Journal of Neural Engineering

RECEIVED
3 August 2020

REVISED
20 January 2021

ACCEPTED FOR PUBLICATION
9 March 2021

PUBLISHED
23 March 2021

PAPER

Data-driven machine learning models for decoding speech categorization from evoked brain responses

Md Sultan Mahmud[1,2,*], Mohammed Yeasin[1,2] and Gavin M Bidelman[2,3,4]

[1] Department of Electrical and Computer Engineering, University of Memphis, 3815 Central Avenue, Memphis, TN 38152, United States of America
[2] Institute for Intelligent Systems, University of Memphis, Memphis, TN, United States of America
[3] School of Communication Sciences and Disorders, University of Memphis, Memphis, TN, United States of America
[4] University of Tennessee Health Sciences Center, Department of Anatomy and Neurobiology, Memphis, TN, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: mmahmud@memphis.edu

Keywords: auditory event-related potentials (ERPs), categorical perception, decision process, behavioral slope, machine learning, stability selection, support vector machine (SVM)

Abstract

Objective. Categorical perception (CP) of audio is critical to understand how the human brain perceives speech sounds despite widespread variability in acoustic properties. Here, we investigated the spatiotemporal characteristics of auditory neural activity that reflects CP for speech (i.e. differentiates phonetic prototypes from ambiguous speech sounds). *Approach.* We recorded 64-channel electroencephalograms as listeners rapidly classified vowel sounds along an acoustic-phonetic continuum. We used support vector machine classifiers and stability selection to determine when and where in the brain CP was best decoded across space and time via source-level analysis of the event-related potentials. *Main results.* We found that early (120 ms) whole-brain data decoded speech categories (i.e. prototypical vs. ambiguous tokens) with 95.16% accuracy (area under the curve 95.14%; *F*1-score 95.00%). Separate analyses on left hemisphere (LH) and right hemisphere (RH) responses showed that LH decoding was more accurate and earlier than RH (89.03% vs. 86.45% accuracy; 140 ms vs. 200 ms). Stability (feature) selection identified 13 regions of interest (ROIs) out of 68 brain regions [including auditory cortex, supramarginal gyrus, and inferior frontal gyrus (IFG)] that showed categorical representation during stimulus encoding (0–260 ms). In contrast, 15 ROIs (including fronto-parietal regions, IFG, motor cortex) were necessary to describe later decision stages (later 300–800 ms) of categorization but these areas were highly associated with the strength of listeners' categorical hearing (i.e. slope of behavioral identification functions). *Significance.* Our data-driven multivariate models demonstrate that abstract categories emerge surprisingly early (∼120 ms) in the time course of speech processing and are dominated by engagement of a relatively compact fronto-temporal-parietal brain network.

1. Introduction

The human brain can map an incredibly large number of stimulus features into a smaller set of groups (Chang *et al* 2010, Holt and Lotto 2010), a process known as categorical perception (CP). Categories allow listeners to extract, manipulate, and precisely respond to sounds (Miller *et al* 2002, 2003, Russ *et al* 2007, Miller and Cohen 2010, Tsunada and Cohen 2014) despite wide variability in their acoustic properties. CP emerges in early life (Eimas *et al* 1971) but is further modified by native language experience (Kuhl *et al* 1992, Xu *et al* 2006, Bidelman and Lee 2015). As such, CP plays an important role in understanding receptive communication and the building blocks of speech perception and language processing across the lifespan.

Some researchers have investigated the role of induced activity in various brain functions. For instance, magnetoencephalography (MEG) studies demonstrate that oscillatory brain activity differs in language vs. non-language stimuli (Eulitz *et al* 1996), suggesting the segmentation and coding of continuous speech relies on cortical oscillations (Gross *et al*

2013). Other studies (Youssofzadeh et al 2020) showed beta power decrements within language processing areas and dominance in left hemisphere (LH) during auditory task processing. Induced activity is releveant in speech categorization studies (Mahmud et al 2021). However, event-related potentials (ERPs) are particularly useful for examining the brain mechanisms of phoneme and speech perception (Celsis et al 1999, Molfese et al 2005) given their excellent temporal resolution and the rapid time course required to process speech signals. Indeed, several neuroimaging studies have documented neural correlates to CP via ERPs (Chang et al 2010, Bidelman 2015, Shen and Froud 2019). In particular, several studies have shown the efficiency of listeners' speech categorization varies in accordance with their underlying brain activity (Perlovsky 2011, Bidelman et al 2013, Bidelman and Alain 2015, Bidelman and Lee 2015). For example, Bidelman et al demonstrated that brain responses in the time frame of 180–320 ms were more robust for phonetic prototypes vs. ambiguous speech tokens, thereby reflecting category-level processing (Bidelman et al 2020). Other studies have shown links between N1-P2 amplitudes of the auditory cortical ERPs and the strength of listeners' speech identification (Bidelman and Walker 2017) and labeling speeds (Al-Fahad et al 2020) in speech categorization tasks (Bidelman et al 2014, Bidelman and Alain 2015). These findings are consistent with the notion that the early N1 and P2 waves of the ERPs are highly sensitive to speech processing and auditory object formation that is necessary to map sounds to meaning (Wood et al 1971, Alain 2007, Bidelman et al 2013).

The neural organization of speech categories also varies spatially, recruiting a widely distributed system across a number of brain regions. Neural responses are elicited by prototypical speech sounds (i.e. those heard with a strong phonetic category) differentially engage Heschl's gyrus and inferior frontal gyrus (IFG) compared to ambiguous speech depending on a listeners perceptual skill level (Bidelman et al 2013, Bidelman and Lee 2015, Bidelman and Walker 2017, Mankel et al 2020). This suggests emergent categorical representations within the early auditory-linguistic pathways. Similarly, Alho et al found that category-specific representations were activated in left IFG (Alho et al 2016) at an early-latency (115–140 ms). Collectively, in terms of the time course of processing, M/EEG (electroencephalogram) studies agree that the neural underpinnings of speech categories emerge within the first few hundred milliseconds after stimulus onset and reflect abstract 'category level-effects' (Toscano et al 2018) and 'phonemic categorization' (Liebenthal et al 2010).

Beyond conventional auditory-linguistic brain regions, neuroimaging also demonstrates a variety of additional areas important to speech perception and language processing (Novick et al 2010, Hickok et al 2011, Lee et al 2012). Among them, superior parietal lobe is associated with writing (Menon and Desmond 2001) and supramarginal gyrus with phonological processing (Deschamps et al 2014, Oberhuber et al 2016) during speech and verbal working memory tasks. Relevant to CP, several studies have found that the left inferior parietal lobe is more activated during auditory phoneme sound categorization (Husain et al 2006, Dufor et al 2007, Desai et al 2008). Indeed, auditory categorical processing has been shown to recruit superior temporal gyrus/sulcus, middle temporal gyrus, premotor cortex, inferior parietal cortex, planum temporal, and IFG (Guenther et al 2004, Bidelman and Walker 2019). Some other neuroimaging and electrocorticography studies have however shown that rostral anterior cingulate cortex is associated with speech control (Paus et al 1993, Sahin et al 2009, Tankus et al 2012) and the orbitofrontal cortex in speech comprehension (Sabri et al 2008). Under some circumstances (e.g. highly skilled listeners), speech categories can even emerge as early as auditory cortex (Chang et al 2010, Bidelman and Lee 2015, Bidelman and Walker 2019).

While category representations seem to emerge early in the time course of speech perception, the task of categorizing sounds can be further separated into pre- and post-perceptual stages of processing (i.e. stimulus encoding vs. decision mechanisms). 'Early' vs. 'late' stage models of category formation have long been discussed in the literature (Fox 1984, McClelland and Elman 1986, Norris et al 2000, Noe and Fischer-Baum 2020). However, few empirical studies have actually separately examined encoding and decision stages of CP. The human brain encodes speech stimuli within ∼250 ms after stimulus onset (Masmoudi et al 2012) and decodes ∼300 ms after stimulus onset (Domenech and Dreher 2010, Mostert et al 2015). Previous studies have largely focused on these specific time windows (e.g. ERP waves) and brain regions when attempting to describe the neural basis of CP. While informative, such hypothesis-based testing can be restrictive and potentially miss the broader and distributed networks associated with speech-language processing that unfold on different time scales (Rauschecker and Scott 2009, Du et al 2016).

In this regard, machine learning (ML) techniques are increasingly used to 'decode' high dimensional neuroimaging data and better understand different states of brain functionality as measured via EEG. ML is a branch of artificial intelligence that *'learns a model'* from the past data to predict future data (Cruz and Wishart 2006). Moreover, data mining approaches in ML identify important properties in neural activity with high accuracy without intervention from human observers. It would be meaningful if brain functioning that has been linked with speech processing (e.g. CP) could be decoded from neural data without, or at least with minimal, *a priori* assumptions on when and where those representation emerge. Indeed, laying the groundwork for the present work, we have recently shown that the speed of listeners' identification in speech categorization tasks can be directly decoded from their full-brain EEGs using an entirely data-drive approach (Al-Fahad *et al* 2020). We have also shown that ML can decode age-related changes in speech processing that occur in older adults (Mahmud *et al* 2020).

