JSLHR

Research Article

Machine Learning Approaches to Analyze Speech-Evoked Neurophysiological Responses

Zilong Xie,[a] Rachel Reetzke,[a] and Bharath Chandrasekaran[b]

Purpose: Speech-evoked neurophysiological responses are often collected to answer clinically and theoretically driven questions concerning speech and language processing. Here, we highlight the practical application of machine learning (ML)–based approaches to analyzing speech-evoked neurophysiological responses.
Method: Two categories of ML-based approaches are introduced: decoding models, which generate a speech stimulus output using the features from the neurophysiological responses, and encoding models, which use speech stimulus features to predict neurophysiological responses. In this review, we focus on (a) a decoding model classification approach, wherein speech-evoked neurophysiological responses are classified as belonging to 1 of a finite set of possible speech events (e.g., phonological categories), and (b) an encoding model temporal response function approach, which quantifies the transformation of a speech stimulus feature to continuous neural activity.
Results: We illustrate the utility of the classification approach to analyze early electroencephalographic (EEG) responses to Mandarin lexical tone categories from a traditional experimental design, and to classify EEG responses to English phonemes evoked by natural continuous speech (i.e., an audiobook) into phonological categories (plosive, fricative, nasal, and vowel). We also demonstrate the utility of temporal response function to predict EEG responses to natural continuous speech from acoustic features. Neural metrics from the 3 examples all exhibit statistically significant effects at the individual level.
Conclusion: We propose that ML-based approaches can complement traditional analysis approaches to analyze neurophysiological responses to speech signals and provide a deeper understanding of natural speech and language processing using ecologically valid paradigms in both typical and clinical populations.

Speech-evoked neurophysiological responses provide information about the neural mechanisms underlying the sensory encoding of the speech signal. They are of great value to researchers and clinicians in the fields of speech, language, and hearing sciences (Martin, Tremblay, & Korczak, 2008; Skoe & Kraus, 2010). Machine learning (ML)–based approaches have become increasing popular, particularly among researchers in the neuroscience of speech and language, in analyzing many different types of speech-evoked neurophysiological data from recordings via invasive techniques such as electrocorticography (Golumbic et al., 2013; Mesgarani & Chang, 2012; Moses, Leonard, & Chang, 2018) and noninvasive techniques such as magnetoencephalography (MEG; e.g., Brodbeck, Presacco, & Simon, 2018; Ding & Simon, 2012a, 2012b) and electroencephalography (e.g., Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Di Liberto & Lalor, 2017; Di Liberto, O'Sullivan, & Lalor, 2015; Di Liberto, Peter, et al., 2018; Khalighinejad, Cruzatto da Silva, & Mesgarani, 2017; Llanos, Xie, & Chandrasekaran, 2017; Xie, Reetzke, & Chandrasekaran, 2018; Yi, Xie, Reetzke, Dimakis, & Chandrasekaran, 2017).

Recent advances in ML-based analysis techniques have led to an emerging paradigm shift in the study of speech and language processing with natural stimuli beyond simplified, controlled stimuli in traditional experimental designs (Hamilton & Huth, 2018). Natural stimuli offer various advantages over simplified, controlled stimuli to shed light on the neural mechanisms underlying speech and language processing (Hamilton & Huth, 2018; Wöstmann, Fiedler, & Obleser, 2017). Nevertheless, the application

[a]Department of Communication Sciences and Disorders, The University of Texas at Austin
[b]Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh
Correspondence to Bharath Chandrasekaran: b.chandra@pitt.edu

Editor-in-Chief: Ryan McCreery
Received June 18, 2018
Revision received October 28, 2018
Accepted November 26, 2018
https://doi.org/10.1044/2018_JSLHR-S-ASTM-18-0244
Publisher Note: This article is part of the Research Forum: Advancing Statistical Methods in Speech, Language, and Hearing Sciences.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

of ML-based approaches for natural stimuli is still limited in the fields of speech, language, and hearing sciences. Therefore, we present the current review to highlight the practical utility of ML-based approaches to analyze speech-evoked neurophysiological responses, with the aim that researchers and clinicians would consider ML-based approaches and natural stimuli in future assessments of speech and language processing.

We will first introduce the major categories of ML-based approaches and outline the general steps for ML-based analysis. Then, we will demonstrate two applications of ML-based approaches to analyze speech-evoked neurophysiological responses. The first application involves electroencephalographic (EEG) responses to simplified, controlled speech stimuli from a traditional experimental design (Xie et al., 2018). Similar designs have dominated research in the fields of hearing, speech, and language (Martin et al., 2008; Skoe & Kraus, 2010; Tremblay, Friesen, Martin, & Wright, 2003). The second application focuses on EEG responses evoked in response to natural continuous speech. To demonstrate each analysis application, we highlight the advantages of ML-based approaches over traditional approaches. Finally, we discuss the clinical utility of ML-based approaches and natural stimuli, summarizing the advantages and limitations of ML-based approaches. We conclude with recommended improvements for the ML-based analysis approaches presented in the current review.

ML-Based Approaches

Two Major Categories of ML-Based Approaches

ML-based approaches to analyze speech-evoked neurophysiological responses fall into two broad categories: decoding models and encoding models. The decoding models generate a speech stimulus output using the features from the neurophysiological responses, whereas the encoding models use speech stimulus features to predict neurophysiological responses. Decoding models can be further delineated into two distinct categories: classification and reconstruction. In a classification approach, speech-evoked neurophysiological responses are classified as belonging to one of a finite set of possible speech events (e.g., phonological categories). In a reconstruction framework, speech-evoked neurophysiological responses are used to reconstruct continuously varying features (e.g., temporal envelope) of the speech stimulus to match the original speech features. For a comprehensive overview of decoding and encoding models used in neurophysiological experiments, see Holdgraf et al. (2017).

General Steps to Implement ML-Based Approaches

Here, we summarize the general steps to implement decoding and encoding models in the context of analyzing speech-evoked neurophysiological responses based on Holdgraf et al. (2017).

Step 1: Extraction of input and output features. For both decoding and encoding models, a representation of the speech stimulus and the neurophysiological signal needs to be estimated. In decoding models, the representation of the neurophysiological signal is used as the input and the speech stimulus representation as the output, and vice versa for encoding models. Examples of speech stimulus representation have been used, including acoustic features such as the amplitude envelope or speech-specific features such as phonemes and phonetic features (Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Di Liberto, Peter, et al., 2018), and lexical–semantic features (Broderick et al., 2018). The representation of neurophysiological signals is often a derivation of the raw neurophysiological signals, such as time-varying amplitude in a frequency band (e.g., Broderick et al., 2018; Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Di Liberto, Peter, et al., 2018), fundamental frequency (F0; Llanos et al., 2017), and spectral magnitude (Yi et al., 2017). For multiple-channel neurophysiological data, the choice of a particular set of channels can also be considered as part of the representation of the neurophysiological signals. The choice of features from the speech stimulus and neurophysiological signal generally underlies assumptions regarding particular levels of speech and language processing that are reflected in the neural responses (Di Liberto et al., 2015).

Step 2: Model selection and cross-validation. A model is selected to quantify the relationship between speech stimulus features and neurophysiological response features. The choice of model determines the type of relationship that can be represented between these features. In the decoding models, the model maps the neurophysiological response features to speech stimulus features, and vice versa for the encoding models. The selected model is fitted with some example data (i.e., training data set) to find an optimal model that yields the least error between the predicted estimates (e.g., predicted neurophysiological responses) and the corresponding original data (e.g., actual neurophysiological responses). Once the model has been estimated, the optimized model is validated to see how well it generalizes to novel example data (i.e., testing data set). The ability of the optimized model to generalize to the testing data set (i.e., model predictive score) is used to index the quality of neural processing of speech signals.

In practice, the model estimation and validation processes are conducted within a data set containing data from a single participant or a group of participants. A proportion of the data (i.e., training data set) is selected for model estimation, and another proportion is "held out" (i.e., testing data set) and used for model validation. These processes are repeated until all the data have been used for model estimation and validation and are often referred to as model cross-validation. The performance of the fitted model is averaged across all instances of validation.

It is critical to evaluate the extent to which a model's predictive score is statistically significant. Permutation tests have often been used to fulfill this purpose (Good,

2013; Ojala & Garriga, 2010). Generally, an empirical null distribution of predictive scores is derived by modeling the relationship between a pseudoversion of the speech stimulus features and the actual neurophysiological response features, or between the actual speech stimulus features and a pseudoversion of the neurophysiological response features. This can be achieved by mismatching the speech stimulus and neurophysiological response features (e.g., assigning incorrect speech stimulus to the neurophysiological responses) or disrupting the inherent structures of the speech stimulus or neurophysiological response features (e.g., randomizing the timing of the speech stimulus or neurophysiological response feature values). Then, the actual model predictive score is tested against the null distribution. The p value can be estimated using the formula: $p = (a + 1) / (n + 1)$ (Phipson & Smyth, 2010), where a is the number of predictive score from the null distribution that exceeds the actual predictive score and n is the total number of predictive scores from the null distribution.

