NeuroImage 226 (2021) 117499

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Single-cell activity in human STG during perception of phonemes is organized according to manner of articulation

Yair Lakertz[a,b,c,1,*], Ori Ossmy[b,d,1], Naama Friedmann[b,c], Roy Mukamel[b,e,2], Itzhak Fried[f,g,2]

[a] Cognitive Neuroimaging Unit, NeuroSpin Center, Gif/Yvette, France
[b] Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel
[c] Language and Brain Lab, Sagol School of Neuroscience and School of Education, Tel-Aviv University, Tel-Aviv, Israel
[d] Department of Psychology and Center for Neural Science, New York University, New York, NY, United States
[e] School of Psychological Sciences, Tel-Aviv University, Tel-Aviv, Israel
[f] Department of Neurosurgery, David Geffen School of Medicine and Semel Institute for Neuroscience, University of California at Los Angeles, Los Angeles, CA, United States
[g] Tel Aviv Medical Center and Sackler Faculty of Medicine, Tel-Aviv University, Tel Aviv, Israel

ARTICLE INFO

Keywords:
Speech perception
Single-unit recordings
Naïve Bayes
Phoneme representation
Superior temporal gyrus

ABSTRACT

One of the central tasks of the human auditory system is to extract sound features from incoming acoustic signals that are most critical for speech perception. Specifically, phonological features and phonemes are the building blocks for more complex linguistic entities, such as syllables, words and sentences. Previous ECoG and EEG studies showed that various regions in the superior temporal gyrus (STG) exhibit selective responses to specific phonological features. However, electrical activity recorded by ECoG or EEG grids reflects average responses of large neuronal populations and is therefore limited in providing insights into activity patterns of single neurons. Here, we recorded spiking activity from 45 units in the STG from six neurosurgical patients who performed a listening task with phoneme stimuli. Fourteen units showed significant responsiveness to the stimuli. Using a Naïve-Bayes model, we find that single-cell responses to phonemes are governed by manner-of-articulation features and are organized according to sonority with two main clusters for sonorants and obstruents. We further find that 'neural similarity' (i.e. the similarity of evoked spiking activity between pairs of phonemes) is comparable to the 'perceptual similarity' (i.e. to what extent two phonemes are judged as sounding similar) based on perceptual confusion, assessed behaviorally in healthy subjects. Thus, phonemes that were perceptually similar also had similar neural responses. Taken together, our findings indicate that manner-of-articulation is the dominant organization dimension of phoneme representations at the single-cell level, suggesting a remarkable consistency across levels of analyses, from the single neuron level to that of large neuronal populations and behavior.

1. Introduction

How are phonemes encoded in the human auditory cortex during speech perception?

Phonemes constitute the basic sound units of language, distinguishing one word from another. However, research in linguistics during the 20th century showed that phonemes are not the ultimate constituents of phonological analysis, and that phonemes can be further factorized into sub-phonemic distinctive features (Jakobson et al., 1951; Chomsky and Halle, 1968; Jakobson, 1968). For example, the phoneme /m/ might be represented by the following features [+sonorant, -continuant, +voice, +nasal, +labial], which indicate its acoustic properties and its place- and manner-of-articulation. Roughly, place-of-articulation describes the point of contact where an obstruction occurs along the vocal tract (labial, velar, etc.), whereas manner-of-articulation describes how airflow is obstructed (nasal, fricative, etc.). The apparently large number of phonemes found in the languages of the world can be thus represented in a compact way by a relatively small number of phonological features (Clements 1985).

Psychophysical studies of speech perception further suggested that these linguistic distinctions regarding phonological features might also have a cognitive reality. For example, in the classic study by Miller and Nicely (1955) it was shown that the underlying psychometric representation of English consonants is related to the representational structure defined by phonological features. Furthermore, by gradually degrading phoneme stimuli with noise, it was shown that some phonological features are more robust to noise compared to others. Specifically, it was

[*] Corresponding author.
E-mail addresses: yair.lakretz@gmail.com, yair.lakretz@cea.fr (Y. Lakertz).
[1] Co-first authorship
[2] Co-senior authorship.

https://doi.org/10.1016/j.neuroimage.2020.117499
Received 18 May 2020; Received in revised form 29 September 2020; Accepted 21 October 2020
Available online 24 October 2020

shown that manner features, such as nasality (distinguishing, e.g., /n/ from /d/ and /m/ from /b/), are more robust compared to place-of-articulation. Subsequently, with advances of neuroimaging and electrophysiological methods, a major effort in research on speech perception focused on whether these linguistic distinctions and psychological representations of speech are also reflected in an underlying neural code.

Numerous neuroimaging studies that examined phoneme perception (Binder et al., 2000; Dehaene-Lambertz et al., 2005; Liebenthal et al., 2005; Möttönen et al., 2006; Desai et al., 2008; Formisano et al., 2008; Liebenthal et al., 2010; DeWitt and Rauschecker 2012; Arsenault and Buchsbaum 2015; Venezia et al., 2019) showed activation in regions that are selective to speech, compared to non-phonemic stimuli. Findings describe a hierarchical organization of regions in the temporal lobe from primary auditory and early posterior auditory areas processing low-level auditory features, to the anterior, ventral Superior Temporal Gyrus (STG) and Superior Temporal Sulcus (STS), processing higher-level phonemic features.