Departing from previous hypothesis-driven studies (Bidelman and Alain 2015, Bidelman and Walker 2017, 2019), the current work used a comprehensive, data-driven approach to examine the neural mechanisms of speech categorization during encoding and decision stages of processing using whole-brain, electrophysiological data. We analyzed speech-evoked ERPs from 64-channel EEG recorded during a rapid speech categorization task in young, normal hearing listeners. Our approach applied state-of-the-art ML techniques including neural classifiers and feature selection methods (i.e. stability selection) to source-level ERPs to investigate the spatiotemporal dynamics of speech categorization. We aimed to determine when and where neural activity from full-brain EEGs differentiated phonetic from phonetically ambiguous speech sounds, and thus showed the strongest evidence of categorical processing using an entirely data-driven, ML approach.

2. Materials and methods

2.1. Participants

Forty-nine young adults (male: 15, female: 34; aged 18–33 years) were recruited as participants from the University of Memphis student body to participate into our ongoing studies on the neural basis of speech perception and auditory categorization (Bidelman and Walker 2017, Bidelman *et al* 2020, Mankel *et al* 2020). All participants had normal hearing sensitivity [i.e. <25 dB hearing level between 500 and 2000 Hz]. All but one listener was right handed according to their Edinburgh Handedness scores (Oldfield 1971) and had achieved a collegiate level of education. None reported any history of neurological disease. All participants were paid for their time and gave informed written consent in accordance with the declaration of Helsinki and a protocol approved by the Institutional Review Board at the University of Memphis.

2.2. Stimuli and task

We used a synthetic five-step vowel token continuum to assess the most discriminating spatiotemporal features while categorizing prototypical vowel speech from ambiguous speech (Bidelman *et al* 2013, 2014). Speech spectrograms are represented in figure 1(A). Each token of the continuum was separated by equidistant steps acoustically based on the first formant frequency ($F1$) and perceived categorically from /u/ to /a/. Each speech token was 100 ms, including 10 ms rise/fall to minimize the spectral splatter in the stimuli. Each speech token contained an identical voice fundamental frequency ($F0$), second ($F2$), and third formant ($F3$) frequencies ($F0$: 150 Hz, $F2$: 1090 Hz, and $F3$: 2350 Hz). To create a phonetic continuum that varied in percept from /u/ to /a/, $F1$ frequency was parameterized over five equal steps from 430 Hz to 730 Hz (Bidelman *et al* 2013).

Stimuli were presented binaurally at an intensity of 83 dB sound pressure level through insert earphones (ER 2; Etymotic Research). Participants heard each token 150–200 times presented in random order. They were asked to label each sound in a binary identification task ('/u/' or '/a/') as fast and accurately as possible. Their response and reaction time were logged. The interstimulus interval was jittered randomly between 400 and 600 ms (20 ms step and rectangular distribution) following listeners' behavioral response to avoid anticipating the next trial (Luck 2005).

2.3. EEG recordings and data pre-procedures

During the behavioral task, EEG was recorded from 64 channels at standard 10-10 electrode locations on the scalp (Oostenveld and Praamstra 2001). Continuous EEGs were digitized using Neuroscan SynAmps RT amplifiers at a sampling rate of 500 Hz. Subsequent preprocessing was conducted in the Curry 7 neuroimaging software suite, and customized routines coded in MATLAB. Ocular artifacts (e.g. eyeblinks) were corrected in the continuous EEG using principal component analysis (Picton *et al* 2000) and then filtered (1–100 Hz bandpass; notched filtered 60 Hz). Cleaned EEGs were then epoched into single trials (−200–800 ms, where $t = 0$ was stimulus onset).

2.4. EEG source localization

To disentangle the sources of CP-related EEG activity, we reconstructed the scalp-recorded responses by performing a distributed source analysis in the Brainstorm software package (Tadel *et al* 2011). All analyses were performed on single-trial data[5]. We used a realistic boundary element head model (BEM) volume conductor and standard low-resolution brain electromagnetic tomography (sLORETA) as

[5] A limitation of this work was that we conducted source localization on single trials which adds noise to the data. Single trial responses were however necessary for bootstrapping and feature selection. Additionally, the use of template (rather than individual) MRI anatomies likely also reduces the precision of source localization and thus underestimates the true source foci. However, this source of 'noise' is the same for all subjects, trials, and tokens so it does not affect our stimulus decoding results. Moreover, any additional noise due to our source localization approach is probably negligible because stability selection works well even when the noise level of data is unknown.

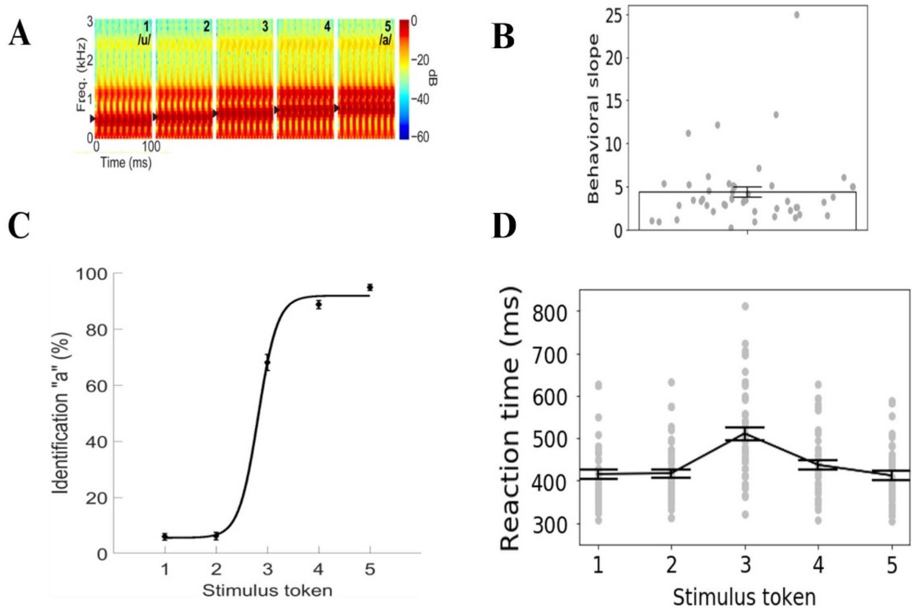

Figure 1. Speech stimuli and behavioral results. (A) Acoustic spectrograms of the speech continuum from /u/ to /a/. (B) Behavioral slope. (C) Psychometric functions showing % 'a' identification of each token. Listeners' perception abruptly shifts near the continuum midpoint, reflecting a flip in perceived phonetic category (i.e. 'u' to 'a'). (D) Reaction time (RT) for identifying each token. RTs are fastest for category prototypes (i.e. Tk1/5) and slow when classifying ambiguous tokens at the continuum midpoint (i.e. Tk3). Silver color dots represent individual participants' data. Errorbars = ±1 s.e.m.

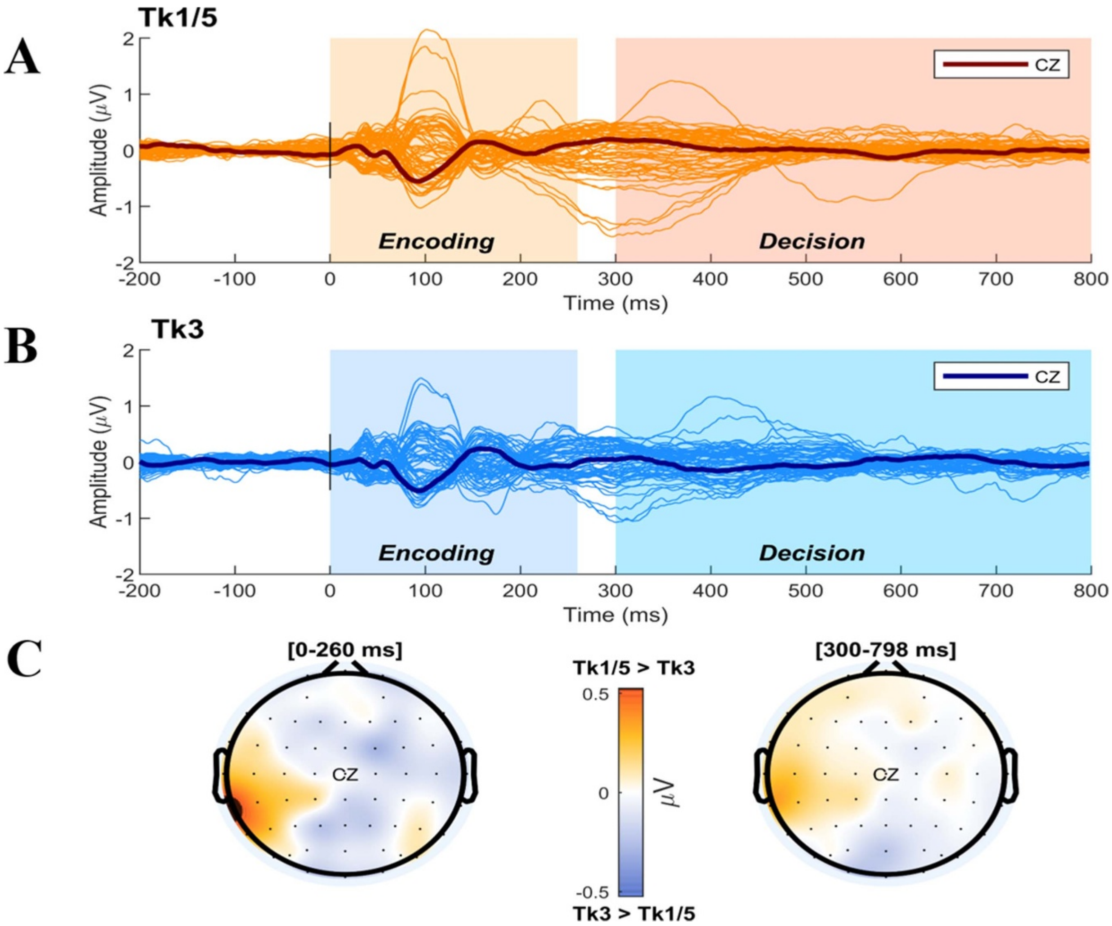

Figure 2. Grand averaged butterfly plots of scalp ERPs (64 channels) to prototypical (A); Tk1/5) vs. category-ambiguous (B); Tk3) vowels. Vertical lines demarcate segments for the stimulus encoding (0–260 ms) and decision period (300 ms–800 ms) analysis windows. $t = 0$ marks stimulus onset. (C) Topographic maps for encoding (left) and decision (right) periods.