Step 3: Model inspection and interpretation. One may evaluate the model parameters to gain insight into the relationship between stimulus speech features and neurophysiological activity (e.g., Di Liberto et al., 2015; Yi et al., 2017). The model parameters may be compared across experimental conditions or participants (Crosse, Butler, & Lalor, 2015; Di Liberto, Crosse, et al., 2018; A. E. O'Sullivan, Crosse, Di Liberto, & Lalor, 2017; J. O'Sullivan et al., 2014). For example, the magnitude of model weights may be used to index the strength of neural responses to the evoking stimulus speech features (e.g., Crosse et al., 2015; Di Liberto, Crosse, et al., 2018).

In the next sections, we illustrate two applications of decoding and encoding models to analyze speech-evoked neurophysiological responses. We focus on the first two steps outlined above, because for Step 3, the model parameters are not always straightforward to interpret and are largely dependent on the chosen model (Holdgraf et al., 2017). However, we want to emphasize to the reader that this is not a trivial step.

Application 1: Speech-Evoked EEG Responses From a Traditional Experimental Design

Speech-Evoked Neurophysiological Responses From Traditional Designs

In many, if not most, neurophysiological studies on speech processing, researchers have typically characterized neural responses to a limited set of repetitive, temporally isolated speech sounds that vary along a limited number of dimensions (Martin et al., 2008; Skoe & Kraus, 2010; Tremblay et al., 2003). This is mainly due to the constraints imposed by noninvasive neurophysiological recordings from humans. Neural responses from noninvasive neuroimaging modalities are susceptible to physiological noise. To overcome the poor signal-to-noise ratio, hundreds (or even thousands) of neural responses to repetitively presented stimuli are averaged together to provide an estimate of the neural response.

Speech-evoked neurophysiological responses from these traditional experimental designs provide invaluable insight on the processing of speech signals throughout the auditory system and are of great value to clinicians in the prevention, diagnosis, and rehabilitation of communicative deficits or disorders (Martin et al., 2008; Skoe & Kraus, 2010). For example, the frequency-following response (FFR), an electrophysiological response that reflects phase-locked activity to the physical properties of acoustic signals (Bidelman, 2015; Chandrasekaran & Kraus, 2010; Marsh, Worden, & Smith, 1970; Moushegian, Rupert, & Stillman, 1973; Skoe & Kraus, 2010; Smith, Marsh, & Brown, 1975; Worden & Marsh, 1968), has been widely adopted by researchers and clinicians. The scalp-recorded FFR provides a noninvasive window into the neural encoding of speech signals along the initial stages of the auditory pathway (Chandrasekaran & Kraus, 2010; Krishnan, 2002; Krishnan, Xu, Gandour, & Cariani, 2004; Skoe & Kraus, 2010). Researchers have used the FFR to index the integrity of early sensory encoding and how the fidelity of the FFR may relate to different speech and language abilities in a variety of populations, including children with developmental disorders (Russo, Nicol, Trommer, Zecker, & Kraus, 2009; Russo et al., 2008; White-Schwoch et al., 2015), older adults with hearing disorders (Anderson, Parbery-Clark, White-Schwoch, Drehobl, & Kraus, 2013), and normal aging adults (Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013).

In the next sections, we present the FFR as an example to demonstrate the application of ML-based approaches to analyze speech-evoked neurophysiological responses from traditional designs. It can be assumed that similar ML-based approaches can be applied to other types of speech-evoked neurophysiological responses from traditional designs.

Advantages of ML-Based Approaches

Prior work has demonstrated the feasibility of decoding models, specifically the classification approach, to characterize FFRs evoked by segmental speech features (e.g., vowels; Sadeghian, Dajani, & Chan, 2015; Yi et al., 2017) and suprasegmental speech features (e.g., linguistically relative pitch patterns; Llanos et al., 2017; Reetzke, Xie, Llanos, & Chandrasekaran, 2018; Xie et al., 2018).

A major motivation to characterize speech-evoked FFRs with ML-based approaches is to improve experimental efficiency by reducing the number of trials needed to evoke a meaningful brain response and, in turn, to shorten overall experimental time and minimize fatigue in participants. Because of the posited brainstem site of origin of the FFR (Bidelman, 2015; Chandrasekaran & Kraus, 2010; Smith et al., 1975), FFRs are small in magnitude and, in turn, have relatively low signal-to-noise ratio at the single-trial level. As a result, the extant studies typically rely on FFRs averaged across thousands of trials (Skoe & Kraus, 2010). In contrast to these prior studies, our recent work

has demonstrated the feasibility of classifying FFRs to vowel stimuli on a single-trial basis. Reliable classification performance can be achieved with as low as 50 trials per vowel stimulus (Yi et al., 2017). Most recently, we have successfully implemented support vector machines (SVMs), a widely used ML algorithm in the EEG literature (Garrett, Peterson, Anderson, & Thaut, 2003; Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007; Subasi & Gursoy, 2010), to classify FFRs to Mandarin lexical tones averaged across less than 100 trials, contributing to a better understanding of the impact of cross-modal attention and online stimulus context on the FFRs (Xie et al., 2018).

Despite the use of less trials, the neural metrics derived from these decoding models demonstrate robust degree of convergence with metrics based on traditional analyses of FFRs (Llanos et al., 2017; Xie et al., 2018). For example, Llanos et al. (2017) utilized a hidden Markov model to classify Mandarin lexical tone categories using the F0 of FFRs in native and nonnative speakers of Mandarin Chinese. Extensive prior work with traditional analyses that quantify the similarity in F0 between FFRs and the evoking Mandarin tones has revealed that the neural tracking of F0 is more robust in native relative to nonnative speakers of Mandarin Chinese (e.g., Bidelman, Gandour, & Krishnan, 2011; Krishnan, Xu, Gandour, & Cariani, 2005; Xie, Reetzke, & Chandrasekaran, 2017). Consistent with these findings, Llanos et al. found higher classification accuracy of the FFRs in native speakers of Mandarin Chinese, compared to nonnative speakers, even when FFRs were averaged across only about 100 trials.

A Demonstration of ML-Based Approach to Analyze FFRs

Sample Data Set

The FFR data for this demonstration are from a study on the effects of visual attention and auditory predictability on the FFRs carried out by Xie et al. (2018). This data set consists of FFRs to three linguistically relevant pitch patterns (Mandarin tones: T1, high-level; T2, low-rising; T4, high-falling) in a group of 20 young adult native speakers of Mandarin Chinese (nine women and 11 men, 19–35 years old). All participants had normal hearing (defined as having pure-tone thresholds of ≤ 25 dB HL for octaves from 250 to 4000 Hz and less than 15 dB difference between the two ears at each frequency) and normal or corrected-to-normal vision. All participants reported having no previous history of hearing problems or neurological disorders.

In this study, each participant completed a visual letter search task with high (target similar to distractors) or low (target dissimilar to distractors) perceptual load, with concurrent Mandarin tones presented in either a predictable (i.e., tones were presented in blocks within which each tone was presented repetitively) or variable (i.e., the tones were presented in a random order) context. Participants were instructed to ignore the sounds and focus their attention on the visual task. They were required to respond to the visual task as quickly and accurately as possible.

To accommodate the constraints of the visual tasks, FFRs to each Mandarin tone were averaged across only about 95 trials (of 96 possible trials) in each condition, the number of which is far fewer compared to prior work (several hundreds to thousands; e.g., Krishnan et al., 2004, 2005; Xie et al., 2017). The FFRs were recorded using a vertical montage of four electrodes (active, ~Fpz; reference, linked mastoids; ground, midforehead) at a sampling rate of 25 kHz. After data collection, the FFR data were bandpass filtered from 80 to 2500 Hz (12 dB/octave, zero phase shift) to predominantly highlight subcortical responses (Bidelman & Alain, 2015; Musacchia, Strait, & Kraus, 2008). The bandpass filtered FFR responses were segmented with a time window of −40 to 150 ms (0 ms corresponds to the onset of the Mandarin tones), and baseline corrected to the prestimulus region (i.e., −40 to 0 ms). The artifact-contaminated trials were rejected, defined as having amplitudes exceeding the range of ± 50 μV. The remaining artifact-free trials were averaged to produce one sample response for each tone in each condition. To improve computational efficiency, the averaged FFRs were down-sampled to 5 kHz.

