Received: 5 October 2022 | Revised: 13 September 2023 | Accepted: 4 October 2023

DOI: 10.1111/psyp.14475

ORIGINAL ARTICLE

PSYCHOPHYSIOLOGY SPR WILEY

Decoding speech sounds from neurophysiological data: Practical considerations and theoretical implications

McCall E. Sarrett[1,2] | Joseph C. Toscano[1]

[1]Department of Psychological and Brain Sciences, Villanova University, Villanova, Pennsylvania, USA

[2]Psychology Department, Gonzaga University, Spokane, Washington, USA

Correspondence
McCall E. Sarrett, Department of Psychological and Brain Sciences, Villanova University, 502 E Boone Ave GU Psychology, AD Box #56, Spokane WA 99258, USA.
Email: sarrett@gonzaga.edu

Funding information
National Science Foundation, Grant/Award Number: 1945069 and 2018933

Abstract

Machine learning techniques have proven to be a useful tool in cognitive neuroscience. However, their implementation in scalp-recorded electroencephalography (EEG) is relatively limited. To address this, we present three analyses using data from a previous study that examined event-related potential (ERP) responses to a wide range of naturally-produced speech sounds. First, we explore which features of the EEG signal best maximize machine learning accuracy for a voicing distinction, using a support vector machine (SVM). We manipulate three dimensions of the EEG signal as input to the SVM: number of trials averaged, number of time points averaged, and polynomial fit. We discuss the trade-offs in using different feature sets and offer some recommendations for researchers using machine learning. Next, we use SVMs to classify specific pairs of phonemes, finding that we can detect differences in the EEG signal that are not otherwise detectable using conventional ERP analyses. Finally, we characterize the timecourse of phonetic feature decoding across three phonological dimensions (voicing, manner of articulation, and place of articulation), and find that voicing and manner are decodable from neural activity, whereas place of articulation is not. This set of analyses addresses both practical considerations in the application of machine learning to EEG, particularly for speech studies, and also sheds light on current issues regarding the nature of perceptual representations of speech.

KEYWORDS

Analysis/Statistical Methods, Auditory Processes, EEG, ERPs, Language/Speech, Machine Learning

1 | INTRODUCTION

Spoken language poses a complex computational problem for listeners, and in turn, for researchers interested in the neurobiology of language. In order to perceive speech, listeners must map continuous acoustic cues onto linguistic categories (Holt & Lotto, 2010), deal with considerable variability across contexts (Hillenbrand et al., 1995; Jongman et al., 2000), integrate multiple cues in real time (Galle et al., 2019; Toscano & McMurray, 2012), and combine those cues with higher-level linguistic information (Andruski et al., 1994; Connine, 1987, 1990). Because there is immense variability in speech at multiple levels, listeners must cope with these issues from the earliest stages of language processing—during initial perceptual encoding.

One of the major challenges in elucidating the mechanisms supporting perceptual encoding of speech sounds is the issue of time. Spoken language comprehension unfolds rapidly, necessitating measures with millisecond-scale precision. Recent work using the event-related potential (ERP) technique has offered a solution to this problem (see Getz & Toscano, 2021, for a recent review).

Toscano et al. (2010) demonstrated that the auditory N1, an early cortical response reflecting sensory processing, indexes graded changes along an acoustic dimension, independently of listeners' phonological categories. Subsequent work has demonstrated that early cortical responses can be influenced by top-down information from context in cases where acoustic cues are ambiguous (Getz & Toscano, 2019; Noe & Fischer-Baum, 2020; Sarrett et al., 2020). This suggests that, at the earliest stages of perceptual processing, listeners' initial encoding of speech sounds reflects continuous acoustic cues, and that this encoding can be flexibly influenced by feedback from higher-level linguistic information.

Despite progress in understanding the neural basis of perceptual encoding, we still do not fully understand how the brain carries out these processes. One reason for this is that conventional ERP analyses (e.g., mean amplitude of specific ERP components, measured at specific electrodes) may only capture some of the perceptual distinctions in speech that researchers are interested in.

The current study aims to address this problem using machine learning techniques to decode the information available in the ERP response across multiple electrode sites and from multiple time points, providing a more sensitive measure than traditional ERP analyses. In doing so, we explore how best to apply decoding techniques to this issue, combining perspectives that have used machine learning to decode information in scalp electroencephalography (EEG; Bae & Luck, 2018), intracranial EEG (iEEG; Rhone et al., submitted), and functional magnetic resonance imaging (fMRI; Haxby et al., 2001). This provides a set of best practices that can be considered in future work studying speech perception and spoken language processing.

In the following sections, we review previous work on speech perception and speech sound encoding that sets up the problems we aim to address, and describe prior work that uses machine learning techniques to decode information from neurophysiological data. We then present a series of analyses using data from Pereira et al. (2018), which investigated speech sound encoding using scalp-recorded EEG across a range of phonological contrasts in natural speech. Results are discussed in terms of their implications for understanding the neurobiology of speech perception and applications of machine learning techniques to future work.

1.1 | Perceptual representations of speech sounds

Speech sounds can be defined along a number of different dimensions (Jakobson et al., 1953; Ladefoged, 1996; Stevens, 2000). For instance, the dimensions of voicing, manner of articulation, and place of articulation are typically used to classify consonant sounds. Voicing refers to whether or not a given sound is produced concurrent with the vibration of the vocal folds (e.g. /b/ is voiced, /p/ is voiceless). Place of articulation refers to the location in the vocal tract where a constriction or obstruction is made to produce a consonant (e.g. /b/ is a voiced bilabial stop consonant, produced with the lips; /d/ is voiced alveolar stop consonant, produced with the tongue hitting the alveolar ridge). Finally, manner is based on the interaction between the articulators and the type of airflow that creates a given speech sound (e.g. /b/ is a stop, which completely obstructs the airflow; /m/ is a nasal, which redirects airflow through the nose).

A longstanding theoretical issue in speech perception concerns the nature of acoustic cues that provide the listener with information about these features. A single cue value can map to multiple different speech sounds, depending on surrounding context. For example, the primary cue to distinguish voicing for stop consonants is Voice Onset Time (VOT; Lisker & Abramson, 1964). Shorter VOTs signal voiced phonemes, such as /b,d,g/, whereas longer VOTs signal voiceless phonemes, like /p,t,k/. For example, in English, voiced bilabial stops (/b/) typically have VOTs around 0 ms and voiceless bilabial stops (/p/) have VOTs around 50 ms. Thus, an intermediate cue value of 25 ms could equally signal either category. In addition, VOT varies depending on speaking rate (Allen & Miller, 1999), such that a VOT of 25 ms might signal a voiced sound when speaking slowly and a voiceless sound when speaking quickly. Moreover, VOT also provides information about place of articulation (Benkí, 2001; Chodroff & Wilson, 2014; Fischer-Jørgensen, 1954; Lehiste & Peterson, 1961; Lisker & Abramson, 1964), further complicating the mapping between cues and linguistic categories.

Early work suggested that listeners perceive changes along these dimensions categorically (Liberman et al., 1957), rather than as continuous acoustic cues. Under this hypothesis, listeners would only encode differences that cross a phoneme category boundary. For example, the 10 ms difference between a 20 ms (/b/) and 30 ms (/p/) VOT would be perceptible, because it crosses a category boundary, whereas the 10 ms difference between a 5 ms (/b/) and 15 ms (also /b/) VOT would be imperceptible. However, subsequent work disproved the categorical hypothesis (Pisoni & Lazarus, 1974). Indeed, it is more advantageous for listeners to retain such subphonemic information in cases when the initial phoneme decision may need to be revised. Moreover, when other relevant information in the speech signal is taken into account—coarticulation, higher-level linguistic information, talker differences, etc.—this favors a perceptual process that inherently relies on noncategorical perception (see McMurray, 2021, for a recent review).

Recent work from cognitive neuroscience supports the idea that listeners encode continuous cues, indicating a flexible neural mechanism subserving gradient acoustic processing (but see Chang et al., 2010, for evidence of category-based representations using iEEG). The majority of evidence shows that the brain encodes acoustic cues gradiently when indexed at the auditory N1 component (Frye et al., 2007; Gwilliams et al., 2018; Sarrett et al., 2020; Toscano et al., 2010, 2018), and the degree of gradiency may vary between individual listeners (Kapnoula & McMurray, 2021). These studies typically present listeners with minimal pairs that vary along a voicing continuum (such as *beach* or *peach*). The VOT at word onset is manipulated to vary between a prototypical /b/ to a prototypical /p/, and participants are asked to categorize the word. Then, the amplitude of the N1 component in response to target word onset is measured.

Shorter (more voiced) VOTs yield a more negative N1 amplitude, whereas longer (more voiceless) VOTs yield a less negative N1. This N1 effect varies linearly with the VOT of the target word, suggesting that early perceptual representations of speech sounds are gradient (not categorical) with respect to the acoustic signal. Moreover, perceptual representations can be influenced by contextual expectations. Orthographic priming (Getz & Toscano, 2019), sentential contexts (Sarrett et al., 2020), and lexical status (Noe & Fischer-Baum, 2020) modulate perceptual encoding, particularly for ambiguous cue values (i.e., when a sound is not clearly a /b/ or a /p/). Taken together, these studies give us insight into the nature of perceptual representations of speech sounds and how perceptual processing interacts with higher-level lexico-semantic processing.

Despite this emerging evidence, details about the neural mechanisms supporting perceptual encoding in the brain remain unclear. For example, while changes in VOT show robustly different responses that are easily observable in ERPs, other relevant dimensions—such as contrasts in place or manner of articulation—yield less clear differences. Pereira et al. (2018) expanded the paradigm described above to use the N1 as an index of other phonetic distinctions, across a range of different phonemes, while participants categorized which phoneme they heard. This work showed that the N1 can be used as an index for many phonetic distinctions, not just voicing.

However, there were some limitations of Pereira et al's results. Some phonemes, such as /s/ and /ʃ/, showed no difference in N1 amplitude, even though participants' classification performance was at ceiling. That is, listeners show behavioral evidence that they are encoding phoneme differences, but traditional ERP methods are unable to detect such effects for some phoneme contrasts. One reason for this may be due to the orientation of dipoles in auditory cortex that respond to differences in acoustic cues. If the dipoles are oriented in such a way that differences cannot be detected at fronto-central electrodes where the N1 component is measured, this implies the need for a different measure. Machine learning offers a number of advantages over traditional ERP analyses: It allows us to detect potentially subtler differences in cortical activity by utilizing information across the entire scalp. This also allows us to reconceptualize how we design EEG experiments altogether, as multidimensional data cannot be studied effectively using univariate designs. Thus, understanding how to best apply machine learning techniques to EEG data will be important for informing future work.

1.2 | Machine learning approaches to decoding neural data

Machine learning[1] may offer a solution to the limitations of traditional ERP analyses. Several different approaches have been used to apply machine learning techniques to neural data, each with its own set of goals, advantages, and limitations.

1.2.1 | Functional MRI

The first approach comes from functional magnetic resonance imaging (fMRI) research. Multivariate pattern analysis (MVPA; previously referred to as multivoxel pattern analysis) has been used by fMRI researchers for decades (Haxby, 2012; Haxby et al., 2001; Norman et al., 2006). This approach was developed as an alternative to conventional fMRI methods, which tended to focus on whether a single voxel or cluster of voxels in a given region showed an increase in activity above a certain threshold during in a particular experiment. MVPA is an umbrella term which encompasses a diverse set of classification algorithms that share a common goal: understanding how the pattern of neural activity (in contrast to simply the *degree* of neural activity) in a given brain region corresponds to different stimulus properties, task demands, or cognitive states. Examples of

[1] For the purposes of the present article, we will use the terms "machine learning", "decoding", and "classifying", to refer to the same type of analysis.

MVPAs include both linear and nonlinear classifiers, such as linear and nonlinear support vector machines (SVMs), neural networks, and correlation-based classifiers (see Duda et al., 2001, for descriptions of these and other algorithms).

The overarching goal of this line of work is to use MVPA as a more sensitive substitute for traditional statistical analyses—to determine if a given brain region is involved in a particular cognitive or perceptual process based on whether or not classification performance in that region is above chance. This approach typically relies on data that are averaged across two or more conditions, though it is also possible to use with single-trial fMRI (Pessoa & Padmala, 2005). MVPA is widely accepted in the fMRI literature and has been used to effectively answer questions about brain function in speech perception (Archila-Meléndez et al., 2018; Correia et al., 2015b; Luthra et al., 2020; Vandermosten et al., 2017), as well as an enormous span of domains, including memory retrieval (Polyn et al., 2005), visual perception (Boynton, 2005), and face and object recognition (O'Toole et al., 2005).

1.2.2 | Intracranial EEG

The second decoding approach comes from iEEG. This line of work has applied machine learning techniques with an eye towards decoding the neural signal for input to brain-computer interfaces (BCIs). As such, researchers have primarily been concerned with decoding individual-trial-level neural activity. This relies on a similar logic: If classification performance is above chance, it means there is some feature of the neural signal that distinguishes the conditions of interest.