the inverse solution within Brainstorm (Tadel *et al* 2011). A BEM model has less spatial errors than other existing head models (e.g. concentric spherical head model). We used Brainstorm's default parameter settings [signal to noise ratio (SNR) = 3.00, regularization noise covariance = 0.1]. From each single-trial sLORETA volume, we extracted the time-courses within 68 functional regions of interest (ROIs) across the LH and right hemisphere (RH) defined by the Desikan-Killiany (DK) atlas (Desikan *et al* 2006) (LH: 34 ROIs and RH: 34 ROIs). Single-trial data were then baseline corrected to the epoch's pre-stimulus interval (−200–0 ms).

To evaluate whether ERPs showed category-related effects, we averaged response amplitudes to tokens at the endpoints of the continuum and compared this combination to the ambiguous token at its midpoint (e.g. Liebenthal *et al* 2010, Bidelman 2015, Bidelman and Walker 2017, 2019). This contrast [i.e. mean (Tk1, Tk5) vs. Tk3] allowed us to assess the degree to which neural responses reflected 'category level-effects' (Toscano *et al* 2018) or 'phonemic categorization' (Liebenthal *et al* 2010). The rationale for this analysis is that it effectively minimizes stimulus-related differences in the ERPs, thereby isolating categorical/perceptual processing. For example, Tk1 and Tk5 are expected to produce distinct ERPs due to exogenous acoustic processing alone. However, comparing the average of these responses (i.e. mean [Tk1, Tk5]) to that of Tk3 allowed us to better isolate ERP modulations related to the process of categorization (Liebenthal *et al* 2010, Bidelman and Walker 2017, 2019). To ensure an equal number of trials and SNR for prototypical and ambiguous stimuli, we considered only 50% of the data from the merged (Tk1/5) samples[6].

2.5. Feature extraction

Previous computational studies have found that ERPs averaged over 100 trials provided the best classification of data while maintaining reasonable signal SNR and computational efficiency (Al-Fahad *et al* 2020, Mahmud *et al* 2020). We quantified source-level ERPs with a mean bootstrapping approach (James *et al* 2013) by randomly averaging over 100 trials (with replacement) 30 times (Al-Fahad *et al* 2020) for each stimulus condition per participant. For each resample and ROI time course, we measured the mean amplitude within a 20 ms sliding window (without overlapping) in the post-stimulus interval (i.e. 0–800 ms). In post hoc analysis, we parsed the epoch into 'encoding' (0–260 ms) and 'decoding/decision process' intervals[7] (>300 ms) to investigate neural decoding related to pre- and post-perceptual processing, respectively. The sliding window resulted in 40 (800 ms/20 ms) ERP features (i.e. mean amplitude per window) for each ROI waveform, yielding a total of 68 × 40 = 2720 features per token (e.g. Tk1/5 vs. Tk3) from each listeners' data. Thus, the encoding and decision windows contained 13 × 68 = 884 (encoding) and 25 × 68 = 1700 (decision) ERP features. ERPs features were then used as input to an support vector machine (SVM) classifier to access the temporal dynamics of the data and determine when in time CP was decodable from brain activity. State-of-the art variable selection (stability selection; see section 2.7) (Meinshausen and Bühlmann 2010) was then applied for identifying where in the brain (e.g. which ROIs) were involved in encoding and decision processes with regard to the categorization of speech. Before submitting to the SVM classifier, the data were z-score normalized to ensure all features were on a common scale range (Casale *et al* 2008).

2.6. SVM classification to identify temporal dynamics of CP

Parameter optimized SVM classifiers provide better performance with small sample sizes data which is common in human neuroimaging studies. Classifier performance is greatly affected by tunable parameters in the SVM model (e.g. kernel, C, γ)[8] (Hsu *et al* 2003). To avoid bias in parameter selection, we used

[6] Our main analyses focused on decoding speech sounds with a clear category (i.e. Tk1 and Tk5) from those which are category ambiguous (Tk3). An interesting question is whether Tk 3 is ambiguous or rather a bistable percept (cf Bidelman *et al* 2013). In attempts to address this question, we analyzed Tk 3 trials split based on listeners' perceptual response [i.e. Tk3(u) and Tk3(a)]. Following our main analyses using a sliding window SVM classifier (e.g. figure 2), we attempted to decode the two percepts induced by the otherwise identical stimulus [e.g. Tk3(u) vs. Tk3(a)]. Maximum decoding of Tk3(u) vs. Tk3(a) was only 64.28%, 63.96%, and 62.98% using whole-brain, LH, and RH source ERPs, respectively (data not shown). Decoding accuracy was equally poor using the entire epoch window with only 62.06% (whole brain), 60.01% (LH), and 59.41% (RH) accuracy, respectively. Thus, performance was essentially at a random chance when decoding the perceptual state via source ERPs. Chance-level performance implies Tk 3 stimuli sounded neither like an /u/ or /a/. It further suggests our main Tk1/5 vs. Tk3 contrast is likely decoding category from category-ambiguous speech activity rather than bistable percepts, *per se*.

[7] There is no clear division between 'encoding' and 'decision/post-processing' stages of perceptual chronometry. The choice of the ∼300 ms mark was motivated by our previous demonstrating categorical coding within the time-frame of the N1-P2 waves of the ERP (< 250 ms) (Bidelman *et al* 2013). We chose to include a subsequent time buffer between the two intervals so as to minimize overlap and therefore what we were decoding in each segment.

[8] Parameters γ and C in the SVM used in this study gives a measure of the influence of training data points on decision boundary and a measure of miss-classification tolerance. The first parameter γ comes from the radial basis function kernel (e.g. $K(x,x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$ or equivalently $K(x,x') = \exp\left(-\gamma||x-x'||^2\right)$ with a parameter γ) where $\gamma = \frac{1}{2\sigma^2}$. In this study, the radial basis kernel is used as a transformation function. A larger value of γ implies smaller σ, which means that the classifier takes into account the effect of samples closer to the decision boundary. On the other hand, smaller γ means that the classifier considers the effect of samples farther from the decision boundary. The C is a parameter of SVM that acts as regularization. It provides the classifier a trade-off between the margin of decision boundary and miss- classification. A larger value of C produces

a grid search approach during the training phase to find optimal kernel, C, and γ values. We randomly split the data into training (80%) and test (20%) sets (Park *et al* 2011). During the training phase (e.g. using 80% data), we fine-tuned the C, and γ parameters using grid search to find the optimal values such that the resulting classifier accurately distinguished prototypical vs. ambiguous speech in the test data that models never seen. The grid search process was conducted with five-fold cross validation, kernels = 'RBF', fine-tune 20 different values of (C and γ) in the following range $C = [1e^{-2}–1e^{3}]$, and $\gamma = [1e^{-4}–1e^{2}]$ (Mahmud *et al* 2020). The SVM learned the support vectors from the training data that comprised the attributes (e.g. ERP features) and class labels (e.g. Tk1/5 vs. Tk3). Then we selected the best model that has maximum margin with the optimal value of C and γ for predicting the unseen test data (only by providing the attributes but no class labels). The classification performance metrics (accuracy, $F1$-score, precision, and recall) are calculated from standard formulas (Saito and Rehmsmeier 2015).

2.7. Stability selection to identify spatial dynamics of CP

Our data included a large number (∼2700) of ERP measurements for each stimulus condition of interest (e.g. Tk1/5 vs. Tk3). Larger numbers of variable/features can lead to overfitting and weak generalization in classification problems since the majority of features from brain activity (i.e. different ROIs, time segments) do not provide discriminative power for decoding CP. Consequently, we aimed to select a limited set of the most salient discriminating features. Stability selection is a feature selection method that works well in high dimensional or sparse data based on the Lasso (least absolute shrinkage and selection operator) (Meinshausen and Bühlmann 2010, Yin *et al* 2017). Over a range of model parameters, stability selection can identify the most stable (relevant) features out of a large number of features.

In stability selection, a feature is considered to be more invariants/relevant if it is more frequently selected over repeated subsampling of the data (Nogueira *et al* 2017). To optimize the model error, the Randomized Lasso randomly subsamples the training data and fits an L1 penalized logistic regression. Over many iterations, feature scores are (re)calculated. The features are shrunk to zero by multiplying the features' co-efficient by zero while the stability score is lower. Remaining non-zero features are considered important variables for classification. Detailed interpretation and mathematical equations of stability selection are explained in Meinshausen and Bühlmann (2010). Stability selection is extremely general and widely used in high dimensional data even when the noise level is unknown (Meinshausen and Bühlmann 2010).