Electrocorticogram (ECoG) research of speech perception shows that the organization of phonemes can significantly differ across brain regions and tasks, depending on whether speech is being produced or perceived (Bouchard et al., 2013; Cheung et al., 2016). Bouchard et al. (Bouchard et al., 2013) showed that during production, phonemes in the ventral sensory-motor cortex (vSMC) are predominantly organized by place-of-articulation features, whereas during listening, the organization was found to be dominated by manner-of-articulation features (Cheung et al., 2016). The same studies also showed that the dominant organizing feature in the STG during perception is also manner-of-articulation. Additionally, other researchers showed that waveform reconstruction from local field potentials in the lateral STG is highest for sound features most critical to speech intelligibility (e.g., low modulation frequencies in both time and frequency; Pasley et al., 2012). Their findings suggest that speech acoustic parameters are encoded in this region (see also, Mukamel et al., 2011; Nourski et al., 2009; and in MEG: Ahissar et al., 2001).

More recently using ECoG, Mesgarani et al. (2014) showed that in the STG, high-gamma activity (75–150 Hz) in response to auditory presentation of phonemes is clustered according to phonetic features such as sonority, nasality and stridency, which remarkably are the same distinctive features defined by linguists (Chomsky and Halle 1968, Grodzinsky and Nelken, 2014). At the neural level, phonemes with common ‘manner-of-articulation’ (such as stridents /s/,/z/,/ʃ/) evoked more invariant responses than phonemes with common ‘place-of-articulation’ (such as alveolars /t/,/d/,/s/,/z/,/n/). This representational structure of phonemes is also supported by scalp EEG recordings (Di Liberto et al., 2015, Khalighinejad et al., 2017, Sankaran et al., 2018), identified in a similar latency window to that found in EcoG studies, of around 150 ms.

However, electrical activity recorded by ECoG or EEG grids reflects average responses of large neuronal populations and is therefore limited in providing insights into activity patterns of single neurons. Whereas some findings suggested that single-unit and LFP activity can differ in some regions (e.g., Donchin et al., 2001), others showed that there is a high correspondence between the two in the auditory cortex (Mukamel et al. 2005; Nir et al. 2007). Moreover, STG neurons were found to be tuned to subsets of phonemes and to have a sparse coding scheme (Creutzfeldt et al., 1989; Chan et al., 2013), consistently with the findings from EcoG and EEG. This suggests that the functional organization of phonemes identified in ECoG recordings (Mesgarani et al., 2014) might be reflected also at the cellular level. However, the exact functional organization of phonemes at the cellular level is yet unknown.

Here, we tested the hypothesis that phoneme representations in the STG described at the single-neuron level are organized based on manner- rather than place-of-articulation features. For this, we studied the functional organization of phonemes as revealed by spiking activity and contrasted the two alternative explanations. We also examined whether the functional organization of phonemes as revealed by single-cell activity matches behavioral responses. To this end, we used the same set of stimuli in both a behavioral experiment with healthy subjects and in the experiment with neurosurgical patients implanted with intracranial depth electrodes.

2. Materials and methods

2.1. Participants and electrophysiological recording

Data was collected from six neurosurgical patients with pharmacologically intractable epilepsy (3 males and 3 females; ages between 21 and 58), implanted with intracranial depth electrodes to identify seizure focus for potential surgical treatment (Mukamel and Fried 2012). The six volunteers were recruited from two centers (UCLA/Tel-Aviv). Electrode location was based solely on clinical criteria. Each electrode terminated in a set of nine 40- μm platinum–iridium microwires (Fried et al., 1999)—eight active recording wires, referenced to the ninth. Signals from these microwires were recorded at 40 kHz using a 64-channel acquisition system. Before surgery each patient underwent placement of a stereotactic headframe, and then a detailed CT and CT-angiogram (CTA) images were obtained and fused to preoperative MRI. Surgical planning was then performed, with selection of appropriate temporal and extra-temporal targets and appropriate trajectories based on clinical criteria. To verify electrode position, CT scans following electrode implantation were co-registered to the preoperative MRI. The participants provided written informed consent to participate in the experiments. The study was approved by and conformed to the guidelines of the Medical Institutional Review Board at UCLA and the Tel-Aviv Sourasky Medical Center.

2.2. Stimuli and behavioral task

The stimuli were constructed of either consonant-vowel (CV) pairs, or vowels /a e i o u/ presented in isolation.[2] The consonants in the CV syllables are listed in Table 1, and the vowel was set to /a/. Patients were presented with 12 repetitions of each CV pair or vowel, presented by 3 different speakers (4 repetitions for each speaker of each stimulus), in a random order (ISI = 1 second).