This approach is suitable both for answering questions about basic brain functions, such as characterizing the neural substrates involved in spoken word recognition (Rhone et al., submitted) or phonetic feature encoding (Mesgarani et al., 2014), and also for advancing technology for BCI applications, such as synthesizing speech from subvocal articulatory representations (Anumanchipalli et al., 2019), writing text from imagined hand movements (Willett et al., 2021), or decoding words from attempted articulations in patients with severe paralysis (Willett et al., 2023). Moreover, iEEG has a much finer temporal resolution than fMRI, which allows researchers to make more precise claims about *when* certain processes are occurring in cortical activity as well.

1.2.3 | **Scalp-recorded M/**EEG

A third approach has addressed this question by developing machine learning tools to decode the contents of scalp-recorded EEG signals (or magnetoencephalography [MEG] signals) as they unfold over time, offering potential advantages over traditional ERP analyses. This approach combines aspects of the fMRI and iEEG approaches and shares many of the same motivations.

One advantage that decoding techniques may offer over traditional ERP analyses is that they can take advantage of the pattern of voltage changes across the entire scalp, in contrast to traditional analyses, which tend to focus on a single electrode or small subset of electrodes. Bae and Luck (2018) conducted a study on the feasibility of applying machine learning to EEG recorded across the scalp. They were primarily interested in decoding the contents of working memory during a visual attention task. Participants saw lines in one of 16 different orientations (e.g., horizontal, vertical, slanted left) in one of 16 different locations on the screen (e.g., top right, bottom left, center). Decoding algorithms were applied to averaged EEG data, and the results revealed that different features of the EEG signal coded for different aspects of the stimulus properties: alpha oscillations carried information about spatial location of the stimulus, whereas sustained ERPs carried information about the orientation of the stimulus. This provided a proof of concept that neural information can be decoded at the scalp level. Since then, follow-up studies have employed similar decoding paradigms with good success (Bae & Luck, 2019). Moreover, other studies have shown that decoding on even a single-trial level is possible (Bayet et al., 2020; Correia, Jansma, Hausfeld, et al., 2015; McMurray et al., 2022; Trammel et al., 2023).

Beach et al. (2021) applied similar techniques to single-trial MEG data, which share many of the same qualities as EEG data. They were interested in how task demands affect the nature of neural representations of speech sounds, using decoding ability as the primary measure. Listeners either actively or passively attended to syllables along a place of articulation continuum (/ba/ to /da/). The active condition required that listeners make a response on each trial by clicking on a picture of a *ball* or a *dog* to indicate which sound they heard, whereas the passive condition had no explicit language-related task. Applying a decoding analysis to the neural data provided information about the degree to which continuous or categorical representations of speech sounds were present during the active versus the passive condition. This was assessed by looking at representational dissimilarity matrices from the output of the classifier. The results showed that listeners maintain both phonemic (categorical) and subphonemic (continuous) speech information during both active and passive tasks, but subphonemic information was maintained in neural activity for a longer duration when a task response was required. This is one of the first studies to look at the decoding of neural representations in speech using scalp-level electrophysiological measures.

1.3 | Goals of the current study

Still, further questions about the nature of cortical representations to speech sounds remain, and the various approaches to machine learning with neural data leave an open question regarding how to best apply these techniques as the field develops. Thus, the first goal of the present study is to unite different decoding traditions: from approaches that use machine learning on averaged data as a substitute for other statistical approaches, to traditions that primarily focus on individual-trial data for BCI or other applications. In particular, we seek to better understand how to maximize machine learning accuracy for scalp-recorded EEG.

Some recent work has had similar aims (Grootswagers et al., 2017; Trammel et al., 2023). For example, Grootswagers et al. (2017) used MEG data from a visual animacy judgment task to compare how decisions in the machine learning pipeline affect classification performance for animacy (animate vs. inanimate). They examined several key researcher decision points, including decisions made at the preprocessing stage (such as resampling and trial averaging) and also at the later classification stage (such as which classification algorithm is used and how classifier cross-validation is handled). Trammel et al. (2023) also examined differences in classification algorithm performance using EEG data during a semantic priming task, and found the support vector machines yield the highest overall classification performance (compared to linear discriminant analysis or random forest algorithms). Due to the high number of individual decisions that a researcher can make throughout the machine learning pipeline, it is critical to better understand how these decisions can be used to improve classification performance, while also mitigating the accompanying risk of researcher bias.

In the present study, we extend previous work to examine the effects of other researcher degrees of freedom in the machine learning pipeline. In particular, our initial analysis (Analysis 1) examines how differences in the input to a classifier (in our case, a support vector machine) affect classification performance. This analysis combines more traditional approaches (which have tended to rely on averaged data) and more recent approaches (which have looked at individual-trial classification) to examine the trade-offs in terms of machine learning accuracy along these dimensions.

In addition, we assess outstanding questions about how human listeners encode speech sounds during early perceptual processing. We use the information from the first analysis to identify parameters to use with the classifiers in two further analyses. As previously outlined, the nature of perceptual representations of speech sounds has not yet been fully characterized. One key example is that in Pereira et al. (2018), several phoneme contrasts did not show significant differences in N1 amplitude. This leaves an open question as to whether and how specific phonemic representations are distinguishable in neural measures. It is possible that more powerful machine learning techniques will be able to distinguish between phonemes where measures of N1 amplitude alone could not. This is assessed in Analysis 2. In addition, theories of speech perception differ in how phonetic features of speech sounds—such as voicing, manner of articulation, and place of articulation—contribute to perceptual representations, and there are open questions about how these representations unfold over time. Thus, in Analysis 3, we examine decoding over time to better understand the contributions of voicing, manner, and place during early stages of speech sound encoding.

After presenting these analyses, we discuss potential directions for future work and provide recommendations for best practices when using decoding analyses to study perceptual encoding of speech sounds.

2 | ANALYSIS 1: OPTIMAL PARAMETERS FOR DECODING VOICING IN SPEECH

2.1 | Introduction

The goal of Analysis 1 was to determine which classifier parameters would yield the highest accuracy for decoding neural representations of speech by examining the classifier parameter space (Pitt et al., 2006). This analysis seeks to find the optimal parameters for classification, given the goals of the specific research question. When doing any sort of parameter optimization, it is important to consider the consequences of certain decisions throughout the machine learning pipeline, particularly in regards to the risk of researcher bias. In the next sections, we discuss both advantages and disadvantages of certain decisions, as cautionary notes to those newly interested in applying these techniques.

In this analysis, we trained a classifier to predict a two-way voicing distinction (voiced vs. voiceless sounds), as previous research has shown that voicing is robustly represented in the neural signal for stop consonants (Frye et al., 2007; Getz & Toscano, 2019; Gwilliams et al., 2018; Noe & Fischer-Baum, 2020; Sarrett et al., 2020; Sharma et al., 2000; Toscano et al., 2010, 2018). Thus, we expect it to be readily decodable. The current analysis also seeks to expand this previous work by testing our ability to decode voicing across all consonants, not just stops.

We varied the input to the classifier across three dimensions. First, we varied the number of trials averaged in the input to the classifier, from individual-trial data (most often used in iEEG/BCI applications) to averaging many trials per condition (stemming from the fMRI/MVPA tradition). Recent EEG work has used both types of input, but trial-level and averaged data have not yet been directly compared on the same classification job in regard to their effect on classification accuracy. Here, we expect that a greater number of trials (resulting in a higher signal-to-noise ratio [SNR]) will yield better classifier performance overall.

Second, we varied the number of timepoints averaged, from very small time windows (2 ms, or a single time sample) to much longer time windows (250 ms). Similar to averaging over trials, averaging over time may be another way to functionally increase the SNR of the data. However, averaging over timepoints sacrifices some degree of temporal precision, and if the time window is *too* long, it may capture segments of the EEG signal that do not contain the information being decoded, thus reducing classification accuracy. By examining the number of trials and number of time points we average across, we are also able to quantify the trade-off between averaging over trials and averaging over time in terms of decoding accuracy.

Third, we varied the type of function that was fit to the data. Traditionally with MVPA, the mean activity of the voxel is used as classifier input, which corresponds to a zero-order polynomial. More temporally precise measures, however, like iEEG (Rhone et al., submitted) or scalp EEG (McMurray et al., 2022), can take into account fluctuations in the neural signal, which may be informative for stimulus decoding. For example, we can consider the slope of the voltage across a certain time window or a quadratic function that encompasses an ERP component. The majority of previous work has been developed to use mean voltage (a zero-order polynomial; Bae & Luck, 2018, 2019; Bayet et al., 2020), but some has used higher-order polynomial fits (McMurray et al., 2022). In order to examine trade-offs across all options, we compared results using the mean voltage (zero-order polynomial), linear fit (first-order polynomial), and quadratic fit (second-order polynomial). These polynomials functioned additively (e.g., the linear classifier comprised the slope plus the mean). This is further described in the Method section below. Connecting these polynomials with trade-offs across time and number of trials yields a more complete picture of the landscape in terms of feature optimization.

In summary, the first analysis explores the parameter space of the classifier along three dimensions by manipulating the number of trials averaged, the number of timepoints averaged, and the type of polynomial fit used for the classifier.

2.2 | Method

Before describing the machine learning techniques used in the current study, we first provide an overview of the data set from Pereira et al. (2018) that is used for our analyses.

2.2.1 | Participants

Participants in Pereira et al. (2018) were 27 members of the Villanova University community. Participants were fluent in English, had self-reported normal hearing, and self-reported normal or corrected-to-normal vision. Twenty three participants were right-handed; four were left-handed. One participant was excluded due to excessive noise in the EEG data. The final sample included 26 participants (11 male, 15 female; with a mean age of 19 years old). All participants gave informed consent before participating in the study, following Villanova IRB protocols.

2.2.2 | Stimuli

Participants heard monosyllabic words, which were minimal pairs and whose word-initial consonant spanned a range of different speech sounds, across voicing, manner, and place (/b, d, g, p, t, k, v, z, f, s, ʃ, ʧ, ʤ, m, n, ɹ, l, w/). Stimuli were natural utterances produced by a female native English speaker, which were edited to remove clicks and pops, and amplitude normalized using Praat (Boersma, 2006). There were 3–8 words for each phoneme and a total of 76 unique words (e.g. for /g/, participants heard the words "gear", "get", and "gill"). The full list of stimuli is in Table 1, and further details about stimulus preparation are available in Pereira et al. (2018).

2.2.3 | Procedure

Participants were fitted with an EEG cap and seated in a sound-attenuated, electrically shielded booth in front of a 22″ monitor. Stimuli were presented over 3 M E-A-RTONE 3a insert earphones using OpenSesame (Mathôt et al., 2012). Participants completed a phoneme categorization task and indicated their response by using a mouse to click on the letter from a display corresponding to the

TABLE 1 List of stimuli.

Phoneme	Words	Voicing	Manner	Place
/b/	bead, beat, beer, bees, bet, bill, bat	Voiced	Stop	Bilabial
/d/	dead, deal, debt, deed, deer, den, dill, dip	Voiced	Stop	Alveolar
/g/	gear, get, gill	Voiced	Stop	Velar
/p/	peas, pen, pet, pin	Voiceless	Stop	Bilabial
/t/	tease, ted, teen, ten, tin, tip	Voiceless	Stop	Alveolar
/k/	Ken, keys, kin	Voiceless	Stop	Velar
/v/	veal, vend, vest	Voiced	Fricative	Labiodental
/z/	zeal, zen, zest, zip	Voiced	Fricative	Alveolar
/f/	fed, feel, feet, fend	Voiceless	Fricative	Labiodental
/s/	said, seal, seat, seen, sin, sip	Voiceless	Fricative	Alveolar
/ʃ/	shed, sheet, shin	Voiceless	Fricative	Postalveolar
/ʤ/	gin, Jeep, Jess	Voiced	Affricate	Postalveolar
/ʧ/	Cheap, chess, chin	Voiceless	Affricate	Postalveolar
/m/	meet, met, mit	Voiced	Nasal	Bilabial
/n/	knit, neat, need, net, nip	Voiced	Nasal	Alveolar
/ɹ/	Red, reed, rip	Voiced	Approximant	Alveolar
/l/	Let, lead, lip, lit	Voiced	Lateral approximant	Alveolar
/w/	Weed, wet, wit	Voiced	Glide	Labiovelar

first phoneme of the word they heard. This display included each of the 18 possible phonemes heard throughout the experiment, arranged in a circle alphabetically, moving clockwise, centered around a fixation point.[2] Each word was presented 10 times, for a total of 760 trials (30–80 repetitions per phoneme). Trial order was randomized for each participant, with breaks every 17 trials. The experimental session lasted approximately two hours.

2.2.4 | EEG **data**

EEG data were collected using a BrainVision 32-channel active electrode setup. Electrodes were placed according to the International 10-20 system. Electrooculograms (EOG) were recorded with the electrode above the center of the left eye (vertical EOG) and two placed near the lateral canthus of each eye (horizontal EOG). Impedances at all electrodes were kept below 10 kΩ.