In our implementation of stability selection, we used a sample fraction = 0.75, number of resamples = 1000, and tolerance = 0.01 (Meinshausen and Bühlmann 2010). In the Lasso algorithm, the feature scores were scaled between 0 and 1, where 0 is the minimum score (i.e. irrelevant feature) and 1 is the maximum score (i.e. most salient or stable feature). We estimated the regularization parameter from the data using the least angle regression algorithm (Efron *et al* 2004, Friedman *et al* 2010). Over 1000 iterations, Randomized Lasso provided the overall feature scores ($0 \sim 1$) based on the number of times a variable was selected. We ranked stability scores to identify the most important, consistent, stable, and invariant features that could decode speech categories via the EEG (i.e. correctly classify Tk1/5 vs. Tk3). We used these ranked features and corresponding class labels to an SVM classifier with different stability thresholds and observed the model performance. We fine-tuned the hyperparameters of SVM classifier using grid search corresponding to different stability thresholds.

a narrower (smaller-margin) hyperplane if that obtains less or no miss-classification. Whereas the smaller value of C allows drawing a wider (bigger-margin) hyperplane even if there are some miss-classifications. The optimal value of γ and C depends on data which is why we used a grid search to tune these parameters in our classification model.

3. Results

3.1. Behavioral results

Behavioral identification (%) functions and reaction time (ms) for speech categorization are depicted in figures 1(C) and (D), respectively. Listeners responses abruptly shifted in speech identity (/u/ vs. /a/) near the midpoint of the continuum, reflecting a change in perceived category. The behavioral speed of speech labeling [e.g. reaction time (RT)] were computed listeners' median response latency for a given condition across the all trials. RTs outside of 250–2500 ms were deemed outliers and excluded from further analysis (Bidelman *et al* 2013, Bidelman and Walker 2017). Listeners spent more time classifying the ambiguous (Tk3) than prototypical speech tokens (e.g. Tk1/5), further confirming categorical hearing (Pisoni and Tash 1974). For each continuum, the identification scores were fit with a two parameters sigmoid function; $P = \frac{1}{[1+e^{-\beta 1(x-\beta 0)}]}$, where P is the proportion of the trial identification as a function of a given vowel, x is the step number along the stimulus continuum, and $\beta 0$ and $\beta 1$ the location and slope of the logistic fit estimated using the nonlinear least-squares regression (Bidelman *et al* 2014, Bidelman and Walker 2017). The slopes of listeners' sigmoidal psychometric function, reflecting the strength of their CP, is presented in figure 1(B).

Table 1. Performance metrics of the SVM classifier corresponding to maximal decoding of prototypical vs. ambiguous vowels from ERPs.

Metric (%)	Whole-brain features	LH features	RH features
Accuracy	95.16	89.03	86.45
AUC	95.14	89.18	86.45
$F1$-score	95.00	89.00	86.00
Precision	95.00	89.00	87.00
Recall	95.00	89.00	86.00

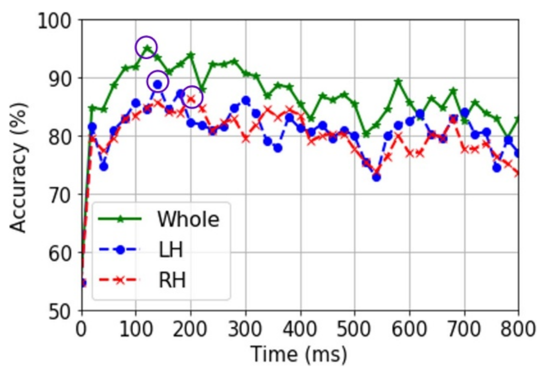

Figure 3. SVM classifier accuracy decoding speech categories from source ERPs. Decoding using whole-brain vs. hemispheres-specific data (LH and RH) across the epoch window. Maximum classification accuracies are marked by circles. Maximum classifier accuracy was observed at ~120 ms suggesting category representations emerge early, ~200 ms before listeners' behavioral categorization decisions (cf figure 1(C)).

3.2. Decoding the time-course of speech categorization from ERPs

We first examined how well categorical speech information could be decoded from whole-brain and individual hemisphere (e.g. LH and RH) ERPs data. During pilot modeling, we carried the grid search approach (mentioned in method). The optimal values of C and γ parameters corresponding to the maximum speech decoding reported in table 1 were: [$C = 10$, $\gamma = 0.05$ for whole-brain data; $C = 20$, $\gamma = 0.01$ for LH data; $C = 20$, $\gamma = 0.01$ for RH data]. We then selected the best model and predicted the class labels (e.g. Tk1/5 vs. Tk3) by feeding the feature vectors only from the unseen test data. The performance metrics were calculated from predicted class labels and true class labels. Time-varying accuracy of the SVM classifier (i.e. distinguishing Tk1/5 vs. Tk3 responses) is shown in figure 3.

Decoding was generally at chance level (54%) at stimulus onset (i.e. $t = 0$) but increased rapidly to a maximum accuracy of 95.16% by 120 ms (marked as circles in figure 3). The individual hemispheres alone were less accurate and decoded speech categories later in time compared to whole-brain data (LH: 89.03% at 140 ms; RH: 86.45% at 200 ms) (marked as circles in figure 3). Other important performance metrics of the SVMs at maximum decoding are reported in table 1. Collectively, the earlier and improved ability of LH compared to RH in decoding phonetic categories is consistent with an LH bias in speech and language processing (Hickok and Poeppel 2000). More critically, the early time course of decoding (120–150 ms) confirms that category level information, an abstract code, emerges very early in the neural chronometry of speech processing and well before listeners' execute their behavioral decision (cf reaction times in figure 1(D)) (Bidelman *et al* 2013, Alho *et al* 2016, De Taillez *et al* 2020).

3.3. Decoding the spatial regions underlying categorization: stimulus encoding vs. decision

We used stability selection to find the most critical brain ROIs that were associated with categorical organization in the encoding (pre-perceptual) vs. decision (post-perceptual) periods of the task structure (see figure 2). ERP features were considered stable (relevant) if they yielded a decoding accuracy performance >80%. The effect of stability threshold selection in the encoding and decision windows is illustrated in figure 4. Each bin of histogram demonstrates the number of features in a range of stability threshold. The x-axis has four labels. The first line represents the stability score (0–1); the second and third line show the number of features and percentage of the selected features in the corresponding bin; line four represents the cumulative unique ROIs up to the lower boundary of the bin. The solid black semi bell-shaped curves of figure 4 represent classification accuracy for the different stability thresholds. In this analysis, the number of unique brain ROIs represents distinct functional brain ROIs of the DK atlas and the number of features represents different time windows extracted from source ERPs. Features selected at each stability threshold were then submitted to an SVM classifier separately for the stimulus encoding and response decision periods.

During stimulus encoding (0–260 ms), 75% of features yielded stability scores 0–0.1. Thus, the majority of spatiotemporal ERP features were selected less than 10% out of 1000 model iterations and therefore carry weak importance in terms of describing categorical speech processing during stimulus encoding. In contrast, at a more conservative stability score of 0.3, 102 (11%) out of 884 ERP features selected from 52 ROIs were able to encode prototypical from ambiguous speech at near-ceiling accuracy (95.8%). Accuracy decreased precipitously with higher (more conservative) stability thresholds resulting in fewer (though more informative) brain ROIs describing category processing. For example, a stability score of 0.6—selecting only the most behaviorally-relevant features—still encoded speech categories well above chance (66.8%) with only five features from five ROIs.

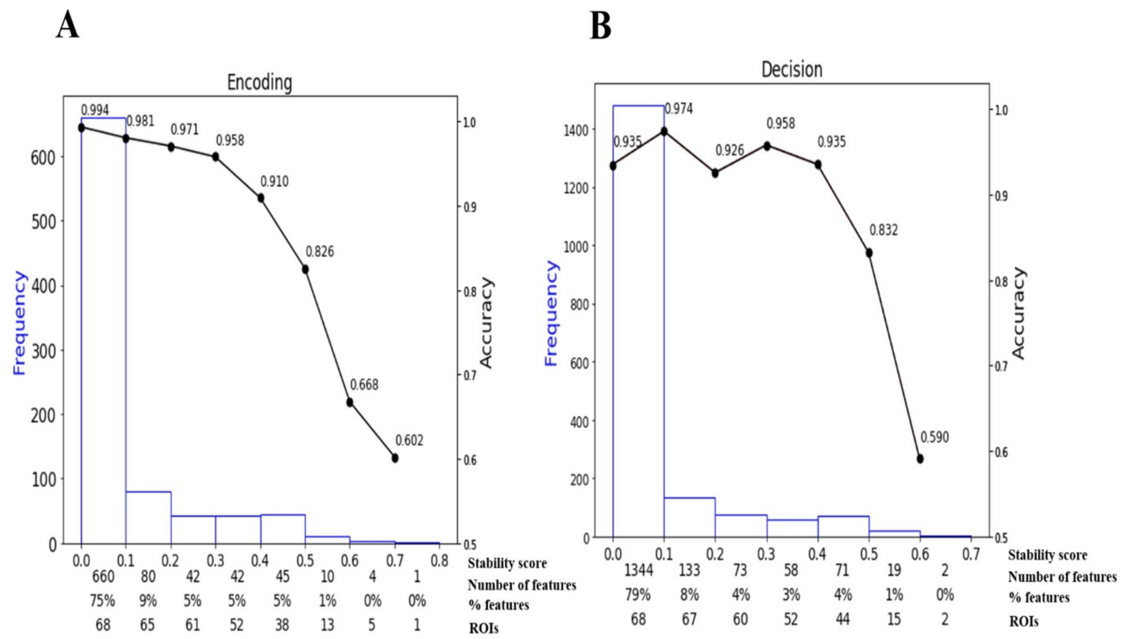

Figure 4. Effect of stability score threshold on model performance during (A) encoding and (B) decision period of the CP task. The bottom of the x-axis has four labels; *Stability score* represents the stability score range of each bin (scores: $0 \sim 1$); *Number of features,* number of features under each bin; *% features,* the corresponding percentage of selected features; *ROIs*, number of cumulative unique brain regions up to the lower boundary of the bin.