A Decoding Model Approach

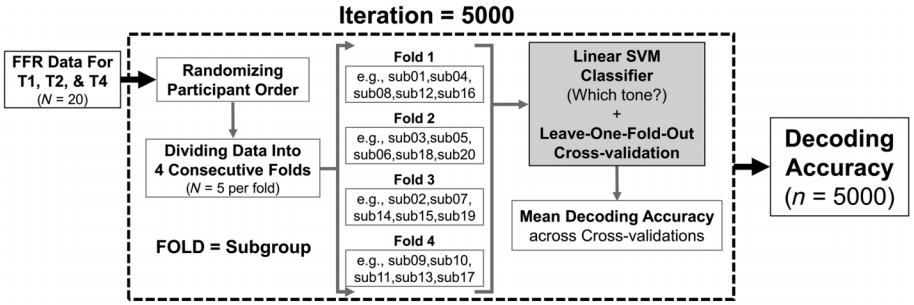
In this demonstration, we chose the classification approach to model speech-evoked FFRs. We used a linear SVM algorithm (model) to classify EEG signals (Garrett et al., 2003; Lotte et al., 2007; Subasi & Gursoy, 2010). There are several advantages associated with SVM: generally better classification performance than many other algorithms (Kotsiantis, Zaharakis, & Pintelas, 2007; Lehmann et al., 2007) and the ability to deal with data with a large number of features but relatively lower number of training examples (Kotsiantis et al., 2007). Note that the linear SVM in its standard form can only classify (or discriminate) data of two classes. To handle N (> 2) classes, the linear SVM was modified using a "one-against-one" approach. Specifically, the linear SVM constructed $N(N - 1) / 2$ classifiers, one for each pairwise combinations of the N classes. Each classifier assigned one vote for its preferred class, and the class with the highest votes across all the classifiers was taken as the classified class. In the following paragraphs, we summarize how SVM is implemented to analyze FFRs to speech stimuli and highlight relevant findings.

The analysis goal is to classify FFRs into the corresponding Mandarin tone categories (T1, T2, and T4). This analysis approach has been reported in Xie et al. (2018).

Step 1: Extraction of input and output features. In this analysis, the input (EEG) features are amplitude values of the FFRs from 10 to 110 ms (after stimulus onset). The output (speech stimulus) features are the Mandarin tone categories (T1, T2, and T4).

Step 2: Model selection and cross-validation. As mentioned above, the SVM was selected to map the FFR amplitude values to the tone categories. The cross-validation procedures are illustrated in Figure 1. A fourfold

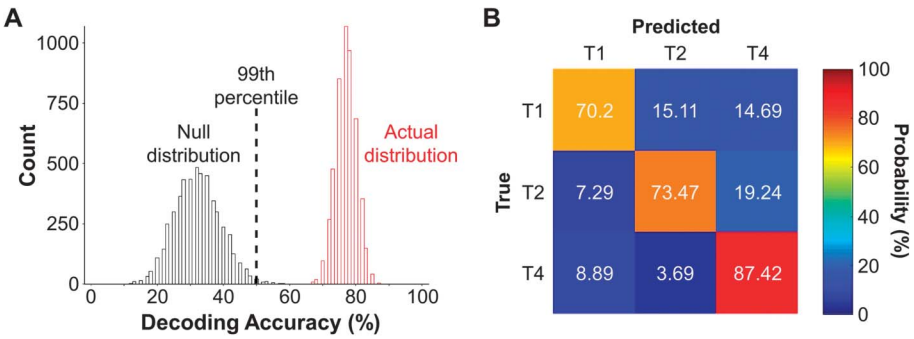
Figure 1. Procedures to implement a decoding model using linear support vector machines (SVMs) to analyze frequency-following responses (FFRs) to linguistically relevant pitch patterns (Mandarin tones): T1, high-level; T2, low-rising; T4, high-falling. This SVM analysis approach was reported in Xie et al. (2018). Leave-one-fold-out: The linear SVM classifier (model) is estimated with three of the four folds to classify FFRs into one of the three tone categories and is validated to see how well it can generalize to FFR data in the held-out fold. Decoding accuracy reflects the percentage that the SVM model correctly identified the tone categories across the four FFR folds.



cross-validation strategy with 5,000 iterations was adopted in the current example. In each iteration, the FFR data were first randomized based on the order of participants and then divided into four consecutive folds (i.e., four subgroups, each contains FFR data from five unique participants). Three of four folds were selected as the training data set for model estimation, and the hold-out fold was selected as the testing data set for model validation. This process was repeated four times until all the four folds had served as the testing data set. The predictive score of the fitted SVM model was quantified as decoding accuracy, which reflects the percentage that the model correctly identified the tone categories across the four FFR testing data set. The whole processes were iterated 5,000 times to derive 5,000 decoding accuracy values. The right (red) histogram in Figure 2A displays the decoding accuracies of a fitted SVM model in one experimental condition. A contingency table (also called *confusion matrix*) of the true versus predicted tone categories from the decoding analysis was also presented in Figure 2B.

We then evaluated the extent to which obtained decoding accuracy was statistically significant. We adopted the permutation test described above. First, an empirical null distribution of decoding accuracies was derived using the same procedures to derive the actual decoding accuracies, except that the SVM model was trained with the FFRs that were associated with the randomly assigned tone categories. The left (black) histogram in Figure 2A displays an example null distribution of decoding accuracies from the same condition as above. Then, the p value was estimated using the formula: $p = (a + 1) / (n + 1)$ (Phipson & Smyth, 2010), where a is the number of decoding accuracies from the null distribution that exceeds the median of the actual distribution of decoding accuracies and n is the total number of decoding accuracies from the null distribution. In the

Figure 2. An example of results for implementing a decoding model using linear support vector machines (SVMs) to analyze frequency-following responses (FFRs) to speech stimuli. The data were taken from Xie et al. (2018). (A) The left (black) histogram represents the null distribution of decoding accuracy (n = 5,000) derived from permutation tests. The right (red) histogram represents the distribution of the actual decoding accuracy (n = 5,000; actual distribution). The black vertical dashed line indicates the 99th percentile in the null distribution. If the median of the actual distribution is higher than the 99th percentile in the null distribution, it indicates that the actual decoding accuracy is statistically significant at an alpha value of .01. (B) A contingency table (also called confusion matrix) of the true versus predicted tone categories from the decoding analysis. Each column corresponds to the true tone category, and each row corresponds to the predicted tone category. The shade and the numbers of a given cell denote the occurrence of a predicted tone category in proportion to the total instances of a given true tone category (i.e., probability).



current example, the p value is 1.9996×10^{-4}, which is below .001, suggesting that the actual decoding accuracy was statistically significant.

Decoding Model Approach Versus Traditional Analysis Approaches

With the above decoding model approach, we obtained decoding accuracies for all four experimental conditions (2 [visual perceptual load: high or low] × 2 [auditory stimulus context: predictable or variable]). As reported in Xie et al. (2018), the decoding accuracies from the four conditions were all significantly above chance. For the predictable auditory context, decoding accuracies were significantly higher in the low visual load condition (Mdn = 76.67%, 99th percentile = 83.33%) relative to the high-load condition (Mdn = 60%, 99th percentile = 68.33%). However, for the variable auditory context, decoding accuracies were significantly lower in the low visual load condition (Mdn = 65%, 99th percentile = 73.33%) relative to the high-load condition (Mdn = 76.67%, 99th percentile = 85%).

As a comparison, we also adopted the traditional approaches to evaluate the fidelity of neural tracking of F0 contour in the Mandarin tones as reflected by the FFRs. As reported in Xie et al. (2018), we calculated two widely utilized metrics to assess the F0 tracking accuracy: stimulus-to-response correlation, which quantifies the similarity of F0 between the stimulus and the evoked FFR, and peak autocorrelation, which quantifies the degree of periodicity in the FFR data. For the stimulus-to-response correlation metric, we found that, for the predictable auditory context, the mean stimulus-to-response correlation was significantly higher in the low-load condition relative to the high-load condition ($p < .05$), but for the variable auditory context, the mean stimulus-to-response correlation was not significantly different between two load conditions ($p > .05$). For the peak autocorrection metric, we did not find any significant effect of visual perceptual load, auditory stimulus context or their interaction (all ps > .05). Thus, results from the decoding model approach and the traditional analyses are partially consistent.

Application 2: EEG Responses to Natural Continuous Speech

Advantages of Natural Continuous Stimuli Over Simplified, Controlled Stimuli

The use of simplified, controlled stimuli has a long tradition of utility and continues to dominate speech and language function research. These stimuli are typically a limited set of isolated speech sounds that differ on a limited number of dimensions (Martin et al., 2008; Skoe & Kraus, 2010; Tremblay et al., 2003). There is currently a surge of interest in the study of speech and language processing using natural stimuli in place of simplified, controlled stimuli. These studies have mainly focused on the cortical processing of speech signals (Broderick et al., 2018; Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Di Liberto, Peter, et al., 2018; Fuglsang, Dau, & Hjortkjær, 2017; Khalighinejad et al., 2017; Kong, Mullangi, & Ding, 2014; Kong, Somarowthu, & Ding, 2015; Mirkovic, Debener, Jaeger, & De Vos, 2015; J. O'Sullivan et al., 2014; Power, Colling, Mead, Barnes, & Goswami, 2016; Power, Foxe, Forde, Reilly, & Lalor, 2012; Puschmann et al., 2017), and some recent work has extended the paradigms to examine processing at auditory subcortical levels (Forte, Etard, & Reichenbach, 2017; Maddox & Lee, 2018).