EEG was collected continuously, referenced online to the left mastoid, and digitized at a sampling rate of 500 Hz. During analysis, the data were band-pass filtered from 0.1 to 30 Hz with a 12 dB/octave rolloff and rereferenced offline to the average of the left and right mastoids (electrodes TP9 and TP10, respectively). Non-stereotypic artifacts were manually rejected via visual inspection. Trials containing stereotypic artifacts (eyeblinks and saccades) were excluded first using a peak-to-peak detection at the EOG channels (vertical: Fp1, horizontal: bipolar channel, calculated from the difference between FT9 and FT10), with a threshold of 75 μV. If a participant had greater than 15% of trials exceeding this threshold, then eyeblinks and saccades were removed using Independent Component Analysis (ICA). The mastoid and EOG channels were removed from the data set, resulting in 26 electrodes per participant (from left to right taking an overhead perspective, rostral to caudal: F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, O2). Data were timelocked to the presentation of the target word, and epoched from 200 ms before target word onset to 800 ms after word onset. Target word onsets were identified by measuring voltage deflections in an additional channel in the EEG data file that contained the sound waveform of the stimulus, recorded using a BrainVision StimTrak. Epochs were baselined to the 200 ms silent period before stimulus onset for each trial.

[2] This arrangement resulted in some classes being clustered together in the response space for the dimensions we attempt to decode: voicing, manner, or place of articulation. For example, both the nasals *M* and *N* occurred in the bottom middle of this arrangement. In theory, if this pattern of a bias to a region of the response space was widespread, we could be concerned that motor planning might contribute to overall decoding accuracy. We verified that, for the majority of classes, there was little or no bias to any part of the response space. Thus, we conclude that it is unlikely for motor response planning to play a role in our ability to decode along these dimensions.

2.2.5 | Machine learning techniques

Data were analyzed using custom MATLAB scripts and the LibSVM package from Chang and Lin (2011). This package uses a radial basis function to optimize classification performance of a nonlinear multiclass support vector machine (SVM).[3] An SVM classifier is optimized to separate input (training) data into distinct sets of classes, and then this model is used to classify a new (testing) set of data. Specifically when using a radial basis function, this optimization is done by tuning two free parameters of the model: cost (C) of the function and width of the kernel (γ). C refers to the degree to which the SVM is penalized for an incorrect guess. A low C parameter (low penalty for misclassification) allows for a higher number of incorrect guesses on training data—for example, by allowing some outliers to be incorrectly classified. This results in a larger margin between the hyperplane that separates classes and the actual data points, and can result in better classification performance on the testing data. In contrast, a high C parameter (large penalty for misclassification) allows for fewer incorrect guesses on the training data. This can result in a smaller margin between the separating hyperplane and the actual data points, and can result in lower classification performance at test. Next, the width of the kernel (γ) determines how far (in a multidimensional feature space) data points in the training set are allowed to be from the separating hyperplane. A low γ can result in underfitting—for example, classifying the whole data set as a single class. In contrast, a high γ can result in overfitting—for example, inability to generalize to new data points.

It is not known which combination of C and γ will yield best classification performance for a given data set. Therefore, we must perform some sort of parameter space search to optimize them. The parameters were optimized for each participant and for each classification job (i.e., for each combination of number of trials × time points × polynomial type). Parameter optimization was done using a hybrid approach, starting with a brute force search and then a finer gradient descent method.[4] The brute force search sampled the linear space from the lower bound (2^{-5} for C, 2^{-19} for γ) to the upper bound (2^{15} for C, 2^{3} for γ) in an 8×8 grid. From this brute force search, the C and γ combination that yielded the best performance was chosen. This combination was then used as the starting point for the gradient descent method, which further refined these parameters until the local minimum was reached.

For Analysis 1, the SVM was trained to perform a voicing distinction. Phonemes were classified as either voiced (/b, d, g, v, z, ʤ, m, n, ɹ, l, w/) or voiceless (/p, t, k, f, s, ʃ, ʧ /). The number of trials in each class (voiced and voiceless) were equalized before inputting the data to the SVM. This was done by identifying which class had fewer trials and indexing that same number of trials from a randomized vector of the other class.[5] Included trials were randomly sampled from the full duration of the experiment.

We varied the three parameters previously described (number of trials averaged, number of time points averaged, and polynomial fit). To manipulate number of trials averaged, we split the data N ways and took the mean voltage of all trials within each number of data splits—ranging from 3 data splits to 25 data splits—or used individual trials for classification. Fewer data splits corresponds to a greater number of trials averaged.

To manipulate number of timepoints averaged, we used a time window centered at 120 ms post-target word onset, which corresponds to the approximate peak of the auditory N1 in this data set, and varied its width by including different numbers of adjacent time points. Time window width ranged from 2 to 150 ms for the zero-order polynomial, and from 10 to 250 ms for the first- and second-order polynomials. We did not fit the higher-order polynomials across the smallest 2 ms time window because the linear

[3] We ran an initial comparison to the SVM package used by Bae and Luck (2018), which uses a linear binary SVM and error-correcting output codes (Dietterich & Bakiri, 1994) for multiclass decoding. Despite the differences in the underlying algorithms, we found that both packages showed comparable classification performance.

[4] There are several ways that parameter optimization can be performed. We used a hybrid approach, though other methods may be useful for other applications. For example, a Bayesian approach (Frazier, 2018) may be particularly useful for classification jobs with a low dimensional space (e.g., less than 20 dimensions). However, because most EEG work uses 32 to 128 electrodes, this approach may not be appropriate for some studies. Nevertheless, the Bayesian approach could be advantageous in developmental work or work with special populations where a longer duration EEG setup is not possible, and thus fewer electrodes are used.

[5] Equalizing the number of trials per class (voiced and voiceless) was necessary for several reasons. Because the experiment was not designed to have an equal number of voiced and voiceless *words*, there were more voiced trials overall in the data set. Thus, there was concern regarding SVM training, both for individual-trial SVMs and averaged-trial SVMs. If trials are extremely imbalanced between conditions, then an individual-trial SVM may adopt a strategy of guessing the more frequently occurring class in its training data, as that guess is more likely to be correct. This could yield above chance performance that is not due to the SVM truly distinguishing between the two classes. The second reason had to do with the use of trial-averaged SVMs. If the number of trials is extremely imbalanced between conditions, this may lead to a difference in the SNR of the ERPs, where the class with the greater number of trials would have a "smoother" ERP or a difference in peak amplitude between conditions (Luck, 2014). The SVM could thus potentially learn to distinguish classes based on this difference, which again could yield above chance performance that is not based on detecting a true difference between conditions. Moreover, individual subjects varied in the number of voiced and voiceless trials in their data set, due to the removal of trials during the artifact rejection stage of data processing.

and quadratic functions cannot be computed for a single data point. We also included two longer time windows for higher-order polynomials, 200 and 250 ms (not used for the zero-order polynomial), as these longer time windows may allow the higher-order functions to better capture the pattern in the underlying data.

Lastly, we varied the type of polynomial used to train the SVM. The zero-order polynomial took the mean across the specified time window at each electrode, resulting in 26 inputs to the SVM (1 parameter ×26 electrodes). The first-order polynomial fit a line to the data across the specified time window, and the slope of this line and its intercept were used as input to the classifier, resulting in 52 features sent to the SVM (2 parameters ×26 electrodes). Similarly, the second-order polynomial fit a quadratic function to the data. This included the value of the quadratic term, the linear slope, and the intercept, resulting in 78 features input to the SVM (3 parameters ×26 electrodes).

SVM cross-validation was performed using a *k*-fold procedure.[6] This involves splitting the data into *k* number of folds: *k*-1 folds are used for training the SVM, and one fold containing withheld (untrained) data is used for testing SVM predictions. For analyses using averaged data, the number of data splits corresponded to the number of *k*-folds. For example, averaging over the greatest number of trials, the data were randomly split 3 ways and then averaged over each third. Two of the folds were used to train the SVM, and the final fold was withheld for testing SVM predictions. For analyses using individual-trial data as input, a 15 *k*-fold procedure was used. Individual-trial SVMs did not average within the *k*-fold, and in theory, any value for *k* could be chosen for cross-validation. Typically, a *k*-value of around 10 will work for most data sets (Karal, 2020). This was run over multiple iterations, such that each fold served as the testing data at least once. For example, the SVM would train on the data in folds 1 through 14, then test on 15th fold. On the next iteration, it would train on folds 2 through 15, and test on 1st fold, and so on (cyclically). Then, the entire *k*-fold procedure was repeated 15 times, so that any idiosyncrasies due to trial assignment to a particular fold would even out. SVM performance over these repetitions was averaged to yield a single value for classification performance. Average classification performance across subjects is reported below.

2.2.6 | Statistical approach

We used one-sample *t*-tests to determine whether a given parameter combination (polynomial order × time window × trials averaged) performed significantly above chance levels, where average SVM performance across participants was compared to numerical chance (i.e., 50%). These *t*-statistics were calculated manually, following the standard formula: $(\bar{X} - \mu)/(S/\sqrt{N})$, where $\bar{X}$ is the sample mean, μ is chance performance, S is standard deviation of the sample, and N is number of participants in the sample.[7]

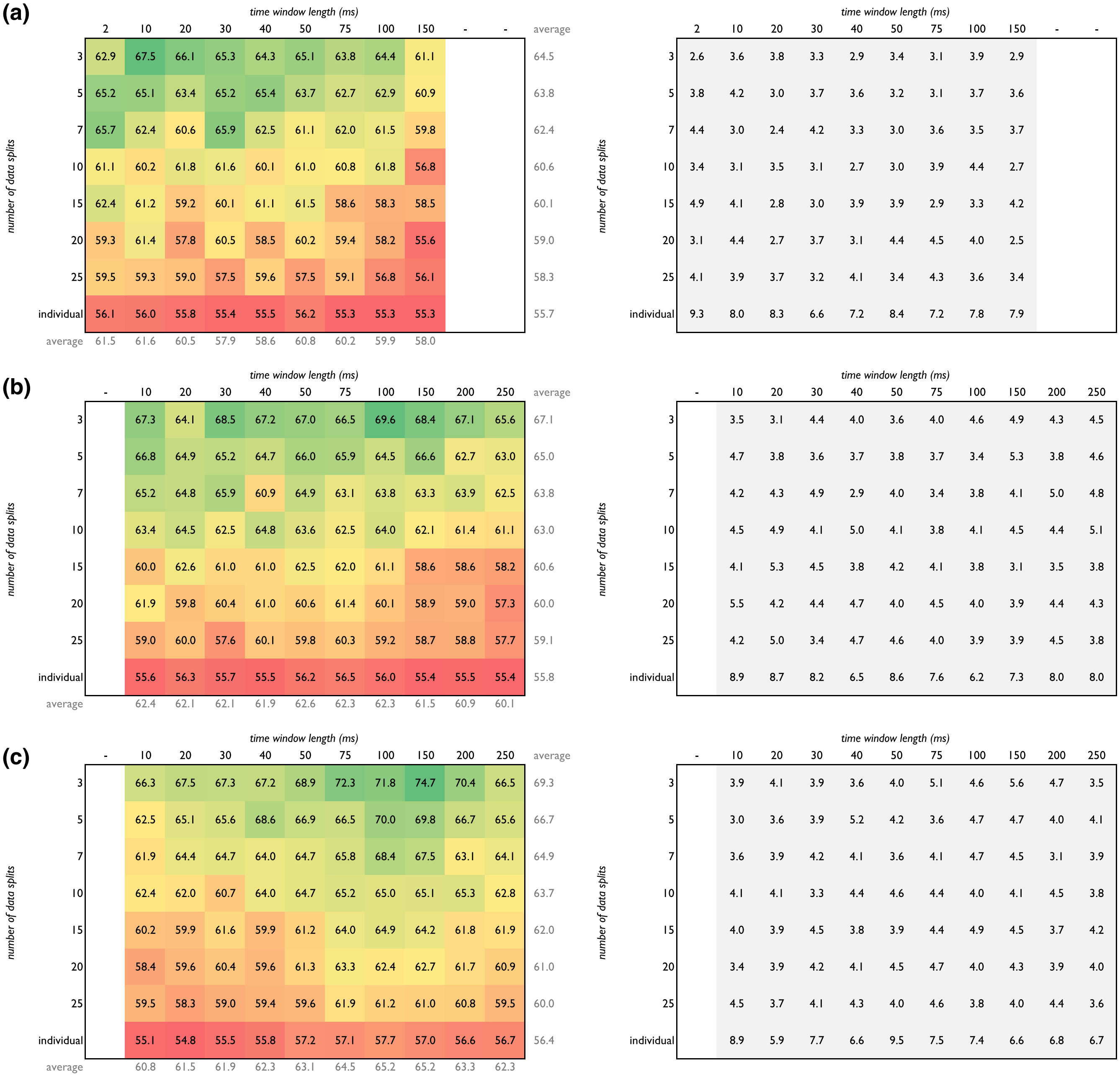
2.3 | Results

Results are shown in Figure 1 on the left with a corresponding table of t-values on the right. Figure 1a shows the classification results from SVMs trained on the zero-order polynomial (mean voltage). The time window width over which the mean was taken is given in the column headers, from 2 to 150 ms. The number of data splits is given in the row labels, from 3 to 25 data splits, as well as individual-trial classification. Each cell shows mean SVM performance across subjects.