At stability score 0.5, speech encoding accuracy 82.6% only using 15 features from 13 unique ROIs. A BrainO visualization (Moinuddin *et al* 2019) of relevant ROIs for the encoding and decision period (threshold stability score $\geqslant 0.5$) are shown in figures 5 and 6 with additional details in table 2.

During the decision period following stimulus encoding (>300 ms), corresponding to the stability score 0.4, only 92 (5%) out of 1700 ERP features were selected, and the classifier showed decoding accuracy of 93.5% (area under the curve 93.6%). At a stability score 0.5 (corresponding to 83.2% accuracy), only 21 (1%) out 1700 ERP features from 15 unique ROIs were needed to describe categorical processing.

3.4. Brain-behavior correspondences

Multivariate regression analysis is widely used to investigate when more than one predictor simultaneously influences an outcome variable (Hanley 1983, Royston and Sauerbrei 2008). To evaluate the behavioral relevance of the brain regions identified via stability selection, we conducted multivariate regression using weighted least squares (WLS) regression (Ruppert and Wand 1994). Regressions were computed between the 15 ROI ERPs identified in the decision interval and listeners' behavioral slopes (figure 1(B)), which indexes their degree of categorical hearing. We computed the mean neural response (i.e. ERP) within each selected region across the stimuli [mean ERP of (Tk1/5 and Tk3)] and then regressed the 15 ROI responses simultaneously against listeners' behavioral slope. The inverse of the absolute error values of the ordinary least squares were used as weights in the WLS to reduce the effect of heteroscedasticity (Seabold and Perktold 2010, Weighted Regression in SAS, R, and Python). The multivariate model robustly predicted listeners' behavioral CP from neural data ($R^2 = 0.85$, $p < 0.00001$; table 3), demonstrating the selected 15 ROIs identified via ML (i.e. stability selection) carried behaviorally relevant information regarding CP.

4. Discussion

We conducted ML analyses on EEG to examine the spatiotemporal dynamics of speech processing during rapid speech sound categorization. We found that speech categories are best decoded via patterned neural activity occurring within 120 ms and no later than 200 ms. We also identified the most relevant brain regions that are involved in encoding and decision stages of categorization. Our findings show a small set of brain areas (15 ROIs) robustly predicts listeners' categorical decisions, accounting for 85.0% of the variance in behavior.

4.1. Speech categories are decoded early (<150 ms) in the time course of perception

We replicate and extend previous work by using whole-brain EEG and SVM neural classifiers to examine the time-course and hemispheric asymmetry as the brain decodes the identity of speech sounds. We found optimal speech decoding in the time frame of the N1 wave (120 ms) of the auditory ERPs using full-brain data. Analysis by hemisphere further showed that LH yielded better and earlier decoding than the

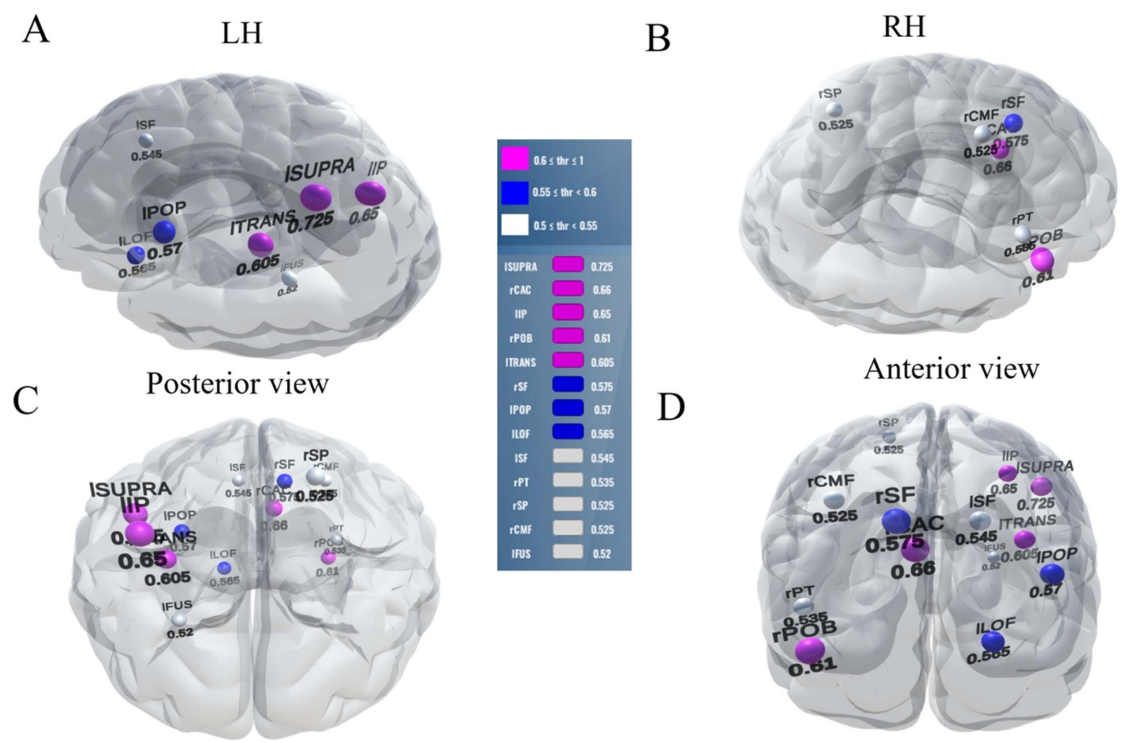

Figure 5. Stable (most consistent) neural network during the *encoding period* of CP. Visualization of brain ROIs corresponding to $\geqslant$0.50 stability threshold (13 top selected ROIs which show categorical organization (e.g. Tk1/5 $\neq$ Tk3) at 82.6%. (A) LH, (B) RH, (C) posterior view and (D) anterior view. Color legend demarcations show high (pink), moderate (blue), and low (white) stability scores. l/r = left/right; SUPRA, supramarginal; CAC, caudal anterior cingulate; IP, inferior parietal; POB, pars orbitalis; TRANS, transverse temporal; SF, superior frontal; POP, pars opercularis; LOF, lateral orbitofrontal; PT, pars triangularis; SP, superior parietal; CMF, caudal middle frontal; FUS, fusiform.

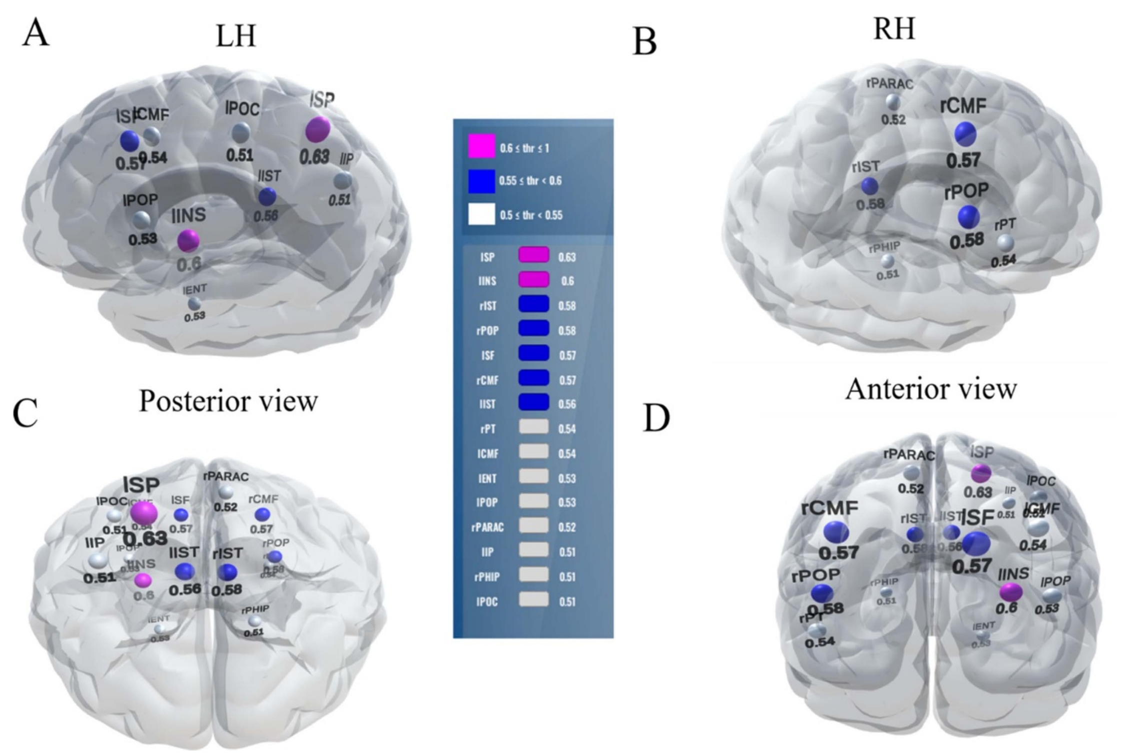

Figure 6. Stable (most consistent) neural network during the *decision period* of CP. Visualization of brain ROIs corresponding to $\geqslant$0.50 stability threshold (15 top selected ROIs which decode Tk1/5 from Tk3 at 83.2%. Otherwise as in figure 5. SP, superior parietal; INS, insula; POP, pars opercularis; SF, superior frontal; CMF, caudal middle frontal; IST, isthmus cingulate; PT, pars triangularis; CMF, caudal middle frontal; ENT, entorhinal; PARAC, paracentral; IP, inferior parietal; PHIP, parahippocampal; POC, postcentral.