Neural responses to natural speech stimuli have often been proposed to be an objective neural measure of speech intelligibility (e.g., Di Liberto, Crosse, et al., 2018; Ding & Simon, 2014; Vanthornhout, Decruy, Wouters, Simon, & Francart, 2018). For example, Vanthornhout et al. (2018) showed that EEG metrics reflecting the neural processing of speech envelope in natural continuous speech correlated with a behavioral measure of speech intelligibility (i.e., speech reception threshold, the signal-to-noise ratio yielding 50% intelligibility) on the same speech stimuli. Furthermore, recent studies have begun to probe what specific psycholinguistic processes are encoded by the neurophysiological responses to natural continuous speech (e.g., Brodbeck et al., 2018; Broderick et al., 2018; Di Liberto et al., 2015). For example, Broderick et al. (2018) estimated an encoding model that maps semantic dissimilarity features from continuous natural speech to the corresponding EEG responses. The estimated model weights shared characteristics (e.g., time course and topographic distribution) with the N400, a component of the event-related potential (ERP) that is thought to reflect semantic processing (Kutas & Federmeier, 2011; Lau, Phillips, & Poeppel, 2008). Furthermore, the model weights significantly correlated with the N400 amplitude. Hence, the neural metric derived from the encoding model may reflect semantic processing of natural speech.

According to Hamilton and Huth (2018), natural stimuli offer at least three advantages over simplified, controlled stimuli from traditional experimental designs. The first advantage is generalizability. Traditional experimental designs typically use a limited set of isolated phonemes/syllables (e.g., Tremblay et al., 2003; Xie et al., 2018; Yi et al., 2017), words (e.g., Galbraith, Arbagey, Branski, Comerci, & Rector, 1995; Marinkovic et al., 2003), or sentences (e.g., Aiken & Picton, 2008; Friederici, Pfeifer, & Hahne, 1993). These stimuli, as well as the tasks that imposed these stimuli (e.g., judging whether a sentence was syntactically correct), are usually uncommon in real-life settings. Hence, in contrast to traditional designs, research with natural stimuli may better generalize to speech and language processing in ethological settings. Consistent with this argument, for example, Bonte, Parviainen, Hytönen, and Salmelin (2006) suggests that neural responses to speech units (e.g., syllables) embedded in continuous speech are different when they are presented in isolation (even though the stimuli are identical).

The second advantage is the ability to directly compare effect sizes across studies on different levels of speech and language features. This can be achieved by instantiating each effect as a model to predict the neurophysiological responses evoked by the same, natural stimulus data set.

The fraction of variance in the neurophysiological responses that can be explained by each effect (model) may reflect the importance of the corresponding effect. This may be difficult to implement in traditional experiments given the variability in methodologies across studies (e.g., different types of stimuli or measures/metrics).

The third advantage is experimental efficiency. Multiple hypotheses regarding different levels of speech and language processing can be tested from a single experiment with natural stimuli. This can be achieved through relating neurophysiological responses with different features in natural continuous speech such as acoustic features (Crosse et al., 2015; Di Liberto et al., 2015; Fuglsang et al., 2017; Kong et al., 2014, 2015; Mirkovic et al., 2015; J. O'Sullivan et al., 2014; Power et al., 2016, 2012; Puschmann et al., 2017), phonemes or phonetic features (Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Di Liberto, Peter, et al., 2018; Khalighinejad et al., 2017), and lexical–semantic features (Brodbeck et al., 2018; Broderick et al., 2018). Traditional experiments, however, are usually unable to address multiple hypotheses simultaneously because they are typically designed to test a specific hypothesis of interest.

A Demonstration of ML-Based Approaches to Analyze EEG Responses to Natural Continuous Speech

ML-Based Approaches to Analyze Neurophysiological Responses to Natural Continuous Speech: A Brief Summary

Decoding models have been used to quantify the relationship between speech stimulus features in continuous speech and neurophysiological responses. Within a reconstruction framework, for instance, an ongoing line of work focuses on reconstructing the temporal envelope of the attended speaker in a multispeaker environment based on EEG responses (e.g., Fuglsang et al., 2017; Mirkovic et al., 2015; J. O'Sullivan et al., 2014; Power et al., 2012; Puschmann et al., 2017), with an eventual clinical goal of developing neurosteered hearing prostheses that can enhance the speaker of interest based on a listener's attention (Das, Van Eyndhoven, Francart, & Bertrand, 2016; J. O'Sullivan, Chen, et al., 2017). Within a classification framework, for example, a recent study extracted EEG responses time-locked to phonemes (phoneme-related potentials [PRPs]) from continuous speech stimuli, as in traditional ERP studies. They found that the PRPs can be reliably classified into phonological categories of plosive, fricative, nasal, and vowel (Khalighinejad et al., 2017).

Researchers have also quantified the forward mapping of a speech feature in continuous speech to neurophysiological responses with encoding models. The temporal response function (TRF) is one type of encoding model often used. The TRF quantifies the transformation of a stimulus representation to continuous neural responses by the brain based on linear regression (Crosse, Di Liberto, Bednar, & Lalor, 2016; Di Liberto & Lalor, 2017; Di Liberto et al., 2015). The neural response, in relation to a stimulus event, does not emerge until a certain time lag (e.g., several tens of milliseconds for cortical responses) and lasts for a certain period (e.g., several hundred milliseconds). Therefore, the TRF is defined as a series of regression weights across a certain set of time lags between stimulus and response (e.g., 0–250 ms in Di Liberto et al., 2015). The value of the TRF at a certain time lag (e.g., 200 ms) indexes the effect of the speech feature on the neural response at that lag (e.g., 200 ms later; Crosse et al., 2016). Like traditional analyses on the ERP, the resulting TRF has often been evaluated in terms of its temporal and spatial (e.g., topographic distribution) dynamics (e.g., Broderick et al., 2018).

Using the TRF, researchers have revealed the neural processing of acoustic features and speech-specific features such as phonemes and phonetic features (Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Di Liberto, Peter, et al., 2018), as well as semantic features (Broderick et al., 2018), as reflected by the EEG responses. Importantly, the TRF is highly related to the ERP components from traditional approaches (Broderick et al., 2018; Maddox & Lee, 2018), suggesting the validity of the TRF in capturing the relationship between speech stimulus features and brain activity. For example, Maddox and Lee (2018) utilized the TRF to model auditory brainstem processing of continuous natural speech using EEG. They found that the TRF demonstrates a high level of morphological similarity (median correlation coefficient of .82) to the standard click-evoked auditory brainstem responses.

Here, we illustrate the utility of SVM as a decoding model, similar to that in Application 1, to classify phonological categories (plosive, fricative, nasal, and vowel) from the PRPs evoked by natural continuous speech and the utilization of TRF as an encoding model to predict EEG responses based on acoustic features (temporal envelope) in natural continuous speech.

Sample Data Set

The EEG responses to continuous speech data come from an unpublished data set in our lab. This data set consists of 62-electrode EEG segments elicited to 15 tracks of story segments (each ~60 s in length) in a group of 16 young adult native speakers of American English (11 women and five men, 18–23 years old). All participants had normal hearing (defined as having pure-tone and bone-conduction thresholds of ≤ 20 dB HL for octaves from 250 to 8000 Hz) and normal or corrected-to-normal vision. To minimize the effect of music training on the EEG response to the speech stimuli (e.g., Bidelman & Alain, 2015; Coffey, Mogilever, & Zatorre, 2017), we recruited participants with either no history or no significant formal music training (≤ 4 years of continuous training, not currently practicing). All participants reported no history of psychological or neurological disorders, no use of neuropsychiatric medication, and no prior history of a hearing deficit.

The story segments were selected from a classic work of fiction, Alice's Adventures in Wonderland (http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-5). The audiobook was narrated in English by an adult male speaker of American English and sampled at 22.05 kHz. In

the study, participants listened to the auditory story segments and ignored concurrent visuospatial stimuli (blue squares at different loci on the screen). To encourage participants to focus on the auditory stimuli, at the end of each segment, they were asked two multiple-choice questions to probe comprehension about the story segments. EEG data were recorded from 64 electrodes (online referenced to TP9; ground at the Fpz electrode site) that are organized in accordance with the extended 10–20 system (Oostenveld & Praamstra, 2001). The EEG data were offline referenced to the average of the electrodes TP9 and TP10, bandpass filtered from 1 to 15 Hz (e.g., Di Liberto et al., 2015), and segmented into epochs that were time-locked to the onset of the auditory story segments. The duration of the epochs matched that of the corresponding story segments. To improve computational efficiency, the segmented EEG data were down-sampled to 128 Hz. An independent component analysis using the restricted Infomax algorithm (Bell & Sejnowski, 1995) was applied to the segmented EEG data to remove ocular artifacts. Finally, the EEG data from each electrode were normalized to ensure zero mean and unit variance.

