Published in final edited form as:
Neuroimage. 2022 October 15; 260: 119457. doi:10.1016/j.neuroimage.2022.119457.

Decoding the temporal dynamics of spoken word and nonword processing from EEG

Bob McMurray[a,*], **McCall E. Sarrett**[b], **Samantha Chiu**[c], **Alexis K. Black**[d], **Alice Wang**[e], **Rebecca Canale**[f], **Richard N. Aslin**[g]

[a]Dept. of Psychological and Brain Sciences, Dept. of Communication Sciences and Disorders, Dept. of Linguistics and Dept. of Otolaryngology, University of Iowa

[b]Interdisciplinary Graduate Program in Neuroscience, Unviersity of Iowa

[c]Dept. of Psychological and Brain Sciences, University of Iowa

[d]School of Audiology and Speech Sciences, University of British Columbia, Haskins Laboratories

[e]Dept. of Psychology, University of Oregon, Haskins Laboratories

[f]Dept. of Psychological Sciences, University of Connecticut, Haskins Laboratories

[g]Haskins Laboratories, Department of Psychology and Child Study Center, Yale University, Department of Psychology, University of Connecticut

Abstract

The efficiency of spoken word recognition is essential for real-time communication. There is consensus that this efficiency relies on an implicit process of activating multiple word candidates that compete for recognition as the acoustic signal unfolds in real-time. However, few methods capture the neural basis of this dynamic competition on a msec-by-msec basis. This is crucial for understanding the neuroscience of language, and for understanding hearing, language and cognitive disorders in people for whom current behavioral methods are not suitable. We applied machine-learning techniques to standard EEG signals to decode which word was heard on each trial and analyzed the patterns of confusion over time. Results mirrored psycholinguistic findings: Early on, the decoder was equally likely to report the target (e.g., *baggage*) or a similar sounding competitor (*badger*), but by around 500 msec, competitors were suppressed. Follow up analyses show that this is robust across EEG systems (gel and saline), with fewer channels, and with fewer

[*]Corresponding Author: Bob McMurray, 278 PBSB, Dept. of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242. Bob-mcmurray@uiowa.edu (B. McMurray).

Author contributions
The project was conceptualized and led by McMurray and Aslin based on a grant written by both. The experiment was designed by McMurray, Aslin, Sarrett, Black and Chiu. Data was collected by Chiu, Sarrett, Black, Canale and Wang. The EEG pipeline was built by Sarrett and Chiu. Machine learning was developed by McMurray, Chiu and Sarrett with consultation from Aslin and Black. Statistical analysis was conducted by McMurray and Sarrett. Results were interpreted by McMurray, Aslin, Black, Sarrett and Chiu. The initial draft of the paper was written by McMurray and Aslin. All authors contributed to the final manuscript.

Classifications
Social Sciences/Psychological and Cognitive Sciences; Biological Sciences/Neuroscience.

Data/code availability
The machine learning code for this manuscript is available at https://osf.io/wa3qr/. Raw data was too large for posting, but a single subject is available for testing the code. The full dataset is available upon request to the first author.

trials. Results are robust within individuals and show high reliability. This suggests a powerful and simple paradigm that can assess the neural dynamics of speech decoding, with potential applications for understanding lexical development in a variety of clinical disorders.

Keywords

Spoken Word Recognition; Speech Decoding; EEG; Machine Learning

1. Introduction

To accurately perceive speech, listeners must solve a fundamental challenge created by the fact that spoken language input unfolds over time. This is true at all levels: the acoustic cues that comprise a phoneme are often spread throughout the word (Galle et al., 2019; McMurray et al., 2008), words comprise multiple phonemes (Marslen-Wilson, 1987), and the meaning of a sentence or section of discourse must be assembled across words and phrases. At the fine-grained level of words and phonemes, this process of temporal integration involves two components. First, the temporal unfolding of the auditory signal creates periods of temporary ambiguity when the earliest portions of a word are insufficient to identify it (e.g., after hearing *be-* the word could be *beaker* or *beetle*). Therefore, word recognition fundamentally requires the auditory system to integrate prior material with some form of memory (e.g., *-ker* only uniquely specifies a word once it is integrated with the prior *bea-*). Second, at each step, fine-grained acoustic information must be used to update the decision set (McMurray et al., 2002). This requires both auditory fidelity and perceptual compensation mechanisms that interpret the variable auditory signal relative to differences across talkers, and contexts (McMurray & Jongman, 2011), as well as a rapid use of this new information to adjust the strength of commitment (activation) to current lexical candidates.

While this extended form of auditory integration is crucial for many aspects of speech and non-speech processing, the present manuscript focuses on the key domain of spoken word recognition, where cognitive science offers clear models and methods (Dahan & Magnuson, 2006). Even in this narrower domain, it is still unclear how the brain solves this auditory integration problem, which is fundamental for recognizing phonemes and words. Advances in multivariate approaches to neuroimaging have identified a network of structures involved in word recognition (Prabhakaran et al., 2006; Righi et al., 2009; Zhuang et al., 2011) and have demonstrated that this process involves both predictive mechanisms, that build and evaluate expectations about upcoming sounds (Blank & Davis, 2016; Gagnepain et al., 2012), and activation mechanisms that accumulate evidence for candidates which compete with each other for recognition (Brodbeck, Hong, et al., 2018; Kocagoncu et al., 2017).

However, despite these advances, existing methods offer only indirect ways to capture the timecourse of auditory integration as they do not directly assess the real-time decisions that unfold in the neural substrate as speech is recognized. As we argue here, this is important because an explosion of work on development (Rigler et al., 2015), clinical language and hearing disorders (Desroches et al., 2006; McMurray et al., 2017; McMurray et al., 2010), and challenging listening conditions (Brouwer & Bradlow, 2015; Hendrickson et al., 2020) suggests that the timecourse of this auditory integration varies along multiple dimensions (see Apfelbaum et al., 2022, for a partial review). This raises the need for a neural assay that can more precisely characterize the temporal dynamics of speech perception.

This manuscript presents a first step in this direction by introducing and validating a machine learning approach that is applied to EEG and fills these criteria. Although similar machine learning methods have been applied to EEG data to decode visual stimuli (Cichy et al., 2015; Bae & Luck, 2018), the specific implementation we provide for decoding speech stimuli is a novel elaboration on these prior demonstrations. Specifically, we introduce a new set of EEG features and a novel approach for decoding – and interpreting the decoding results – as the speech signal unfolds over time. While we do not argue that this fully captures word recognition, the processes of auditory evidence accumulation and decision making that we are attempting to capture is well worked out for spoken words, as are many of the developmental and clinical concerns that raise the need for this approach. We thus start with a short discussion of the cognitive science and cognitive neuroscience of spoken word recognition.

1.1. The Cognitive Science of Auditory Integration in Word Recognition

Word recognition fundamentally requires listeners to build representations across large swaths of time in the auditory input and can be characterized by a dynamically unfolding decision among multiple candidates. Cognitive science offers clear mechanistic models of this process (Hannagan et al., 2013; McClelland & Elman, 1986). Such models agree that word recognition is characterized by competition (Fig. 1A): as soon as any portion of the input is heard, listeners activate a variety of candidates to the degree that they match the unfolding input. These may include onset competitors (cohorts, such as *beetle* while hearing *beaker*) and offset competitors (rhymes, such as *speaker*) (Allopenna et al., 1998). The strength with which these candidates are activated is affected by higher level and contextual factors such as word frequency (Marslen-Wilson, 1987), or sentential context (Dahan & Tanenhaus, 2004). Candidates inhibit each other (Dahan et al., 2001) until a winner emerges and the word is ultimately recognized. Such competition models are relevant to all areas of language comprehension (Elman & McClelland, 1986; MacDonald et al., 1994), suggesting word recognition can serve as a model system for understanding these time extended integration processes.

Competition accounts of word recognition have been built in part on results from psycholinguistic methods like the Visual World Paradigm (VWP) (Tanenhaus et al., 1995). In this task, listeners hear words and select the corresponding picture from a small array of visually depicted options (usually pictures) representing candidates that may compete for recognition (e.g., for a target word, *beaker,* pictures may include *beetle* and *speaker*). Listeners must execute a series of eye-movements to locate the correct picture. These eye-movements are launched during processing, and thus can reveal the degree to which specific classes of candidates are considered with millisecond precision (Fig. 1B).

The VWP has proven to be critical for revealing subtle patterns of deficits associated with language and hearing disorders (Desroches et al., 2006; McMurray et al., 2017; McMurray et al., 2010; Smith & McMurray, in press), for characterizing the auditory integration in special populations such as bilinguals (Spivey & Marian, 1999), second language learners (Sarrett et al., in press), and aging adults (Revill & Spieler, 2012), and for understanding how even typical adults alter processing in the face of challenging listening conditions (Brouwer & Bradlow, 2015; Hendrickson et al., 2020). This work has highlighted the astonishing diversity of approaches to integrating auditory input to recognize words (for a partial review, see McMurray et al., 2022). For example, children with language disorders show changes in the asymptote of the functions in Fig. 1B – they do not fully commit to the target and maintain consideration of competitors. In contrast, younger typically developing children reach the same asymptotic level of fixations, fully fixating the target and suppressing the competitor. However, they are slower to do so, and competitors take longer to be suppressed (Rigler et al., 2015). An even more dramatic departure from the typical pattern is shown by pre-lingually deaf children (and by adults facing severely degraded input or very quiet input): they appear to delay lexical access until more information has arrived, with significantly slowed fixations to the target (by as much as 200 msec). Consequently, they show *less* competition (since by the time they begin lexical access, the target can be disambiguated from the competitor). Many of these kinds of patterns have also been observed for individuals hearing speech under various forms or degrees of degradation (e.g., in noise, in quiet, vocoded) (Ben-David et al., 2011; Brouwer et al., 2012; Farris-Trimble et al., 2014; Hendrickson et al., 2020), a critical issue in work on hearing loss.

This explosion of clinical and applied work raises the need for better measures that are less constrained by task demands and more revealing of the underlying neural substrate. In this regard, three limitations to the VWP may render interpretations more difficult and limit its clinical utility. First, the VWP relies on slow eye-movement responses that lag behind the true on-line comprehension process (by upwards of 200 msec, and these delays can compound to lead to substantial noise over a trial;McMurray, in press). Second, the VWP relies on picturable objects; it cannot easily assess more abstract words such as *democracy* or *patience*. Third, the VWP may not be suitable when cognitive or neurological disorders create deficiencies in eye-movement control, visual attention, or picture recognition (e.g., ADD, agnosia).

However, the most important concern is that the “read out” of the word-recognition system in the VWP is via semantic processing, as names must be matched to visual/semantic features of the pictured objects. Consequently, apparent differences in word recognition (e.g., across individuals or as a function of experimental conditions) could derive from differences in lexical, semantic, or even visual/attentional processes rather than the fundamental auditory integration process itself. Nevertheless, a crucial first step for assessing spoken word recognition in any subject population is to establish the integrity of the auditory/cognitive system that integrates auditory information. That is, if a listener delays lexical access in quiet speech or background noise, is this because the auditory system is slower to accumulate evidence, or is this because the lexical/semantic system is slower to access meaning from the speech signal? This needs to be examined separately from the downstream consequences of the overall auditory integration process for language understanding.

These limitations could be overcome with a spatio-temporal *neural* index of auditory integration and spoken word recognition (see Getz & Toscano, 2021, for an analogous argument). Neuroimaging methods would allow for minimal tasks that are less confounded by higher level cognitive processes. They may also be able to isolate mechanisms of SWR (and any deficits) at the level of auditory encoding. Clinically, this could help reveal how peripheral auditory deficits (e.g., cochlear implants) or deficits in the early auditory system (neuropathy, central auditory processing disorder) impact cortical mechanisms of language processing such as how the brain accumulates auditory information into meaningful chunks.

1.2. Neural Measures of Auditory Integration and Word Recognition

Classic neuroimaging approaches to the study of lexical access and spoken word recognition focused on identifying brain regions whose activity was modulated by phonetic category "goodness" or presence/absence of lexical competitors. For example, Blumstein et al. (2005) showed gradient activity in the left IFG as a phonetic distinction varied from its prototypical value to the category boundary. Prabhakaran et al. (2006), Righi et al. (2009), Luthra et al. (2019) and Zhuang et al. (2011) all reported modulations in frontal and temporal cortical areas as words varied in frequency or neighborhood density (a metric of lexical competition). Unfortunately, these fMRI studies were not able to characterize the timecourse of lexical competition because of the sluggish hemodynamic response function. Thus, the use of neuroimaging methods with response times in the msec range – EEG and MEG – have been employed to address this question.

The primary advantage of EEG/MEG is its excellent temporal resolution, but it suffers from poor spatial resolution unless sophisticated cortical source modeling is utilized. Moreover, although traditional Event Related Potential (ERP) approaches to EEG have identified components in the average waveform associated with a range of language processing operations including speech cue encoding (Getz & Toscano, 2021), phonemic categorization (Kazanina et al., 2006), and semantic integration (e.g., the N400; Kutas & Federmeier, 2011), there is no unique ERP signature of *lexical competition*. However, recent machine-learning methods have been applied to EEG/MEG signals to capitalize on multivariate patterns of activity as a vehicle for building sophisticated models of the neural correlates of spoken language processing (Xie et al., 2019). These models fall into two complementary categories – encoding and decoding – that establish reliable relationships between the dimensions of the speech signal and features embedded in the EEG/MEG responses during listening epochs.

1.2.1. Encoding models—

The fundamental logic of encoding models (DiLiberto et al., 2015) is to map a set of properties in the speech signal (e.g., amplitude envelope, phonemes, semantic features) to the EEG/MEG signals. This mapping is performed iteratively by seeking a weighting function (or temporal filter) for each channel of the EEG/MEG signal that best predicts the EEG/MEG response for a given property of the speech signal (from low-level acoustics to high-level semantics). If successful, this multivariate temporal response function (mTRF) can then be applied to withheld (or novel) speech signals to predict the expected EEG/MEG responses. Thus, an encoding model is evaluated by how accurately the mTRF performs; that is, the evaluation-metric is the correlation between the actual properties of the speech signal and the predicted properties of the speech signal based on the EEG/MEG responses convolved with the relevant mTRF.

A powerful aspect of encoding models is that they can be deployed in the context of naturalistic (i.e., continuous) speech. That is, the mTRF can be fit to a speech signal of any length, including an entire narrative, and the resultant encoding model can then be evaluated over similarly lengthy speech passages. The metric for evaluating the encoding model is how well it predicts the sequence of linguistic properties (e.g., phonemes or semantic features) or other characteristics (e.g., emotional valence) in novel passages. Encoding models have been used to evaluate multiple linguistic levels, including phonemes (Brandmeyer et al., 2013; Xie et al., 2019), phonemic and lexical surprisal (Donhauser & Baillet, 2020; Gillis et al., 2021; Weissbart et al., 2020), semantic surprisal (Broderick et al., 2021), and clarity of speech in noise and its resultant comprehension accuracy (Etard & Reichenbach, 2019). The superior source localization accuracy of MEG over EEG has enabled these encoding models to be mapped onto underlying brain regions, thereby providing further insights about the neural circuits involved in speech comprehension, including the number and ordering of competitors (Brodbeck, Presacco, et al., 2018; Gwilliams et al., 2020; Gwilliams et al., 2018; Kocagoncu et al., 2017), and the predictability of a segment from prior context (Blank & Davis, 2016; Choi et al., 2020; Gagnepain et al., 2012).