Table 2. Most important brain regions describing speech categorization during stimulus encoding (13 ROIs) and response decision (15 ROIs) at a stability threshold ⩾ 0.5.

	Encoding (82.6% total accuracy)			Decision (83.2% total accuracy)		
Rank	ROI name	ROI abbrev.	Stability score	ROI name	ROI abbrev.	Stability score
1	Supramarginal L	lSUPRA	0.73[a]	Superior parietal L	lSP	0.63
2	Caudal anterior cingulate R	rCAC	0.66	Insula L	lINS	0.60
3	Inferior parietal L	lIP	0.65	Isthmus cingulate R	rIST	0.58
4	Pars orbitalis R	rPOB	0.61	Pars opercularis R	rPOP	0.58
5	Transverse temporal L	lTRANS	0.61	Superior frontal L	lSF	0.57
6	Superior frontal R	rSF	0.58	Caudal middle frontal R	rCMF	0.57
7	Pars opercularis L	lPOP	0.57	Isthmus cingulate L	lIST	0.56
8	Lateral orbitofrontal L	lLOF	0.57	Pars triangularis R	rPT	0.54
9	Superior frontal L	lSF	0.55	Caudal middle frontal L	lCMF	0.54
10	Pars triangularis R	rPT	0.54	Entorhinal L	lENT	0.53
11	Superior parietal R	rSP	0.53	Pars opercularis L	lPOP	0.53
12	Caudal middle frontal R	rCMF	0.53	Paracentral R	rPARAC	0.52
13	Fusiform L	lFUS	0.52	Inferior parietal L	lIP	0.51
14				Parahippocampal R	rPHIP	0.51
15				Postcentral L	lPOC	0.51

[a] A score of 0.73, for example, means that out of 1000 iterations, the ERP feature of this ROI was selected 730 times by stability selection.

Table 3. WLS regression results describing how individual brain ROIs predict behavioral CP.

	ROI name	ROI abbrev.	Coefficient	*t*-value	*p*-value
1	Superior parietal L	lSP	−0.2163	−3.008	0.004920
2	Insula L	lINS	0.1808	5.188	0.000010
3	Isthmus cingulate R	rIST	−0.2679	−3.764	0.000633
4	Pars opercularis R	rPOP	0.1231	4.429	0.000093
5	Superior frontal L	lSF	−0.1726	−3.190	0.003055
6	Caudal middle frontal R	rCMF	0.1544	2.367	0.023774
7	Isthmus cingulate L	lIST	0.2259	2.792	0.008545
8	Pars triangularis R	rPT	−0.0214	−0.679	0.501925
9	Caudal middle frontal L	lCMF	0.0153	0.345	0.732223
10	Entorhinal L	lENT	0.1170	5.009	0.000013
11	Pars opercularis L	lPOP	0.1475	3.892	0.000441
12	Paracentral R	rPARAC	0.2223	3.308	0.002226
13	Inferior parietal L	lIP	−0.1017	−1.364	0.181508
14	Parahippocampal R	rPHIP	−0.0422	−2.097	0.043540
15	Postcentral L	lPOC	0.1809	2.749	0.009512

RH, where optimal decoding occurred 20–80 ms later (LH: 140 ms; RH: 200 ms). These latencies are compatible with the N1-P2 waves of the auditory ERPs and suggest a rapid speed to phonetic categorization (Bidelman *et al* 2013, Alho *et al* 2016, De Taillez *et al* 2020). Our results are consistent with previous neuroimaging studies that have shown the N1 and P2 ERPs are sensitive to auditory perceptual object identification (Wood *et al* 1971, Alain 2007, Bidelman *et al* 2013). The better decoding by LH as compared to RH activity is consistent with the dominance of LH in phoneme discrimination and speech sound processing (Zatorre *et al* 1992, Frost *et al* 1999, Tervaniemi and Hugdahl 2003, Bidelman and Howell 2016, Bidelman and Walker 2019). Our neural decoding results also corroborate previous hypothesis-driven work (Chang *et al* 2010, Bidelman *et al* 2013, 2014) by confirming speech sounds are converted to an abstract, categorical representation within the first few hundred milliseconds after stimulus onset.

4.2. Differential brain-networks involved in encoding and decision processing

Our results help identify the most stable, relevant, and invariant functional brain ROIs that support the brain-networks involved in encoding and decision processes of speech categorization using an entirely data-driven approach (stability selection coupled with SVM). During stimulus encoding, stability selection have identified 13 consistent ROIs that differentiate speech categories (82.6% accuracy; 0.5 stability threshold). Out of these 13 regions, eight of the ROIs are critically involved in the dorsal-ventral pathway for speech-language processing (Hickok and Poeppel 2004). These included areas in frontal lobe including IFG [BA 44, (i.e. pars opercularis L, pars triangularis R), i.e. 'Broca's area'], three regions from parietal and two regions from temporal lobe including primary auditory cortex (i.e. transverse temporal L). For later decision stages of the task, the same criterion of decoding performance (83.2% @ 0.5 stability threshold) have identified 15 ROIs that showed categorical neural organization. Out of these 15 regions, eight areas are from inferior frontal areas including BA 44 (i.e. pars opercularis L, pars opercularis R) and BA 45 (i.e. pars triangularis R), four regions from parietal lobe, and three regions from temporal lobe. Our data reveal two, relatively sparse, and partially overlapping neural networks that support different stages of speech categorization process.

Among the encoding and decision networks identified from our EEG data, five regions were common between the two topologies. Notably were the inclusion of BA44/45 that are heavily involved in speech-language processing (Novick *et al* 2010, Hickok *et al* 2011, Lee *et al* 2012). Early activation of IFG (during encoding) could be due to higher order speech centers exerting an inhibitory influence on auditory representations in order to prevent interference from nonlinguistic cues (Liberman *et al* 1981, Dehaene-Lambertz *et al* 2005) and optimize categorization, particularly under states of uncertainty (Carter and Bidelman 2021). The left inferior parietal lobe also appears as a common hub among the two networks. Superior parietal areas have been linked with auditory, phoneme, sound categorization, particularly when listeners are asked to resolve context or ambiguity (Dufor *et al* 2007, Myers and Blumstein 2008, Feng *et al* 2018). Involvement of superior frontal lobe in both networks is perhaps consistent with its role in higher cognitive functions and working memory (Klingberg *et al* 2002, Nyberg *et al* 2003). The fact that these extra-sensory regions can decode category structure even during stimulus encoding (<150 ms) suggests that the formation of speech categories might operate nearly in parallel within lower-order (sensory) and higher-order (cognitive-control) brain structures (Toscano *et al* 2018). However, these category representations need not be isomorphic across the brain. For example, category formation might reflect a cascade of events where speech units are reinforced and further discretized by a recontact of acoustic-phonetic with lexical representation of the speech category (Myers and Blumstein 2008).

Notable among the non-overlapping regions between stages were left primary auditory cortex (transverse temporal) and supramarginal gyrus, both of which were exclusive to the stimulus encoding period. Both regions have been implicated in the early acoustic analysis of the speech signal and related phonological processing (Zatorre *et al* 1992, Hickok *et al* 2000, Geiser *et al* 2008, Whitwell *et al* 2013, Deschamps *et al* 2014, Oberhuber *et al* 2016). Intuitively, their absence during the decision stage further suggests the categorical representation of speech, while present early in time (<150 ms), might take different forms in auditory-sensory cortex before being broadcast to decision mechanisms downstream.

Left postcentral gyrus is also exclusive during decision. Activation of this area proximal to the behavioral response execution most probably reflects motor planning and/or speech reconstruction (Martin *et al* 2014). Additional non-overlapping ROIs included pars opercularis in the RH. Right IFG has been implicated in attentional control and response imbibition (Hampshire *et al* 2010), which is consistent with its exclusive involvement in the decision stage of our task. Presumably, the other non-overlapping regions identified via stability selection (superior parietal L, insula L, Isthmus cingulate (l/rIST), caudal middle frontal L, entorhinal L, paracentral R, parahippocampal R) are also involved in decision processes, though as of yet, in an unknown way. Minimally, the involvement parahippocampal regions implies putative memory and retrieval processes. Still, more detailed localization studies (e.g. using functional magnetic resonance imaging) are needed to validate our EEG data, which offers a much coarser spatial resolution.

It is noticeable that during encoding, 7 out of 13 ROIs are from LH; for decoding, 9 out of 15 ROIs. The LH bias in our decoding data is perhaps expected given the LH dominance in auditory language processing (Caplan 1994, Tzourio *et al* 1998, Hull and Vaid 2006). Moreover, our results support previous studies by confirming a bilateral fronto-parietal network involved in auditory attentional, working memory (Belin *et al* 2002, Schneiders *et al* 2012), sound discrimination tasks (Hickok and Poeppel 2000), and phoneme categorization (Lee *et al* 2012, Loui 2015, Bidelman and Walker 2019). Interestingly, our study shows that only 15 brain regions (during decision) are needed to predict listeners' behavior CP with 85.0% accuracy.