We found that voicing is decodable above chance given a wide range of feature inputs to the SVM. The results also indicate that both averaging over trials and averaging over timepoints has an effect on SVM accuracy. Generally speaking, a greater number of trials averaged corresponds to higher SVM accuracy. This is true across a wide range of time window widths, but beyond 50 ms, averaging over more timepoints does not generally yield greater SVM accuracy. The effect of time window width can be explained in part by the pattern of the underlying data (i.e., the duration of the auditory N1 component, which captures the information relevant to the voicing distinction that the classifiers were trained on). As the time window width increases past 50 ms, it is likely that there is some averaging of positive voltages from the surrounding P50 and P2 components. This could obscure the differences in N1

[6] There are a number of different ways that cross-validation can be performed. Grootswagers et al. (2017) compares a few of these, as well as how they affect classification performance. They show similar performance between *k*-fold and leave-one-trial-out cross-validation and suggest that the particular implementation of cross-validation is study-specific. For the purposes of the present analyses, we used *k*-fold cross-validation throughout.

[7] For Analysis 1, we were unable to run all the permutations needed for a true SVM validation, due to computational constraints. Ideally, a statistical validation would be performed, consisting of running a new SVM on data with randomized trial codes. This allows researchers to estimate the general variation in classification performance due to random chance. However, the results (to be discussed) from Analyses 2 and 3 indicate that comparing classification to numerical chance yields similar—and in some cases, more conservative—estimates of statistical significance. Moreover, because Analysis 1 was largely an exploration of the parameter space, we found this singular statistical approach to suffice.



(a)

time window length (ms)

number of data splits	2	10	20	30	40	50	75	100	150	-	-	average
3	62.9	67.5	66.1	65.3	64.3	65.1	63.8	64.4	61.1			64.5
5	65.2	65.1	63.4	65.2	65.4	63.7	62.7	62.9	60.9			63.8
7	65.7	62.4	60.6	65.9	62.5	61.1	62.0	61.5	59.8			62.4
10	61.1	60.2	61.8	61.6	60.1	61.0	60.8	61.8	56.8			60.6
15	62.4	61.2	59.2	60.1	61.1	61.5	58.6	58.3	58.5			60.1
20	59.3	61.4	57.8	60.5	58.5	60.2	59.4	58.2	55.6			59.0
25	59.5	59.3	59.0	57.5	59.6	57.5	59.1	56.8	56.1			58.3
individual	56.1	56.0	55.8	55.4	55.5	56.2	55.3	55.3	55.3			55.7
average	61.5	61.6	60.5	57.9	58.6	60.8	60.2	59.9	58.0			

time window length (ms)

number of data splits	2	10	20	30	40	50	75	100	150	-	-
3	2.6	3.6	3.8	3.3	2.9	3.4	3.1	3.9	2.9		
5	3.8	4.2	3.0	3.7	3.6	3.2	3.1	3.7	3.6		
7	4.4	3.0	2.4	4.2	3.3	3.0	3.6	3.5	3.7		
10	3.4	3.1	3.5	3.1	2.7	3.0	3.9	4.4	2.7		
15	4.9	4.1	2.8	3.0	3.9	3.9	2.9	3.3	4.2		
20	3.1	4.4	2.7	3.7	3.1	4.4	4.5	4.0	2.5		
25	4.1	3.9	3.7	3.2	4.1	3.4	4.3	3.6	3.4		
individual	9.3	8.0	8.3	6.6	7.2	8.4	7.2	7.8	7.9		

(b)

time window length (ms)

number of data splits	-	10	20	30	40	50	75	100	150	200	250	average
3		67.3	64.1	68.5	67.2	67.0	66.5	69.6	68.4	67.1	65.6	67.1
5		66.8	64.9	65.2	64.7	66.0	65.9	64.5	66.6	62.7	63.0	65.0
7		65.2	64.8	65.9	60.9	64.9	63.1	63.8	63.3	63.9	62.5	63.8
10		63.4	64.5	62.5	64.8	63.6	62.5	64.0	62.1	61.4	61.1	63.0
15		60.0	62.6	61.0	61.0	62.5	62.0	61.1	58.6	58.6	58.2	60.6
20		61.9	59.8	60.4	61.0	60.6	61.4	60.1	58.9	59.0	57.3	60.0
25		59.0	60.0	57.6	60.1	59.8	60.3	59.2	58.7	58.8	57.7	59.1
individual		55.6	56.3	55.7	55.5	56.2	56.5	56.0	55.4	55.5	55.4	55.8
average		62.4	62.1	62.1	61.9	62.6	62.3	62.3	61.5	60.9	60.1	

time window length (ms)

number of data splits	-	10	20	30	40	50	75	100	150	200	250
3		3.5	3.1	4.4	4.0	3.6	4.0	4.6	4.9	4.3	4.5
5		4.7	3.8	3.6	3.7	3.8	3.7	3.4	5.3	3.8	4.6
7		4.2	4.3	4.9	2.9	4.0	3.4	3.8	4.1	5.0	4.8
10		4.5	4.9	4.1	5.0	4.1	3.8	4.1	4.5	4.4	5.1
15		4.1	5.3	4.5	3.8	4.2	4.1	3.8	3.1	3.5	3.8
20		5.5	4.2	4.4	4.7	4.0	4.5	4.0	3.9	4.4	4.3
25		4.2	5.0	3.4	4.7	4.6	4.0	3.9	3.9	4.5	3.8
individual		8.9	8.7	8.2	6.5	8.6	7.6	6.2	7.3	8.0	8.0

(c)

time window length (ms)

number of data splits	-	10	20	30	40	50	75	100	150	200	250	average
3		66.3	67.5	67.3	67.2	68.9	72.3	71.8	74.7	70.4	66.5	69.3
5		62.5	65.1	65.6	68.6	66.9	66.5	70.0	69.8	66.7	65.6	66.7
7		61.9	64.4	64.7	64.0	64.7	65.8	68.4	67.5	63.1	64.1	64.9
10		62.4	62.0	60.7	64.0	64.7	65.2	65.0	65.1	65.3	62.8	63.7
15		60.2	59.9	61.6	59.9	61.2	64.0	64.9	64.2	61.8	61.9	62.0
20		58.4	59.6	60.4	59.6	61.3	63.3	62.4	62.7	61.7	60.9	61.0
25		59.5	58.3	59.0	59.4	59.6	61.9	61.2	61.0	60.8	59.5	60.0
individual		55.1	54.8	55.5	55.8	57.2	57.1	57.7	57.0	56.6	56.7	56.4
average		60.8	61.5	61.9	62.3	63.1	64.5	65.2	65.2	63.3	62.3	

time window length (ms)

number of data splits	-	10	20	30	40	50	75	100	150	200	250
3		3.9	4.1	3.9	3.6	4.0	5.1	4.6	5.6	4.7	3.5
5		3.0	3.6	3.9	5.2	4.2	3.6	4.7	4.7	4.0	4.1
7		3.6	3.9	4.2	4.1	3.6	4.1	4.7	4.5	3.1	3.9
10		4.1	4.1	3.3	4.4	4.6	4.4	4.0	4.1	4.5	3.8
15		4.0	3.9	4.5	3.8	3.9	4.4	4.9	4.5	3.7	4.2
20		3.4	3.9	4.2	4.1	4.5	4.7	4.0	4.3	3.9	4.0
25		4.5	3.7	4.1	4.3	4.0	4.6	3.8	4.0	4.4	3.6
individual		8.9	5.9	7.7	6.6	9.5	7.5	7.4	6.6	6.8	6.7

FIGURE 1 Results of parameter space exploration in Analysis 1. Average SVM performance on a two-way voicing classification (chance = 50%, $N = 26$ subjects). Red colors correspond to lower SVM accuracy; green colors correspond to higher accuracy. Column headers are time window width in milliseconds, centered at 120 ms post-target word onset. Row headers are the number of data splits, which is inversely proportional to number of trials averaged. Row and column means are shown in gray in the margins. SVM performance on all tested feature inputs was significantly above chance. (a) Results for the zero-order polynomial (i.e., mean voltage with a total of 26 features). (b) Results for the first-order polynomial (i.e., linear function with a total of 52 features). (c) Results for the second-order polynomial (i.e., quadratic function with a total of 78 features).

amplitude between voicing conditions, resulting in a decrease in performance.

Figure 1b shows SVM results using the first-order polynomial (linear function) as input to the classifier. Time windows were centered at 120 ms (as in Figure 1a). The results for the first-order polynomial differed somewhat from the results for the zero-order polynomial. Performance overall tended to increase by adding another feature (slope) to the SVM input. Mid-range time window widths (50–150 ms) tended to yield better performance than shorter time windows, especially for data that were averaged over many trials.

Figure 1c shows SVM performance using the second-order polynomial (quadratic function) as input. For this polynomial, the longer time windows (75–200 ms) showed the highest SVM performance. This polynomial also yielded the best performance across all SVMs—74.7%—using SVM input that averaged across 3 data splits (the largest number of trials) and across a large number of timepoints (150 ms). Moreover, the individual-trial SVMs showed robust above chance classification using the second-order polynomial, particularly at wider time windows (reaching 57.7% with a 100 ms time window) This may suggest that the SVM makes better use of the greater number of features available in the quadratic function for decoding individual trials, particularly when compared to the first- and zero-order individual-trial SVMs, which showed lower performance overall.

2.4 | Discussion

The results of Analysis 1 reveal that all three dimensions affect classifier performance, though the relationship between them is complex. In general, averaging over a larger number of trials yielded better performance, as expected. The optimal time window width, however, depended on the type of polynomial fit to the data. For the mean voltage, time windows from 10 to 50 ms produced the best performance; for the linear function, time windows from 20 to 150 ms produced the best performance; and for the quadratic function, time windows of 75–200 ms produced the best performance.

In addition, the patterns observed are consistent with what we know about the underlying data. For this initial exploration of the parameter space, we used a speech sound contrast that produces a known effect in scalp-recorded ERPs: voiced sounds evoke a more negative N1 than voiceless sounds. This is true for stop consonants (Frye et al., 2007; Sharma et al., 2000; Toscano et al., 2010), as well as fricatives and affricates (Pereira et al., 2018). Given the pattern in the ERP data in these previous studies, we would expect good performance for a classifier trained on the parameters of a quadratic function, fit with a large number of trials, over a sufficiently large time window that encompasses the N1 component—this is precisely the combination of features that produced the highest classification accuracy.

Interestingly, when using higher-order polynomials, such as the quadratic in Figure 1c, we see that individual-trial SVMs are able to perform similarly to some averaged-trial SVMs, particularly for the lower-order polynomials. This suggests that using a greater number of features of the EEG signal can enhance SVM performance, particularly on individual-trial data. This is an important consideration for applications that would necessitate trial-level data, such as BCIs, or for correlating brain responses with trial-level behavior. However, it should be noted that while individual-trial performance was similar numerically, it never reached as high a level of accuracy as averaged-trial performance (consistent with Grootswagers et al., 2017).

Overall, these analyses underscore the need for the experimenter to evaluate what to prioritize when choosing the input to the SVM. For example, if precision in the timecourse of decoding is most critical, then it may be best to average over many trials and use a shorter time window with mean voltage as input. If individual-trial level decoding is most important, then it may be best to use a higher-order polynomial over a longer time window. If maximizing classifier accuracy is most crucial, it may be best to use a higher-order polynomial on data averaged over many trials. Experimenters must be cautious, as decisions made at this stage can seriously influence the interpretation of later results. All of the feature combinations we tested resulted in above chance decoding. However, it is possible that if we were attempting to decode a contrast that is less robustly represented in the neural signal, some of these feature combinations with lower performance may not have been statistically above chance. This means that experimenters may draw errant conclusions that some contrast or representation is not decodable from the neural signal, when perhaps the features given to the classifier simply did not capture it. As such, these parameter choices must be made mindfully.

3 | ANALYSIS 2: PHONEME DISTINCTIONS

3.1 | Introduction

Analysis 2 seeks to apply these principles to examine whether specific phoneme pairs can be reliably decoded from the EEG signal. Previous work has shown that certain distinctions, such as /b/ vs. /p/, show a robust effect on N1 amplitude. However, other phonemic distinctions do not show an N1 effect, despite being perceptually distinct. In Pereira et al. (2018), for example, some phonemes, such as /s/ and /ʃ/, were statistically indistinguishable at the N1. However, participants' phoneme classification performance during the task was near ceiling. Thus, listeners were able to perceive the difference in these phonemes, even though differences in perceptual encoding could not be detected at the N1 with conventional analysis approaches. This suggests that the N1 component is only sensitive to certain acoustic distinctions in speech, possibly due to differences in the orientation of dipoles coding different types of acoustic cues. For example, neuronal populations coding VOT may be oriented in such a way that differences are observable in the N1 at frontal electrodes, while those coding spectral mean (one of the primary cues distinguishing /s/ and /ʃ/) might be oriented differently, such that differences are not observed in the N1, even though listeners do perceive this acoustic distinction.