While encoding models have been powerful at illuminating the levels of speech processing, there are several limitations as currently implemented. First, because encoding models predict the *amount of neural activity*, they can directly reveal what conditions cause a particular brain region to work more or less hard, but they may not be able to reveal the unfolding of the speech-based information or the lexical decision itself (as does the VWP) (c.f., Gagnepain et al., 2012) as individual lexical items are presumably not represented by localized neural regions. Second, encoding models are typically based on group estimates (relating activity to lexical statistics), and do not provide assessments of SWR for individual participants or for specific words. This is important for ultimately meeting the clinical goals described above.

Third, and most importantly, missing so far from the implementation of encoding models for speech is a detailed estimate of the time-course of lexical competition. There is ample evidence that lexical competition plays a role in encoding models. However, the precision with which timecourse information has been estimated remains rather coarse. In principle, the phoneme-level mTRF could be evaluated for each word in the speech stream to address this timecourse question, but the mTRF in current encoding models is based on aggregating across relatively few exemplars of each word. In fact, that is one of the powerful aspects of encoding models – they are designed to generalize across all of the acoustic/phonetic and talker variability contained in natural speech corpora. Thus, unless the training data fed into the encoding model is more constrained (or more voluminous), it is not clear that timecourse information about lexical competition derived from the mTRF will have sufficient fidelity to answer the kinds of questions that have already been revealed at the behavioral level using eye-tracking data from the Visual World Paradigm.

The critical distinction in achieving such a measure is that the profile of lexical competition entails more than just encoding *accuracy*. Rather, the hallmark of lexical competition is that the system goes through states in which multiple options are briefly entertained (e.g., Fig. 1) before many are suppressed with some characteristic timecourse. While this is not inconsistent with encoding models, it has not been attempted, and doing so may require a large number of repetitions of specific words so that their encoding patterns can be identified.

1.2.2. Decoding models—As with encoding models, the key logic of decoding models is to find a link between the multivariate patterns present in the EEG/MEG signal and some relevant linguistic property of the speech signal. However, rather than predicting the EEG/MEG signal from the model and comparing predicted to actual EEG/MEG, a decoding model operates in the opposite direction by predicting the linguistic property from the EEG/MEG signal itself. That is, a set of features from each EEG channel provides a pattern of neural activity that, with an appropriate weighting function (much like the mTRF in encoding models), is used to predict the likelihood that a given word elicited the multivariate pattern of EEG activity. Then the trained model is evaluated on a withheld (or novel) set of trials to determine how accurately the EEG pattern predicts the stimulus on each trial.

Importantly, decoding models can be trained at each time-point post stimulus onset, thereby providing precise temporal resolution about the magnitude of lexical competition. Moreover, with a finite set of words or nonwords, each of which has a unique pattern of elicited EEG activity, the relative decoding accuracy of all items in that trained set can be assessed at each time point. Critically, by focusing on the pattern of confusion – not just the pattern of accuracy – decoding models can in principle track the partial decision-states that are the hallmark of lexical competition.

Decoding models have not been extensively applied to speech perception. Thus, the present study deploys the decoding approach to harness the power of machine learning techniques applied to multivariate patterns of EEG activity to estimate the strength of evidence that a listener has heard a given speech stimulus at each time-point after the onset of a word. As mentioned above, the choice of EEG features is critical. Unlike encoding models where the critical features in the EEG signal are discovered, in a decoding model they must be specified in advance. Consequently, a decoding model that fails to capture the relevant features from the EEG signal that map reliably onto linguistic events will result in decoding accuracy that does not exceed change levels (established by permutation tests). Our goal is to provide a metric of how speech-related neural activity is integrated after word onset and builds incrementally as the speech signal unfolds during word recognition to reveal lexical competition (e.g., Fig. 1). Importantly, our decoding approach offers the promise of providing evidence of lexical competition at the level of individual participants, which is precisely what is needed to characterize the kinds of variations observed in typical development and in clinical populations.

1.3. The Present Approach

The present project combines recent advances in machine learning with standard EEG techniques to develop a method to estimate the dynamics of real-time auditory integration and decision making in SWR. To be clear, there are limits to what can be concluded from any neural measure of SWR because SWR consists of at least two fundamental levels of information − the acoustic/phonetic level and the meaning/semantic level. Any familiar acoustic event (e.g., the sound of a bell) can be retained in memory for later recognition, but words allow the mapping of that sound-based memory onto the referential meaning associated with that sound. For example, an infant who has not yet learned that the sound *ball* refers to a solid round object could nevertheless recognize the familiar sound pattern (*ball*) and discriminate it from the similar sound *fall*. Thus, a sound-based recognition process could be sufficient for SWR as long as the acoustic/phonetic analysis of the words had sufficient sensitivity and a robust representation in memory (unlikely for unfamiliar non-native phonetic categories) (Goldinger, 1998). Words, of course, add the possibility of semantics. However, even recognizing a word at the acoustic/phonetic level requires listeners to integrate material over time, and to sort out competing sound patterns.

As described in earlier sections, the VWP has a long history of addressing this mapping of spoken words to picturable referents. It therefore circumvents the problem faced by neural measures which do not rely on picture-based matching of the spoken words, and therefore do not unambiguously tap into the semantic product of word recognition. However, the concerns raised above about the limitations of the VWP raise the need for a complementary decoding paradigm that can identify the auditory precursors to word recognition. Thus, the present EEG-based paradigm has no referential component except the internal mapping of sounds to meanings that is already established for known words (such a mapping is absent in the case of unknown words or nonce words). Nevertheless, it is important to obtain a measure of purely sound-based neural decoding because a necessary component of SWR involves the decoding of the acoustic/phonetic information that defines a word-form over time. This could be crucial for identifying auditory integration deficits that could underlie a variety of clinical disorders and real-world challenges (e.g., speech in noise). Thus, the present approach evaluates how well an EEG-based paradigm in a non-referential context can assess the time-course of SWR even if the meaning-based component of SWR is not necessarily engaged.

As a first step, our approach focuses on sensor-space (scalp-based) EEG signals. We begin by asking whether there is a paradigm with sufficient sensitivity and selectivity to decode spoken words without attempting to determine how these scalp-based signals map onto the underlying neural substrate. When this computational approach is eventually coupled with source-localization, MEG, intercranial EEG (iEEG), or MRI, it could reveal not just what brain areas are involved in integrating auditory information to support word recognition, but what cognitive functions they perform. Such approaches could also reveal how these neural networks emerge over development or differ with communicative impairment. Furthermore, by isolating auditory cortical mechanisms we could address fundamental questions such as how high-level context shapes auditory perception (Gow & Olson, 2016), or how lower level auditory processes cascade to enable language understanding (Sarrett et al., 2020). This is crucial for revealing the causal mechanisms of a variety of clinical disorders and applied situations in which word recognition differs, as it may help reveal the degree to which differences are due to auditory integration or to downstream lexico-semantic processes.

The primary goal of the present project, therefore, was to develop tools that use EEG to recover the dynamics of auditory information integration and decision-making even if a meaning-based level of information was not engaged. We used straightforward – and publicly available – machine learning tools to create a paradigm that can be easily deployed in the laboratory or potentially in the clinic. While machine learning and multivariate techniques have long been used with fMRI (Norman et al., 2006), their application to continuous time varying signals is fairly novel (Bae & Luck, 2018; Grootswagers et al., 2017; King & Dehaene, 2014), and only a handful of studies have deployed such techniques with human speech (Beach et al., 2021; Brandmeyer et al., 2013). To date, these approaches have largely focused on the *overall accuracy* of classification after the entire word has been heard; however, as Fig. 1 illustrates, the primary issue we investigate is not decoding accuracy *per se*, but the pattern of partial confusions as the word unfolds over time due to lexical competition. These confusions are inferred in the VWP by aggregating probabilities of eye-movements to pictured referents across multiple repetitions of trials with the same spoken word. But because eye-movements are not a continuous variable (i.e., fixations can only occur every 200 msec), on any given trial one cannot determine the level of confusion between the target word and its cohort. In contrast, in the present EEG paradigm, we can train multivariate "templates" for each word and then ask, on each individual trial, precisely how confusable the target and cohort templates are at each msec as the auditory word-form unfolds in real-time.

As noted above, we cannot be certain in the absence of pictured referents whether we are tapping into the meaning-based level of word representation. But regardless, SWR must rely on a lower-level acoustic/phonetic decoding process of auditory integration to enable the downstream recognition of high-level meaning/semantics. Here, a critical marker of this kind of integration is evidence of co-activation when items overlap (e.g., *baggage* and *badger*). That is, during the onset period of an item (e.g., the *ba-* in *badger*), would the classifier report evidence for both the target and a competitor (*baggage*)? And how does that competition resolve over time? This is relevant for any auditory stimulus (both words and nonwords) in a speeded task that must rely on a form of auditory memory.

Thus, the present approach provides an essential first step in evaluating the neural correlates of the SWR process, with two key advantages over psycholinguistic methods: (a) the neural paradigm is entirely passive and does not require control over an overt behavior (e.g., eye gaze), and (b) decoding of multiple word candidates can be assessed in parallel on a msec-by-msec basis for a given trial. The long-term goal of the methods introduced here is to develop a *neural paradigm* that could be used with a variety of populations – from infancy to elderly adults, as well as people with communicative impairments. Thus, we focus on straightforward technologies and a minimal set of task demands to establish a robust proof-of-concept.

In our neural paradigm, participants performed a simple task that was designed to keep them minimally attentive to a small set of words while EEG was recorded. Stimuli included eight word-pairs (cohorts) that overlapped at onset and would be expected to create a brief period of ambiguity or competition (e.g., *badger/baggage, mushroom/muscle;* see Table 1); these were matched to eight nonwords (e.g., *babbid/baddow, musheme/muspil*). We then trained a support vector machine (SVM) classifier to identify which of the eight words (for that individual subject) was the stimulus at each time window over the epoch, on each trial. The classifier was trained and tested at consecutive 20 msec increments over the post stimulus-onset epoch. At each step we recorded the proportion of time the classifier chose the target word (e.g., *badger*), its cohort competitor (*baggage*), or one of the six unrelated words (*mushroom*) to construct classification curves analogous to Fig. 1. To test the extensibility of the procedure we tested participants in both a 64-channel, low impedance, active electrode system (N = 16 from the University of Iowa) and others on a 128-channel high impedance EGI system (N = 15 from the Haskins Laboratory). Thus, our analyses focus on the validity, utility, and reliability of the method with respect to factors like the number of trials, channel configuration and the EEG features that support categorization.

2. Methods

2.1. Participants

All subjects were right-handed, monolingual, native English speakers between 18 and 30 years old. All participants had normal or corrected to normal vision. The Iowa sample consisted of 16 subjects (7 male, 9 female); the Haskins sample consisted of 15 subjects (1 male, 14 female). One additional Haskins participant was dropped from analysis due to poor quality EEG.

2.2. Design and Items

Items consisted of pairs of bisyllabic words and nonwords with overlapping onset phonemes (Table 1). Pairs overlapped at onset by at least the initial two phonemes in order to elicit robust lexical competition. Half of the stimulus pairs were words and half were nonwords. Each pair had an onset phoneme that was unique from all other pairs, and that differed from all other pairs in multiple features. For example, the phoneme /b/ was only used in the word pair *badger*/*baggage* and the corresponding nonword pair *babbid*/*baddow*.

Over the course of the experiment, participants heard one of two sets of four word-pairs and four nonword-pairs (List A or List B in Table 1). Lists were counterbalanced such that each list had either a word- or a nonword-pair from the entire inventory of onset phonemes, but that any given subject was not tested on both the words and their matched non-words. For example, if *badger/baggage* served as a word pair for a given subject, *babbid/baddow* did not occur as a non-word pair for that subject. The lists did not differ in their average positional phoneme probability (M_{ListA} = .243, M_{ListB} = .23; t(30) = .621, p = .539) or biphone probability (M_{ListA} = .014, M_{ListB} = .010; t(24.212) = 1.564, p = .131), computed using Vitevitch and Luce (2004).

We used five unique auditory exemplars of each stimulus item to ensure that the machine learning classifier did not rely on the unique acoustic properties (including background noise and pitch variations) of individual exemplars of a stimulus. Each of the 4 word-pairs and 4 nonword-pairs (16 items total) was presented 60 times (5 exemplars × 12 repetitions), for 960 trials.

2.3. Stimuli

Stimuli were recorded by a male native English speaker with a Mid-western dialect recorded at a sampling rate of 44,100 Hz. For each item we recorded 10–15 exemplars in a carrier sentence (*He said* *badger*) that was designed to ensure a more uniform prosody and speaking rate. We then selected the 5 clearest exemplars for use in the study.

Stimuli first underwent noise reduction in Audacity (Audacity Team, 2015). For this, we estimated the spectrum of the noise from a 1 second silent interval, and then subtracted this from the whole recording. Stimuli were then cut from the onset phoneme of the target word to the release of the final phoneme at the nearest zero crossing. Clicks were manually removed in Praat (Boersma & Weenink, 2009) at the nearest zero crossing. Finally, amplitudes were normalized using Praat and 0.100 sec of silence was added to the start and end of each stimulus to avoid artifacts from the sound card turning on. The average duration was 594 msec (SD = 90).

2.4. Procedure

Upon arrival in the lab, participants gave informed consent and completed a short demographic questionnaire. Then, participants were fitted with an electrode cap and moved to the EEG recording booth or room (see below for details on Iowa and Haskins EEG setups, respectively). Participants sat approximately 80 cm from the center of the display monitor. Target words were played over Etymotic ER1 insert earphones. Fourteen Iowa participants used an Acer monitor with a 1960 × 1080 display, and two Iowa participants used a Dell monitor with 1680 × 1050 display (both with a 60 Hz refresh rate). The Haskins participants used a 19" Dell monitor operating at 1280 × 1024 resolution (60 Hz refresh).

During EEG recording, participants completed a word identification task in which they reported the word they heard via a key press. Subjects heard a spoken word and matched it to one of two words (presented in text) that appeared about 1300 msec later, well after the word was complete. Participants used the left and right arrow keys to indicate their choice. Visual feedback ("Correct!" or "Incorrect") was given after each response, and then the trial advanced.

On each trial, a black fixation cross on a gray background appeared on average 800 msec before the target word, to allow a "silent" period for later baselining of the EEG signal. This time was jittered by ±100 msec to avoid anticipatory effects on EEG between trials. When the audio file ended, two orthographic response options (the word and its foil) were presented. On half of the trials the response options were the target (e.g., *badger*) and its cohort competitor (*baggage*). On the other half of trials, the options were the target (e.g., *badger*) and an unrelated word (*mushroom*) from another set. These were randomized throughout the experiment. The interval between the start of the target word to the response options was an average of 1350 msec, with a 100 msec jitter to avoid anticipatory effects in the data prior to response. Response options appeared 443 pixels to the left and right of the fixation cross. Visual feedback appeared 50 msec after the response. After a correct response, "Correct!" appeared in the center of the display for 500 msec and the experiment advanced automatically to the next trial. After an incorrect response the subjects saw "Incorrect, press the space bar to continue" and advanced to the next trial when the participant was ready. This allowed participants to take an optional break if needed. The intertrial interval was 700 msec with a 200 msec jitter. A mandatory break lasting 60 seconds was inserted after every 240 trials.