A Decoding Model Approach

The analysis goal is to classify PRPs from continuous speech stimuli into the corresponding phonological categories (plosive, fricative, nasal, and vowel) based on Khalighinejad et al. (2017). The procedures are illustrated in Figures 3A–3C.

Step 1: Extraction of input and output features. In this analysis, the input (EEG) features are amplitude values of the PRPs. The output (speech stimulus) features are the phonological categories (plosive, fricative, nasal, and vowel). The PRPs are extracted as follows: As illustrated in Figure 3A, to obtain a time-locked EEG response to each phoneme, the EEG data were segmented and aligned to phoneme onset with a predefined time window (e.g., 0–600 ms). The computation of phonemes and onset information can be achieved via a combination of automatic tools and manual correction (Di Liberto et al., 2015; Mesgarani, Cheung, Johnson, & Chang, 2014). The PRPs are calculated by averaging all the instances of the segmented EEG responses to each phoneme. Examples of the PRPs are displayed in Figure 3B. Khalighinejad et al. (2017) suggests that PRPs from the frontocentral electrodes provide the best distinction of the phonological categories. Therefore, one analysis option is to focus on PRPs from this subset of electrodes. In the current example, we focus on all the 62 electrodes for illustration purposes.

Step 2: Model selection and cross-validation. As mentioned above, the SVM was selected to map the PRP amplitude values to the phonological categories. The cross-validation procedures are illustrated in Figure 3C. The PRP data from 15 of the 16 participants were used as the training data set for model estimation and the hold-out participant as the testing data set for model validation. The predictive score of the fitted SVM model was quantified as decoding accuracy, which reflects the percentage that the model correctly identified the phonological category labels of the PRP data in the testing data set. Figure 3D displays the results from an example participant. This plot shows the decoding accuracies across all the 62 electrodes.

Note that, in natural speech, phonemes are not evenly distributed across phonological categories. This may cause the SVM model to bias toward the phonological categories with the highest number of phonemes. Techniques to handle such issue have been extensively explored in the ML literature (e.g., Batuwita & Palade, 2013). In our example, we chose to balance the number for each phonological category using the *cosmo_balance_partitions* function from the *CoSMoMVPA* toolbox (Oosterhof, Connolly, & Haxby, 2016) in MATLAB (The MathWorks). This function generated multiple pairs of training and testing data set, with PRP samples from each phonological category occurring at least once across the training data set. The decoding accuracy was calculated by averaging the accuracies across all the pairs.

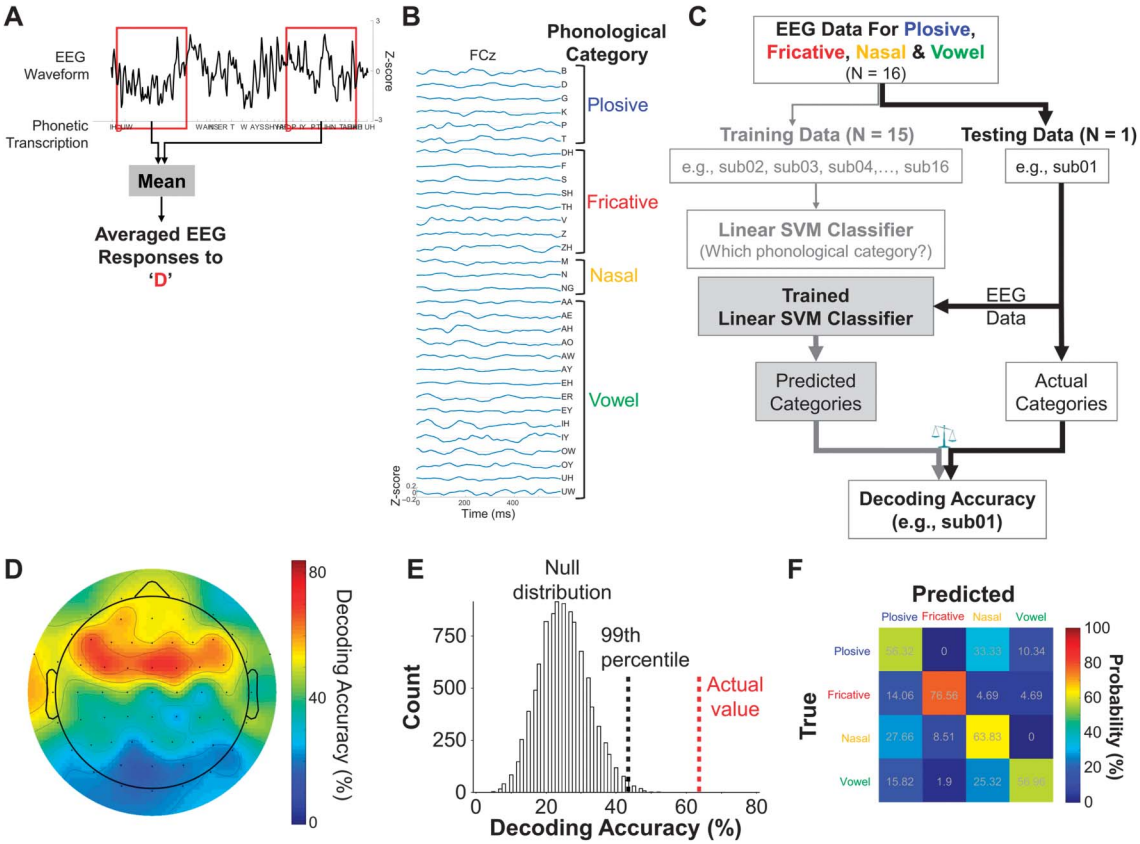
Furthermore, we adopted a permutation test to determine whether the obtained decoding accuracy was statistically significantly. We randomly assigned the labels of the phonological category in the training data set and estimated the SVM model with the shuffled PRP data and then predicted the phonological category labels of testing data set. This label shuffling and model cross-validation was iterated 10 times for each participant ($n = 16$) and each electrode ($n = 62$), and a null distribution of the decoding accuracy ($n = 10 \times 16 \times 62 = 9{,}920$) was obtained. The left (black) histogram in Figure 3E shows an example null distribution of the decoding accuracies. Then, the actual decoding accuracy averaged across seven frontocentral electrodes (FC5, FC3, FC1, FCz, FC2, FC4, FC6; indicated by the red vertical dashed line in Figure 3E) was tested against the null distribution. The p value was calculated using the formula: $p = (a + 1) / (n + 1)$ (Phipson & Smyth, 2010), where a is the number of decoding accuracies from the null distribution that exceeds the actual decoding accuracy and n is the total number of decoding accuracies from the null distribution (i.e., 9,920). In the example here, the p value is 1.008×10^{-4}, which is below .001, suggesting that the actual decoding accuracy was statistically significant.

An Encoding Model Approach

In this example, we present a TRF model approach based on regularized linear regression (Crosse et al., 2016) to analyze EEG responses to continuous speech. In the following sections, we summarize the main procedures and highlight relevant findings from an example participant.

Step 1: Extraction of input and output features. In this analysis, the input (speech stimulus) features are amplitude envelope, the most studied speech feature in the literature focusing on continuous speech (Ding & Simon, 2014). However, other speech features have also been studied including phoneme/phonetic features (Di Liberto, Crosse, et al., 2018; Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Di Liberto, Peter, et al., 2018) and semantic features (Broderick et al., 2018). The use of different speech features is assumed to reflect hierarchical levels of speech processing (Brodbeck