Because decoding analyses are more sensitive, they may be better at detecting weaker, but nevertheless informative dipoles, using information distributed across the entire scalp. To assess this, we examined decoding performance for specific phoneme pairs. There were substantially fewer trials for the classifier to train on for each phoneme individually, compared to when phonemes were grouped by whether they were voiced or voiceless (as in Analysis 1). Thus, informed by Analysis 1, we selected the features that would maximize classification accuracy (see Method below for details).

We were interested in testing a range of phonemes to determine which pairs were significantly decodable during early perceptual processing. These were selected to sample along three relevant phonological dimensions in the data set: voicing, manner, and place. For voicing, this included the pairs /b/ vs. /p/, /z/ vs. /s/, and /ʤ/ vs. /ʧ/. For manner, this included /b/ vs. /m/, /z/ vs. /n/, and /t/ vs. /s/. Finally, for place, this included /b/ vs. /d/, /z/ vs. /v/, and /ʃ/ vs. /s/. Given the phonemic inventory of English and the phonemes available in the data set, it was not possible to achieve perfect balance among the three contrasts, such that each pair differed in only one feature and each phoneme appeared in each list. However, we chose the pairs we used deliberately, such that within a given feature dimension, there was variation across the other features. For example, the voicing contrasts included a bilabial stop consonant pair (/b/ vs. /p/), an alveolar fricative pair (/z/ vs. /s/), and a postalveolar affricate pair (/ʤ/ vs. /ʧ/). We also chose pairs that both showed significant differences in N1 amplitude in Pereira et al.'s (Pereira et al., 2018) original analyses, as well as those that did not.

The stop consonant voicing contrast (/b/ vs. /p/) served as a baseline condition, as Analysis 1 and previous work has shown that voicing across stop consonants is readily detectable at the N1. Therefore, it should also be easily decodable. If decoding analyses are a sensitive enough alternative to traditional analyses and if there exist dipoles that code for such phoneme differences in neural activity, we predict that the other phoneme pairs should also be decodable significantly above chance regardless of whether or not they showed N1 differences in Pereira et al. (2018). We know that there must be neural representations that code for these phoneme differences, as they are perceptually distinct, and participants in Pereira et al. performed at ceiling during the phoneme categorization task. Therefore, the key question is whether or not machine learning will be able to detect these differences by taking advantage of the neurophysiological activity across the entire scalp.

3.2 | Method

3.2.1 | Machine learning techniques

As in Analysis 1, we used custom MATLAB scripts and LibSVM (Chang & Lin, 2011) to train the classifiers. We ran nine separate SVMs: one for each phoneme pair (see the Phoneme contrast column in Figure 2). We were interested in detecting differences in the neural signal that are likely very small, and we had a predefined time window in which we were interested in decoding (during early perceptual processing). For this analysis, we did not require individual-trial level data.

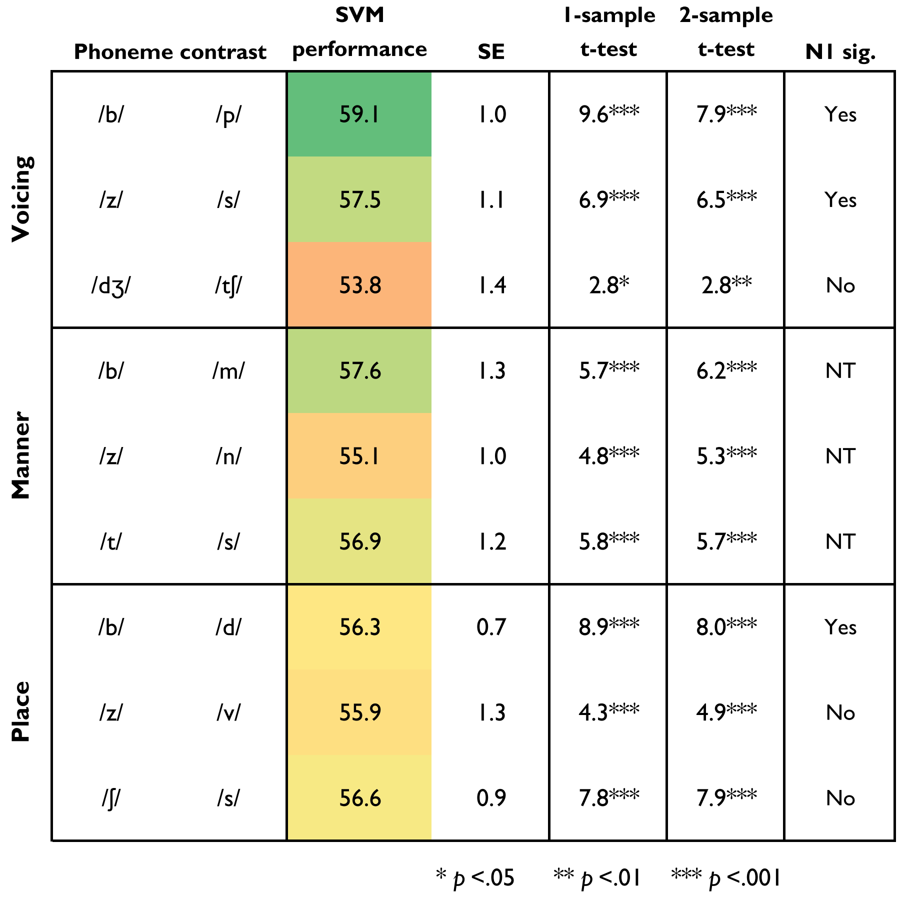
Thus, we chose SVM parameters that sought to maximize SVM performance. Based on Analysis 1, we used a 3-fold average and fit a second-order polynomial to the data over a 150 ms time window, centered at 120 ms. Free parameters C and γ were determined for each subject and for each classification job, and these were optimized using the hybrid brute force search and gradient descent approach, as previously described.

Cross-validation was performed with a 3k-fold procedure; this was repeated 15 times for each classification job. Mean classification performance was calculated across all repetitions.

3.2.2 | Statistical approach

We adopted a two-part statistical approach for this analysis: (1) a one-sample t-test of average SVM performance across participants compared to numerical chance (the same approach used in Analysis 1), and (2) a two-sample t-test comparing SVM performance on each job to a 100-repetition SVM classification on randomly shuffled data.

This second approach was part of a broader methodological goal. The strongest evidence that information is decodable in the neural signal (and not due to some other random variation) is by running a comparison SVM that is trained and tested on data with shuffled trial codes. This allows us to examine what the baseline SVM performance is for a given data set, and we would expect the SVM to perform at or around chance. Typically this procedure uses tens—if not hundreds—of repetitions, making it computationally costly. Therefore, we wanted to compare this procedure using shuffled data across many repetitions with a less costly alternative (one-sample t-test against numerical chance) to understand how these two statistical



	Phoneme contrast		SVM performance	SE	1-sample t-test	2-sample t-test	N1 sig.
Voicing	/b/	/p/	59.1	1.0	9.6***	7.9***	Yes
	/z/	/s/	57.5	1.1	6.9***	6.5***	Yes
	/dʒ/	/tʃ/	53.8	1.4	2.8*	2.8**	No
Manner	/b/	/m/	57.6	1.3	5.7***	6.2***	NT
	/z/	/n/	55.1	1.0	4.8***	5.3***	NT
	/t/	/s/	56.9	1.2	5.8***	5.7***	NT
Place	/b/	/d/	56.3	0.7	8.9***	8.0***	Yes
	/z/	/v/	55.9	1.3	4.3***	4.9***	No
	/ʃ/	/s/	56.6	0.9	7.8***	7.9***	No

* *p* <.05 ** *p* <.01 *** *p* <.001

FIGURE 2 Results of phoneme classification in Analysis 2. Average SVM performance on a two-way phoneme classification (chance = 50%, $N = 26$ subjects). Redder colors correspond to lower SVM accuracy; greener colors correspond to higher accuracy. Each row represents a separate classification job. Time window width was 150 ms, centered at 120 ms post-target word onset. Data were averaged across 3 folds, and a second-order polynomial was fit to the averaged data. SVM performance and corresponding t-values (from both 1-sample *t*-tests against numerical chance and 2-sample *t*-tests against randomly shuffled data) are shown in the columns. The results of Pereira et al. (2018)'s original analyses are shown in the final column. "Yes" indicates there was a significant difference in N1 amplitude for a given contrast, "No" indicates that there was not, and "NT" indicates that the specific contrast was not tested. The present analyses show that all phoneme contrasts were decodable significantly above chance.

approaches relate. This will be reported for the current analysis and for Analysis 3 (detailed below).

3.3 | Results

Results are shown in Figure 2. We found that all of the tested phoneme contrasts were decodable significantly above chance, regardless of whether or not N1 amplitude showed significant differences in Pereira et al. (2018). For the three phoneme pairs that differ in voicing, we found that /b/ vs. /p/ showed average SVM classification accuracy of 59.1% (*SE*: 1.0, $t_1(25) = 9.6$; $p < .001$; $t_2(25) = 7.9$, $p < .001$), /z/ vs. /s/ showed average SVM accuracy of 57.5% (*SE*: 1.1, $t_1(25) = 6.9$; $p < .001$; $t_2(25) = 6.5$, $p < .001$), and /ʤ/ vs. /ʧ/ showed average SVM accuracy of 53.8% (*SE*: 1.4, $t_1(25) = 2.8$; $p < .05$; $t_2(25) = 2.8$, $p < .01$). Thus, although decoding accuracy was not particularly high for these contrasts, it was consistently above chance, even for the phoneme contrast that did not yield a significant difference in mean N1 amplitude (/ʤ/ vs. /ʧ/).

Next, we examined the three phoneme pairs that differed in manner of articulation. Manner of articulation differences were not examined in Pereira et al. (2018). /b/ vs. /m/ had an average SVM accuracy of 57.6% (*SE*: 1.3, $t_1(25) = 5.7$; $p < .001$; $t_2(25) = 6.2$, $p < .001$), /z/ vs. /n/ had an average accuracy of 55.1% (*SE*: 1.0, $t_1(25) = 4.8$; $p < .001$; $t_2(25) = 5.3$, $p < .001$), and /t/ vs. /s/ had an average accuracy of 56.9% (*SE*: 1.2, $t_1(25) = 5.8$; $p < .001$; $t_2(25) = 5.7$, $p < .001$). Thus, as with phoneme contrasts differing in voicing, the overall accuracy for these phoneme pairs was consistently above chance.

Lastly, we ran classifiers for the three pairs that differed in place of articulation. /b/ vs. /d/ had an average accuracy of 56.3% (*SE*: 0.7; $t_1(25) = 8.9$; $p < .001$; $t_2(25) = 8.0$, $p < .001$), /z/ vs. /v/ had an average accuracy of 55.9% (*SE*:

1.3; $t_1(25) = 4.3$; $p < .001$; $t_2(25) = 4.9$, p < .001), and /ʃ/ vs. /s/ had an average accuracy of was 56.6% (*SE*: 0.9; $t_1(25) = 7.8$; $p < .001$; $t_2(25) = 7.9$, $p < .001$). Thus, again, accuracy was consistently above chance for all three phoneme contrasts, including the two contrasts that did not differ in mean N1 amplitude (/z/ vs. /v/ and /ʃ/ vs. /s/).

3.4 | Discussion

In this analysis, we trained SVMs to classify different pairs of phonemes. We show that we are able to decode phoneme differences significantly above chance during early speech perception. This was true regardless of whether or not conventional ERP methods detected a difference between these phonemes. In cases where differences were not detected by conventional methods, our results suggest that this is not because the information is unavailable in the neural signal. Rather, the information may be better reflected in more subtle, distributed patterns of voltages across the scalp. This indicates that machine learning offers a powerful alternative for examining neural responses to speech sounds, and potentially other stimuli, and that it can detect effects otherwise not seen in traditional analyses.

4 | ANALYSIS 3: TIMECOURSE OF PHONETIC FEATURE DECODING

4.1 | Introduction

Analysis 3 sought to expand on the previous analyses by characterizing the timecourse of decoding, running multiple classifiers over the epoch, in contrast to Analyses 1 and 2, which evaluated classifiers centered at a single time point. In the current analysis, we examine how decoding unfolds along the dimensions of voicing, manner, and place. Information along each of these dimensions must be quickly extracted from the acoustic signal for speech to be accurately understood. Analysis 2 showed that we can decode contrasts for specific phoneme pairs that differ along each of these dimensions. However, it remains unclear whether voicing, manner, and place are decodable as categories themselves. For example, manner and place of articulation differences might be decodable when examining specific phoneme contrasts, but may not be decodable as phonetic features more broadly. Moreover, the timecourse of feature decoding might be similar across these articulatory dimensions, or some features might be decoded earlier than others. Thus, Analysis 3 asks whether perceptual representations of speech are organized according to articulatory features along these dimensions more generally, and if so, how the timecourse of processing may differ among the three features.

Specifically, Analysis 3 examines differences in (1) how long voicing, manner, and place are represented in the neural signal, (2) how early in time we can decode each of these dimensions, and (3) how the timing of peak decoding accuracy differs among them.