All stimuli were recorded in an anechoic chamber with a RØDE NT2-A microphone and a Metric Halo MIO2882 audio interface, at a sampling rate of 44.1 kHz. Stimuli were generated by two male and one female Hebrew speakers. The total number of stimuli was 63 (21 phonemes * 3 speakers). Since some participants were native English speakers and some were native Hebrew speakers, we chose phoneme stimuli that are approximately similar across English and Hebrew (verified in a perceptual task with native English speakers; see Phoneme perception experiment). Length and pitch (by semi-tone intervals) were compared across recorded tokens to choose the most highly comparable stimulus-types. This was done by looking at differences in timeline arrangement, using built-in pitch tracker in a commercial software (Logic Pro-X). Further cleaning of noise residues in high-resolution mode was done using Waves X-Noise software. Figure S1 shows an example of the waveform of the syllable /ʃa/ (top), with the corresponding spectrogram (bottom), articulated by one of the male speakers. The participants were only requested to listen carefully to the syllables.

2.3. Data preprocessing

To detect spiking activity, the data was band-pass filtered offline between 300 and 3000 Hz and spike sorting was performed using WaveClus (Quiroga et al., 2004), similar to previous publications (Quiroga et al., 2005; Ossmy et al., 2015). This process allows determining whether the data was recorded from a single- or multi-unit (for

[2] Stimulus files are available at (https://github.com/yairlak/phonemes_single_unit_STG) and additional acoustic properties are presented in the Supplemental Materials.

Table 1
Stimuli details. List of phonemes used in the experiment and their corresponding features.

		a	e	o	i	u	n	m	l	j	f	v	s	z	ʃ	ʒ	p	b	t	d	k	g
	Sonorant	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−
	Vowel	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Manner	Nasal	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−
	Approximant	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−
	Fricative	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+	−	−	−	−	−	−
	Plosive	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+
Place	Labial	−	−	−	−	−	−	+	−	−	+	+	−	−	−	−	+	+	−	−	−	−
	Coronal	−	−	−	−	−	−	−	+	−	−	−	+	+	+	+	−	−	+	+	−	−
	Dorsal	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	+	+
	Alveolar	−	−	−	−	−	+	−	+	−	−	−	+	+	−	−	−	−	+	+	−	−
	Palatal	−	−	−	−	−	−	−	−	+	−	−	−	−	+	+	−	−	−	−	−	−
	Velar	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+

full technical details see Quiroga et al., 2005), and yields for each detected unit (single or multi) a vector of time stamps (1 ms resolution) during which spikes occurred. To assess responsiveness of each neuron to the phonemes, we computed a *t*-test between the spike-count distribution before stimulus onset (−500–0 ms) and after (0–500 ms). Neurons with statistically-significant responses ($p < 0.05$) to at least one phoneme were included in subsequent analyses.

2.4. Similarity of neural and behavioral responses

To test whether the similarity of phonemes at the behavioral level corresponds with the similarity of population spiking activity in STG, we compared two phoneme-similarity matrices - a behavioral and a neural one. The Behavioral Similarity (BS) is calculated from phoneme confusability according to:

$$BS_{ij} = \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}} \tag{1}$$

where p_{ij} is the proportion of times that phoneme *i* was called phoneme *j*. p_{ii} is the hit rate for phoneme *i*, and $\sum_j p_{ij} = 1$. Thus, BS_{ij} is high if subjects frequently confused phoneme *i* with phoneme *j* (high similarity).

The Neural Similarity (NS) is based on spiking activity in the following way: first, we z-scored the spike-count activity in the response window across all responsive neurons from all patients. Then, for each pair of phonemes *i* and *j*, we calculated their Euclidean distance d_{ij} in the response space, and neural similarity was defined according to the following (monotonic) function NS_{ij}= exp(-d_{ij}).

Finally, we performed Spearman rank correlation between the two matrices. The result is therefore not affected by the exact shape of the function.

2.5. Naïve Bayes model

We modeled the observed spike counts from all units assuming that the number of spikes follows a Poisson distribution. Formally, when observed spike-count x_i in unit *i* follows a Poisson distribution $x_i \sim Poisson(\lambda_i)$, the probability of observing *k* spikes in a time bin, generated by the unit in response to the presentation of stimulus type *s*, is:

$$p(x_i|s) = e^{-\lambda_{i,s}} \frac{\lambda_{i,s}^{k}}{k!} \tag{2}$$

where $\lambda_{i,s}$ is the firing rate of unit *i* in response to stimulus type *s*. We modeled the joint spiking activity across units using a Naïve Bayes model. The Naïve Bayes model typically assumes that given a stimulus type (a phoneme or a phonological feature), the observed spike counts across units are independent of each other. We evaluated this assumption for our case by first calculating pairwise correlations among unit activities in response to the various stimuli and found that all pairwise correlations are low (|r| <0.1; Figure S4; panels A & B). Since low pairwise correlation does not ensure independence, we further estimated the mutual-information between pairs of units (see, e.g., Rolls and Treves, 2011) and found that the mutual information among all pairs is low (Figure S4; panels C & D), specifically for unit pairs recorded in the same patient. This suggests that the independence assumption of the model holds for our data. Given that, the joint probability of stimulus and responses can be conveniently factorized. Formally, the probability of observing a spike-count pattern $x \in R^n$ across units in response to the presentation of a stimulus type *s* is:

$$p(x|s) = \prod_{i=1}^{n} p(x_i|s) = \prod_{i=1}^{n} e^{-\lambda_{i,s}} \frac{\lambda_{i,s}^{k_i}}{k_i!} \tag{3}$$

where k_i is the number of observed spikes in unit *i*, and *n* is the number of units. Because the classes were not equal in size (e.g., in classification according to manner, there were six phonemes in the fricative and plosives classes but only five phonemes in the vowels class and four in the Nasal-Approximants class), we randomly sampled a subset from the larger classes to match the size of the smallest class in a 5-fold cross-validation procedure. We then split the samples of each class into a training and a test-set according to a 80%−20% ratio, respectively.