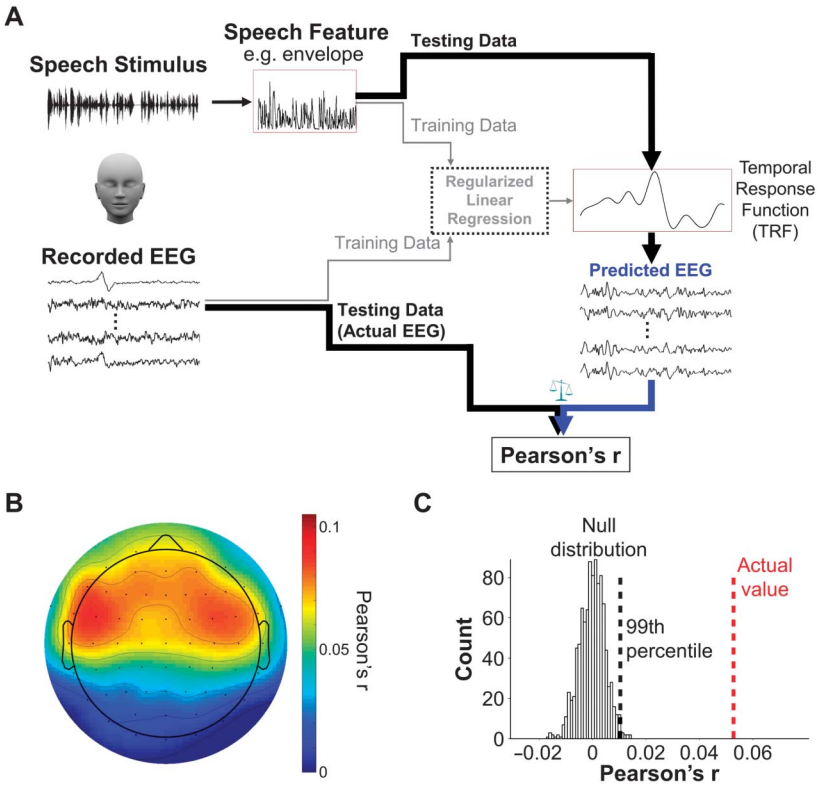
Figure 3. (A–C) Procedures to implement a decoding model using linear support vector machines (SVMs) to analyze EEG responses to continuous speech stimuli. (a) Extraction of EEG responses time-locked to phoneme onset (PRPs) with a predefined time window (e.g., 0–600 ms). (B) Examples of the PRPs at electrode FCz from one participant (from an unpublished data set in our lab). The corresponding phonemes are grouped into phonological categories plosive, fricative, nasal, and vowel. (C) Procedures to classify PRPs into one of the four phonological categories using linear SVM classifier (model). The SVM model is estimated with PRP data from 15 of the 16 participants and is validated to see how well it can generalize to PRP data in the held-out participant. Decoding accuracy reflects the percentage that the SVM model correctly identified the phonological categories in the held-out participant. (D) Results from one example participant (from an unpublished data set in our lab). This plot shows the topographic distribution of decoding accuracies across 62 electrodes. (E) The left (black) histogram represents the null distribution of decoding accuracies (n = 9,920) derived from permutation tests. The red vertical dashed line represents the actual decoding accuracy averaged across seven frontocentral electrodes (FC5, FC3, FC1, FCz, FC2, FC4, FC6). The black vertical dashed line indicates the 99th percentile in the null distribution. If the actual decoding accuracy is higher than the 99th percentile in the null distribution, it indicates that the actual decoding accuracy is statistically significant at an alpha value of .01. (F) A contingency table (also called confusion matrix) of the true versus predicted phonological categories from the decoding analysis. Each column corresponds to the true phonological category, and each row corresponds to the predicted phonological category. The shade and the numbers of a given cell denote the occurrence of a predicted phonological category in proportion to the total instances of a given true phonological category (i.e., probability).



et al., 2018; Di Liberto et al., 2015). The output (EEG) features are time-varying amplitude in a frequency band from 1 to 15 Hz. Previous studies have demonstrated that a subset of frontocentral electrodes shows the highest model performance (e.g., Di Liberto & Lalor, 2017; Di Liberto et al., 2015). Hence, we may choose to focus on EEG from this subset of electrodes. In this example, we will focus on all the 62 electrodes for illustration purposes.

Step 2: Model selection and cross-validation. The procedures are illustrated in Figure 4A. The analysis was implemented using the multivariate TRF in MATLAB (The MathWorks) toolbox (Crosse et al., 2016). As mentioned above, the TRF was chosen to map the speech stimulus amplitude to the EEG features at each electrode (i.e., predicting EEG features from the speech features). A leave-one-trial-out cross-validation strategy was adopted in the current example. Data from 14 of the 15 trials were chosen as the training data set for model estimation and the hold-out trial as the testing data set for model validation. This process was repeated until all the trials had served as the testing data set. The predictive score of the fitted model was quantified as EEG prediction accuracy, which reflects the Pearson correlation coefficient between the predicted and actual EEG features across all the 15 trials. Higher Pearson's r value is taken as reflective of better neural representation of the corresponding speech feature (Di Liberto et al., 2015). Figure 4B displays the results from an example participant. This plot shows the Pearson's r values across all the 62 electrodes.

Figure 4. (A) Procedures to implement an encoding model approach (temporal response function, TRF) to analyze EEG responses to continuous speech stimuli. The TRF model is estimated with data from 14 of the 15 trials to predict EEG features from the speech stimulus features and validated to see how well it can predict the EEG features of the held-out trial. Pearson's *r* between the predicted and actual EEG features reflects how well the TRF model can predict EEG features from speech stimulus features. (B) Results from one example participant (from an unpublished data set in our lab) when temporal envelope was used as the speech feature. This plot shows the topographic distribution of Pearson's *r* values across 62 electrodes. (C) The left (black) histogram represents the null distribution of Pearson's *r* values (n = 1,000) derived from permutation tests. The red vertical dashed line represents the actual Pearson's *r* value averaged across all 62 electrodes. The black vertical dashed line indicates the 99th percentile in the null distribution. If the actual Pearson's *r* is higher than the 99th percentile in the null distribution, it indicates that the actual Pearson's *r* is statistically significant at an alpha value of .01.



Furthermore, we adopted a permutation test to determine whether the obtained EEG prediction accuracy (Pearson's *r* value) was statistically significantly. We shuffled the timing of the speech stimulus feature values and modeled the relationship between the shuffled speech stimulus feature and the actual EEG responses (not shuffled) at each electrode. This shuffling and modeling analyses were iterated 1,000 times, and a null distribution of EEG prediction accuracies was obtained. The left (black) histogram in Figure 4C shows an example null distribution of EEG prediction accuracies that were averaged across all the electrodes. Then, the actual EEG prediction accuracy (indicated by the red vertical dashed line in Figure 4C) that was averaged across all the electrodes was tested against the null distribution. The p value was calculated using the formula: $p = (a + 1) / (n + 1)$ (Phipson & Smyth, 2010), where a is the number of EEG prediction accuracies from the null distribution that exceeds the actual EEG prediction accuracy and n is the total number of EEG prediction accuracies from the null distribution (i.e., 1,000). In the example here, the p value is 9.99×10^{-4}, which is below .001, suggesting that the actual EEG prediction accuracy was statistically significant.

Discussion

Summary of the Review

In the current review, we presented a step-by-step guide of two applications of ML-based approaches to analyze speech-evoked neurophysiological responses with empirical results. In Application 1, we showed the utility of a decoding model using the SVM to classify speech-evoked neurophysiological responses from traditional designs with simplified, controlled stimuli (i.e., FFRs to Mandarin lexical tones; Xie et al., 2018). In Application 2, we demonstrated the utility of a decoding model (SVM) to classify phonological categories (plosive, fricative, nasal, and vowel) from PRPs evoked by natural continuous speech and an encoding model (i.e., TRF) in predicting EEG responses based on acoustic features (temporal envelope) in natural continuous speech.

Clinical Utility of ML-Based Approaches: An Example With Developmental Dyslexia

A major focus of speech-evoked neurophysiological responses is to characterize aberrant neural processing in a range of communication disorders. ML-based approaches paired with natural stimuli may be a potent tool to identify and quantify these aberrant neural processes. For example, in developmental dyslexia, a neurological disorder affecting reading and spelling (Démonet, Taylor, & Chaix, 2004), difficulties have been evidenced in phonemic and phonological processing (Di Liberto, Peter, et al., 2018; Ramus et al., 2003). Di Liberto, Peter, et al. (2018) demonstrated the utility of ML-based approaches, specifically encoding models similar to the one described in the current review, to assess phonological processing deficit in developmental dyslexia by characterizing EEG responses to natural continuous speech stimuli. Consistent with prior work, they found that children with developmental dyslexia, relative to age-matched and reading level–matched typically developing children, exhibited reduced neural encoding of phonetic features in natural continuous speech stimuli. The robustness of phonetic feature encoding was related to psychometric measures of phonological skills.

Several aspects of their methodologies support the clinical utility of ML-based approaches. First, EEG responses were collected to an audio story of only 9 min, not requiring participants to sit for an extended period of time. Second, EEG responses were recorded with no explicit responses to the speech stimuli required from the participants, eliminating task demands from participants. Third, the study adopted a cross-group strategy for model cross-validation, such that the encoding models were trained with data from a subset of typically developing children and tested on data from the remaining typically developing children or that from children with dyslexia.

Advantages of ML-Based Approaches to Study Speech and Language Processing

The ML-based approaches facilitate the investigation of speech and language processing with ecologically valid paradigm (e.g., continuous natural speech) while moving away from traditional designs with simplified, controlled speech sounds (Martin et al., 2008; Skoe & Kraus, 2010; Tremblay et al., 2003). As discussed earlier, the use of natural stimuli offers at least three advantages (Hamilton & Huth, 2018). The first advantage is being more generalizable to speech and language processing in ethological settings. The second advantage is the ability to directly compare effect sizes across studies on different levels of speech and language features.

The third advantage is that multiple hypotheses surrounding speech and language processing can be tested within a single experiment using continuous, natural stimuli. For example, our second application demonstrated that the encoding of acoustic features (e.g., temporal envelope) and the encoding of phonological categories could be studied from the same EEG recording to natural continuous speech stimuli. The advantage of experimental efficiency with natural stimuli can also be achieved with a shorter assessment time. For example, Di Liberto and Lalor (2017) demonstrated that, to obtain reliable estimate of phoneme-level processing using continuous speech, the needed EEG data can be substantially reduced from at least 30 min to only 10 min of recording by adopting a cross-group strategy for model cross-validation (i.e., the model is estimated using data from a proportion of participants in a group and validated using data from the held-out participants). In the above-mentioned clinical example, Di Liberto, Peter, et al. (2018) adopted a similar method for model cross-validation and used EEG data from only 9 min of recording.