Based on previous work (Toscano et al., 2010), we predict that the maximum decoding accuracy for voicing will occur at the peak of the N1 component, where we see the largest differences in ERP amplitude. Moreover, other work (Sarrett et al., 2020; Toscano et al., 2018) predicts that decoding duration may persist in time outside of the canonical N1 component peak.

The predicted timecourse for manner of articulation is less clear. There are six different manners of articulation for English consonants: stops (or plosives), nasals, fricatives, affricates, approximants, and lateral approximants (Reetz & Jongman, 2020). Manner is defined by articulatory differences in how the constrictions in the vocal tract interact with the airflow. However, differences in manner also have very distinct acoustic realizations. For example, fricatives typically have broad, aperiodic high frequency energy that is sustained for at least 50 ms (Jongman et al., 2000). Stops, on the other hand, consist of a brief burst (<10 ms) of energy that may be followed by a period of weaker aspiration (Lisker & Abramson, 1964). A systematic pattern of differences across manner at the auditory N1 is not readily apparent (Pereira et al., 2018). However, applying machine learning techniques may help reveal the degree to which differences in manner are detectable during perceptual encoding.

Finally, we investigate the timecourse of processing for place of articulation. In English, consonants vary among bilabial, labiodental, alveolar, postalveolar, and velar places of articulation (Ladefoged & Johnson, 2014). However, the acoustic features of the speech signal that are important for distinguishing place of articulation remain unclear. Listeners' perception of place distinctions has informed models of speech perception (Diehl & Kluender, 1989; Liberman & Mattingly, 1985; Nearey, 1990; Stevens & Blumstein, 1978; Sussman et al., 1991), and the nature of perceptual representations for place distinctions has long been a topic of debate. However, there has been little ERP work examining differences in place of articulation during early speech perception. This makes it difficult to predict whether place will be decodable from the EEG signal, and if so, what its timecourse may be.

To address these issues, we ran a different classifier for each dimension. We are most interested in characterizing the timecourses of perceptual encoding and identifying temporal differences in maximal decoding accuracy as well as duration of decoding.

Thus, based on Analysis 1, we used a three-way data split (averaging over many trials) and mean voltage at a small time window (10 ms) as input to the SVM.

4.2 | Method

4.2.1 | Machine learning techniques

As in Analyses 1 and 2, we used custom MATLAB scripts and the LibSVM package from Chang and Lin (2011). To decode voicing, we used the same two classes (voiced vs. voiceless) that were used in Analysis 1. The voiced class included trials from words beginning with any voiced consonants (/b, d, g, v, z, ʤ, m, n, ɹ, l, w/). The voiceless class included trials from words beginning with any voiceless consonants (p, t, k, f, s, ʃ, ʧ/). Chance level for this classifier was 50%. To decode manner, the data were split into four classes: stops, fricatives, affricates, and nasals.[8] The stops class included trials from words beginning with /b, d, g, p, t, k/, fricatives included /v, z, f, s, ʃ/, affricates included /ʤ, ʧ/, and nasals included /m, n/. Chance level was 25%. To decode place, the data were split into five classes: bilabial, labiodental, alveolar, postalveolar, and velar.[9] The bilabial class included trials from words beginning with /b, p, m/, labiodental included /v, f/, alveolar included /d, t, z, s, ɹ, l, n/, postalveolar included /ʤ, ʧ, ʃ/, and velar included /g, k/. Chance level was 20%.

For each of these three classifiers, the data at each time point were split into three sections (for *k*-fold cross-validation) and averaged. The number of trials in each class was equalized, as in Analysis 1. For the voicing distinction, this included averaging over approximately[10] 100 trials per class and per subject in each *k*-fold. For the manner and place distinctions, this involved averaging over approximately 20 trials per class and per subject for each *k*-fold. Input to the SVM included the zero-order polynomial (mean voltage) in a 10 ms time window across 26 electrodes.

As in Analyses 1 and 2, the free parameters of the SVM (C and γ) were optimized for each participant and for each classification job. This was done in two phases, since we were running multiple SVMs over the epoch to examine the timecourse. In the first phase, we determined the optimal C and γ parameters from a representative timepoint. This timepoint was 120 ms after target word onset (the peak of the N1 component). We then used the same hybrid approach to optimize C and γ as in Analysis 1, starting with a brute force search and then a finer gradient descent method. In the second phase, the C and γ from the representative timepoint were held constant, and an SVM was fit with these parameters across the full timecourse in 10 ms increments, from 100 ms pretarget word onset to 250 ms post-target word onset.

Cross-validation involved a 3 *k*-fold procedure as described in Analysis 1. Trials were randomly assigned to one of three folds and averaged. Each fold served equally as a training and testing set. The *k*-fold procedure was repeated 15 times for each classification job, to remove any idiosyncrasies in SVM performance due to random trial assignment. Average SVM performance across these repetitions is reported.

4.2.2 | Statistical approach

As in Analysis 2, we took two complementary statistical approaches: (1) one-sample *t*-tests compared to numerical chance, and (2) two-sample *t*-tests from a 100-repetition SVM on randomly shuffled data.

With this set of analyses, we were primarily interested in characterizing the timecourse of decoding. Thus, it was necessary to run multiple *t*-tests (at each timepoint) to determine when SVM classification performance was above chance. When running multiple comparisons, it is critical to control for family-wise error. Traditionally, this is done with a Bonferroni correction (Bonferroni, 1936), which assumes independent significance tests. However, significance tests in a timeseries are not entirely independent of each other, as test statistics over time are highly autocorrelated. This would make applying a Bonferroni correction overly conservative. Another common method is cluster-based permutation analysis (Maris & Oostenveld, 2007), which is less conservative than Bonferroni, but is unable to make precise claims about the timing of significant effects (Sassenhagen & Draschkow, 2019).

Here, we adopt an approach from Oleson et al. (2017), which takes into account the autocorrelation of the test statistic over time on a millisecond-by-millisecond basis to define a new alpha level (for applied examples, see Seedorff et al., 2018, using eye-tracking data, Sarrett et al., 2020, using ERP data, and McMurray et al., 2022, using machine learning performance over time). This allows us to make more precise claims about the timing of our effects of interest without being statistically over-conservative. For each comparison (voicing,

[8] Words beginning with approximates, lateral approximates, and glides (/w, ɹ, l/) were excluded from the manner analysis, as there was not a sufficient number of items per phoneme for these classes to be decoded.

[9] Words beginning with /w/ were excluded from the place analysis, as /w/ requires constrictions at multiple locations (bilabial and velar).

[10] Numbers are approximate, as the exact number of trials per class per subject varied due to artifact rejection.

manner, and place), we first smoothed each subject's classification performance over time using a 50 ms triangular window. This was done to remove any idiosyncrasies in the pattern of SVM performance from the data. Then, we ran *t*-tests (either one-sample against numerical chance, t_1, or two-sample against randomly shuffled data, t_2) at each time point from -100 to 250 ms seconds in 10 ms increments (for a total of 36 comparisons). The autocorrelation of the test statistics was calculated using the bdots package (version 1.0.1; Nolte et al., 2020) in R (version 4.1.0; R Core Team, 2022). Autocorrelation (ρ) and corrected alpha (α) are reported for each analysis below.

4.3 | Results

Figure 3 shows SVM accuracy over time for decoding a voicing distinction, from 100 ms before target word onset to 250 ms post-target word onset. SVM performance peaks at 66.88% at 140 ms post-target word onset. Both statistical approaches indicated that decoding was significantly above chance from 110 to 220 ms post-target word onset, after correcting for multiple comparisons (t_1: $\rho = .952$, corrected $\alpha = .009$; t_2: $\rho = .957$, corrected $\alpha = .008$). These results suggest that information about voicing is available in the neural signal for at least 100 ms during the perceptual encoding of speech sounds, consistent with previous work (Toscano et al., 2018).

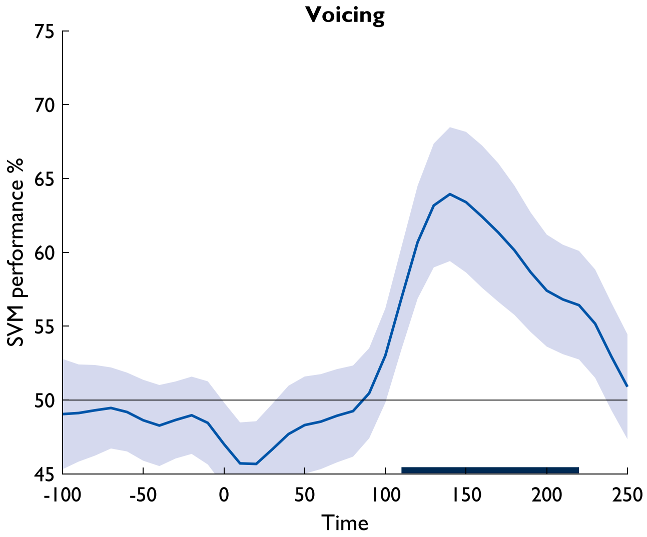


FIGURE 3 Average SVM accuracy over time for decoding voicing distinctions (voiced vs. voiceless). Chance level is 50%, marked with a black horizontal line. Standard error of the mean is shown by the shaded region. Time windows in which decoding is significantly above chance using a two-sample *t*-test against randomly shuffled data are marked by the heavy weighted blue line at the bottom of the figure. Note that the y-axis scale differs between this figure and Figures 4 and 5; in each figure the maximum y-axis value is 1.5× chance performance for that analysis.

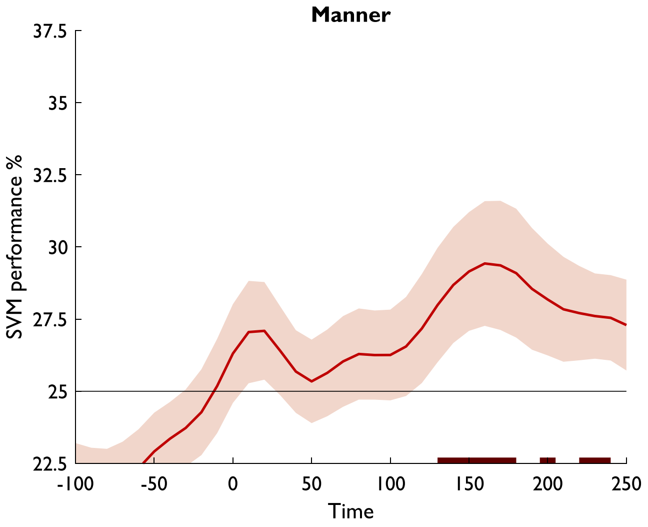


FIGURE 4 Average SVM accuracy for decoding manner. Figure properties are the same as Figure 3.

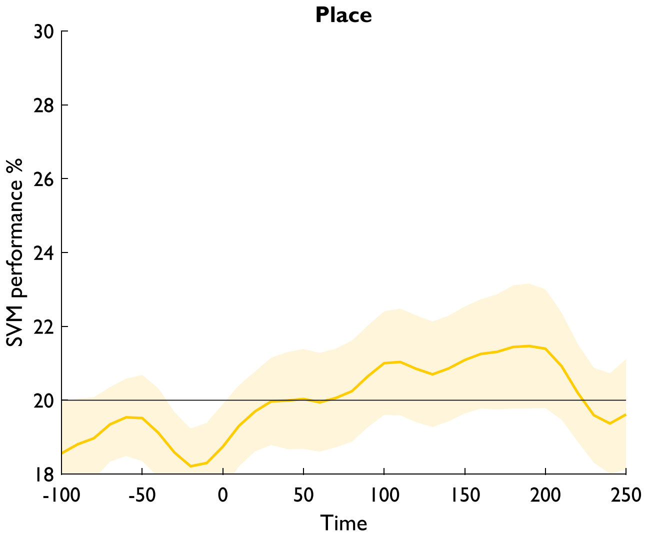


FIGURE 5 Average SVM accuracy for decoding place. Figure properties are the same as Figure 3.

Figure 4 shows SVM accuracy for decoding manner of articulation. Manner was decodable from roughly 140 to 240 ms post-target word onset, and maximal classification accuracy of 31.06% occurred at 170 ms. For the one-sample *t*-tests, decoding was above chance from 140 to 170 ms and from 220 to 240 ms ($\rho = .918$, corrected $\alpha = .007$). The two-sample *t*-tests against shuffled data showed significant decoding from 140 to 180 ms, at 200 ms, and from 220 to 240 ms ($\rho = .921$, corrected $\alpha = .008$). These results indicate that manner of articulation is represented in neural activity for at least 100 ms during speech perception.

Figure 5 shows average SVM accuracy over time for place of articulation. Place was not decodable significantly above chance for either statistical approach, after correcting for multiple comparisons (t_1: $\rho = .899$, corrected $\alpha = .007$; t_2: $\rho = .895$, corrected $\alpha = .007$).

4.4 | Discussion

Analysis 3 aimed to characterize the timecourse of decoding accuracy for voicing, manner, and place. We showed that both voicing and manner are decodable from scalp-level ERPs, but place was not. It is difficult to directly compare among conditions due to differences in the SVMs themselves (e.g., different number of classes in each job). Nevertheless, we discuss some potential implications of the pattern of results.