We estimated the firing rate parameters $\lambda_{i,s}$ from the training data using maximum likelihood. That is, for each stimulus type s and unit *i*, we found the firing-rate parameter $\lambda_{\mathrm{i,s}}$ that maximizes the likelihood of observing the spike counts in the training-set trials: $\prod_{t \in training-set} e^{-\lambda_{i,s}} \frac{\lambda_{i,s}^{k_i^t}}{k_i^t!}$, where k_i^t is the number of observed spikes in unit *i* in trial *t*. For the Poisson distribution, as in this case, the maximum-likelihood estimator can be shown to be equal to the mean spike-count.

2.6. Inference

Having estimated all firing-rate parameters $\lambda_{\mathrm{i,s}}$, we now describe inference in the model. Given an observed activity pattern across all units $\mathrm{x_t}$, we infer for each trial t in the test-set the most probable stimulus type s. Using Bayes rule, the posterior distribution is:

$$p(s|x^t) \propto p(x^t|s)p(s) = \prod_{i=1}^{n} e^{-\lambda_{i,s}} \frac{\lambda_{i,s}^{k_i}}{k_i^t!} p(s) \tag{4}$$

where $p(s)$ is the prior probability of the stimulus type, which was set as uniform. The mode of the posterior distribution indicates the most probable stimulus type given the firing pattern across units.

2.7. Model evaluation

The model is evaluated by comparing the predictions of the model from the inference stage and the ground-truth labels. For binary classification tasks, we use the area under the curve (AUC) as a measure for model performance, with posterior probabilities as scores. For multiclass classification, the full posterior distribution provides additional

Table 2
Recording details. Distribution of recorded spiking activity in STG across hemispheres and patients (total responsive STG units = 14).

	Left STG	Right STG
Patient1	No units recorded	**Responsive: 3 multi-unit** Not Responsive: None
Patient2	**Responsive: 1 single-unit; 1 multi-unit** Not Responsive: 1 single-unit; 4 multi-unit	No units recorded
Patient3	**Responsive: 1 multi-unit** Not Responsive: 1 single-unit; 2 multi-unit	No units recorded
Patient4	No units recorded	**Responsive: 2 multi-units; 3 single-units** Not Responsive: None
Patient5	Responsive: None Not Responsive: 2 single unit; 2 multi-unit	**Responsive: 1 multi-unit** Not Responsive: 6 multi-unit
Patient6	No units recorded	**Responsive: 2 single-unit** Not Responsive: 4 single-unit; 9 multi-unit

information compared to its mere mode. For each stimulus type, we calculate the average posterior distribution across all trials in the test set, and use this to construct for each classification task a confusion matrix, in which rows correspond to average posterior distributions. In all cases, statistical significance is determined from the distribution of values across test sets.

3. Results

3.1. Basic characteristics of the neural responses

We recorded spiking activity from a total of 45 units in six patients implanted with intracranial depth electrodes, while they listened to a variety of phonemes (See Materials and Methods). Of the 45 units, 14 exhibited significant increases in firing rate following stimulus onset and were taken for further analysis (see Materials and Methods, Table 2, and Fig. 1). Fig. 2 depicts rasters and peri-stimulus time histograms (PSTH) plots of spiking activity from one unit in the right STG of patient 4 (see Fig. 1). In most neurons, increases in firing rate were observed ~180 ms following stimulus onset, likely due to conductance delays until the signal reaches STG. Some responses contained two activity peaks (e.g., the PSTHs of /b p d s/ in Fig. 2), which may be a result of the structure of the stimuli—a consonant followed by the vowel. For some phonemes, there is a sparse response before the stimulus onset that is not locked to stimulus presentation and was considered as spontaneous activity or noise.

To identify periods for which the neural response is most informative with respect to phoneme identity, we defined a ‘response window’ — the time window for which spiking activity is most separable across phonemes. To that end, we defined a separability index based on the ratio of spike-count variability across trials of different phonemes and trials in which a single phoneme was presented. Spike counts were calculated in 200 ms windows, and the separability index was calculated in the range of −100 ms to +500 ms relative to stimulus onset in steps of 1 ms. Fig. 3A shows the average of the separability index across all units. The center of the most informative time window is around 179 ms after stimulus onset and was used in subsequent analysis (similar to (Chan et al., 2013); Changing the time window for calculating spike counts in the range of 100–300 ms instead of 200 ms did not substantially change the profile of the separability index). This period is similar to the P2 component during phonemic and non-phonemic processing reported in EEG studies, with an activity that peaks at a similar range of time delays from sound onset (Liebenthal et al., 2010).