Indeed, our first application suggests that ML-based approaches may also improve experimental efficiency of traditional experiments with simplified, controlled stimuli. In this application, a decoding model can classify FFRs to Mandarin lexical tones with averages of less than 100 trials. This contrasts with traditional analysis approach guidelines that have called for averaged FFRs over thousands of trials (Skoe & Kraus, 2010). The number of FFR trials needed for the ML-based approaches is within a range for traditional approaches to study auditory cortical processing with EEG. That means that the ML-based approaches open up the opportunity to noninvasively study multiple levels of speech and language processing (i.e., subcortical and cortical auditory processing) truly simultaneously. This would further improve the experimental efficiency of traditional experiments. Indeed, Xie et al. (2018) not only analyzed the FFRs as the example presented in this review to examine early encoding of speech signals but also examined cortical responses to the speech signals using the same EEG recordings.

ML-Based Approach as a Potent Tool in the Clinical Diagnosis of Communication Disorders

Communication disorders are often multifactorial and broad based (Baum, Stevenson, & Wallace, 2015; Bishop & Leonard, 2014; Dronkers & Baldo, 2010; Goswami, 2015), affecting different subprocesses ranging from acoustic encoding to semantic/syntactic processes. Often, these deficits can span different hierarchical levels of speech and language processing. For example, in developmental dyslexia, difficulties have been evidenced in acoustic processing (Goswami et al., 2002; Power et al., 2016) and phonemic and phonological processing (Di Liberto, Peter, et al., 2018; Ramus et al., 2003). Current clinical diagnostic protocols require the implementation of many different types of speech and language standardized assessments (e.g., tests of phonological processing, receptive vocabulary, auditory comprehension) to gain a full picture of an individual's communication ability (Bishop, 2004; Johnson & Myers, 2007; Tomblin, Records, & Zhang, 1996). This poses a challenge for difficult-to-test populations such as young children and individuals with poor attentional abilities, as diagnostic sessions tend to last for several hours and usually multiple days.

ML-based approaches combined with EEG methodology may be a potent tool to complement standard behavioral assessments of speech and language ability and provide a critical link between clinical observations of communication deficits with their underlying neurobiology. As discussed above, ML-based approaches allow the examination of hierarchical levels of speech and language processing from a single recording and with shorter recording time (e.g., only EEG responses to only 9 min of natural continuous speech in Di Liberto, Peter, et al., 2018). Such efficiency from ML-based approaches may greatly cut down the time needed for assessments of speech and language ability. A reduction in the assessment time and efforts could benefit test populations such as young children and individuals with poor attentional abilities. Moreover, as highlighted in the clinical example, the EEG recordings do not require explicit responses from participants. This may provide a diagnostic tool to assess hard-to-test populations (e.g., nonverbal children) that may be limited in the ability to initiate (e.g., verbal or motor) responses to the test materials. Furthermore, the cross-group strategy of model cross-validation in abovementioned clinical example suggests that, like behavioral assessments, we may build a normative database and utilize the normative data to predict whether a new client falls into the "typical" range or is at risk for communication deficits.

Refining ML-Based Approaches

ML-based approaches are not limited to those highlighted in the current review and are constantly evolving. Several refinements may be applied to the existing ML-based approaches to better characterize speech-evoked neurophysiological responses. First, more advanced methods may be used for feature selection from the EEG data and improve the interpretability of the EEG features. For example, Yi et al. (2017) described an application of ML-based approaches to decode FFRs to vowel stimuli. In that study, they first constructed a spectral feature space from a database of vowel stimuli using principle component analysis. The vowels were the same as the stimuli used for FFR recordings but were produced by different speakers. The spectral feature space contains 12 spectral components that explain 80% of the variance of the vowel database. They then projected single-trial FFRs onto the spectral feature space and derive 12 spectral features that were used for the decoding analysis. The decoding performance of single-trial FFRs were significant above chance even when only 50 trials per stimulus were used for training the decoding model. Interestingly, the spectral feature most relevant to the decoding analysis contains three extrema corresponding with the first three formants of the vowel stimuli.

Second, other more sophisticated ML-based approaches may be adopted to capture the relationship between the evoking speech stimuli and evoked neurophysiological responses and yield better performance. For example, de Cheveigné et al. (2018) compared models based on canonical correlation analysis and an encoding model similar to that implemented in the current review in the ability to quantify EEG responses to continuous speech. They found that the canonical correlation analysis models yielded significantly higher correlation values between stimulus and EEG features relative to the simple encoding model.

Finally, in the context of decoding analysis, we may include EEG data from multiple electrodes to yield higher decoding performance relative to single electrodes. The benefit to decoding performance from a larger number of electrodes has been shown in various applications of decoding models to EEG data, such as decoding the attended speech stream from a mixture of two simultaneous talkers (Mirkovic et al., 2015) and decoding motor imagery hand movements (Zich, De Vos, Kranczioch, & Debener, 2015). However, decoding performance may asymptote beyond a certain number of electrodes even if more electrodes are included, because only data from a subset of electrodes are informative of the processes of interest while other electrodes seem redundant (e.g., Mirkovic et al., 2015; Zich et al., 2015).

The informativeness of different EEG electrodes is related to the underlying neural sources of the processes of interest. To more directly utilize source information to decode speech-evoked neurophysiological responses, we may collect these responses using MEG, which provides more accurate source localization of the speech-evoked neurophysiological responses than EEG. Consistently, a recent study collected neural responses to continuous speech stimuli using MEG and suggest that the processing of acoustic, lexical, and semantic features in continuous speech involves different brain areas with different temporal dynamics (Brodbeck et al., 2018).

Limitations of ML-Based Approaches

Although the ML-based approaches hold many advantages, there are still several limitations. First, it may be difficult to interpret the stimulus features that the models used to relate to neural activity. Recent work has made an effort to fill this gap by implementing models where features can be preselected. For example, the encoding models highlighted in this review (i.e., TRF) require for the experimenter to preselect the features of interest as a critical component of the model (e.g., temporal envelope vs. phonetic features; Di Liberto et al., 2015). As discussed above, our previous work on developing ML-based approaches to classify FFRs to vowel stimuli has further shown that interpretable features (i.e., spectral cues) can be derived from ML-based models (Yi et al., 2017). The other main limitation is that the ML-based approaches are computationally intensive for both the machine and the human operator. To implement the ML-based approaches, a certain degree of expertise on programming and ML is required, which many practicing clinicians may not possess. In turn, this may limit the clinical translation of such analysis approaches in the context of providing supplemental information for diagnostic evaluations.

Conclusions

In conclusion, current ML-based approaches are useful complements to traditional approaches to analyze neurophysiological responses to speech signals. These analysis approaches allow for a more efficient examination of different aspects of natural speech and language processing using ecologically valid paradigms in both typical and clinical populations.

Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grants R01DC013315 and R01DC015504 to B. C. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing, 29*(2), 139–157.

Anderson, S., Parbery-Clark, A., White-Schwoch, T., Drehobl, S., & Kraus, N. (2013). Effects of hearing loss on the subcortical representation of speech cues. *The Journal of the Acoustical Society of America, 133*(5), 3030–3038.

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). Reversal of age-related neural timing delays with training. *Proceedings of the National Academy of Sciences, 110*(11), 4357–4362.

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. In H. He, & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications* (pp. 83–100). Hoboken, NJ: Wiley.

Baum, S. H., Stevenson, R. A., & Wallace, M. T. (2015). Behavioral, perceptual, and neural alterations in sensory and multisensory function in autism spectrum disorder. *Progress in Neurobiology, 134,* 140–160.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.

Bidelman, G. M. (2015). Multichannel recordings of the human brainstem frequency-following response: Scalp topography, source generators, and distinctions from the transient ABR. *Hearing Research, 323,* 68–80.

Bidelman, G. M., & Alain, C. (2015). Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception. *Journal of Neuroscience, 35*(3), 1240–1249.

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience, 23*(2), 425–434.

Bishop, D. V. M. (2004). Specific language impairment: Diagnostic dilemmas. In *Classification of developmental language disorders: Theoretical issues and clinical implications* (pp. 309–326). Mahwah, NJ: Erlbaum.

Bishop, D. V. M., & Leonard, L. (2014). *Speech and language impairments in children: Causes, characteristics, intervention and outcome*. Psychology Press.

Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex, 16*(1), 115–123. https://doi.org/10.1093/cercor/bhi091

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage, 172,* 162–174. https://doi.org/10.1016/j.neuroimage.2018.01.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology, 28*(5), 803.e3–809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Chandrasekaran, B., & Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology, 47*(2), 236–246.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research, 352,* 49–69.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience, 35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (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., Van Eyndhoven, S., Francart, T., & Bertrand, A. (2016). *Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses*. 2016 IEEE 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 77–80.