First, we found differences in the duration over which information about voicing and manner was represented in cortical activity. It is not entirely clear what differences in duration of decoding or timing of maximum accuracy across these dimensions may mean. Minimally, they indicate that information along a given dimension is represented in neural activity for differing amounts of time, but the reason underlying these differences remains unclear. One potential interpretation of this pattern is that duration of decoding may link to the robustness of the representation of a given dimension in neural activity. Because both voicing and manner were decodable for approximately 100 ms in cortical activity, this suggests that each is robustly represented during speech perception.

In contrast, place of articulation was not significantly decodable above chance. Although all three dimensions are defined by their articulatory differences, manner and voicing both have clear acoustic distinctions that map onto their different classes. Place, however, has less clear acoustic correlates (though see Stevens & Blumstein, 1978, which maps spectral shape to place differences for stop consonants). This suggests that place of articulation, defined in terms of articulatory differences, may be only weakly represented in cortical activity, if at all. Longstanding debates have centered on the nature of perceptual representations and whether they are primarily auditory or primarily articulatory in nature (Diehl & Kluender, 1989; Liberman & Mattingly, 1985). The current results show that place of articulation, which has the weakest link to reliable auditory features, was also not reliably decoded. This may indicate that perceptual representations of speech are more closely linked to aspects of auditory encoding than to aspects of gestural features, but further work is needed to clarify these results.

Second, the dimensions differed slightly in the times at which they were decodable from neural activity. Manner was decodable starting at 140 ms, whereas voicing was decodable earlier, at 100 ms. One possible interpretation of this incongruency is that the timing of window onset and offset may reflect when cues to each dimension are available in the acoustic signal. The primary cue to voicing for stop consonants—VOT—unfolds early, which may lead to the earlier decoding of voicing. Cues to manner, however, are more varied, both acoustically and temporally. Some spectral cues, such as periodicity, are available essentially immediately in the acoustic signal, but other timing-based cues, such as duration of noise, arrive later (Reetz & Jongman, 2020). This may require cue integration across time before a reliable interpretation of the acoustic signal can be made. As a result, these cues may be computed and reflected in neural activity more slowly relative to cues to voicing.

Third, these dimensions differed in when maximal decoding accuracy occurred. Voicing peaked around 140 ms; manner peaked around 170 ms. These differences in the timing of peak accuracy suggest some temporal asynchrony in neural processing of these dimensions, though the timecourses did show substantial overlap. Such differences in timing could occur because different dimensions of the speech signal are coded by different populations of neurons, which may be actively processing perceptual representations during overlapping, but not congruent, time windows. It is also possible that maximum decoding is influenced by the reliability of cues along a given dimension, though accuracy is difficult to directly compare among these classification jobs for a number of reasons. We return to this point in the General Discussion.

5 | GENERAL DISCUSSION

Machine learning techniques, while not new in cognitive science, have been relatively underutilized for analyzing EEG data. These techniques offer exciting new avenues for research and novel ways of designing experiments to address outstanding issues, not just about the neural processing of spoken language, but questions in other domains as well. We presented three analyses that assess methodological considerations involved in the application of such techniques to EEG data. We sought to unite approaches from different lines of work by systematically manipulating parameter input to the SVM to characterize how to best maximize machine learning accuracy (Analysis 1). We also applied these techniques to unresolved questions about the nature of perceptual representations of speech sounds, measuring differences in the neural signal that are not detectable with conventional approaches (Analysis 2) and assessing the timecourse of processing (Analysis 3). Below, we discuss the results from each analysis, synthesize the overall implications, and suggest directions for future research.

5.1 | Methodological considerations for machine learning analyses

In Analysis 1, we explored how the features of the EEG signal that the classifier is trained on affect classification performance. The goal of this analysis was to bring together different traditions in the application of machine learning techniques to neural data. These traditions differ in how a classifier is trained on a given data set, such as whether to use individual-trial data or averaged-trial data, data from individual time points or data averaged over a large time window, and whether to fit a polynomial to the data or not. Each of these factors (averaging more trials, averaging more time points, or fitting a polynomial to capture the data pattern) could influence the ability to detect differences in the neural signal.

We found that each dimension affected classification accuracy. Generally, averaging over a greater number of trials increased SVM performance, particularly for lower-order polynomials. The time window that resulted in the best performance depended on the order of the polynomial: the zero-order polynomial performed best with the shortest time windows, the first-order polynomial with the mid-length time windows, and the second-order polynomial with the longest time windows. This is consistent with the way in which the polynomials captured patterns in the data.

The dimension of time is particularly important for speech perception work, as spoken language unfolds quickly over time, and many debates in speech perception hinge on *when* in time different stages of linguistic processing unfold (e.g., when and at what level does linguistic information feedback during speech perception; Getz & Toscano, 2019; Noe & Fischer-Baum, 2020; Sarrett et al., 2020). Thus, an important consideration for researchers will be choosing the time window over which a classifier is trained. Using longer time windows, which is necessary to achieve above chance performance for some higher-order polynomials and for individual-trial SVMs, may make it more challenging to make precise claims about the millisecond-level timing of effects of interest. Thus, such parameter choices will need to be made carefully and appropriately for each research question.

Some researchers may be interested in not just the ERP responses from EEG, but also in event-related band power (ERBP) changes. These machine learning techniques can, in principle, be applied to time-frequency analyses as well, and there are a number of studies that have included various ERBP changes in their feature input to classifiers.

However, these studies have shown mixed results as to whether ERBP measures increase overall classification accuracy. McMurray et al. (2022), for example, did not find that adding mean ERBP increased classification accuracy for auditory words in any of the tested bands (delta, theta, alpha, beta, gamma), though this may depend heavily on the stimuli or task decisions that are being classified. In contrast, Bae and Luck (2019) did show classification accuracy significantly above chance when using alpha band power as input, and alpha may reflect a dissociable representation from ERPs in visual perception. It is important to note that filtering settings used during data preprocessing will affect which frequencies are available at the scalp. Similarly, while the high gamma band has been shown in intracranial work to carry critical speech information, there is still debate over whether meaningful high gamma information is measurable at the scalp or whether it is largely blocked by the dura, skull, and skin (though some have attempted to measure it, with some success; see Synigal et al., 2020).

Researchers may also be interested in determining the maximum level of classification accuracy that these machine learning methods can achieve. Across all three analyses, we found that maximum accuracy for a given classification job was around 70%. There are several potential reasons for this. One possibility is that there may have been an insufficient amount of data for the classifier to achieve any higher accuracy. Designing an experiment with a greater number of trials may allow for gains in classification accuracy (as suggested by McMurray et al., 2022). In addition, the spatial smearing of the neural signal in scalp EEG could have resulted in there not being enough information available in the signal at the scalp level to achieve a higher accuracy.

At present, we cannot make claims about why we observed such a ceiling in classification performance, and it may be due to either or both of these factors. However, even studies using iEEG data—which avoids the issue of spatial smearing—do not always show 100% classification accuracy. For example, Rhone et al. (submitted) showed a maximum SVM accuracy of ≈40% for decoding the word a listener was hearing, based on recordings from auditory cortex, with a chance level of 10%.

While classification accuracy in the present study did not reach 100%, we did achieve classification accuracy significantly above chance. For some applications, this may not be adequate. However, adequate performance critically depends on the goals of a particular machine learning application. For BCIs, 100% classification accuracy may be the goal (e.g., for synthesizing movements of an artificial limb from a neural prosthesis). However, when using machine learning as a replacement for traditional hypothesis testing, the goal may simply be to decode significantly above chance. In this case, 70% accuracy on a binary classification indicates that some information about the stimulus is present in the neural signal (and therefore decodable). Because the data set used in this analysis was not designed specifically for machine learning techniques, it is also possible that we can increase classification accuracy with a more tailored design. For example, the number of trials per class could be increased to at least 50, or the number of electrodes per subject could be increased to 64, both of which have been shown to yield gains classification performance (McMurray et al., 2022). That is, while the current data set showed a ceiling of around 70%, this may not be the limit for speech sound decoding. This exploration should be a goal of future work.

Finally, we also compared two different statistical approaches in Analyses 2 and 3, looking at (1) SVM performance compared to numerical chance, and (2) SVM performance compared to randomly shuffled data. The latter is the more accepted and more technically appropriate comparison. However, the former comes with a much lower computational cost. We found that comparing SVM performance to numerical chance yielded generally similar results, thus making it a potentially suitable alternative. As machine learning becomes a more widely used technique, it is important to ensure that it is accessible for researchers at a range of institutions who wish to work with this method, particularly for those who might not have a high-performance computing cluster. The present analyses indicate that one way to cut down on computing costs is by using numerical chance as a baseline for statistical comparisons.

5.2 | Theoretical considerations for models of speech perception

The current results also provide insights into the neural representations used to perceive speech. In Analysis 2, we looked at classification accuracy for specific phoneme pairs. We found that machine learning offers a more sensitive measure than traditional ERP analyses: Phoneme pairs that did not differ in N1 amplitude were decodable significantly above chance when taking advantage of the full pattern of activity across the scalp.

Such phoneme classification analyses could be expanded in future work to relate individual-trial classification to participant task responses, allowing us to expand the types of questions that we can address. Some studies have looked at classifier confusions in passive listening tasks (Mesgarani et al., 2014), but few have related these directly to participant behavior (Beach et al., 2021). A critical future direction will involve examining the link between decoding of perceptual representations and category-level decisions, including errors (e.g., by using a phoneme categorization task and relating participant response confusion matrices to SVM confusion matrices). Some phonemes are more difficult to recognize than others (e.g., /θ/; Toscano & Allen, 2014). For these sounds, and in cases when there is greater difficulty in perceiving phoneme distinctions (e.g., in speech-in-noise tasks or for listeners with hearing loss), listeners make more errors in their categorization responses. It is possible that these response errors correlate with less precise perceptual encoding of these sounds, which may show up as a greater number of SVM confusions, and lower overall SVM accuracy. Thus, machine learning may be able to help us compare differences in perceptual representations across different phonemes, between varying task types, and among clinical populations.

Finally, in Analysis 3, we examined the timecourse of decoding for voicing, manner of articulation, and place of articulation. The goal here was to apply what was learned in Analysis 1 to address questions about the nature of perceptual representations of speech sounds as they unfold over time. We found that voicing and manner—but not place—were decodable from the EEG signal during early speech perception. There are several reasons why we may not have observed an effect of place in this analysis. First, in contrast to Analysis 2, this analysis did not compare specific phoneme pairs, but attempted to decode an overall effect of place. Thus, the effects observed in Analysis 2 may be due to differences in acoustic cues that distinguish particular pairs of phonemes, rather than differences in place of articulation per se. Second, this analysis did not attempt to control for differences in manner when decoding place, which could have diminished our ability to detect an effect of place, as these two dimensions are not completely independent of each other; other studies have found effects of place by looking at cross-generalization across manner of articulation (Archila-Meléndez et al., 2018; Correia, Jansma, & Bonte, 2015). Third, our analysis focused on early perceptual representations. Other work (Archila-Meléndez et al., 2018; Correia, Jansma, & Bonte, 2015) has found that place is decodable using fMRI, but these effects were most consistent in higher-level associative language processing regions. The results were more mixed when decoding from early perceptual processing regions, such as Heschl's gyrus, planum temporale, and superior temporal gyrus. Thus, there may be representations organized around place of articulation during spoken language comprehension, but these representations may become available at a later stage of processing.

For the effects of manner and voicing, we found that maximum accuracy, onset of decodability, and decoding duration varied between these two dimensions. Differences in these measures could arise for a number of reasons. One possibility is that the reliability of cues along a given dimension contributes to how well that dimension is decoded.

There are often multiple acoustic cues that signal category-level differences (Lisker, 1986). Some cues are more informative than others, and listeners learn to weight these cues accordingly (Toscano & McMurray, 2010). More reliable cues may be more decodable from cortical activity, as the perceptual system has become attuned to using information along those dimensions. Future work could test effects of cue reliability by looking at perceptual processing in cases where the reliability of certain cues differs, for example, across talkers (Clayards, 2017; Kleinschmidt, 2019).

The timing of above chance performance is another measure that could shed light on perceptual representations along these phonetic feature dimensions. We observed that voicing peaked earlier than manner and that voicing was decodable earlier in the epoch. This pattern may be explained in a number of ways. One possibility is that differences in timing of maximal accuracy reflect when cues to each dimension are available in the acoustic signal, or alternatively, how quickly information along each dimension is extracted from the signal.

It remains difficult to compare results across different phonological feature dimensions directly, particularly maximum accuracy differences. Because there are differences in the number of classes, chance levels are not the same for each type of distinction. This could affect SVM performance overall (i.e., distinguishing a greater number of classes may be a more difficult classification job). Indeed, it may have been a partial contributor to the lack of significant decoding for place of articulation in Analysis 3. These issues could be addressed to some extent, for example, by only examining two-way distinctions for each feature or by examining specific pairs of phonemes, as in Analysis 2 (which did find that certain place distinctions are decodable). However, there are also limitations imposed by the naturally-occurring linguistic features of the speech sounds, such as which phoneme contrasts exist in a given language. Future work may be able to address these issues further by comparing contrasts in other languages that have more complex phonetic category structures along some of these dimensions (e.g., with Korean, Hindi, or some indigenous languages of the Americas, which each have more voicing contrasts than English; Davis, 1994; Gordon et al., 2000, 2001; Kim & Lotto, 2002; Lisker & Abramson, 1964; McDonough et al., 1992). Further studies will be necessary to better characterize what the various aspects of the timecourse of decoding—maximum decoding accuracy, duration of decoding, and decoding time window—reflect in terms of cognitive processing.