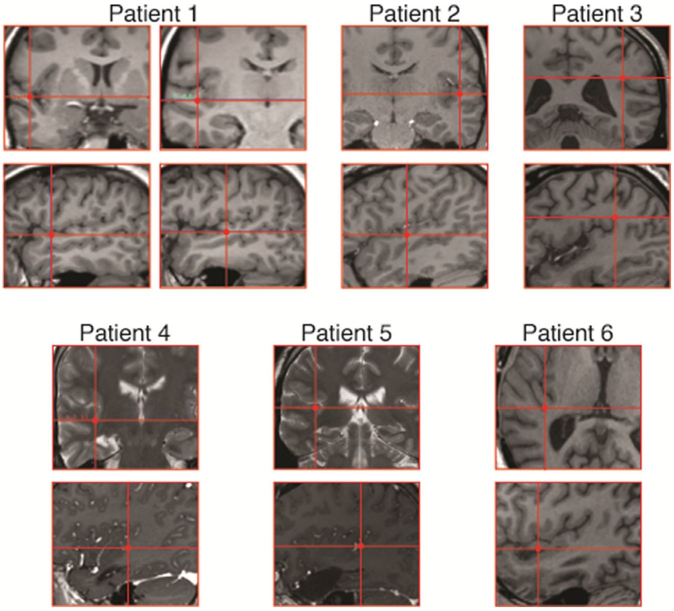

Fig. 1. MRI localization of microelectrodes for the 6 patients with responsive units. Patients 1–5: coronal (top) and sagittal (bottom) T1 weighted views (sagittal 4,5 –post Gadolinium). Patient 6: axial (top) and sagittal (bottom) T1-weighted views. Location of microwire depicted with a single cross or dot on the sagittal views and as the most distal point on electrode trajectory on coronal or axial views (in 1a,b also by blue crosshair). Other red contacts along the electrode shaft (green in 1b) depict the macro contacts for recording of intracranial EEG (iEEG). Note that in patient #3 microwire is in quite posterior in the peri-Sylvian gray matter as superior temporal gyrus turns into supramarginal gyrus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. The functional organization of phonemes

To examine whether neural responses in STG are functionally clustered, we represented each phoneme as a vector of firing-rate values. To capture the temporal dynamics of the neural response, each phoneme was represented by mean firing rates across trials in four 50 ms consecutive bins (Chan et al., 2013; Mesgarani et al., 2014; Ossmy et al., 2015) in the response window (79ms-279 ms) for all fourteen units, giving 56 dimensions in total. Fig. 3B shows the firing rate for all units in response to all phoneme stimuli, evaluated within the entire response window (200 ms).

Next, we applied principal component analysis (PCA) to project the neural representation to a lower dimensional space, spanned by two principal components of the data, where the variance explained by the first and second PCs was 22.73% (p < 0.001; permutation test) and 12.13% (p = 0.057; permutation test), respectively. We found that the sonorant and obstruent phonemes have relatively distinct neural representations, as each group encompasses a different region of the plane (Fig. 4A). Based on Euclidean distances among the neural representations of the phonemes, we generated a similarity matrix among the phonemes (Fig. 4B, top panel) and performed an unsupervised hierarchical clustering on the similarity matrix. We found a central cluster of obstruents (except for /k/, and including /e/), separated from most sonorants - the vowels /a o i u/ and nasal approximants /n m l j/ (Fig. 4B, bottom panel). In addition, the obstruent cluster is further divided into a sub-cluster containing all stridents /s ʃ z ʒ/. These results point to a functional organization based on manner-of-articulation features, since clustering tends to separate obstruents from sonorants,