2.5. EEG Methods

2.5.1. Iowa Equipment and Procedures—EEG signals were recorded in either an electrically shielded sound-attenuated booth (N = 14) or in a quiet room (N = 2), dimly lit by battery powered lights. EEG was recorded via a 64-channel Brain Vision actiSlim system, placed according to the International 10–20 system. Impedances at electrodes were less than 5 kOhms prior to recording. EEG was recorded at 500 Hz and amplified using a Brain Vision actiChamp system. Electrodes were referenced offline to the average of all electrodes for each subject. Horizontal and Vertical electrooculogram (EOG) recordings were recorded using Fp1 and Fp2, the two frontal-most electrodes. In the Iowa sample, EEG was synchronized to the auditory stimulus by recording audio data simultaneously to a separate channel of the EEG via a BrainVision StimTracker.

2.5.2. Haskins Equipment and Procedures—EEG was acquired in a quiet room with a testing area and a control area separated by a partial wall. The Electrical Geodesics Inc. (EGI) net amps 300 high-impedance amplifier EEG system and experiment presentation computers are located in one area, and the participant wearing the EEG net was located in the other area. EEG was collected at a 1000 Hz sampling rate via a 128-electrode geodesic sensor net. Online recordings were referenced to the vertex (Cz) and were later re-referenced to the average of all electrodes for each subject. The maximal impedance was kept under 40 kΩ (impedances were rechecked periodically through the testing session). EEG was continuously recorded using Netstation 5.4 on a MacPro. Synchronization was performed by sending triggers directly from the subject computer to the EEG system at sound onset.

2.5.3. EEG preprocessing—Both data collection teams used an identical custom preprocessing pipeline based on EEGLab functions (Delorme & Makeig, 2004) and implemented in Matlab. First, we excluded bad channels from the continuous data with bad impedances that were identified by the experimenter during recording. Second, we sequentially high-pass and low-pass filtered the continuous signal from 0.1 Hz to 30 Hz, both with an 8 dB/octave rolloff. Non-stereotypic artifacts were then manually removed from the signal. Eye movement artifacts were removed using Independent Component Analysis.

Trials were then time-locked to the onset of the target word. For the Iowa sample, this was detected by identifying the first sample in the secondary audio channel that crossed a predetermined threshold. In the Haskins data, this was identified by the stimulus triggers sent by ExperimentBuilder, the experimental software, to NetStation, the EEG recording software. Epochs started at −0.5 sec relative to target word onset and ended 1 sec after the target word onset. EEG was baseline-corrected based on a 300 msec window (from −300 msec to 0). After epoching, additional trials containing artifacts with an absolute value greater than 150 microvolts were excluded (3.13% of trials).

2.6. Machine Learning Analyses

Machine learning was performed on individual subjects on the basis of individual trials (not averaged data), using techniques largely similar to those we have pioneered with iEEG (Nourski et al., 2015). This consisted of three steps: feature selection, SVM parameter setting, and validation. Completely commented Matlab code along with data for one single subject are available on our OSF site at https://osf.io/wa3qr/.

2.6.1. General Methods—All machine learning analyses were conducted using a Support Vector Machine (SVM) framework implemented in LibSVM (Chang & Lin, 2011). The SVM used a radial basis function transformation of the data. Consequently, there were two free parameters, the *cost* parameter (C) and the width of the basis function (λ). SVMs were trained on two eight-alternative tasks to identify which of the 8 words (or 8 non-words) the subject heard on that trial. We also explored a full 16AFC training task (among all words and nonwords simultaneously). While performance was above chance, it was too low to reflect meaningful dynamics.

All analyses were implemented with a 15-fold cross validation procedure in which the SVM was trained on 14/15$^{\text{ths}}$ of the trials and tested on the remaining 1/15th. This was repeated 15 times, such that each trial served as test trial once. Decoding accuracy (or identification proportions) were then computed as the average across all trials. Assignment of trials to each fold was random with the constraint that we attempted to equalize the number of trials from each word in each fold. This procedure was then repeated 30 times to allow for sampling noise created by the assignment of trials to folds. Data used to train the classifier represented several features computed separately for each trial for each channel. Each feature was Z-scored across trials (within subject) prior to entering the machine learning analyses.

2.6.2. Feature Selection—The goal of feature selection was to determine which properties of the EEG signal were most useful in decoding wordform identity. Prior work using ECoG has systematically explored the space of both electrophysiological and time/frequency parameters to find optimal properties for speech decoding (Nourski et al., 2015). However, given the spatial imprecision of EEG signals compared to cortical electrodes, as well as the fact that some frequency bands are severely attenuated by the skull/scalp/dura mater, these could not be assumed for EEG.

To avoid overfitting the data, our strategy was to systematically explore a full space of features for the first five Iowa participants. We then locked these features for all further subjects (including the participants run at Haskins on a different EEG system). Feature selection focused on the following properties of the data: First, we considered the mean EEG voltage in each channel over a given time window. Second, we added higher order polynomial terms (slope, quadratic, cubic, etc.) reflecting the change in voltage at that channel over one of several time windows. This captures something akin to the phase of the signal, though unlike phase-based time/frequency approaches it makes no assumptions about the frequency band. Whenever a higher order polynomial term (e.g., a cubic) was added, all lower order terms (quadratic, linear, intercept) were also retained. Third, we considered the mean power within frequency bands for each channel, within that time window. We explored five frequency bands: delta (δ, 0 – 3 Hz), theta (θ, 4 – 7 Hz), alpha (α, 8–15 Hz), beta (β, 15 – 30 Hz), and gamma (λ,30 −70 Hz)[1] .

At this stage, each feature was extracted for each trial, for each channel at a fixed time window starting at 200 msec (which was suggested by exploratory work showing maximum performance near that range). We explored three time-window lengths (75, 125, and 250 msec). This led to about 100 permutations of these basic features (e.g., EEG slope + θ over a 75 msec window). Within each particular set of features, we performed a 16 × 16 brute force search of possible settings for C and λ (the free parameters of the SVM). C ranged from 2^{-2} to 2^{18} and λ from 2^{-19} to 2^{-3} . This was done in a 15-fold cross-validation procedure, with the accuracy of the held out trials serving as the performance for that feature set. Each feature set was then run 30 times to smooth out variation due to the random foldings. Finally, test performance across each run was averaged and saved for that feature. The maximum performance across this matrix of possible features was then saved.

We found substantially better performance at longer time windows in the first five subjects. We also found that coding the EEG as a cubic (plus the lower order quadratic, slope and mean terms) within that time window yields the best results (there were few further gains for quartic and quintic terms). None of the time/frequency measures were useful by themselves, or in addition to the polynomial terms. This was later confirmed with the entire sample (see Results).

2.6.3. SVM Parameter Selection—With the features selected, we next set the two free parameters of the SVM. This was done using a hybrid search approach. For each subject, we extracted the features identified from our initial search (the cubic polynomial of the EEG over a 250 msec window) for each channel, starting at 200 msec. The SVM was then trained and tested at a particular combination of C and λ. Again, this was done in a 15-fold cross-validation procedure with 30 runs to smooth out the effects of the random assignments of trials to foils. Accuracy for that particular combination of C and λ was the average of the test-trials across 30 runs .

The optimal C and λ for that subject were initially based on an 8 × 8 brute force search using the same ranges described above. After this coarse search, we used the maximum values as the starting point for a constrained gradient descent method using a GPS/pattern search approach. This was done separately for words and nonwords to find the optimal parameters for each iteration through the analysis pipeline for each subject.

[1] Preliminary analyses also attempted a msec-by-msec decoder (no time window) that has been used in decoding studies of visual stimuli (Bayet et al., 2018, 2020; Cichy et al., 2015), but that set of features (the voltage from each electrode after low-pass filtering) did not result in significant decoding accuracy.

2.6.4. Final Time-course Analyses—The final analysis deployed the SVM-classification pipeline using the features and free parameters identified from the initial feature-selection process on the first five subjects, but now generalized across successive time windows. At each starting time (ranging from −.5 sec to +1 sec), the EEG was extracted and the polynomial was fit to extract the four parameters that served as features. Next, a new SVM was trained to identify which of the 8 words or nonwords was the target. Rather than extracting a parametric value for accuracy, on each trial we determined which of the 8 words or nonwords was the most likely response. This was then classified as target (the SVM reported the correct word), cohort (the SVM reported the onset competitor, e.g., *baggage* when the target was *badger*) or one of the other six unrelated words. These were averaged across trials for the target word and further averaged across the six non-targets to compute the response at each timepoint. The time window was moved in 20 msec increments to compute the overall decoding accuracy function across the entire timecourse.

2.7. Statistical Approach for Analyzing Decoding Performance

Identification responses (e.g., proportion target/cohort/unrelated identification at a given time) served as the dependent variable in all analyses. This can serve as a common metric that allows for pooling data across subjects with different numbers of electrodes, features etc.

2.7.1. Figures—For visualization, data were averaged across trials, within subject, and smoothed with a 0.1 sec triangular window. Error bars reflect standard error of the mean across subjects.

2.7.2. Detailed analysis of the timecourse of competition (mixed models)— To characterize the timecourse of competition, we asked when the SVM identification of different competitor types (i.e., target, cohort, unrelated) differed significantly from each other. We ran a set of linear mixed effects (LME) models every 20 msec over the full timecourse epoch, from 100 msec before the onset of the target word to 1150 msec after the target word began. This model predicted SVM classification performance from a set of contrast codes designed to capture key variables of interest (e.g., target vs. cohort, cohort vs. unrelated). Models were run using the lme4 (v. 1.1–23; Bates & Sarkar, 2011) in R (v. 4.0.3).

Separate models were run for the word and nonword results. Data were smoothed using a 100 msec triangular window. The DV in each model was the proportion of SVM classification of a specific word-type (e.g., target, cohort and unrelated). The fixed effects in each model were two contrast codes. The first captured a specific planned comparison (e.g., Target vs. unrelated); the second was orthogonal to it (Target and Unrelated vs. cohort). This latter contrast was obviously not of scientific interest, but it was included in the model as an orthogonal contrast code to the intended contrast (in this example, the Target vs. Unrelated contrast) so that the error variance reflected the full dataset rather than just the two conditions in the intended contrast (since all datapoints were relevant to either the primary or secondary contrast).

The first model quantified whether SVM classification of target items differed from the phonologically unrelated item. This model included the fixed effect of Target vs. Unrelated as the contrast of interest, coded as Target (+0.5), Unrelated (−0.5), and Cohort (0). The second contrast of this model was Target + Unrelated vs. Cohort (Target or Unrelated: +0.33; Cohort: −0.66). The second model quantified whether SVM classification of the cohort differed from the unrelated as the primary contrast (Cohort: +0.5; Unrelated: −0.5; Target: 0). The secondary contrast was Cohort + Unrelated (+0.33) v. Target (−0.66). The third model asked if target identification (+0.5) significantly differed from cohorts (−0.5, with unrelated at 0). This model included the fixed effect of Target (+0.5) v. Cohort (−0.5, with unrelated at 0). The secondary contrast in this model was Target + Cohort (+0.33) v. Unrelated (−0.66).

Potential random effects in these models included Subject and Item. The random effects structure for each of these models was chosen using the model space approach developed by Seedorff et al. (submitted). In this approach, one tests all possible random effects structures, and chooses the model with the lowest Aikake's Information Criterion (AIC). This approach has been shown to hold Type I error constant at 0.05 while maximizing statistical power, so as not to be overly conservative. Because we ran multiple models across time, the random effects structure was determined by using this model space approach across all three models and at multiple representative timepoints throughout the epoch. Then, the distribution of AICs by random effects structure was inspected, and the model which had the lowest AIC at the highest number of timepoints was selected to run across the full timecourse. For each model for both Words and Nonwords, this resulted in the maximal random effects structure: with both contrast codes serving as random slopes for both subject and item.

Finally, due to the large number of models (e.g., across time points), it was critical to control for family wise error. However, the results of significance tests over a timeseries are autocorrelated, and therefore are not truly independent tests. We thus used the familywise error correction of the Bootstrapped Difference of Timeseries (Oleson et al., 2017) analysis package in R to correct for family error. It computes the auto-correlation (ρ) of the t-statistic over the timeseries, and then computes a corrected significance level (α*).

2.7.3. Peak Detection—

For analyses examining permutations of the basic paradigm (e.g., fewer trials, channels, etc), we employed simpler statistics based on the "peak" (maxima) identification responses. Here, better decoding performance should yield higher target and cohort identification rates and lower unrelated rates. To perform this analysis, we first extracted the data for each subject for each condition (e.g., 64 vs. 32 vs. 16 channels). We next averaged performance on words and nonwords (unless otherwise specified). Target, Cohort and Unrelated identification rates were then smoothed with a 160 msec triangular window[2] . Lastly, we extracted the peak identification rates for each competitor type. These were manually checked against the full timecourse for a subset of subjects and then compared against the baseline in a series of paired t-tests.

[2] Note that this smoothing window was larger than the 100 msec window used for most visualizations because indices like peak needed to be identified from individual subjects and was highly susceptible to noise; in contrast visualizations were usually averaged across subjects and therefore required less smoothing.

3. Results

3.1. The Timecourse of Auditory Integration

Fig. 2 shows the results of our primary analysis. Results for both words (Panels A,C) and nonwords (Panels B,D) showed clear evidence of integrating auditory information over time, and of partial parallel activation of the competing wordform (compare to Fig. 1). Within about 100 msec of word onset, classification responses favored the target (*baggage*) and cohort (*badger*) over the unrelated words (*mushroom*). These in turn were indistinguishable from each other until about 300 msec when the cohort began to be suppressed. There was remarkably similar performance across the two EEG systems.

To identify the time periods in which these curves differed, we used the mixed model approach described in the Methods. This tested the contrast between targets and cohorts, cohorts and unrelated and targets and unrelated at each 20 msec timeslice. Alpha was adjusted for the large number of contrasts (corrected for family-wise error using: Oleson et al., 2017), and the relevant statistics are shown in Fig. 2. This was done separately for words and nonwords for each of the two samples.

Table 2 shows a summary of the significant time windows for each contrast and for each sample. Across all four analyses (word/nonword × EEG system), the target deviated from the unrelated item from about 150 msec to 850 msec. Cohorts deviated from unrelated items at a similar point (roughly 150 msec) but persisted for a shorter period of time, ending at about 500–600 msec. Targets did not differ from cohorts until later around 350–400 msec, and generally stayed separated until about 700–800 msec. Words and nonwords did not show substantial differences; however, results (particularly for cohort vs. unrelated identification) were somewhat less robust for the high impedance system. This suggests the additional channels afforded by that system may not fully offset the loss of signal fidelity (Kappenman & Luck, 2010), though we note that differences were not substantial.

As a follow-up analysis, we asked whether there were differences in competition dynamics between words and nonwords. Thus, following the approach described above, we combined the data from the word and nonword classifications and added a factor indicating which type of stimulus was heard (this was done separately for each of the two samples). We then ran three additional models which started with the same contrasts as in the model described above and added Word/Nonword (coded as Word: +1; Nonword: −1), along with its interactions with the contrasts (e.g., Target v. Cohort). No significant main effect of Word/Nonword nor any significant interactions with the three contrasts of interest were found.