In this work, we pooled Tk1 (i.e., /u/) and Tk 5 (i.e., /a/) stimuli since they are categorically *un*ambiguous vowels and examined their decoding relative to Tk 3, which is categorically ambiguous (Bidelman *et al* 2013). This approach partly assumes categorical responses of Tk1 and Tk5 are similar to one another. In contrast, Tk3 might represent a mixture of ambiguous responses, plus categorical responses to the perception of Tk1 or Tk5 (i.e. bistable perception). Though we do not find strong support for this notion in decoding source-level ERPs see Footnote (#2). Nevertheless, future work could examine decoding as a function of listeners' labeling speeds (e.g. Al-Fahad *et al* 2020) or listeners' trial-to-trial phonetic perception (Bidelman *et al* 2013) of speech tokens to unpack these alternate possibilities.

Acknowledgments

Requests for data and materials should be directed to G M B (gmbdlman@memphis.edu). This work was supported by the National Institutes of Health (NIH/NIDCD R01DC016267) and department of Electrical and Computer Engineering at the University of Memphis.

Author contributions

G M B designed the experiment, M S M, G M B, and M Y analyzed the data and wrote the paper.

ORCID iDs

Md Sultan Mahmud https://orcid.org/0000-0002-7361-122X
Gavin M Bidelman https://orcid.org/0000-0002-1821-3261

References

Al-Fahad R, Yeasin M and Bidelman G M 2020 Decoding of single-trial EEG reveals unique states of functional brain connectivity that drive rapid speech categorization decisions *J. Neural Eng.* **17** 016045

Alain C 2007 Breaking the wave: effects of attention and learning on concurrent sound perception *Hear. Res.* **229** 225–36

Alho J, Green B M, May P J, Sams M, Tiitinen H, Rauschecker J P and Jääskeläinen I P 2016 Early-latency categorical speech sound representations in the left inferior frontal gyrus *Neuroimage* **129** 214–23

Belin P, McAdams S, Thivard L, Smith B, Savel S, Zilbovicius M, Samson S and Samson Y 2002 The neuroanatomical substrate of sound duration discrimination *Neuropsychologia* **40** 1956–64

Bidelman G M 2015 Induced neural beta oscillations predict categorical speech perception abilities *Brain Lang.* **141** 62–69

Bidelman G M and Alain C 2015 Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception *J. Neurosci.* **35** 1240–9

Bidelman G M, Bush L and Boudreaux A 2020 Effects of noise on the behavioral and neural categorization of speech *Front. Neurosci.* **14** 153

Bidelman G M and Howell M 2016 Functional changes in inter- and intra-hemispheric cortical processing underlying degraded speech perception *Neuroimage* **124** 581–90

Bidelman G M and Lee -C-C 2015 Effects of language experience and stimulus context on the neural organization and categorical perception of speech *Neuroimage* **120** 191–200

Bidelman G M, Moreno S and Alain C 2013 Tracing the emergence of categorical speech perception in the human auditory system *Neuroimage* **79** 201–12

Bidelman G M and Walker B S 2017 Attentional modulation and domain-specificity underlying the neural organization of auditory categorical perception *Eur. J. Neurosci.* **45** 690–9

Bidelman G M and Walker B 2019 Plasticity in auditory categorization is supported by differential engagement of the auditory-linguistic network *Neuroimage* **201** 116022

Bidelman G M, Weiss M W, Moreno S and Alain C 2014 Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians *Eur. J. Neurosci.* **40** 2662–73

Caplan D 1994 *Language and the Brain* (New York: Academic) pp 1023–53

Carter J A and Bidelman G M 2021 Auditory cortex is susceptible to lexical influence as revealed by informational vs. energetic masking of speech categorization *Brain Res.* **1759** 147385

Casale S, Russo A, Scebba G and Serrano S 2008 Speech emotion classification using machine learning algorithms *2008 IEEE Int. Conf. Semantic Computing* pp 158–65

Celsis P, Doyon B, Boulanouar K, Pastor J, Démonet J-F and Nespoulous J-L 1999 ERP correlates of phoneme perception in speech and sound contexts *Neuroreport* **10** 1523–7

Chang E F, Rieger J W, Johnson K, Berger M S, Barbaro N M and Knight R T 2010 Categorical speech representation in human superior temporal gyrus *Nat. Neurosci.* **13** 1428

Cruz J A and Wishart D S 2006 Applications of machine learning in cancer prediction and prognosis *Cancer Inform.* **2** 117693510600200030

De Taillez T, Kollmeier B and Meyer B T 2020 Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech *Eur. J. Neurosci.* **51** 1234–41

Dehaene-Lambertz G, Pallier C, Serniclaes W, Sprenger-Charolles L, Jobert A and Dehaene S 2005 Neural correlates of switching from auditory to speech perception *Neuroimage* **24** 21–33

Desai R, Liebenthal E, Waldron E and Binder J R 2008 Left posterior temporal regions are sensitive to auditory categorization *J. Cogn. Neurosci.* **20** 1174–88

Deschamps I, Baum S R and Gracco V L 2014 On the role of the supramarginal gyrus in phonological processing and verbal working memory: evidence from rTMS studies *Neuropsychologia* **53** 39–46

Desikan R S *et al* 2006 An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest *Neuroimage* **31** 968–80

Domenech P and Dreher J-C 2010 Decision threshold modulation in the human brain *J. Neurosci.* **30** 14305–17

Du Y, Buchsbaum B R, Grady C L and Alain C 2016 Increased activity in frontal motor cortex compensates impaired speech perception in older adults *Nat. Commun.* **7** 12241

Dufor O, Serniclaes W, Sprenger-Charolles L and Démonet J-F 2007 Top-down processes during auditory phoneme categorization in dyslexia: a PET study *Neuroimage* **34** 1692–707

Efron B, Hastie T, Johnstone I and Tibshirani R 2004 Least angle regression *Ann. Stat.* **32** 407–99

Eimas P D, Siqueland E R, Jusczyk P and Vigorito J 1971 Speech perception in infants *Science* **171** 303–6

Eulitz C, Maess B, Pantev C, Friederici A D, Feige B and Elbert T 1996 Oscillatory neuromagnetic activity induced by language and non-language stimuli *Cogn. Brain Res.* **4** 121–32

Feng G, Gan Z, Wang S, Wong P C and Chandrasekaran B 2018 Task-general and acoustic-invariant neural representation of speech categories in the human brain *Cereb. Cortex* **28** 3241–54

Fox R A 1984 Effect of lexical status on phonetic categorization *J. Exp. Psychol. Hum. Percept. Perform.* **10** 526

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

Frost J A, Binder J R, Springer J A, Hammeke T A, Bellgowan P S, Rao S M and Cox R W 1999 Language processing is strongly left lateralized in both sexes: evidence from functional MRI *Brain* **122** 199–208

Geiser E, Zaehle T, Jancke L and Meyer M 2008 The neural correlate of speech rhythm as evidenced by metrical speech processing *J. Cogn. Neurosci.* **20** 541–52

Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P and Garrod S 2013 Speech rhythms and multiplexed oscillatory sensory coding in the human brain *PLoS Biol.* **11** e1001752

Guenther F H, Nieto-Castanon A, Ghosh S S and Tourville J A 2004 Representation of sound categories in auditory cortical maps *J. Speech Lang. Hear. Res.* **47** 46–57

