lstm based architecture to relate speech stimulus to eeg pp. 941–945. IEEE.

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party". *Journal of Neuroscience* 30:620–628.

Kriegeskorte N, Douglas PK (2019) Interpreting encoding and decoding models. *Current Opinion in Neurobiology* 55:167–179.

Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G (2013) The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE* 8:e53398.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology* 102:349–359.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236.

Molloy K, Griffiths TD, Chait M, Lavie N (2015) Inattentional Deafness: Visual Load Leads to Time-Specific Suppression of Auditory Evoked Responses. *Journal of Neuroscience* 35:16046–16054.

Montoya-Martínez J, Bertrand A, Francart T (2019) Optimal number and placement of eeg electrodes for measurement of neural tracking of speech. *bioRxiv* .

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. *NeuroImage* 56:400–410.

Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human superior temporal gyrus. *SCIENCE ADVANCES* p. 14.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex* 25:1697–1706.

Owens A, Efros AA (2018) Audio-Visual Scene Analysis with Self-Supervised Multisensory Features In Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors, *Computer Vision – ECCV 2018*, Vol. 11210, pp. 639–658. Springer International Publishing, Cham Series Title: Lecture Notes in Computer Science.

O'Sullivan J, Herrero J, Smith E, Schevon C, McKhann GM, Sheth SA, Mehta AD, Mesgarani N (2019) Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron* 104:1195–1209.e3.

Power AJ, Lalor EC, Reilly RB (2011) Endogenous Auditory Spatial Attention Modulates Obligatory Sensory Activity in Auditory Cortex. *Cerebral Cortex* 21:1223–1230.

Puvvada KC, Simon JZ (2017) Cortical Representations of Speech in a Multitalker Auditory Scene. *The Journal of Neuroscience* 37:9189–9196.

Scheer M, Bülthoff HH, Chuang LL (2018) Auditory Task Irrelevance: A Basis for Inattentional Deafness. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 60:428–440.

Synigal SR, Teoh ES, Lalor EC (2020) Including Measures of High Gamma Power Can Improve the Decoding of Natural Speech From EEG. *Frontiers in Human Neuroscience* 14:130.

Teoh ES, Cappelloni MS, Lalor EC (2019) Prosodic pitch processing is represented in delta?band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience* 50:3831–3842.

Tian Y, Ma L (2020) Auditory attention tracking states in a cocktail party environment can be decoded by deep convolutional neural networks. *Journal of Neural Engineering* 17:036013.

Tibshirani S, Friedman H, Hastie T (2017) *The Elements of Statistical learning* Springer Series in Statistics.

Treder MS, Purwins H, Miklody D, Sturm I, Blankertz B (2014) Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *Journal of Neural Engineering* 11:026009.

Wong DDE, Fuglsang SA, Hjortkjær J, Ceolini E, Slaney M, de Cheveigné A (2018) A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding. *Frontiers in Neuroscience* 12:531.

Zhuang X, Yang Z, Cordes D (2020) A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping* 41:3807–3833.

Zuk NJ, Teoh ES, Lalor EC (2020) EEG-based classification of natural sounds reveals specialized responses to speech and music. *NeuroImage* 210:116558.