3.2. The Timecourse of Auditory Integration in Individual Subjects

Fig. 3 shows the same results for 16 randomly selected subjects (8 from each sample). The pattern of competition observed in Fig. 2 at the group level is robustly observed in each individual subject across both the low impedance (top 8 panels) and high impedance (bottom 8 panels) samples. The between-subject variability is broadly consistent with what is often observed in the VWP (McMurray et al., 2010).

These visualizations suggest that the machine learning approach here can yield data that is interpretable at the level of single subjects. This raises the potential of using this technique for studies of development, special populations, or individual differences. To set the stage for this work, it is important to compute the reliability of this neural paradigm to determine how stable these differences are.

Testing for this work was done during the Covid-19 pandemic when it was important to limit contact with subjects, and there were some institutional constraints on data collection. Consequently, it was not feasible to bring subjects in twice to assess test-retest reliability. We considered using split-half reliability by training the classifier on two separate sets of trials and correlating the results. However, our analysis of the number of trials needed for accurate decoding (presented shortly) suggested that limiting decoding to 50% of the data would yield less reliable results.

Thus, we adopted a hybrid split-half approach that computes a measure of reliability that assesses the coherence of the data across trials akin to Cronbach's α. First, the classifier was trained on the entire dataset (the same primary classification analysis as in Fig. 2). Next, trials were randomly assigned to set A or set B. Random assignment was constrained to have approximately equal numbers of trials from each item. From these subsets of trials, the time-course of identification (e.g., Fig. 2) was saved for each subset. We then averaged the data from each set (A or B) and smoothed it with a 0.2 sec window.

Second, we extracted summary indices from these time-course functions. These indices describe critical aspects of these time-course functions that have proven useful indicators in prior VWP studies. These include:

- *Peak Target, Cohort:* estimated using the same procedures described above. These have been linked to DLD (McMurray et al., 2010)
- *Time of Target and Cohort Peak, Unrelated Minimum:* Time post-stimulus onset at which the peak or minimum was first detected. Target and cohort peak times have been linked to hearing loss (McMurray et al., 2017).
- *Minimum unrelated:* estimated using similar procedure to peak.
- *Slope of Target at 50%:* linear slope of target identification as a function of time over 0.1 sec surrounding the point where the target crossed 50% of its maximum. This has been linked to typical development (Rigler et al., 2015; Zangl et al., 2005).
- *Time when Target+Cohort deviates from Unrelated:* First, we computed the sum of the target and cohort identification; this was normalized to be a proportion of its maximum value; then saved the time at which this crossed 40% of its maximum. Measures like these have also been used to assess development (Rigler et al., 2015).
- *Time when Target deviates from Cohort:* First we computed the target – cohort identification curve. This was normalized to be a proportion of its maximum

value. Then saved the time at which this crossed 40% of its maximum. Measures like these have also been used to assess development (Rigler et al., 2015).

These indices were identified for the A and B sets of trials for each subject. We then computed a correlation coefficient across the 16 subjects to estimate the reliability of the measure across the two subsets. This was then repeated 12 times (with different random assignments of trials to sets) to compute the average correlation among sets. This measure of reliability assesses the coherence of the data across trials (rather than across sessions).

Results are shown in Table 3. Most of the reliability estimates (expressed as correlation coefficients) were moderate to large and reliability was very high (r>.90) for peak target and cohort identification for both types of EEG systems. Other metrics listed in Table 3 showed mixed results. The peak time was highly reliable for the low impedance system for words, but less so for the high impedance systems and for non-words for both systems. Slope and deviation indices were modest across words/nonwords and both systems. Virtually all of these correlations were significant. The lower correlations may reflect the difficulty of estimating some of these parameters from noisy data (rather than the coherence of the measure), and future work should explore more sophisticated analytic approaches for obtaining precise estimates of temporal properties of the decoding results (Oleson et al., 2017).

3.3. An overpowered classifier?

One concern with any machine learning approach is whether the paradigm is “too powerful” – perhaps this paradigm can yield above chance performance even when the data do not show an underlying separation between the categories? The analyses described previously took several steps to avoid this. First, we used a 15-fold cross validation procedure (so that the “test” data is never part of the training set). Second, we selected features based on only a small set of subjects and then fixed them for all subsequent analyses. Finally, while the parameters of the SVM (C and λ) are fit to each subject, they are only fit at a single time bin and then locked for the entire timecourse. Nonetheless this last step raises the possibility that we could be overfitting the data by optimizing these two free parameters for each subject’s data.

To rule out this possibility, we ran permutation tests on a subset of the data. In these tests, the assignment of words to trials was randomly shuffled. This should disrupt any relationship between the EEG patterns and its source in the auditory stimulus. We then repeated the entire analysis pipeline. If above chance performance was observed, this would be problematic.

These permutation tests were conducted on three subjects whose data were used in the initial round of feature selection (as these subjects would be most susceptible to overfitting), and an additional three subjects from outside of this select group of five. All subjects came from the low impedance / Iowa sample. For each these six subjects, we randomly shuffled the assignment of words to trials. Next, we optimized the free parameters of the SVM for that shuffled dataset, using the same procedure described above: a short brute-force search, followed by a gradient descent algorithm. As before, this was conducted only at a time bin centered at 200 msec. Next, we used those free parameters to assess the full timecourse of processing. This was then repeated 100 times, and the full timecourse at each repetition was saved. To compute 95% confidence intervals, we extracted the results for each repetition and computed the average target, cohort and unrelated identifications at each time bin (averaging across the words). This was smoothed with a 100 triangular window. Finally, at each time bin, we sorted the resulting identification performance and identified the range at which 95% of the observed data fell.

Results are shown in Fig. 3. For each subject, the mean of the permutation tests (the magenta bar) hovered close to chance (0.125). This suggests that on average there was no systematic bias. Moreover, confidence intervals included chance at every point, and at many points, target cohort and unrelated fixations were outside of this range. This validates that the procedure used here does not artificially inflate performance.

In summary, the foregoing analyses establish the basic viability of the paradigm. The decoding analysis shows a dynamic pattern of classification that fits a lexical competition profile, featuring early consideration of both the correct item (target) and overlapping ones (cohort), but suppression of the incorrect ones later. This was seen for both words and nonwords, and nearly identical across different EEG recording systems. This pattern was observed in individual listeners and appears reliable. It is not an artifact of the classification approach.

With these basic properties established we next turn to several analyses that examine the extensibility of this paradigm and what the classification response means.

3.4. Electrode configurations

We next asked whether equivalent classification performance is obtained with different electrode configurations. We first addressed this question by reducing the number of channels contributing to the analysis. This is important for future applications to children or clinical populations where high-density arrays may not be feasible (due to setup time or cost). We used the same machine-learning approach with smaller arrays of channels. For the low impedance sample, this was done for 32 and 16 channels; for subjects tested with a high impedance EEG system this was done at 64 and 32. Channels were retained along the standard 10/20 grid, covering the full scalp but with lower density by sub-sampling.

Fig. 4 shows the results, averaged across words and nonwords for each recording system. In these plots, the original results using all available channels are shown in black; colored lines reflect the new analysis with fewer channels. For the low impedance systems, we observed no differences when the number of channels was reduced from the full 64 to 32 (Panel A)—the black lines (all channels) are indistinguishable from the reduced (colored) lines. Statistically, there was no difference in peak identification for targets ($t(15) = 0.19$, $d = -0.05$, $p = 0.85$), cohorts ($t(15) = 0.18$, $d = -0.05$, $p = 0.86$) or unrelateds ($t(15) = 0.12$, $d = 0.03$, $p = 0.91$). Further reducing the number of channels to 16 channels (Panel B), however, showed a significant performance decrement for all three word types (Target: $t(15) = 3.39$, $d = 0.85$, $p = 0.0041$; Cohort: $t(15) = 2.50$, $d = 0.62$, $p = 0.0247$; Unrelated: $t(15) = 4.20$, $d = -1.05$, $p = 0.0008$).

A similar pattern was observed for the high impedance (128 channel) system. As before, when the number of channels was reduced by half (64 channels, Fig. 5C), performance was again indistinguishable from the original analysis, with no difference for target peak (t(14) = 1.00, d = 0.26, p = 0.33), cohort peak (t(14) = 0.83, d = −0.21, p = 0.42) or unrelated peak (t(14) = 0.25, d = −0.06, p = 0.81). However, again when the number of channels was reduced to a quarter of the original size (32, Fig. 5D), significant performance decrements were noted for targets (Target: t(14) = 4.73, d = 1.22, p = 0.0003) and unrelateds (t(14) = 2.85, d = −0.74, p = 0.013), but not for cohorts (t(14) = 1.06, d = 0.27, p = 0.31).

Thus, the number of channels can easily be reduced by half in either EEG system. More importantly, the robustness can be plainly seen in individual subjects (Fig. 5F, H for the two systems respectively). The low impedance system may be slightly more robust to the absolute number of channels (32 channels showed no decrement with this system, but a decrement for the high impedance system).

We next asked if decoding performance was driven by whole head coverage or was primarily based on responses from auditory areas. Early auditory ERP components such as the P50, N1 and P2 are typically strongest at fronto-central electrodes. This reflects the origin of these components in Heschl’s Gyrus and the Superior Temporal Gyrus, which have dipoles oriented toward the top of the head. We thus compared the analyses with half of the channels (32 or 64) distributed evenly across the scalp to a new analysis with the same number of channels but centered in the fronto-central region (Fig. 5E,G for the two EEG systems). This showed a significant reduction in decoding performance for both low impedance (target: t(15) = 4.20, d = 1.05, p = 0.0008; cohort: t(15) = 2.85, d = 0.71, p = 0.0121; unrelated: t(15) = 4.68, d = −1.17, p = 0.0003) and high impedance systems for targets (t(14) = 3.84, d = 0.99, p = 0.0018) and unrelateds (t(14) = 3.36, d = −0.87, p = 0.0047), but not cohorts (t(14) = 1.29, d = 0.33, p = 0.22). This performance decrement cannot be attributed merely to the loss of channels since no decrement was observed with 32 channels across the full scalp, suggesting that the neural basis of decoding performance requires contributions from electrode locations that tap into neural systems beyond early auditory processing areas.

3.5. How many trials are needed?

The long-term goal of this project was to develop a method that could be used with children and clinical populations for evaluating the integrity of cortical processes that integrate auditory input over time. The current experiment used 960 trials (60 repetitions/word). This could be completed in about an hour by a typical adult, but this is too long for other populations.

Thus, we replicated the analysis with fewer trials to determine if a robust response could be obtained with less data. The analysis was repeated using the first 75% (45 repetitions/item), 50% (30 repetitions/item) or 25% (15 repetitions/item) of trials. For each subject, this subset of the data was first extracted. We then estimated the two free parameters for the SVM using the same hybrid brute-force/pattern-search procedure at a single time (0.2 sec). Lastly, we repeated the timecourse analysis using these parameters. Note that at 25% of trials we could not use our standard 15-fold cross-validation which required at least one stimulus in each “fold” of the data – a single rejected trial (e.g., due to eye-blink or muscle artifact) would make this impossible. Thus, we backed off to a 10-fold cross-validation for this analysis of the first 25% of trials.

Fig. 6 shows the results. At 75% of trials (45 repetitions / item; Panel A, D), performance was barely reduced and not significantly different from baseline for subjects tested with both low impedance (Target: t(15) = 1.14, d = 0.28, p = 0.27; Cohort: t(15) = 0.59, d = −0.15, p = 0.56; Unrelated: t(15) = 1.08, d = −0.27, p = 0.30) and high impedance EEG (Target: t(14) = 1.48, d = −0.38, p = 0.16; Cohort: t(14) = 1.28, d = −0.33, p = 0.22; Unrelated: t(14) = 1.31, d = 0.34, p = 0.21). This can also be seen in individual subjects. Fig. 7 (top row) shows 6 representative subjects at 75% of trials and shows nearly identical performance to that with all trials (black curves).

With only 50% of trials (30 repetitions, Fig. 6B, E), we found a slight reduction in identification rates compared to 100% of trials for targets which was significant for low impedance EEG (t(14) = 2.23, d = 0.58, p = 0.043) but not high impedance: t(14) = 1.60, d = 0.41, p = 0.131). There was no significant effect for cohorts (Low Impedance: t(15) = 1.45, d = 0.36, p = 0.17; High Impedance: t(14) = 0.02, d = − 0.01, p = 0.98) and a moderate increase for unrelated items for low impedance: t(15) = 2.70, d = −0.68, p = 0.0163) but not high impedance (t(14) = 0.89, d = −0.23, p = 0.39). This was coupled with a breakdown in the pattern for some subjects (Fig. 7, subjects 1,2,23), but not others (e.g., 24, 25).

Finally, at 25% of trials (Fig. 6, C,F), the overall pattern still showed a competition profile, but it was dramatically reduced, with large differences for targets (Low Impedance: t(15) = 5.99, d = 1.50, p<0.001; High Impedance: t(14) = 3.71, d = 0.96, p = 0.0023), cohorts (High Impedance: t(15) = 3.00, d = 0.75, p = 0.0090; Low Impedance: t(14) = 2.12, d = 0.55, p = 0.0525), and unrelated items (High Impedance: t(15) = 6.64, d = −1.66, p<0.001; Low Impedance: t(14) = 3.83, d = −0.99, p = 0.0018). Few subjects showed the canonical profile (Fig. 7, bottom row).

Thus, good performance at the individual level can be obtained with 75% of the repetitions (45 reps / item) and group-level patterns are robust at 50% (30 repetitions). Note that this is on par with the number of repetitions used in machine learning approaches with electrocorticography (Nourski et al., 2015), which has much lower noise. While 640 trials are still substantial for many applications, experimenters may be able to reduce the number of items to further reduce the scale of the test, or possibly test across multiple sessions. A purely passive listening paradigm may also enable this larger number of trials feasible as the participant does not need to be actively engaged, and there is no time needed for an inter-trial interval or a behavioral response.

At a very small number of repetitions (25% / 15 repetitions), we found a weak pattern. These should be interpreted with caution, as not only is there less absolute data, but also because we could not perform a 15-fold validation with this data and had to back off to a 10-fold – a reduction in the amount of data actually used to train any individual SVM by about a third. While in principle one could employ less rigorous constraints on the cross validation, SVMs are sensitive to differences in the base frequency of individual words so we are cautious to recommend such a reduced number of trials per item.

3.6. Training Effects: Are the nonwords words?

The primary analysis found little evidence that words and nonwords were treated differently by the decoding approach. This suggests that decoding performance may be tapping something more akin to auditory memory and integration processes that serve spoken word recognition (and lexical access) rather than word recognition itself. An alternative hypothesis however is that with 60 repetitions, the non-words may have rapidly become lexicalized. Prior work has shown rapid integration of novel words into the lexicon using fewer repetitions and with minimal task demands such as used here (Kapnoula et al., 2015). One can address this by performing classification analyses separately on the first and second halves of the trials. If nonwords were rapidly lexicalized, then one might expect differences between words and non-words during the first half of the experiment but not the second.

To address this possibility, we conducted a new analysis in which subjects' trials were split in half, and the classification analysis was repeated separately for the early and late halves of the experiment. Because this hypothesis does not concern the kinds of measurement issues described above, and because our prior analyses show few differences in performance across the two EEG systems, we pooled across the two samples for a larger total sample size (N = 31). While this would not normally be done in a standard EEG analysis, this illustrates the power of using machine learning to convert the raw EEG to a performance metric which can now be appropriately pooled across systems. We note that our analysis of the number of trials needed for decoding (Fig.s 6,7) suggests that 50% (30 repetitions) is likely the minimum number of trials needed, and even this is likely only usable at the group level, not the individual level. Thus, we were anticipating an underpowered analysis and any conclusions drawn must be tentative.