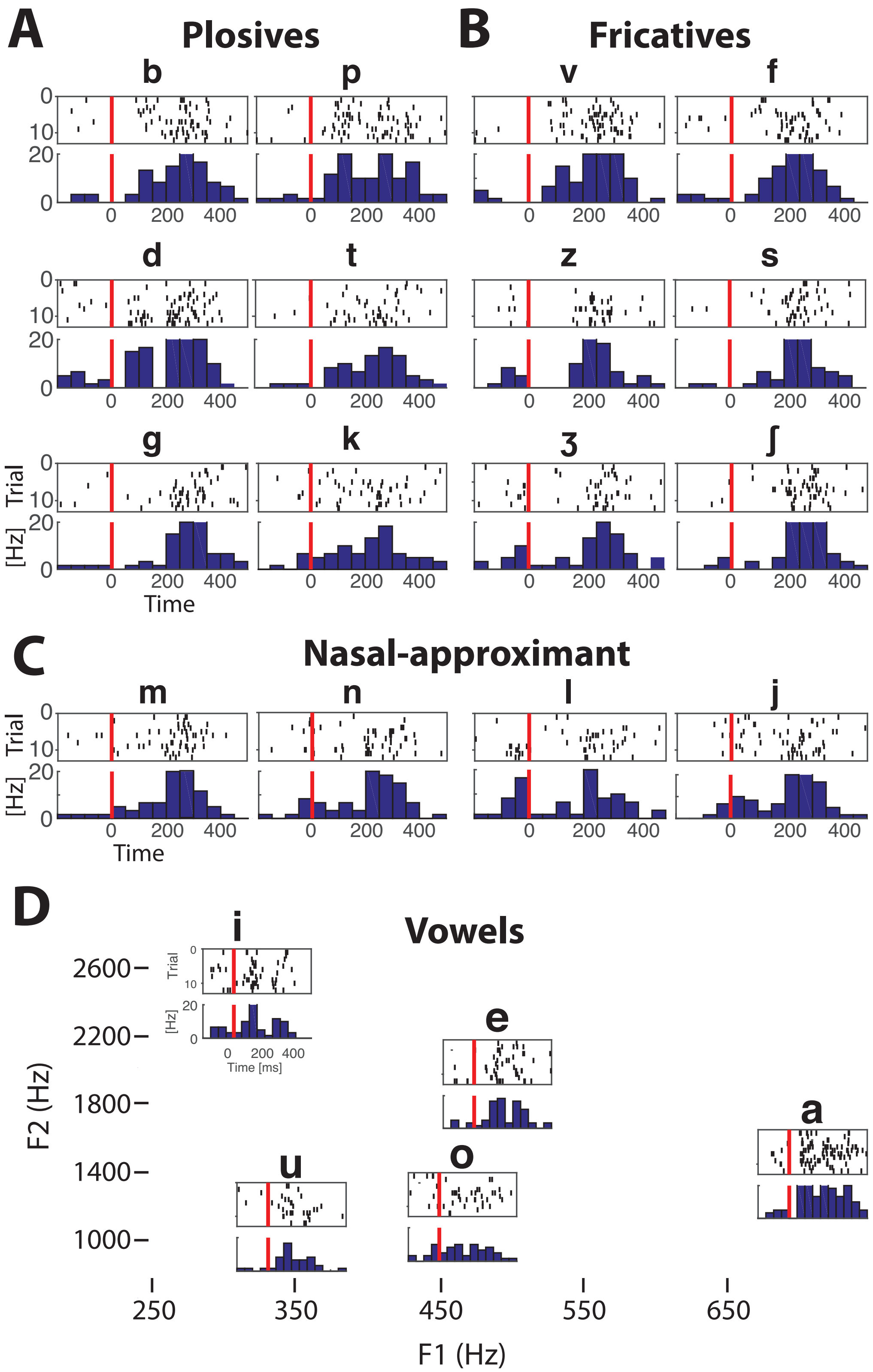

Fig. 2. Rasters and Peri-Stimulus Time Histogram plots for one example unit, from patient 4, located at the right hemisphere (see Fig. 1). Consonants are grouped into three groups: plosives, fricatives, and nasal-approximant. **(A)** Voiced (left) and unvoiced (right) plosives. **(B)** Voiced (left) and unvoiced (right) fricatives. **(C)** nasal-approximant (left) and affricate (right) phoneme. **(D)** Vowel rasters are embedded in approximate locations in the formant space.

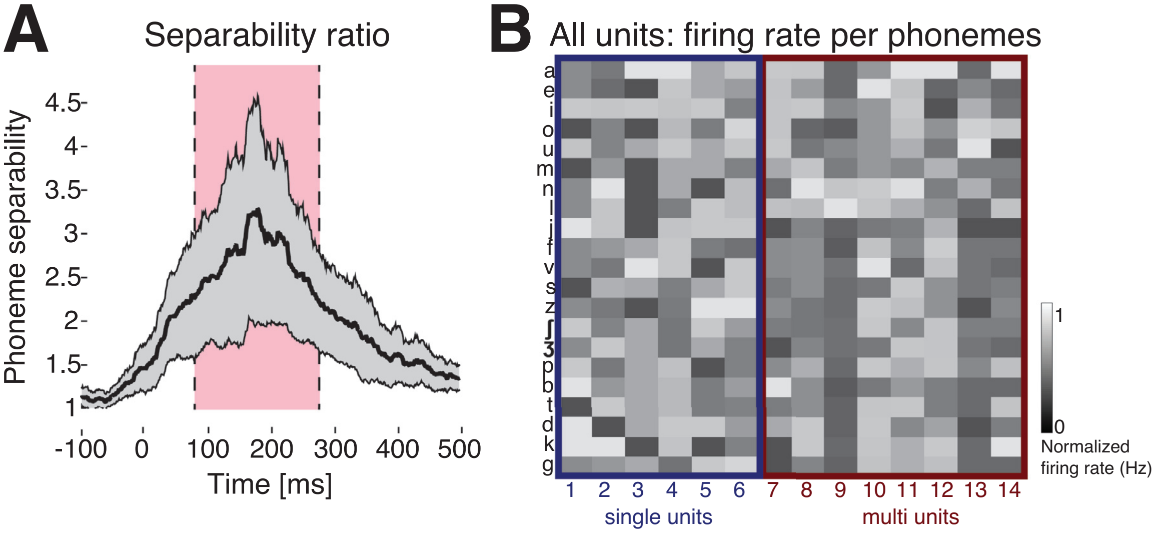

Fig. 3. (A) Response window. The between-phoneme to within-phoneme variability ratio of the spike-count, for a running window of 200 ms (calculated between −100 ms to 500 ms relative to stimulus onset in steps of 1 ms), averaged across responsive units (error bars represent SEM across units). Ticks on the abscissa represent the center of the time window. Response window (79ms-279 ms, shaded area) had the maximal value of phoneme separability index. **(B)** Mean firing rates in the entire 200 ms response window (all four 50 ms bins), for all units, in response to all phoneme stimuli. Color scale represents mean firing-rates for each unit in the response window, normalized by the peak response of each unit.