5.3 | Conclusions

The current study addresses methodological considerations in applying machine learning techniques to neurophysiological data and also offers theoretical insights into the nature of perceptual representations of speech. We show the importance of selecting which features of the EEG signal to use as input to the classifier (Analysis 1), we demonstrate that some limitations of traditional ERP analyses can be overcome using machine learning approaches (Analysis 2), and we show that perceptual encoding is sensitive to the phonetic features of voicing and manner, characterizing the timecourse of processing for these features (Analysis 3). Taken together, these analyses provide the groundwork for the continuing development and application of machine learning techniques with EEG data.

AUTHOR CONTRIBUTIONS

McCall E. Sarrett: Conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing. **Joseph C. Toscano:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; writing – review and editing.

ACKNOWLEDGMENTS

This material is based on upon work supported by the National Science Foundation under Grant No. 1945069. This work used the Augie High-Performance Computing cluster, funded by NSF Grant No. 2018933, at Villanova University. We would like to thank Bob McMurray for valuable discussions regarding code implementation and Mark Pitt for helpful comments on the initial version of the manuscript.

DATA AVAILABILITY STATEMENT

Code will be made available via Open Science Framework at osf.io/2mby6.

ORCID

McCall E. Sarrett https://orcid.org/0000-0002-1842-7307

REFERENCES

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, *106*(4), 2031–2039. https://doi.org/10.1121/1.427949

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187. https://doi.org/10.1016/0010-0277(94)90042-6

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493–498. https://doi.org/10.1038/s41586-019-1119-1

Archila-Meléndez, M. E., Valente, G., Correia, J. M., Rouhl, R. P., van Kranen-Mastenbroek, V. H., & Jansma, B. M. (2018). Sensorimotor representation of speech perception. Cross-decoding of place of articulation features during selective attention to syllables in 7t fMRI. *Eneuro*, *5*(2), e0252-17.2018 1-12. https://doi.org/10.1523/ENEURO.0252-17.2018

Bae, G.-Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bae, G.-Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage*, *184*, 242–255. https://doi.org/10.1016/j.neuroimage.2018.09.029

Bayet, L., Zinszer, B. D., Reilly, E., Cataldo, J. K., Pruitt, Z., Cichy, R. M., Nelson, C. A., III, & Aslin, R. N. (2020). Temporal dynamics of visual representations in the infant brain. *Developmental Cognitive Neuroscience*, *45*, 100860. https://doi.org/10.1016/j.dcn.2020.100860

Beach, S. D., Ozernov-Palchik, O., May, S. C., Centanni, T. M., Gabrieli, J. D., & Pantazis, D. (2021). Neural decoding reveals concurrent phonemic and subphonemic representations of speech across tasks. *Neurobiology of Language*, *2*, 1–62. https://doi.org/10.1162/nol_a_00034

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect1084 perception of voicing in English. *Journal of Phonetics*, *29*(1), 1–22. https://doi.org/10.1006/jpho.2000.0128

Boersma, P. (2006). Praat: Doing phonetics by computer. http://www.praat.org/.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3–62.

Boynton, G. M. (2005). Imaging orientation selectivity: Decoding conscious perception in V1. *Nature Neuroscience*, *8*(5), 541–542. https://doi.org/10.1038/nn0505-541

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432. https://doi.org/10.1038/nn.2641

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *Journal of the Acoustical Society of America*, *136*(5), 2762–2772.

Clayards, M. (2017). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, *75*(1), 1–23. https://doi.org/10.1121/1.4896470

Connine, C. (1990). Effects of sentence context and lexical knowledge in speech processing. In *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 281–294). The MIT Press.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, *26*(5), 527–538. https://doi.org/10.1016/0749-596X(87)90138-0

Correia, J. M., Jansma, B., Hausfeld, L., Kikkert, S., & Bonte, M. (2015). EEG decoding of spoken words in bilingual listeners: From words to language invariant semantic-conceptual representations. *Frontiers in Psychology*, *6*, 71. https://doi.org/10.3389/fpsyg.2015.00071

Correia, J. M., Jansma, B. M., & Bonte, M. (2015). Decoding articulatory features from fmri responses in dorsal speech regions. *Journal of Neuroscience*, *35*(45), 15015–15025. https://doi.org/10.1523/JNEUROSCI.0977-15.2015

Davis, K. (1994). Stop voicing in Hindi. *Journal of Phonetics*, *22*(2), 177–193. https://doi.org/10.1016/S0095-4470(19)30192-5

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144. https://doi.org/10.1207/s15326969eco0102_2

Dietterich, T. G., & Bakiri, G. (1994). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, *2*, 263–286. https://doi.org/10.1613/jair.105

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. *International Journal of Computational Intelligence and Applications*, *1*, 335–339.

Fischer-Jørgensen, E. (1954). Acoustic analysis of stop consonants. *Le Maître Phonétique*, *32*, 42–59. https://doi.org/10.1121/1.1908634

Frazier, P. I. (2018). A tutorial on bayesian optimization, arXiv preprint. https://doi.org/10.48550/arXiv.1807.02811

Frye, R. E., Fisher, J. M., Coty, A., Zarella, M., Liederman, J., & Halgren, E. (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience*, *19*(9), 1476–1487. https://doi.org/10.1162/jocn.2007.19.9.1476

Galle, M. E., Klein-Packard, J., Schreiber, K., & McMurray, B. (2019). What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. *Cognitive Science*, *43*(1), e12700. https://doi.org/10.1111/cogs.12700

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science*, *30*(6), 830–841. https://doi.org/10.1177/0956797619841813

Getz, L. M., & Toscano, J. C. (2021). The time-course of speech perception revealed by temporally-sensitive neural measures. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(2), e1541. https://doi.org/10.1002/wcs.1541

Gordon, M., Munro, P., & Ladefoged, P. (2000). Some phonetic structures of Chickasaw. *Anthropological Linguistics*, *42*(3), 366–400. https://www.jstor.org/stable/30028763

Gordon, M., Potter, B., Dawson, J., De Reuse, W., & Ladefoged, P. (2001). Phonetic structures of Western Apache. *International Journal of American Linguistics*, *67*(4), 415–448. https://www.jstor.org/stable/1265755

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227. https://doi.org/10.3758/APP.72.5.1218

Jakobson, R., Fant, C. G., & Halle, M. (1953). Preliminaries to speech analysis: The distinctive features and their correlates. *Language*, *29*, 472–481.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, *108*(3), 1252–1263. https://doi.org/10.1121/1.1288413

Kapnoula, E. C., & McMurray, B. (2021). Idiosyncratic use of bottom-up and top-down information leads to differences in speech perception flexibility: Converging evidence from ERPs and eye-tracking. *Brain and Language*, *223*, 105031. https://doi.org/10.1016/j.bandl.2021.105031

Karal, Ö. (2020). Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1–5). IEEE.

Kim, M. R., & Lotto, A. J. (2002). An investigation of acoustic characteristics of Korean stops produced by non-heritage learners. *The Korean Language in America*, *7*, 177–187.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language Cognition and Neuroscience*, *34*(1), 43–68. https://doi.org/10.1121/1.4787210

Ladefoged, P. (1996). *Elements of acoustic phonetics*. University of Chicago Press.

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America*, *33*(4), 419–425. https://doi.org/10.1121/1.1908681

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/versus /p/ in trochees. *Language and Speech*, *29*(1), 3–11. https://doi.org/10.1177/002383098602900102

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT Press.

Luthra, S., Correia, J. M., Kleinschmidt, D. F., Mesite, L., & Myers, E. B. (2020). Lexical information guides retuning of neural patterns in perceptual learning for speech. *Journal of Cognitive Neuroscience*, *32*(10), 2001–2012. https://doi.org/10.1162/jocn_a_01612

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McDonough, J., Ladefoged, P., & George, H. (1992). Navajo vowels and universal phonetic tendencies. *Journal of the Acoustical Society of America*, *92*(4), 2416. https://doi.org/10.1121/1.404686

McMurray, B. (2021). Categorical perception: Lessons from an enduring myth. *Journal of the Acoustical Society of America*, *149*(4), A33. https://doi.org/10.1121/10.0016614

McMurray, B., Sarrett, M. E., Chiu, S., Black, A. K., Wang, A., Canale, R., & Aslin, R. N. (2022). Decoding the temporal dynamics of spoken word and nonword processing from EEG. *NeuroImage*, *260*, 119457. https://doi.org/10.1016/j.neuroimage.2022.119457

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/sicence.1245994

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*(3), 347–373. https://doi.org/10.1016/S0095-4470(19)30379-1

Noe, C., & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition*, *197*, 104162. https://doi.org/10.1016/j.cognition.2019.104162

Nolte, C., Seedorff, M., Oleson, J., Brown, G., Cavanugh, J., & McMurray, B. (2020). *bdots: Bootstrapped differences of timeseries*. R package version 1.0.1. https://doi.org/10.1016/j.jml.2018.05.004

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, *26*(6), 2708–2725. https://doi.org/10.1177/0962280215607411

O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience*, *17*(4), 580–590. https://doi.org/10.1162/0898929053467550

Pereira, O., Gao, Y. A., & Toscano, J. C. (2018). Perceptual encoding of natural speech sounds revealed by the N1 event-related potential response. *Auditory Perception & Cognition*, *1*(1–2), 112–130. https://doi.org/10.1080/25742442.2018.1545106

Pessoa, L., & Padmala, S. (2005). Quantitative prediction of perceptual decisions during near-threshold fear detection. *Proceedings* *of the National Academy of Sciences*, *102*(15), 5612–5617. https://doi.org/10.1073/pnas.0500566102

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, *55*(2), 328–333. https://doi.org/10.1121/1.1914506

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83. https://doi.org/10.1037/0033-295X.113.1.57

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966. https://doi.org/10.1126/science.1117645

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reetz, H., & Jongman, A. (2020). *Phonetics: Transcription, production, acoustics, and perception*. John Wiley & Sons.

Rhone, A., Farris-Trimble, A., Nourski, K., Kawasaki, H., Howard, M. A., III, & McMurray, B. (submitted). Neural decoding reveals the functional anatomy of auditory integration and competition in speech perception. https://doi.org/10.31234/osf.io/bd6eh

Sarrett, M. E., McMurray, B., & Kapnoula, E. C. (2020). Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and sentential expectations. *Brain and Language*, *211*, 104875. https://doi.org/10.1016/j.bandl.2020.104875

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), e13335. https://doi.org/10.1111/psyp.13335

Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries differ: Using the bootstrapped differences of timeseries (BDOTS) to analyze visual world paradigm data (and more). *Journal of Memory and Language*, *102*, 55–67. https://doi.org/10.1016/j.jml.2018.05.004

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *Journal of the Acoustical Society of America*, *108*(6), 3030–3035. https://doi.org/10.1121/1.1320474

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. *30*). MIT Press.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368. https://doi.org/10.1121/1.382102

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, *90*(3), 1309–1325. https://doi.org/10.1121/1.401923

Synigal, S. R., Teoh, E. S., & Lalor, E. C. (2020). Including measures of high gamma power can improve the decoding of natural speech from EEG. *Frontiers in Human Neuroscience*, *14*, 130. https://doi.org/10.3389/fnhum.2020.00130

Toscano, J. C., & Allen, J. B. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2293–2307. https://doi.org/10.1044/2014_JSLHR-H-13-0244

Toscano, J. C., Anderson, N. D., Fabiani, M., Gratton, G., & Garnsey, S. M. (2018). The time-course of cortical responses to speech revealed by fast optical imaging. *Brain and Language*, *184*, 32–42. https://doi.org/10.1016/j.bandl.2018.06.006

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301. https://doi.org/10.3758/s13414-012-0306-z

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*(10), 1532–1540. https://doi.org/10.1177/0956797610384142

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. *NeuroImage*, *277*, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

Vandermosten, M., Correia, J., Vanderauwera, J., Wouters, J., Ghesquière, P., & Bonte, M. (2017). Brain activity patterns of phonemic representations are atypical in beginning readers with family risk for dyslexia. *Developmental Science*, *23*(1), e12857.

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, *593*(7858), 249–254. https://doi.org/10.1038/s41586-021-03506-2

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., & Henderson, J. M. (2023). A high-performance speech neuroprosthesis. *Nature*, *620*, 1–6. https://doi.org/10.1038/s41586-023-06377-x

How to cite this article: Sarrett, M. E., & Toscano, J. C. (2024). Decoding speech sounds from neurophysiological data: Practical considerations and theoretical implications. *Psychophysiology*, *61*, e14475. https://doi.org/10.1111/psyp.14475