Fig. 8 shows the results. When we compare the performance for *words* between early and late trials, we see few differences – targets and cohorts deviate from each other around 0.2 sec in both conditions, and the target remains higher than the cohort until a little before 1 second. However, in nonwords we see a subtly different pattern. During the early trials, targets and cohorts do not deviate until much later (after 400 msec) and show an overall smaller difference. In contrast the *late* nonword trials look quite similar to the words. This can be seen in Fig. 9, which shows the *difference* between target and cohort identification for words and nonwords separately in the early trials (Panel A) and late trials (Panel B).

To characterize these subtle differences, we used a simplified version of the analysis applied to the primary data set, conducting a significance test at each time window and correcting for multiple comparisons using the family-wise error correction of (Oleson et al., 2017). First, we compared targets to cohorts within each of the four conditions (early/late × words/nonwords). For the early *word* trials (Fig. 9A), the target deviated from the cohorts from 330 to 810 msec (σ = 0.962, α^* = .0082). In contrast for nonwords, the deviation did not begin until later and was significant from 390 to 770 msec (σ = .965, α^* = .0084). This suggests that during the early trials nonwords were not as robustly differentiated between targets and cohorts. When we turned to the late trials, however, we saw a different pattern in that both words (significant from 330 to 830 msec, σ = 0.966, α^* = .0085) and nonwords (significant from 330 to 770 msec; σ = 0.953, α^* = .0076) showed target and cohort onset at the same time. This supports the idea that the non-words may have been lexicalized over repeated presentations. Importantly, however, these conclusions should be tempered by the fact when we subtracted the target/cohort difference (e.g., Fig. 9), this metric did not differ between words and nonwords (e.g., comparing the purple and green bars in each panel) in either the early or late phase of the experiment. This is likely because the experiment was underpowered for detecting effects, particularly when only 50% of the data contributed to decoding.

Thus, there is some evidence that at early trials non-words behave differently than words, and in a way that is consistent with the idea that words more stably engage lexical competition. However, evidence is mixed, and we cannot rule out the possibility that words and non-words behave similarly at both early and late trials within the one-hour exposure to 60 tokens/item.

3.7. What features of the EEG support performance?

Finally, we asked which features of the EEG signal were essential for the excellent decoding seen in Fig.s 2 and 3, and to ask whether other theoretically driven features could be relevant. To do this, we used an approach similar to the feature selection that was initially run on the first five Iowa subjects, examining overall performance at a time window starting at 200 msec (where peak performance was observed; see Fig. 2), and testing a large set of features and other parameters. For each set of features, we saved the accuracy of target identification, after optimizing the cost and width of the basis function (the free parameters of the SVM). This was examined separately for each type of EEG system.

We began this examination by varying the duration of the time window using only the EEG features from our preliminary set (Fig. 10, left side of each panel). These were analyzed in a three-way ANOVA with impedance (high vs. low, between), duration (75, 125, 250 msec, within) and stimulus type (word/nonword, within). This showed a significant main effect of duration (F(2,58) = 34.1, p<.0001). This did not interact with lab (F(2,58) = 1.096, p = .34) or stimulus type (F<1). Follow-up comparisons showed that reducing the window size from 250 to 125 msec (starting in both cases at 200 msec) led to significantly lower performance (averaged across words and nonword (t(30) = 2.16, d = 0.39, p = 0.039) and a further reduction to 75 msec showed an even more substantial decrease (t(30) = 9.42, d = 1.69, p<0.0001). There was no main effect of impedance (F<1), impedance × mapping interaction (F<1), or three-way interaction (F<1). This confirms that our selection of EEG features from the first five subjects was robust for the performance of all subsequent subjects. Longer time windows may be necessary to accurately decode words, particularly when higher order temporal properties of the signal are extracted (e.g., slope, quadratic, cubic).

Next, we investigated the contribution of time-frequency features. On each trial, for each channel, we estimated the band-pass filtered power over time in five frequency bands: delta (0–3 Hz), theta (4–7 Hz.), alpha (8–15 Hz.), beta (15–30 Hz.) and gamma (30–70 Hz). Theta, was of particular interest given claims that it is involved in grouping acoustic input for speech perception (Giraud & Poeppel, 2012). We trained the classifier on the average power from 200 to 450 msec in a single frequency band, or by combining that band with the ERP features (3rd order polynomial) (Nourski et al., 2015).

Fig. 10 (right side of each panel) shows the results. These were analyzed with an ANOVA assessing impedance (high/low, within), EP (EP+TF vs. TF only, within), frequency band (δ, θ, α, β, γ, within), and item type (word/nonword, within). Frequency-band features alone showed performance that was slightly (~2%) above chance. Moreover, there was no main effect of frequency band (F(3,87) = 1.083, p = .361) suggesting a similar effect across all bands. While frequency-band information was significantly above chance overall (p<.001 by one sample t-tests), this was within the limits of the permutation analyses in Fig. 4. Finally, exploratory smaller scale permutation analyses on just the frequency-band information confirmed that there was no significant evidence that these frequency bands contribute to decoding performance. In contrast, decoding accuracy improved dramatically when ERP features were added (F(1,29) = 255.4, p<.0001), and this did not interact with impedance (F<1), or frequency band (F<1). There was no effect of impedance (F(3,87) = 1.359, p = .26).or of word/nonword status (F<1) and no interaction with any other factors (all p>.2).

Finally, adding the frequency-band information to the ERP features showed a small but significant *decrement* compared to the ERP features alone for all four frequency bands (θ: t(30) = 5.97, p<.0001; α: t(30) = 4.64, p = .0038; β: t(30) = 4.32, p = .0006; λ: t(15) = 6.15, p<.0001). This probably reflects the fact that SVM classifiers performs more poorly in general with too many variables. Thus, we see little evidence that time-frequency information carries any useful information for decoding the auditory input, and substantially more information is carried by the ERP.

4. Discussion

This project sought to develop a robust neural paradigm that can reveal the detailed timecourse of auditory integration and phonological competition using standard EEG procedures combined with off-the-shelf machine-learning techniques. Cognitive science largely using the Visual World Paradigm has shown distinct differences in the detailed timecourse of lexical competition associated with a variety of language and cognitive disorders (Desroches et al., 2006; McMurray et al., 2017; McMurray et al., 2010). However, the VWP may be of limited use with special populations who cannot follow instructions or who have impairments of attention/eye-movement control, and it can only test a limited subset of words (cf., those that are picturable). Moreover, the VWP relies on eye-movements that exhibit intrinsic delays compared to the actual rate of processing spoken language (McMurray, in press). A neural measure (EEG) could not only more directly tap into the underlining SWR process, but also would not require visual referents, opening up the entire lexicon for detailed scrutiny.

Our results provide compelling evidence that EEG can yield a highly reliable assay of the competition processes that undergird word recognition. In fact, there was a remarkable level of similarity between results from computational models (Fig. 1A), canonical VWP studies (Fig. 1B) and our neural paradigm (Fig. 2). Our decoding analyses confirmed the same characteristic pattern in which the target word and its phonological (cohort) competitor are active immediately after word onset (at levels that were greater than unrelated items). This is followed by a further rise in target activation and a fall-off in cohort activation. These results using non-invasive EEG signals are similar to those obtained using ECoG (Rhone et al., 2022), lending support to the inference that they are measuring a similar neural substrate, despite the spatial filtering that occurs from scalp-based electrodes.

These EEG measures of speech decoding were remarkably sensitive at the individual subject level. Decoding performance for individual subjects accurately reflected the group patterns (Fig. 3), and cross-trial reliability was very high for many aspects of the decoding functions (Table 1). While more detailed psychometric work is necessary, our analyses suggest that our neural paradigm has great promise for developmental and clinical applications, an essential milestone for diagnostic work. Specifically, a variety of clinical disorders (delayed language development, developmental language disorder, dyslexia, hearing loss, auditory neuropathy) have been associated with impairments in phonological processing, auditory integration, and lexical competition (McMurray et al., 2022). This suggests that many of these disorders may show distinct profiles when we examine the timecourse of speech processing using our neural paradigm.

Also encouraging was the fact that our neural paradigm was robust across several factors (see Table 4 for a summary). First, two different EEG systems (gel-based and saline-based) provided nearly identical results despite different numbers of electrodes and different levels of impedance. This latter factor is important as traditional ERPs are highly sensitive to this variable (Kappenman & Luck, 2010). This cross-platform consistency was impressive since the specific EEG features and time-windows that were optimized for one system were applied directly to the other. Second, the number of channels did not affect decoding accuracy unless they were reduced below 32 or were limited to the fronto-central electordes. This suggests that our neural paradigm is amenable to less expensive low-density EEG systems used for clinical and developmental applications. Third, the number of stimulus repetitions per item required for reliable decoding of the timecourse of SWR is not so high as to prevent its utility for special populations (e.g., reducing the set of items from 16 to 8 and the number of repetitions from 60 to 30 would result in the entire training and testing protocol being completed in 30 minutes).

An unresolved interpretive issue is what the classifier is using as the “neural code” for words. Our classifier performed best with longer time windows, and with a higher-order polynomial description of the EEG signal over time. Thus, there is substantial information in the dynamics of the EEG signal over a temporal window of 200 msec. This may reflect in part some form of neural entrainment to the envelope of the signal (Brodbeck & Simon, 2020) and this kind of information may be embedded in analyses based on the phase of oscillations in specific frequency bands. However, we note that the absolute magnitude of entrainment detectable in scalp EEG is often quite low (cross correlations ~ 0.1), suggesting this may not be sufficient. Moreover, decoding accuracy was remarkably reduced if only fronto-central electrodes (which respond to auditory areas) were used (Fig. 5E-H). This suggests the need for a broadly distributed set of neural generators to maximize SWR performance in our neural paradigm, rather than just the fronto-central electrodes that typically have the strongest association with auditory areas. In addition, we found poor performance with features based on power in traditional EEG frequency bands alone, and they did not improve performance over and above the ERP features alone. While these oscillations are undoubtedly involved in language, this suggests they do not carry information about the content of auditory input, but may play a role as modulators of this content (e.g., participating in word segmentation).

One final question is whether our neural measure is tapping lexical processing or the prelexical auditory analysis of speech signals. Our goal was not to distinguish these two levels, but to focus on the prelexical auditory analysis where there is a distinct methodological need. Nonetheless, this remains an important question. In this regard, overall performance did not differ between words and nonwords (Fig. 2). Critically, we note that our decoding approach was not asked to discriminate words from nonwords (indeed, this was not possible[3]). Rather, we asked if word/non-word status modulated the decision dynamics and found no difference. There are several possibilities for interpreting the absence of this lexical/nonlexical effect.

First, our measure may actually reflect some aspect of word recognition, but with 60 repetitions per item, the nonwords may have become rapidly lexicalized over the course of the experiment (Kapnoula et al., 2015), and Fig. 8 shows limited evidence for such a learning effect. Supporting this, we note that performance was maximal with electrodes across the whole head. This suggests that our response does not merely reflect low-level auditory encoding.

Second, there may have been a subtle word/non-word difference but we were unable to detect it. Gwilliams et al. (2020), present an encoding model that shows much more robust effects of surprisal in sentence context than isolated words (as was used here). This seems unlikely to account for the differences between our results and those of Gwilliams et al. as our measure was not surprisal but the efficiency of activating words, and recent work with the VWP suggests little effect of sentence contexts on this measure (Smith & McMurray, in press). However prior work using encoding models also suggests that word/nonword effects may only be detectable for words with late points of disambiguation (Di Liberto et al., 2019), while most of our words could be disambiguated after the 2nd or 3rd phoneme. This seems like a more likely explanation for the failure to find word/nonword differences in our results.

Third, because the encoding approach has been implemented with continuous speech containing sequences of words, rather than isolated words as in the present work, measures of surprisal (at both the phonemic and lexical levels) can be obtained (Brodbeck, Hong, et al., 2018; Gwilliams et al., 2018). These surprisal results provide clear evidence that listeners are sensitive to the ordering of linguistic events, which differ between sequences of words and nonwords. In contrast, because our nonwords were highly phonotactically legal and repeated many times, there may have been little surprisal. Future work using our decoding approach could be expanded to include words and nonwords embedded in sentences (potentially eliciting surprisal effects) to determine whether effects of phoneme or lexical sequencing alter decoding accuracy. The fact that it did not suggests that at some

[3] The set of words and nonwords for a given subject were highly distinct (e.g., a subject who heard *badger/baggage* as words did not also hear *babbid/baddow* as nonwords). Consequently, attempts to use the classifier to decode word vs. nonword performance were not attempted as good performance could be driven entirely by phonemic dissimilarity of the words and nonwords.

level the processes of integrating auditory information over time to support a decision may be similar for both words with a rich semantic structure and meaningless (but with familiar wordforms) non-words.

Finally, an intriguing possibility is that we found no differences because the words were treated as nonwords. Our words were heard in isolation, in a task that required no semantic analysis, and were repeated many times. As a result, the words may not have fully engaged the language network, and subjects may have been somewhat inattentive to higher level lexical factors by the end of the experiment. Consequently, the words may have been processed as efficiently as the nonwords.

In summary, while our results do not unambiguously index *word* recognition, they provide clear evidence of tracking a host of processes that are relevant to it: rapid encoding of the auditory input, integration and accumulation of input over time, and dynamic decision making. Some of these processes may be seen as a form of “speech tracking” in which the EEG signal passively reflects the auditory input. Indeed the longer time window for analysis, for example, may be crucial for picking up things like the morphology of the N1/P2 complex which lasts several hundred msec and has been linked to specific phonetic features like voicing (Frye et al., 2007; Toscano et al., 2010). Such processes are likely critical to the broader set of processes needed to support word recognition. What is critical here is the ability to see how those processes are integrated to lead to the temporal dynamics of word recognition, a picture that may be crucial to understanding a variety of communication disorders. Thus, the paradigm we have described here may serve as a robust indicator of the integrity of a constellation of critical complex auditory functions that are essential for spoken word recognition and cannot be assessed in other ways in subject populations who are unable to perform standard psycholinguistic tasks.

Acknowledgements

The authors would like to thank Ariane Rhone for critical assistance in developing the machine learning pipeline and Hiroyuki Oya and Matthew Howard for support in developing these techniques. This project was supported by R21DC017596 awarded to BM and RNA, DC008089 to BM, and DC 004290 awarded to Matthew Howard III.

References

Allopenna P, Magnuson JS, Tanenhaus MK, 1998. Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. Journal of Memory and Language 38 (4), 419–439.

Apfelbaum KS, Goodwin C, Blomquist C, McMurray B, 2022. The development of lexical competition in written and spoken word recognition. Quarterly Journal of Experimental Psychology

Audacity Team, 2015. Audacity: Free Audio Editor and Recorder

Bae G-Y, Luck SJ, 2018. Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. Journal of Neuroscience 38 (2), 409–422. [PubMed: 29167407]

Bates D, Sarkar D, 2011. lme4: Linear mixed-effects models using S4 classes. GNU Public License

Beach SD, Ozernov-Palchik O, May SC, Centanni TM, Gabrieli JDE, Pantazis D, 2021. Neural Decoding Reveals Concurrent Phonemic and Subphonemic Representations of Speech Across Tasks. Neurobiology of Language 2 (2), 254–279. doi:10.1162/nol_a_00034. [PubMed: 34396148]

Ben-David BM, Chambers CG, Daneman M, Pichora-Fuller MK, Reingold EM, Schneider BA, 2011. Effects of aging and noise on real-time spoken word recognition: Evidence from eye movements. Journal of Speech, Language, and Hearing Research 54 (1), 243–262.