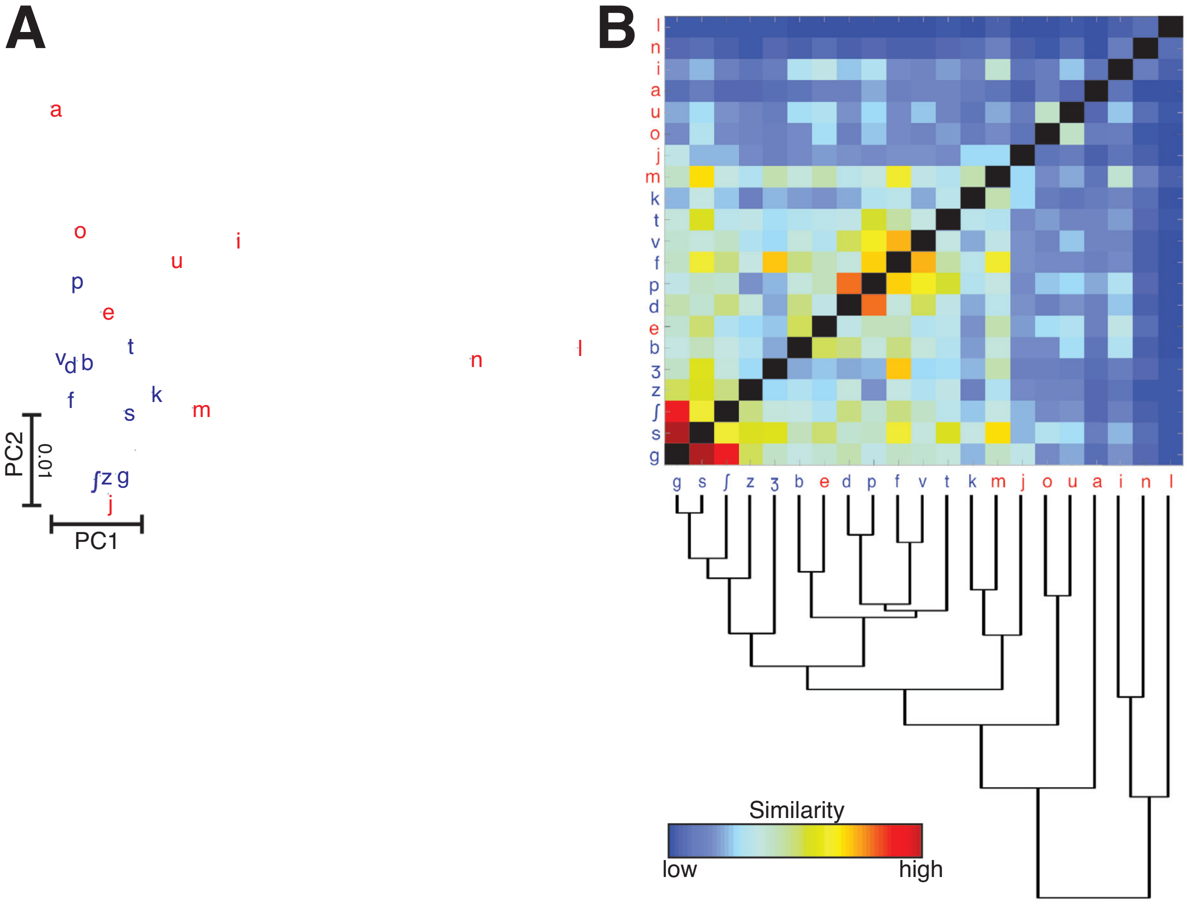

Fig. 4. (A) Neural representations of phonemes along the first two principal components of the data. Phonemes are represented using the International Phonetic Alphabet (IPA) notation. Colors: sonorant phonemes (red), obstruent phonemes (blue). **(B)** Hierarchical clustering. Top panel depicts the similarity matrix based on the neural population responses. Similarity metric is based on Euclidean distances among the neural representations of the phonemes in 56 dimensions (14 units x 4 time bins; see Materials and Methods). Colors: sonorant phonemes (red), obstruent phonemes (blue). Bottom panel depicts hierarchical clustering of the same data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and to group strident phonemes together. Therefore our next analysis focused on quantifying and comparing response invariances to manner- and place-of-articulation features directly, using a Naïve Bayes model for spike generation (see Materials and Methods for details).

If manner is a more dominant organizing principle than place, we expect the model to achieve better decoding performance for manner- compared to place-of-articulation features. The confusion errors made by the model are also informative regarding the functional organization of phonemes — higher confusion rate between two classes indicates a higher similarity between their neural representations. If manner is a more dominant dimension at the single-cell level, we expect to observe lower confusion rates of the model among phonemes with different manners of articulation and higher confusion rates among phonemes that share the same manner of articulations.