Hampshire A, Chamberlain S R, Monti M M, Duncan J and Owen A M 2010 The role of the right inferior frontal gyrus: inhibition and attentional control *Neuroimage* **50** 1313–9
Hanley J A 1983 Appropriate uses of multivariate analysis *Annu. Rev. Public Health* **4** 155–80
Hickok G, Costanzo M, Capasso R and Miceli G 2011 The role of Broca's area in speech perception: evidence from aphasia revisited *Brain Lang.* **119** 214–20
Hickok G, Erhard P, Kassubek J, Helms-Tillery A K, Naeve-Velguth S, Strupp J P, Strick P L and Ugurbil K 2000 A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: implications for the explanation of conduction aphasia *Neurosci. Lett.* **287** 156–60
Hickok G and Poeppel D 2000 Towards a functional neuroanatomy of speech perception *Trends Cogn. Sci.* **4** 131–8
Hickok G and Poeppel D 2004 Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language *Cognition* **92** 67–99
Holt L L and Lotto A J 2010 Speech perception as categorization *Atten. Percept. Psychophys.* **72** 1218–27
Hsu C-W, Chang -C-C and Lin C J 2003 A practical guide to support vector classification technical report department of computer science and information engineering (Taipei: National Taiwan University)
Hull R and Vaid J 2006 Laterality and language experience *Laterality* **11** 436–64
Husain F T, Fromm S J, Pursley R H, Hosey L A, Braun A R and Horwitz B 2006 Neural bases of categorization of simple speech and nonspeech sounds *Hum. Brain Mapp.* **27** 636–51
James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning* Vol. 112 (Berlin: Springer)
Klingberg T, Forssberg H and Westerberg H 2002 Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood *J. Cogn. Neurosci.* **14** 1–10
Kuhl P K, Williams K A, Lacerda F, Stevens K N and Lindblom B 1992 Linguistic experience alters phonetic perception in infants by 6 months of age *Science* **255** 606–8
Lee Y-S, Turkeltaub P, Granger R and Raizada R D 2012 Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis *J. Neurosci.* **32** 3942–8
Liberman A M, Isenberg D and Rakerd B 1981 Duplex perception of cues for stop consonants: evidence for a phonetic mode *Percept. Psychophys.* **30** 133–43
Liebenthal E, Desai R, Ellingson M M, Ramachandran B, Desai A and Binder J R 2010 Specialization along the left superior temporal sulcus for auditory categorization *Cereb. Cortex* **20** 2958–70
Loui P 2015 A dual-stream neuroanatomy of singing *Music Percept.* **32** 232–41
Luck S J 2005 *An Introduction to the Event-related Potential Technique* (Cambridge, MA: MIT Press) pp 45–64
Mahmud M S, Ahmed F, Al-Fahad R, Moinuddin K A, Yeasin M, Alain C and Bidelman G 2020 Decoding hearing-related changes in older adults' spatiotemporal neural processing of speech using machine learning *Front. Neurosci.* **14** 1–15
Mahmud M S, Yeasin M and Bidelman G M 2021 Speech categorization is better described by induced rather than evoked neural activity *J. Acoust. Soci. Am.* **149** 1644–56
Mankel K, Barber J and Bidelman G M 2020 Auditory categorical processing for speech is modulated by inherent musical listening skills *Neuroreport* **31** 162–6
Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone N E, Rieger J, Schalk G, Knight R T and Pasley B N 2014 Decoding spectrotemporal features of overt and covert speech from the human cortex *Front. Neuroeng.* **7** 14
Masmoudi S, Dai D Y and Naceur A 2012 *Attention, Representation, and Human Performance: Integration of Cognition, Emotion, and Motivation* (New York: Psychology Press)
McClelland J L and Elman J L 1986 The TRACE model of speech perception *Cogn. Psychol.* **18** 1–86
Meinshausen N and Bühlmann P 2010 Stability selection *J. R. Stat. Soc. Series B Stat. Methodol.* **72** 417–73
Menon V and Desmond J E 2001 Left superior parietal cortex involvement in writing: integrating fMRI with lesion evidence *Cogn. Brain Res.* **12** 337–40
Miller C T and Cohen Y E 2010 Vocalizations as auditory objects: behavior and neurophysiology *Primate Neuroethology* ed M L Platt and A A Ghazanfar (Oxford: Oxford University Press) pp 237–55
Miller E K, Freedman D J and Wallis J D 2002 The prefrontal cortex: categories, concepts and cognition *Phil. Trans. R. Soc.* B **357** 1123–36
Miller E K, Nieder A, Freedman D J and Wallis J D 2003 Neural correlates of categories and concepts *Curr. Opin. Neurobiol.* **13** 198–203
Moinuddin K A, Yeasin M and Bidelman G M 2019 *BrainO* (available at: https://github.com/cvpia-uofm/BrainO) Accessed 9 September
Molfese D, Key A P F, Maguire M, Dove G O and Molfese V J 2005 Event-related evoked potentials (ERPs) in speech perception *The Handbook of Speech Perception* (Oxford: Blackwell) p 99121
Mostert P, Kok P and De Lange F P 2015 Dissociating sensory from decision processes in human perceptual decision making *Sci. Rep.* **5** 18253
Myers E B and Blumstein S E 2008 The neural bases of the lexical effect: an fMRI investigation *Cereb. Cortex* **18** 278–88
Noe C and Fischer-Baum S 2020 Early lexical influences on sublexical processing in speech perception: evidence from electrophysiology *Cognition* **197** 104162
Nogueira S, Sechidis K and Brown G 2017 On the stability of feature selection algorithms *J. Mach. Learn. Res.* **18** 174–1
Norris D, McQueen J M and Cutler A 2000 Merging information in speech recognition: feedback is never necessary *Behav. Brain Sci.* **23** 299–325
Novick J M, Trueswell J C and Thompson-Schill S L 2010 Broca's area and language processing: evidence for the cognitive control connection *Lang. Linguist. Compass* **4** 906–24
Nyberg L, Marklund P, Persson J, Cabeza R, Forkstam C, Petersson K M and Ingvar M 2003 Common prefrontal activations during working memory, episodic memory, and semantic memory *Neuropsychologia* **41** 371–7
Oberhuber M, Hope T M H, Seghier M L, Parker Jones O, Prejawa S, Green D W and Price C J 2016 Four functionally distinct regions in the left supramarginal gyrus support word processing *Cereb. Cortex* **26** 4212–26
Oldfield R C 1971 The assessment and analysis of handedness: the Edinburgh inventory *Neuropsychologia* **9** 97–113
Oostenveld R and Praamstra P 2001 The five percent electrode system for high-resolution EEG and ERP measurements *Clin. Neurophysiol.* **112** 713–9
Park Y, Luo L, Parhi K K and Netoff T 2011 Seizure prediction with spectral power of EEG using cost-sensitive support vector machines *Epilepsia* **52** 1761–70
Paus T, Petrides M, Evans A C and Meyer E 1993 Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses: a positron emission tomography study *J. Neurophysiol.* **70** 453–69
Perlovsky L 2011 Language and cognition interaction neural mechanisms *Comput. Intell. Neurosci.* **2011** 454587
Picton T W, Van Roon P, Armilio M L, Berg P, Ille N and Scherg M 2000 The correction of ocular artifacts: a topographic perspective *Clin. Neurophysiol.* **111** 53–65
Pisoni D B and Tash J 1974 Reaction times to comparisons within and across phonetic categories *Percept. Psychophys.* **15** 285–90

Rauschecker J P and Scott S K 2009 Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing *Nat. Neurosci.* **12** 718

Royston P and Sauerbrei W 2008 *Multivariable Model-building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables* vol 777 (New York: Wiley)

Ruppert D and Wand M P 1994 Multivariate locally weighted least squares regression *Ann. Statist.* **22** 1346–70

Russ B E, Lee Y-S and Cohen Y E 2007 Neural and behavioral correlates of auditory categorization *Hear. Res.* **229** 204–12

Sabri M, Binder J R, Desai R, Medler D A, Leitl M D and Liebenthal E 2008 Attentional and linguistic interactions in speech perception *Neuroimage* **39** 1444–56

Sahin N T, Pinker S, Cash S S, Schomer D and Halgren E 2009 Sequential processing of lexical, grammatical, and phonological information within Broca's area *Science* **326** 445–9

Saito T and Rehmsmeier M 2015 The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets *PLoS One* **10** e0118432

Schneiders J, Opitz B, Tang H, Deng Y, Xie C, Li H and Mecklinger A 2012 The impact of auditory working memory training on the fronto-parietal working memory network *Front. Hum. Neurosci.* **6** 173

Schultz E, Tan H and Hao S Weighted Regression in SAS, R, and Python (available at: https://jbhender.github.io/Stats506/F17/Projects/Abalone_WLS.html) (Accessed 27 May 2020)

Seabold S and Perktold J 2010 Statsmodels: econometric and statistical modeling with Python. *Proc. 9th Python Science Conf.* vol 57 p 61

Shen G and Froud K 2019 Electrophysiological correlates of categorical perception of lexical tones by English learners of Mandarin Chinese: an ERP study *Bilingualism* **22** 253–65

Tadel F, Baillet S, Mosher J C, Pantazis D and Leahy R M 2011 Brainstorm: a user-friendly application for MEG/EEG analysis *Comput. Intell. Neurosci.* **2011** 8

Tankus A, Fried I and Shoham S 2012 Structured neuronal encoding and decoding of human speech features *Nat. Commun.* **3** 1–5

Tervaniemi M and Hugdahl K 2003 Lateralization of auditory-cortex functions *Brain Res. Rev.* **43** 231–46

Toscano J C, Anderson N D, Fabiani M, Gratton G and Garnsey S M 2018 The time-course of cortical responses to speech revealed by fast optical imaging *Brain Lang.* **184** 32–42

Tsunada J and Cohen Y E 2014 Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits *Front. Neurosci.* **8** 161

Tzourio N, Crivello F, Mellet E, Nkanga-Ngila B and Mazoyer B 1998 Functional anatomy of dominance for speech comprehension in left handers vs right handers *Neuroimage* **8** 1–16

Whitwell J L, Duffy J R, Strand E A, Xia R, Mandrekar J, Machulda M M, Senjem M L, Lowe V J, Jack C R Jr and Josephs K A 2013 Distinct regional anatomic and functional correlates of neurodegenerative apraxia of speech and aphasia: an MRI and FDG-PET study *Brain Lang.* **125** 245–52

Wood C C, Goff W R and Day R S 1971 Auditory evoked potentials during speech perception *Science* **173** 1248–51

Xu Y, Gandour J T and Francis A L 2006 Effects of language experience and stimulus complexity on the categorical perception of pitch direction *J. Acoust. Soc. Am.* **120** 1063–74

Yin Q-Y, Li J-L and Zhang C-X 2017 Ensembling variable selectors by stability selection for the Cox model *Comput. Intell. Neurosci.* **2017** 1–10

Youssofzadeh V, Stout J, Ustine C, Gross W L, Conant L L, Humphries C J, Binder J R and Raghavan M 2020 Mapping language from MEG beta power modulations during auditory and visual naming *Neuroimage* **220** 117090

Zatorre R J, Evans A C, Meyer E and Gjedde A 1992 Lateralization of phonetic and pitch discrimination in speech processing *Science* **256** 846–9