Blank H, Davis MH, 2016. Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLoS Biology 14 (11), e1002577. doi:10.1371/journal.pbio.1002577. [PubMed: 27846209]

Blumstein SE, Myers EB, Rissman J, 2005. The perception of voice onset time: an fMRI investigation of phonetic category structure. Journal of Cognitive Neuroscience 17 (9), 1353–1366. [PubMed: 16197689]

Boersma P, Weenink D, 2009. Praat: doing phonetics by computer In (Version Version 5.1.05).

Brandmeyer A, Farquhar JDR, McQueen JM, Desain PWM, 2013. Decoding Speech Perception by Native and Non-Native Speakers Using Single-Trial Electrophysiological Data. PLoS ONE 8 (7), e68261. doi:10.1371/journal.pone.0068261. [PubMed: 23874567]

Brodbeck C, Hong LE, Simon JZ, 2018. Rapid transformation from auditory to linguistic representations of continuous speech. Current Biology 28 (24), 3976–3983 e3975. [PubMed: 30503620]

Brodbeck C, Presacco A, Simon JZ, 2018. Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. Neuroimage 172, 162–174. [PubMed: 29366698]

Brodbeck C, Simon JZ, 2020. Continuous speech processing. Current Opinion in Physiology

Broderick MP, Di Liberto GM, Anderson AJ, Rofes A, Lalor EC, 2021. Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. Scientific reports 11 (1), 4963. doi:10.1038/s41598-021-84597-9. [PubMed: 33654202]

Brouwer S, Bradlow AR, 2015. The Temporal Dynamics of Spoken Word Recognition in Adverse Listening Conditions [journal article]. Journal of Psycholinguistic Research 1–10. doi:10.1007/s10936-015-9396-9.

Brouwer S, Mitterer H, Huettig F, 2012. Speech reductions change the dynamics of competition during spoken word recognition. Language and Cognitive Processes 27 (4), 539–571.

Chang C-C, Lin C-J, 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Choi HS, Marslen-Wilson WD, Lyu B, Randall B, Tyler LK, 2020. Decoding the Real-Time Neurobiological Properties of Incremental Semantic Interpretation. Cerebral Cortex 31 (1), 233–247. doi:10.1093/cercor/bhaa222.

Clayards M, Tanenhaus MK, Aslin RN, Jacobs RA, 2008. Perception of speech reflects optimal use of probabilistic speech cues. Cognition 108 (3), 804–809. [PubMed: 18582855]

Dahan D, Magnuson JS, 2006. Spoken-word recognition. In: Traxler MJ, Gernsbacher MA (Eds.), Handbook of Psycholinguistics. Academic Press, pp. 249–283.

Dahan D, Magnuson JS, Tanenhaus MK, Hogan E, 2001. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes 16 (5/6), 507–534.

Dahan D, Tanenhaus MK, 2004. Continuous Mapping From Sound to Meaning in Spoken-Language Comprehension: Immediate Effects of Verb-Based Thematic Constraints. Journal of Experimental Psychology: Learning Memory and Cognition 30 (2), 498–513. [PubMed: 14979820]

Delorme A, Makeig S, 2004. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134, 9–21. [PubMed: 15102499]

Desroches AS, Joanisse MF, Robertson EK, 2006. Phonological deficits in dyslexic children revealed by eyetacking. Cognition 100, B32–B42. [PubMed: 16288732]

Di Liberto GM, Wong D, Melnik GA, de Cheveigné A, 2019. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage 196, 237–247. [PubMed: 30991126]

DiLiberto GM, O’Sullivan JA, Lalor EC, 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology 25 (19), 2457–2465. [PubMed: 26412129]

Donhauser PW, Baillet S, 2020. Two Distinct Neural Timescales for Predictive Speech Processing. Neuron 105 (2), 385–393. doi:10.1016/jh.neuron.2019.10.019, e389. [PubMed: 31806493]

Elman JL, McClelland JL, 1986. Exploiting lawful variability in the speech wave. In: Perkell JS, Klatt D (Eds.), Invariance and variability in speech processes Erlbaum, pp. 360–380.

Etard O, Reichenbach T, 2019. Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. The Journal of Neuroscience 39 (29), 5750–5759. doi:10.1523/jneurosci.1828-18.2019. [PubMed: 31109963]

Farris-Trimble A, McMurray B, Cigrand N, Tomblin JB, 2014. The process of spoken word recognition in the face of signal degradation: Cochlear implant users and normal-hearing listeners. Journal of Experimental Psychology: Human Perception and Performance 40 (1), 308–327. [PubMed: 24041330]

Frye RE, McGraw Fisher J, Cody A, Zarella M, Liederman J, Halgren E, 2007. Linear coding of voice onset time. Journal of Cognitive Neuroscience 19, 1476–1487. [PubMed: 17714009]

Gagnepain P, Henson RN, Davis MH, 2012. Temporal predictive codes for spoken words in auditory cortex. Current Biology 22 (7), 615–621. [PubMed: 22425155]

Galle ME, 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. doi:10.1111/cogs.12700.

Getz LM, Toscano JC, 2021. The time-course of speech perception revealed by temporally-sensitive neural measures. Wiley Interdisciplinary Reviews: Cognitive Science 12 (2), e1541. [PubMed: 32767836]

Gillis M, Vanthornhout J, Simon JZ, Francart T, Brodbeck C, 2021. Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics. The Journal of Neuroscience 41 (50), 10316–10329. doi:10.1523/jneurosci.0812-21.2021. [PubMed: 34732519]

Giraud A-L, Poeppel D, 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience 15 (4), 511. [PubMed: 22426255]

Goldinger SD, 1998. Echoes of Echos? An episodic theory of lexical access. Psychological Review 105, 251–279. [PubMed: 9577239]

Gow DW, Olson B, 2016. Sentential influences on acoustic-phonetic processing: A Granger causality analysis of multi-modal imaging data. Language, Cognition and Neuroscience 31 (7), 841–855. [PubMed: 27595118]

Grootswagers T, Wardle SG, Carlson TA, 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. doi:10.1162/jocn_a_01068. [PubMed: 27779910]

Gwilliams L, King JR, Marantz A, Poeppel D, 2020. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. BioRxiv 10.1101/2020.04.04.025684v2.abstract.

Gwilliams L, Linzen T, Poeppel D, Marantz A, 2018. Spoken Word Recognition, the Future Predicts the Past. The Journal of Neuroscience 38 (35), 7585–7599. doi:10.1523/jneurosci.0065-18.2018. [PubMed: 30012695]

Hannagan T, Magnuson J, Grainger J, 2013. Spoken word recognition without a TRACE [Original Research]. Frontiers in Psychology (563) 4. doi:10.3389/fpsyg.2013.00563. [PubMed: 23378839]

Hendrickson K, Spinelli J, Walker E, 2020. Cognitive processes underlying spoken word recognition during soft speech. Cognition 198, 104196. doi:10.1016/j.cognition.2020.104196. [PubMed: 32004934]

Kapnoula E, Packard S, Gupta P, McMurray B, 2015. Immediate lexical integration of novel word forms. Cognition 134 (1), 85–99. [PubMed: 25460382]

Kappenman ES, Luck SJ, 2010. The effects of electrode impedance on data quality and statistical significance in ERP recordings. Psychophysiology 47 (5), 888–904. doi:10.1111/j.1469-8986.2010.01009.x. [PubMed: 20374541]

Kazanina N, Phillips C, Idsardi W, 2006. The influence of meaning on the perception of speech sounds. Proceedings of the National Academy of Sciences 103 (30), 11381–11386. doi:10.1073/pnas.0604821103.

King JR, Dehaene S, 2014. Characterizing the dynamics of mental representations: the temporal generalization method. Trends in Cognitive Sciences 18 (4), 203–210. doi:10.1016/j.tics.2014.01.002. [PubMed: 24593982]

Kocagoncu E, Clarke A, Devereux BJ, Tyler LK, 2017. Decoding the Cortical Dynamics of Sound-Meaning Mapping. The Journal of Neuroscience 37 (5), 1312–1319. doi:10.1523/jneurosci.2858-16.2016. [PubMed: 28028201]

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

Luthra S, Guediche S, Blumstein SE, Myers EB, 2019. Neural substrates of subphonemic variation and lexical competition in spoken word recognition. Language, Cognition and Neuroscience 34 (2), 151–169. [PubMed: 31106225]

MacDonald MC, Pearlmutter NJ, Seidenberg MS, 1994. Lexical nature of syntactic ambiguity resolution. Psychological Review 101, 676–703. [PubMed: 7984711]

Marslen-Wilson WD, 1987. Functional parallelism in spoken word recognition. Cognition 25 (1–2), 71–102. [PubMed: 3581730]

McClelland JL, Elman JL, 1986. The TRACE model of speech perception. Cognitive Psychology 18 (1), 1–86. [PubMed: 3753912]

McMurray B, 2022. I'm not sure that curve means what you think it means: Toward a more realistic understanding of eye-movement control in the Visual World Paradigm Psychonomic Bulletin & Review. https://psyarxiv.com/pb2c6/.

McMurray B, Apfelbaum KS, Tomblin JB, 2022. The slow development of real-time processing: Spoken Word Recognition as a crucible for new about thinking about language acquisition and disorders. Current Directions in Psychological Science doi:10.1177/09637214221078325. https://psyarxiv.com/uebfc/.

McMurray B, Clayards M, Tanenhaus MK, Aslin RN, 2008. Tracking the time course of phonetic cue integration during spoken word recognition. Psychonomic Bulletin and Review 15 (6), 1064–1071. [PubMed: 19001568]

McMurray B, Farris-Trimble A, Rigler H, 2017. Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. Cognition 169, 147–164. [PubMed: 28917133]

McMurray B, Jongman A, 2011. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review 118 (2), 219–246. [PubMed: 21417542]

McMurray B, Samelson VS, Lee SH, Tomblin JB, 2010. Individual differences in online spoken word recognition: Implications for SLI. Cognitive Psychology 60 (1), 1–39. [PubMed: 19836014]

McMurray B, Tanenhaus MK, Aslin RN, 2002. Gradient effects of within-category phonetic variation on lexical access. Cognition 86 (2), B33–B42. [PubMed: 12435537]

Norman KA, Polyn SM, Detre GJ, Haxby JV, 2006. Beyond mind-reading: multivoxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10 (9), 424–430. [PubMed: 16899397]

Nourski KV, Steinschneider M, Rhone AE, Oya H, Kawasaki H, Howard MA 3rd, McMurray B, 2015. Sound identification in human auditory cortex: Differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. Brain and Language 148, 37–50. doi:10.1016/j.bandl.2015.03.003. [PubMed: 25819402]

Oleson JJ, Cavanaugh JE, 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. doi:10.1177/0962280215607411. [PubMed: 26400088]

Prabhakaran R, Blumstein SE, Myers EB, Hutchison E, Britton B, 2006. An event-related fMRI investigation of phonological–lexical competition. Neuropsychologia 44 (12), 2209–2221. doi:10.1016/j.neuropsychologia.2006.05.025. [PubMed: 16842827]

Revill KP, Spieler DH, 2012. The effect of lexical frequency on spoken word recognition in young and older listeners. Psychology and aging 27 (1), 80. [PubMed: 21707175]

Rhone AE, Farris-Trimble A, Nourski KV, Kawasaki H, Howard MA III, McMurray B, Neural decoding reveals the functional anatomy of auditory integration and competition in speech perception, 2022, https://psyarxiv.com/bd6eh/.

Righi G, Blumstein SE, Mertus J, Worden MS, 2009. Neural Systems underlying Lexical Competition: An Eye Tracking and fMRI Study. Journal of Cognitive Neuroscience 22 (2), 213–224. doi:10.1162/jocn.2009.21200.

Rigler H, Farris-Trimble A, Greiner L, Walker J, Tomblin JB, McMurray B, 2015. The slow developmental timecourse of real-time spoken word recognition. Developmental Psychology 51 (12), 1690–1703. [PubMed: 26479544]

Salverda AP, Dahan D, McQueen J, 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition 90 (1), 51–89. [PubMed: 14597270]

Sarrett M, McMurray B, Kapnoula E, 2020. Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and semantic expectations. Brain and Language 211, 104875. [PubMed: 33086178]

Sarrett ME, Shea C, McMurray B, 2022. Within-and between-language competition in adult second language learners: implications for language proficiency. Language, Cognition and Neuroscience 37 (2), 165–181.

Seedorff M, Oleson J, McMurray B, 2022. Maybe maximal: Good enough mixed models optimize power while controlling type I error

Smith FX, McMurray B, 2022. Lexical Access Changes Based on Listener Needs: Real-Time Word Recognition in Continuous Speech in Cochlear Implant User. Ear and Hearing https://psyarxiv.com/wyaxd/.

Spivey MJ, Marian V, 1999. Cross talk between native and second languages: Partial activation of an irrelevant lexicon. Psychological Science 10 (3), 281–284.

Strauss TJ, Harris HD, Magnuson JS, 2007. jTRACE: a reimplementation and extension of the TRACE model of speech perception and spoken word recognition. Behav Res Methods 39 (1), 19–30. [PubMed: 17552468]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC, 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634. [PubMed: 7777863]

Toscano JC, McMurray B, Dennhardt J, Luck S, 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. [PubMed: 20935168]

Vitevitch MS, Luce PA, 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, and Computers 36, 481–487.

Weissbart H, Kandylaki KD, Reichenbach T, 2020. Cortical Tracking of Surprisal during Continuous Speech Comprehension. Journal of Cognitive Neuroscience 32 (1), 155–166. doi:10.1162/jocn_a_01467. [PubMed: 31479349]

Xie Z, Reetske R, Chandrasekaran B, 2019. Machine learning approaches to analyze speech-evoked neurophysiological responses. Journal of Speech, Language, and Hearing Research 62, 597–601.