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

Démonet, J.-F., Taylor, M. J., & Chaix, Y. (2004). Developmental dyslexia. *The Lancet, 363*(9419), 1451–1460.

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. *ENeuro, 5*(2), ENEURO-0084.

Di Liberto, G. M., & Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hearing Research, 348,* 70–77.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology, 25*(19), 2457–2465.

Di Liberto, G. M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., & Lalor, E. C. (2018). Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. *NeuroImage, 175,* 70–79.

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, 109*(29), 11854–11859.

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology, 107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience, 8,* 311.

Dronkers, N. F., & Baldo, J. V. (2010). Language: Aphasia. In *Encyclopedia of neuroscience* (pp. 343–348). Cambridge, MA: Academic Press.

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *ELife, 6,* e27203.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research, 1*(3), 183–192.

Fuglsang, S. A., Dau, T., & Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage, 156*(April), 435–444. https://doi.org/10.1016/j.neuroimage.2017.04.026

Galbraith, G. C., Arbagey, P. W., Branski, R., Comerci, N., & Rector, P. M. (1995). Intelligible speech encoded in the human brain stem frequency-following response. *Neuroreport, 6*(17), 2363–2367.

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

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party.". *Neuron, 77*(5), 980–991.

Good, P. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York, NY: Springer Science & Business Media.

Goswami, U. (2015). Sensory theories of developmental dyslexia: Three challenges for research. *Nature Reviews Neuroscience, 16*(1), 43–54.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences, 99*(16), 10911–10916.

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience,* 1–10.

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience, 11,* 61.

Johnson, C. P., & Myers, S. M. (2007). Identification and evaluation of children with autism spectrum disorders. *Pediatrics, 120*(5), 1183–1215.

Khalighinejad, B., Cruzatto da Silva, G., & Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *The Journal of Neuroscience, 37*(8), 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

Kong, Y., Mullangi, A., & Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research, 316,* 73–81.

Kong, Y., Somarowthu, A., & Ding, N. (2015). Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. *Journal of the Association for Research in Otolaryngology, 16*(6), 783–796. https://doi.org/10.1007/s10162-015-0540-x

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering, 160,* 3–24 [Informatica, 31, 249–268].

Krishnan, A. (2002). Human frequency-following responses: Representation of steady-state synthetic vowels. *Hearing Research, 166*(1–2), 192–201.

Krishnan, A., Xu, Y., Gandour, J. T., & Cariani, P. A. (2004). Human frequency-following response: Representation of pitch contours in Chinese tones. *Hearing Research, 189*(1–2), 1–12.

Krishnan, A., Xu, Y., Gandour, J. T., & Cariani, P. A. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research, 25*(1), 161–168. https://doi.org/10.1016/j.cogbrainres.2005.05.004

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62,* 621–647.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience, 9*(12), 920–933.

Lehmann, C., Koenig, T., Jelic, V., Prichep, L., John, R. E., Wahlund, L.-O., … Dierks, T. (2007). Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *Journal of Neuroscience Methods, 161*(2), 342–350.

Llanos, F., Xie, Z., & Chandrasekaran, B. (2017). Hidden Markov modeling of frequency-following responses to Mandarin lexical tones. *Journal of Neuroscience Methods, 291,* 101–112.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering, 4*(2), R1–R13.

Maddox, R. K., & Lee, A. K. C. (2018). Auditory brainstem responses to continuous natural speech in human listeners. *ENeuro, 5*(1), ENEURO-0441.

Marinkovic, K., Dhond, R. P., Dale, A. M., Glessner, M., Carr, V., & Halgren, E. (2003). Spatiotemporal dynamics of modality-specific and supramodal word processing. *Neuron, 38*(3), 487–497.

Marsh, J. T., Worden, F. G., & Smith, J. C. (1970). Auditory frequency-following response: Neural or artifact. *Science, 169*(3951), 1222–1223.

Martin, B. A., Tremblay, K. L., & Korczak, P. (2008). Speech evoked potentials: From the laboratory to the clinic. *Ear and Hearing, 29*(3), 285–313.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*(7397), 233–236.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343,* 1006–1010. https://doi.org/10.1126/science.1245994.

Mirkovic, B., Debener, S., Jaeger, M., & De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications. *Journal of Neural Engineering, 12*(4), 046007.

Moses, D. A., Leonard, M. K., & Chang, E. F. (2018). Real-time classification of auditory sentences using evoked cortical activity in humans. *Journal of Neural Engineering, 15*(3), 36005.

Moushegian, G., Rupert, A. L., & Stillman, R. D. (1973). Scalp-recorded early responses in man to frequencies in the speech range. *Electroencephalography and Clinical Neurophysiology, 35*(6), 665–667.

Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hearing Research, 241*(1–2), 34–32.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research, 11*(June), 1833–1863.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology, 112*(4), 713–719.

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of

neuroimaging data in MATLAB/GNU Octave. *Frontiers in Neuroinformatics, 10,* 27.

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Frontiers in Human Neuroscience, 10*(January), 679. https://doi.org/10.3389/fnhum.2016.00679

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of Neural Engineering, 14*(5), 56001.

O'Sullivan, J., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., . . . Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex, 25*(7), 1697–1706.

Phipson, B., & Smyth, G. K. (2010). Permutation *p*-values should never be zero: Calculating exact *p*-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology, 9*(1). https://doi.org/10.2202/1544-6115.1585

Power, A. J., Colling, L. J., Mead, N., Barnes, L., & Goswami, U. (2016). Neural encoding of the speech envelope by children with developmental dyslexia. *Brain and Language, 160,* 1–10. https://doi.org/10.1016/j.bandl.2016.06.006

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience, 35*(9), 1497–1503.

Puschmann, S., Steinkamp, S., Gillich, I., Mirkovic, B., Debener, S., & Thiel, C. M. (2017). The right temporoparietal junction supports speech tracking during selective listening: Evidence from concurrent EEG-fMRI. *Journal of Neuroscience, 37*(47), 11505–11516.

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain, 126*(4), 841–865.

Reetzke, R., Xie, Z., Llanos, F., & Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology, 28*(9), 1419.e4–1427.e4.

Russo, N. M., Nicol, T., Trommer, B., Zecker, S., & Kraus, N. (2009). Brainstem transcription of speech is disrupted in children with autism spectrum disorders. *Developmental Science, 12*(4), 557–567.

Russo, N. M., Skoe, E., Trommer, B., Nicol, T., Zecker, S., Bradlow, A., & Kraus, N. (2008). Deficient brainstem encoding of pitch in children with autism spectrum disorders. *Clinical Neurophysiology, 119*(8), 1720–1731.

Sadeghian, A., Dajani, H. R., & Chan, A. D. C. (2015). Classification of speech-evoked brainstem responses to English vowels. *Speech Communication, 68,* 69–84.

Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: A tutorial. *Ear and Hearing, 31*(3), 302–324.

Smith, J. C., Marsh, J. T., & Brown, W. S. (1975). Far-field recorded frequency-following responses: Evidence for the locus of brainstem sources. *Clinical Neurophysiology, 39*(5), 465–472.

Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems With Applications, 37*(12), 8659–8666.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*(6), 1284–1294.

Tremblay, K. L., Friesen, L., Martin, B. A., & Wright, R. (2003). Test–retest reliability of cortical evoked potentials using naturally produced speech sounds. *Ear and Hearing, 24*(3), 225–232.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology, 19,* 181–191.

White-Schwoch, T., Carr, K. W., Thompson, E. C., Anderson, S., Nicol, T., Bradlow, A. R., . . . Kraus, N. (2015). Auditory processing in noise: A preschool biomarker for literacy. *PLoS Biology, 13*(7), e1002196.

Worden, F. G., & Marsh, J. T. (1968). Frequency-following (microphonic-like) neural responses evoked by sound. *Electroencephalography and Clinical Neurophysiology, 25*(1), 42–52.

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience, 32*(7), 855–869.

Xie, Z., Reetzke, R., & Chandrasekaran, B. (2017). Stability and plasticity in neural encoding of linguistically relevant pitch patterns. *Journal of Neurophysiology, 117*(3), 1407–1422.

Xie, Z., Reetzke, R., & Chandrasekaran, B. (2018). Taking attention away from the auditory modality: Context-dependent effects on early sensory encoding of speech. *Neuroscience, 384,* 64–75. https://doi.org/10.1016/J.NEUROSCIENCE.2018.05.023

Yi, H. G., Xie, Z., Reetzke, R., Dimakis, A. G., & Chandrasekaran, B. (2017). Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach. *Brain and Behavior, 7*(6), e00665.

Zich, C., De Vos, M., Kranczioch, C., & Debener, S. (2015). Wireless EEG with individualized channel layout enables efficient motor imagery training. *Clinical Neurophysiology, 126*(4), 698–710.