Zangl R, Klarman L, Thal D, Fernald A, Bates E, 2005. Dynamics of Word Comprehension in Infancy: Developments in Timing, Accuracy, and Resistance to Acoustic Degradation. Journal of Cognition and Development 6 (2), 179–208. [PubMed: 22072948]

Zhuang J, Randall B, Stamatakis EA, Marslen-Wilson WD, Tyler LK, 2011. The Interaction of Lexical Semantics and Cohort Competition in Spoken Word Recognition: An fMRI Study. Journal of Cognitive Neuroscience 23 (12), 3778–3790. doi:10.1162/jocn_a_00046. [PubMed: 21563885]

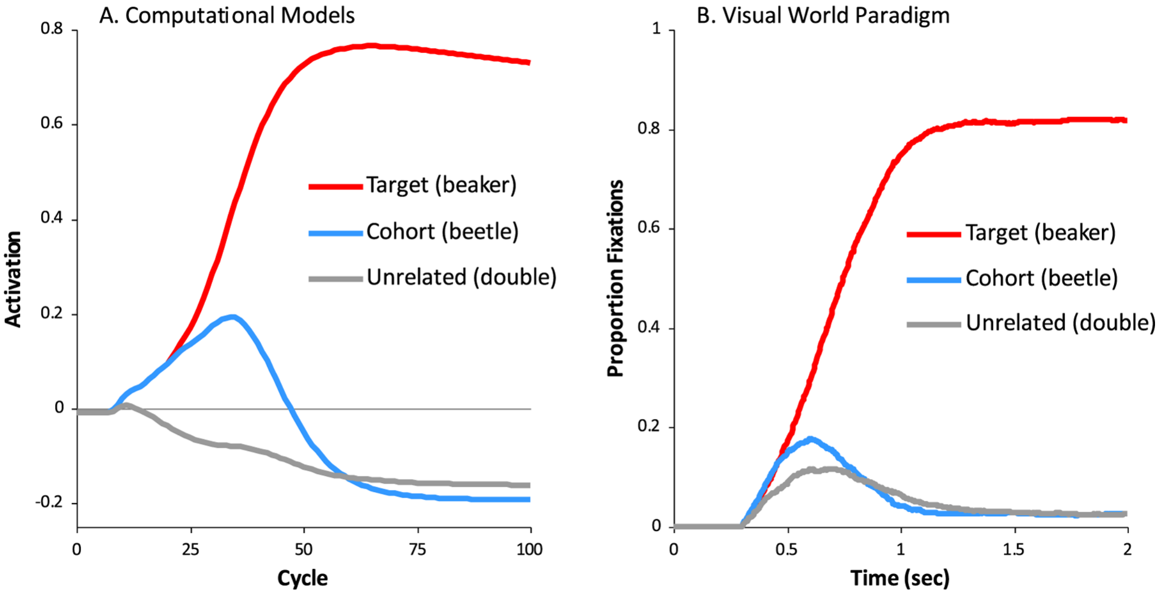

Fig. 1. The dynamics of spoken word recognition: model and human behavior.

A) activation as a function of time in the TRACE model (McClelland & Elman, 1986) as a target word (here, *beaker*) is heard. Words are differentially active depending on match to the input. Activations generated with TRACE (Strauss et al., 2007). B) Competition dynamics can be characterized in humans using the Visual World Paradigm (Allopenna et al., 1998; Salverda et al., 2003) in which eye movements to pictures representing various lexical candidates are monitored while the subject hears a target word (here, *wizard*) Shown is the likelihood of fixating each object over time after hearing wizard for 40 typical adolescents (Clayards et al., 2008; McMurray et al., 2010).

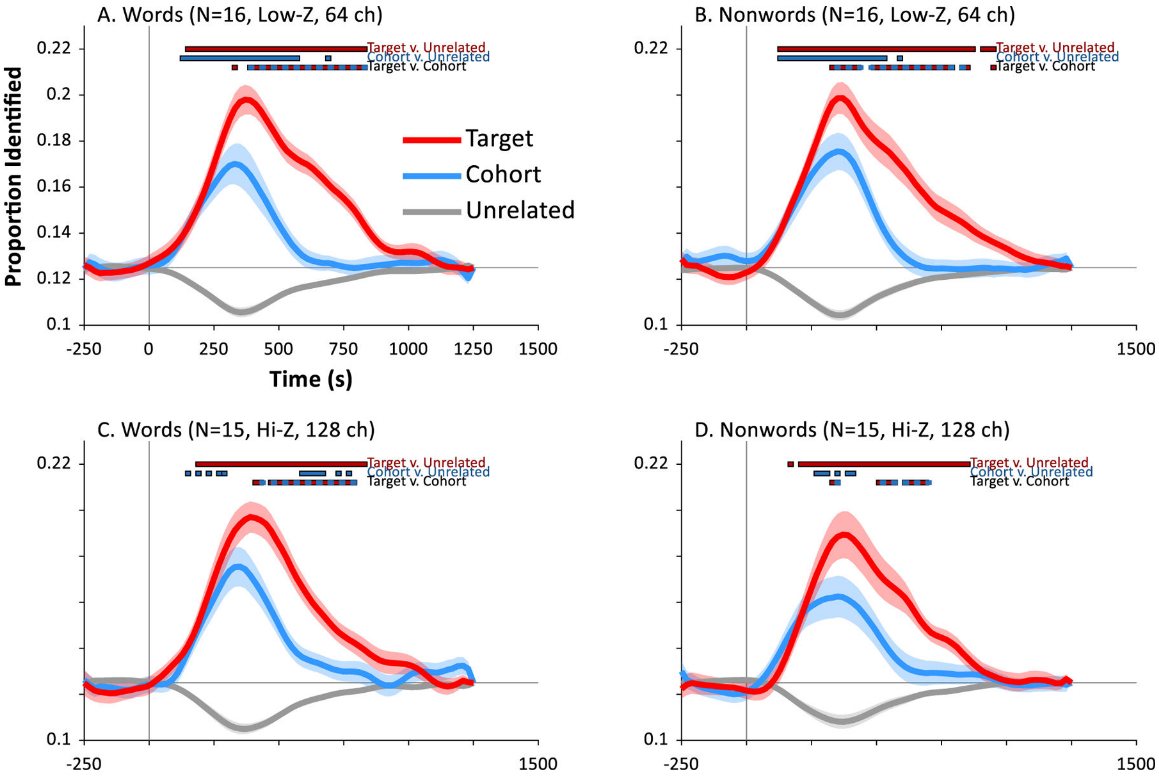

Fig. 2.

Results of machine learning analysis. Each panel shows the likelihood of the classifier choosing the target (the word that was heard, e.g., *badger*), its cohort (an onset competitor, e.g., *baggage*) or an unrelated word (e.g., *mushroom*) as a function of time. Word onset is at 0 sec. Each point marks the onset of a bin (e.g., data at .25 sec represents a classification analysis using EEG data from .250 to .500 sec). Chance (thin gray line) is 0.125. A) Results for classification among the 8 words in subjects tested on a 64-channel low impedance EEG at Iowa (N = 16). B) Results for classification among the 8 nonwords in the Iowa sample. C) Results for classification among the words for subjects tested in a 128-channel high impedance EEG at Haskins (N = 15). D) Results for classification among nonwords for Haskins sample.

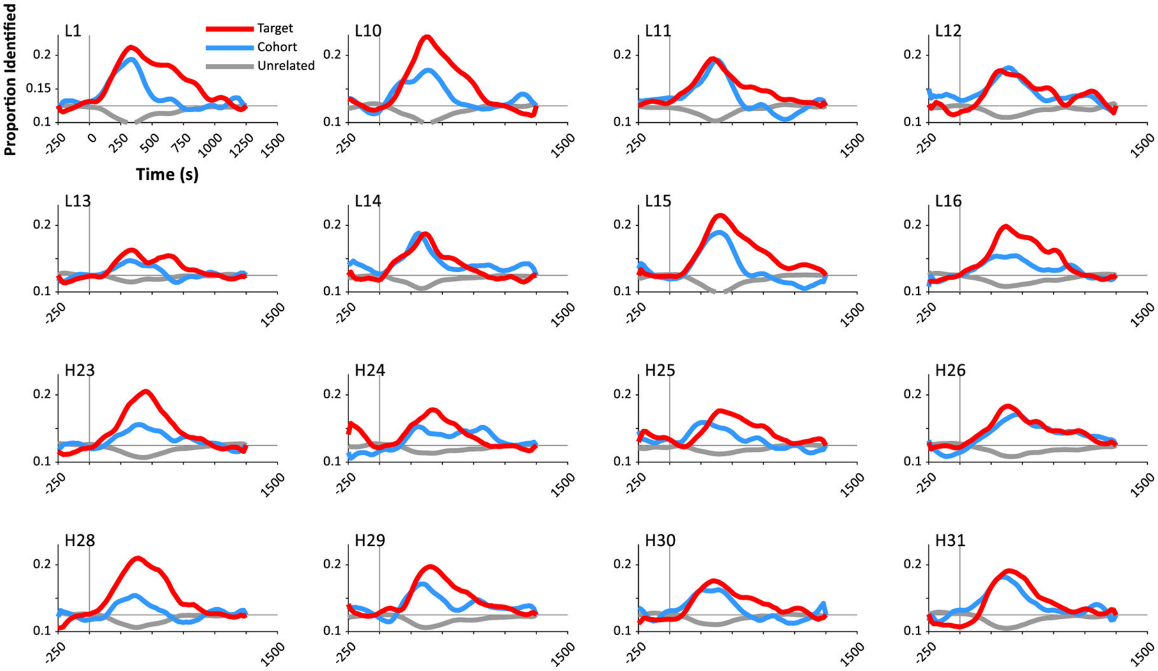

Fig. 3.

Performance on the classifier over time for representative individual subjects. Each panel shows the likelihood of selecting the target (the word the subject heard, e.g., *baggage*), its onset competitor or cohort (*badger*) or an unrelated word at each time. Each point represents the start of the time-window used for the training data. Chance (marked in black) – is 0.125. The first 8 subjects are from the low impedance / Iowa sample; the next 8 are from the high impedance / Haskins sample.

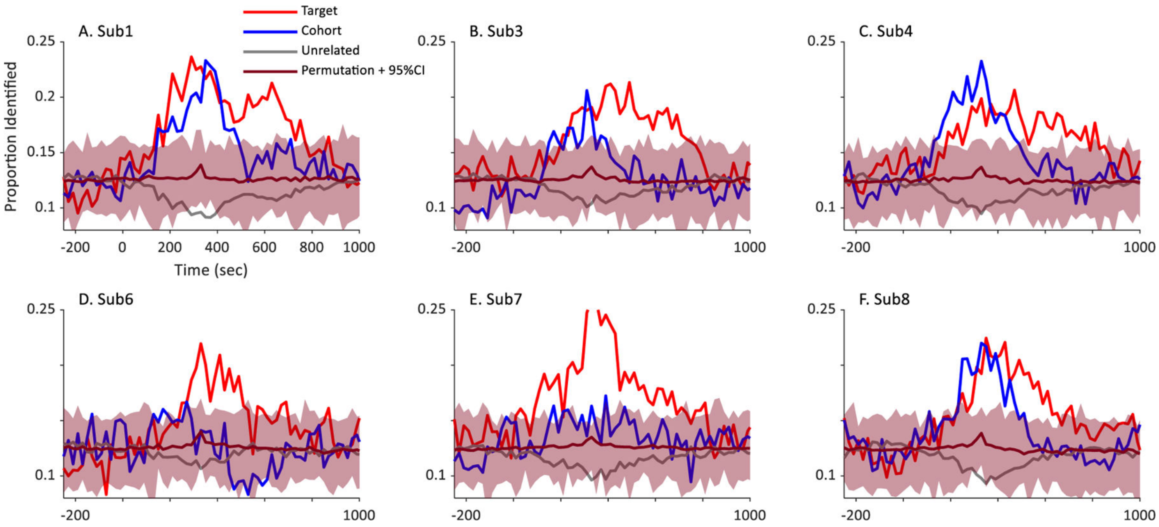

Fig. 4.

Results of a permutation test conducted on fix individual subjects. Shown is each subject’s proportion of target, cohort, and unrelated identifications over time (e.g., Fig. 3, main text). The magenta bar represents the mean of 100 runs in which the assignment of stimuli to trials was randomized; the confidence intervals represent the values between which 95% of the observed results fell. Note that subjects 1,3 and 4 (top row) were subjects on which the initial feature identification was conducted; subjects 6–8 (bottom row) were not.

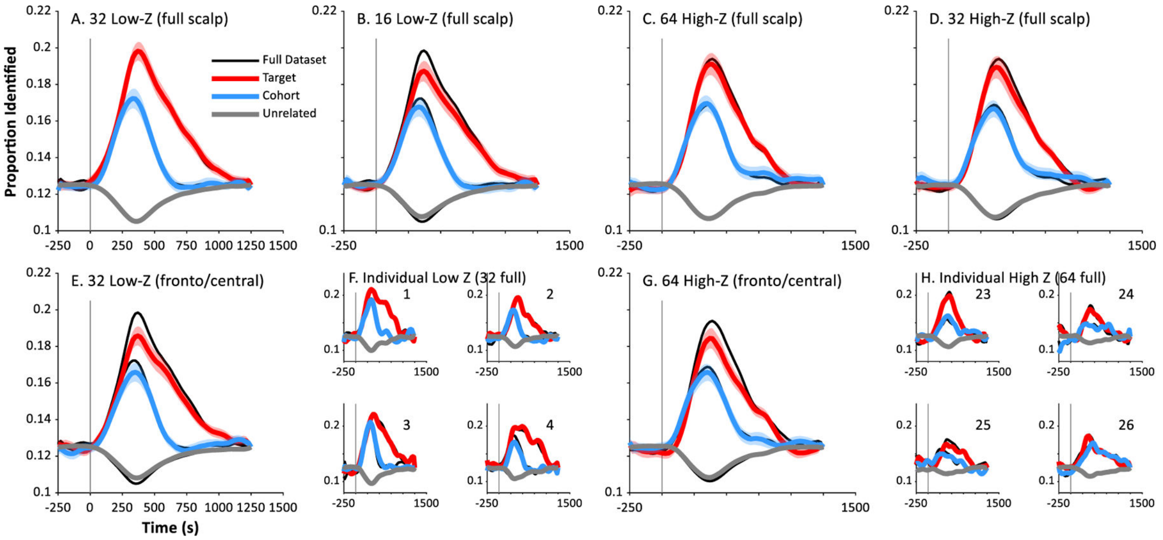

Fig. 5.

Performance of the classifier as a function of channel configuration averaged across words and nonwords. In each panel, the original data (e.g., the average of words and nonwords) is shown in black; red, blue and gray lines depict results of an identical analysis with a reduced number of channels. A) For low impedance EEG, 32 channels (sampled from across the scalp) yields identical performance to 64 (the black curves are behind the colored); B) For low impedance systems, 16 channels yield significant reductions in both peak target and cohort identification. C) High impedance systems with 64 channels show little performance decrement relative to the full 128 channels: D) High impedance EEG with 32 channels shows noticeable reductions. E) In contrast to A, the use of only the 32 fronto-central channels shows large drops in performance. F) Four representative low impedance subjects for analyses with 32 channels (full scalp), matching A. G) In contrast to C, the use of only the 64 fronto-central channels shows a decrement in performance for high impedance systems. H) Four representative high impedance subjects for analyses with 64 channels (full scalp), matching C.

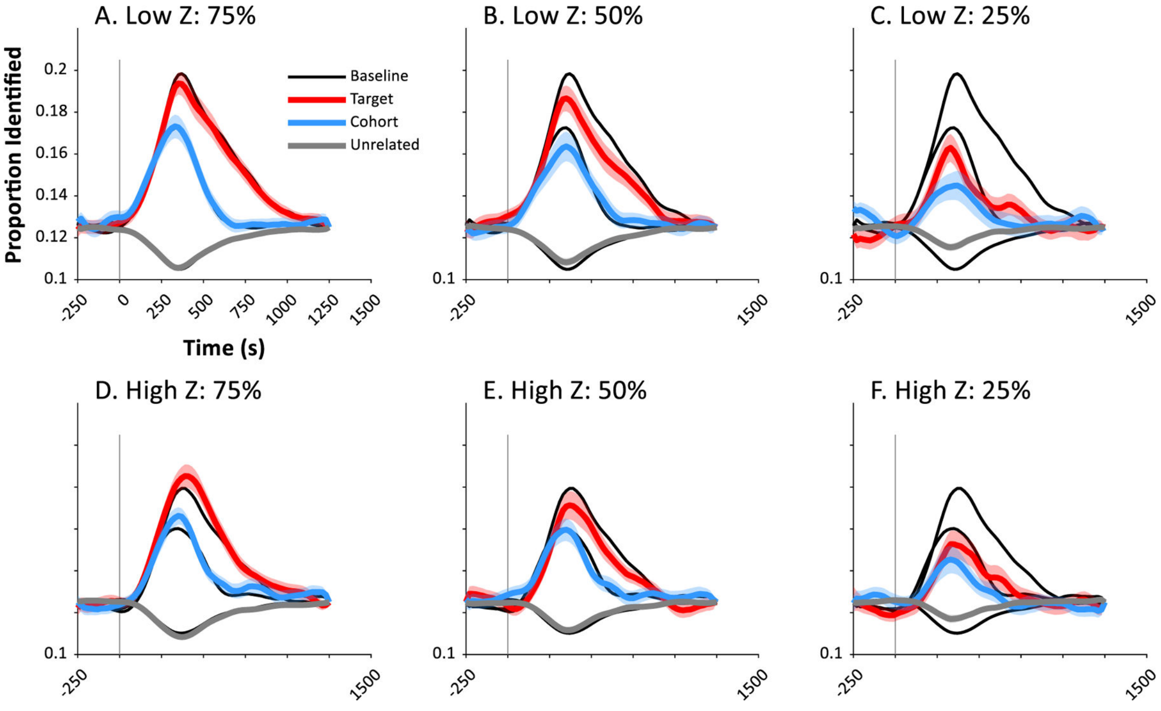

Fig. 6.

Effect of number of trials on performance. In each curve, the black represents the original data trained on 100% of trials (e.g., Fig. 2, main text). Colored curves are the same results trained on the first 75%, 50% or 25% of trials. Top row: Low impedance systems; Bottom row: High impedance.

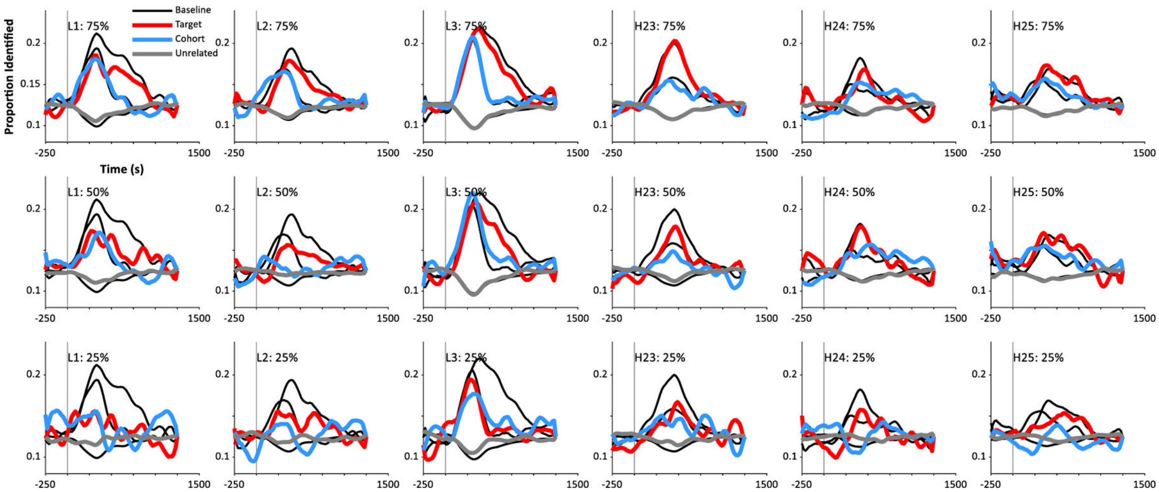

Fig. 7.

Effect of number of trials on performance in individual subjects. Each column represents a subject. Top row: 75% of trials; Middle: 50% of trials, Bottom: 25% of trials. The left 3 subjects were tested with the low impedance EEG; the right three with a high impedance system.

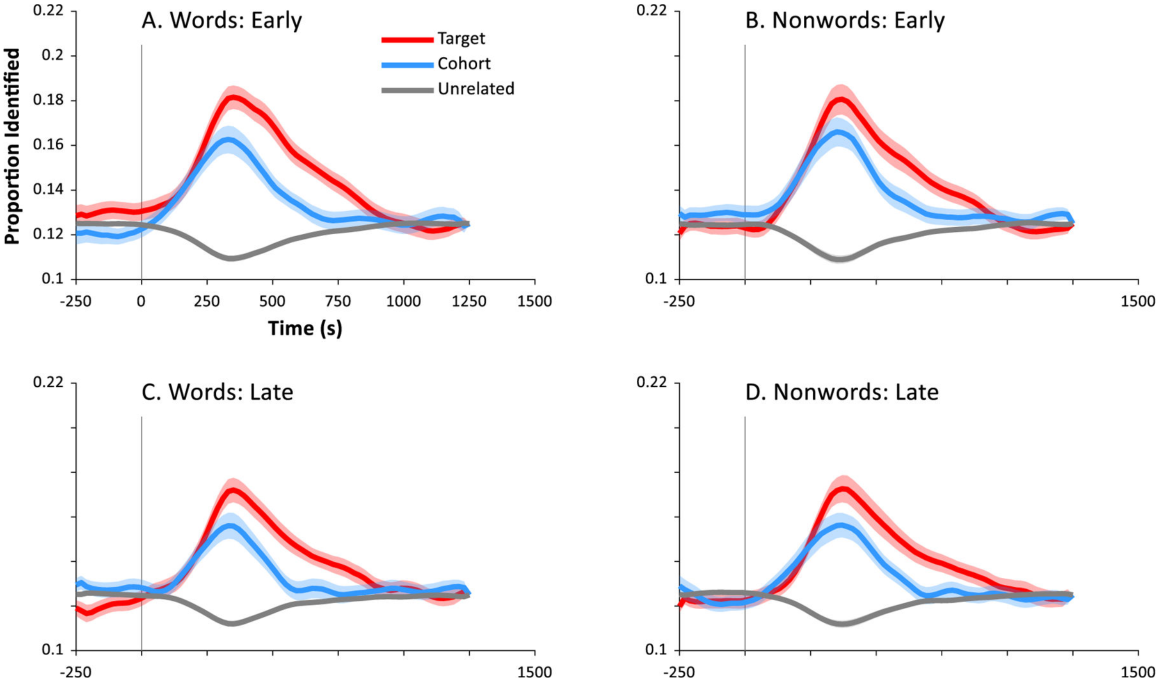

Fig. 8.

Effect of word/nonword status and early vs. late trials on performance. Each curve shows the proportion of identifications as the target, cohort and unrelated at each 0.020 sec bin.

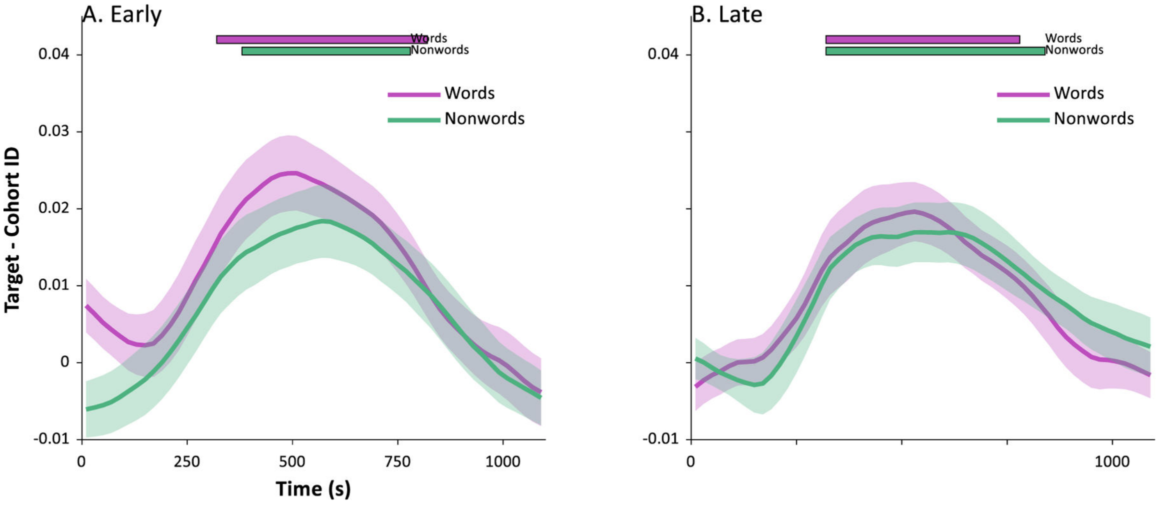

Fig. 9.

Target minus competitor identification rates as a function of time and time in the experiment. Results are pooled over both samples. Here, a value of 0 indicates that the target did not differ from the cohort. A) Early trials (1st half of trials). B) Late trials (2nd half of trials). Significance bars test the difference between target and cohort at each time (corrected for family wise error), asking if the curve deviates from 0. At no point did the curves significantly differ from each other. (early vs. late trials) on performance.

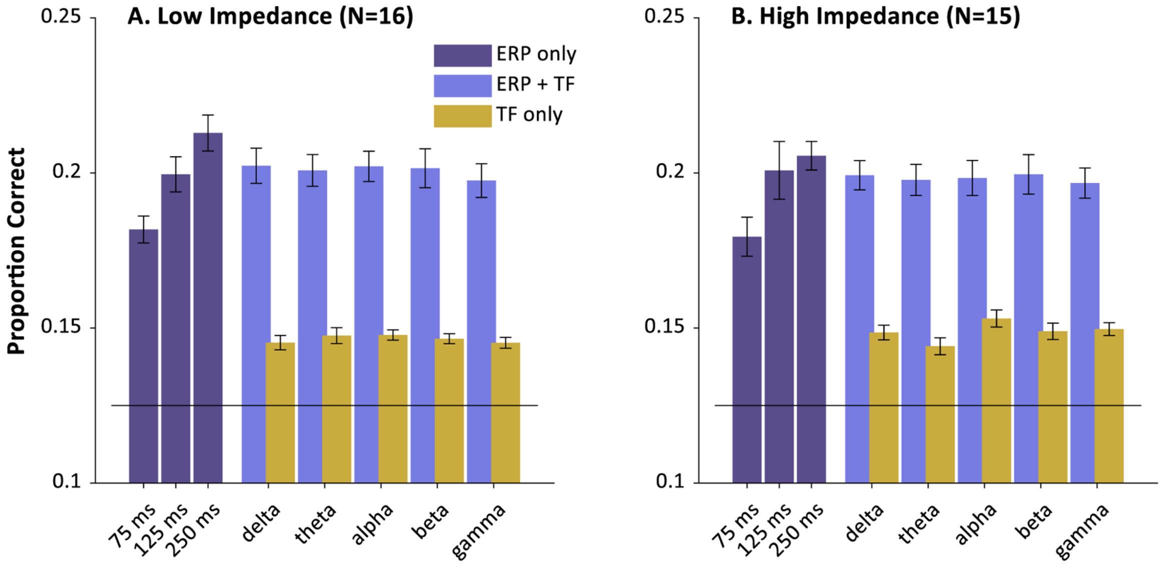

Fig. 10.

Performance (accuracy) of the classifier at 200 msec for different feature sets. For each model, the maximum performance across a range of C and λ (free parameters of the SVM) was saved and averaged across subjects. Error bars represent SEM. Purple bars on the left of each panel show performance of classifiers trained on the ERP only for a range of window lengths. In all three window lengths, the features consisted of the mean, linear slope, quadratic and cubic over that time window. In the right portion of each panel is shown a 2 × 5 analysis crossing the mean power in five frequency bands and whether the ERP was also included. The time-window was always 250 msec and the ERP (if present) was coded as the mean, linear slope and quadratic within the time -window.

Table 1

Items used in the experiment. Note that each participant only heard four words and four nonwords. Words in set A were paired with non-matching nonwords from set B (or vice versa) to avoid confusion.

Words			**Nonwords**		
List	**Items**		**List**	**Items**	
A	badger	baggage	B	babbid	baddow
A	muscle	mushroom	B	muspil	musheme
A	desert	devil	B	dethin	dezhune
A	waffles	washer	B	wathind	wassa
B	captive	cashew	A	cathrung	caffo
B	lobster	lodging	A	lodrum	logort
B	peaches	peacock	A	peatash	peapung
B	sunburn	sundae	A	sungoom	sunjee

Table 2

Significant time windows and details of the family wise error correction. Note that Low-Z refers to the 64-channel Iowa sample (N = 16) run on a low impedance EEG system; High-Z refers to the 128 high impedance system at Haskins (N = 15). Time windows with single time were significant for one timestep (20 msec).

Predictor	Mapping	EEG	ρ	α*	Significant Time window(s)			
Target vs. Unrelated	Words	Low-Z	.945	.017	150–830	850		
		High-Z	.926	.017	190–830			
	Nonwords	Low-Z	.941	.017	130–870	910–950		
		High-Z	.928	.018	170	210–850		
Cohort vs. Unrelated	Words	Low-Z	.928	.016	130–570	690		
		High-Z	.937	.018	150 190 230 270	290 590–670 730 770		
	Nonwords	Low-Z	.930	.016	130–530	590		
		High-Z	.945	.019	270–310	350	390–410	
Target vs. Cohort	Words	Low-Z	.862	.013	330	390–830		
		High-Z	.790	.013	410–430	470–790		
	Nonwords	Low-Z	.885	.014	330–430	490–790	830–850	950
		High-Z	.851	.015	330–350	510–570	610–690	

Table 3

Hybrid split half reliability for several measures extracted from the dynamic identification curves. Reliability is expressed as a pearson correlation after trials are randomly split into two groups. Correlations are averaged across 12 random splits. With 16 subjects, the threshold for significance at $\alpha = .05$ (marked with *) is r>.498, and at $\alpha = .10$ (marked with +) is r>.426.

Measure	**Words**		**Nonwords**	
	Low	**High**	**Low**	**High**
Peak *Target* Identification Rate	0.872*	0.928*	0.933*	0.963*
Peak *Cohort* Identification Rate	0.940*	0.951*	0.954*	0.948*
Minimum *Unrelated* Identification Rate	0.963*	0.971*	0.978*	0.980*
Time of *Target* Peak	0.834*	0.535*	0.581*	−0.001
Time of *Cohort* Peak	0.876*	0.028	0.734*	0.517*
Time of *Unrelated* Minimum	0.923*	0.476+	0.442+	−0.174
Slope of *Target* at 50%	0.645*	0.710*	0.677*	0.873*
Time when *T+C* deviates from Unrelated	0.924*	0.715*	0.656*	0.352*
Time when *Target* deviates from Cohort	0.248	0.018	0.368	0.664*

Table 4

Summary of decoding performance across methodological factors. Rows in gray are the baseline for comparison. √ indicates that a given level of the factor showed no significant departure from that baseline; ~ indicates mixed evidence with one or more significant departures; ✗ indicates a significant deparature in almost all levels. n/a: not tested. F/C: fronto-central electrodes only.

Factor	Level	Low-Z (Iowa)	Hi-Z (Haskins)
Number of Electrodes	128	√	√
	64	√	√
	32	√	✗
	16	✗	n/a
Electrode Geometry Reps/Item	64 – F/C	n/a	✗
	32 – F/C	✗	n/a
	60	√	√
	45	√	√
	30	~	√
	15	✗	✗