We examined the performance of the model on two multi-class classifications, for each of the two cases: manner- and place-of-articulation features. For each classification, we labeled the phonemes according to the corresponding phonological features. For manner, we label /a e i o u/ as 'vowel', /n m l j/ as 'nasal-approximant', /f v s z ʃ ʒ / as 'fricative', /b d g p k/ as 'plosives'; and for place-of-articulation, /b p f v m/ as 'labial', /t d s z n/ as 'alveolar', /ʃ ʒ/ as palatal and /k g/ as velar. We then generated a confusion matrix per classification according to the inferences of the model. Fig. 5 shows the significant mean posterior distribution for all phonological features ($p < 0.05$; t-test compared to chance level), organized in a confusion matrix. Classification according to manner-of-articulation (Fig. 5A) resulted in a diagonal structure with higher values on the diagonal, compared to the place-of-articulation classification (Fig. 5B). We quantified the extent to which each matrix is diagonal by computing the ratio between the mean of diagonal values and the mean of non-diagonal values. We found a significant difference between the two matrices (manner = 2.89±0.43, place = 1.22±0.49, $p < 0.001$; t-test).

To establish the dominance of manner-of-articulation features in distinguishing phonemes, we performed a third classification task. For each phonological feature (e.g., [nasal]), we labeled all phonemes as either + or - ([+nasal] or [-nasal] respectively), and calculated the area under curve (AUC) value for each binary classification. Fig. 5C depicts AUC values for all phonological features in descending order. AUC values in all four manner-of-articulation features are significant ($p < 0.05$; compared to chance level, AUC = 0.5) whereas for place-of-articulation, only the labial feature is significantly above chance level.

3.3. A comparison between neural and behavioral similarity

Finally, we directly compared neural and perceptual similarities of phonemes. Traditionally, perceptual phoneme similarity is estimated using behavioral tasks, assuming that confusion between two phonemes is correlated with perceptual similarity (Miller and Nicely 1955; Tversky 1977; Shepard 1987). We tested whether phoneme similarity, as estimated in a previous behavioral task (Lakretz et al., 2018), is reflected in neural activity in the STG during listening to the same set of phoneme stimuli.

To that end, we generated one behavioral and one neural similarity matrix. The behavioral similarity matrix is estimated from confusion errors made by thirty-two healthy human subjects, and the neural similarity matrix is derived from the neural representations of the STG responses obtained in the neurosurgical subjects (see Materials and Methods). Since behavioral tasks are limited in generating confusions between consonants and vowels, we focused on the confusion between consonant phonemes only (averaged across subjects). We found a significant correlation (with large span of the scatter) between the behavioral and the neural similarity matrices (Fig. 5D; $\rho = 0.45$, $p < 0.001$, Spearman correlation). To test whether this correlation can be partly explained by acoustic properties of the stimuli, we also calculated the correlation between the neural and acoustic similarity matrices, but found no significant correlation (Figure S3; $\rho = -0.15$, $p = 0.1$, Spearman correlation). Taken together, this suggests that perceptual similarity observed in behavioral tasks can be represented at the level of spiking activity of small population of neurons in STG.

3.4. Phoneme perception test

Some of the patients in the current study were native speakers of English and some were native speakers of Hebrew. Differences in native language can affect how phonemes are processed and perceived. Our stimuli were generated by Hebrew speakers only, we therefore selected a subset of phoneme stimuli that are similar in English and Hebrew. To verify that English speakers perceive the selected phonemes, which were generated by Hebrew speakers, in a similar way to Hebrew speakers, we tested the extent to which the selected phoneme stimuli used in the experiment are indeed similar across English and Hebrew speakers. To that end, we performed a phoneme perception test. Eighteen native English speakers (age range 18.2–35, 12 females; monolingual) sat in front of a screen with headphones and listened to the phoneme stimuli used in the study. After each phoneme, subjects were presented with 21 phonemes on the screen and were asked to select the phonemes they heard. In addition, they were asked to rate their level of confidence in the phoneme selection. Order of played phonemes and options on the screen were randomized across subjects. Participants identified the phonemes with 79% accuracy, significanlt higher than chance ($p < 0.05$; t-test) and high confidence levels (9.2/10 averaged across subjects).

4. Discussion

A fundamental question in the research of speech perception concerns the functional representation of phonemes in the auditory cortex. Recently, this question has been addressed by studies in neuroscience using invasive ECoG recordings (Mesgarani et al., 2014). These recordings provide a precious glimpse into the neural representations of linguistic entities, such as the objects of speech perception, with high temporal resolution and spatial localization compared to non-invasive recording techniques. Invasive techniques can record extracellular electrical activity either at the level of LFPs or at the level of action potentials generated by single cells (Mukamel and Fried 2012). So far, evidence from invasive recordings regarding the representation of phonemes was based on activity of large populations of neurons, thus leaving open the question regarding the representation of phonemes at the single-unit level. We characterized seemingly distributed, yet possibly clustered, response patterns (14 neurons) to different vowels and consonant-vowel syllables. We directly inquired whether in STG, (a) the organization of phoneme representation at the level of single-cell activity is dominated by manner or by the place-of-articulation; and (b) perceptual representation of phonemes at the behavioral level matches the neural representation at the cellular level.

We found that the structure of the neural representations of phonemes in a relatively small population of neurons demonstrates a separation between sonorant and obstruent phonemes. These findings are in agreement with previous ECoG studies (Mesgarani et al., 2014) that examined the organization of phonemes in the STG and found that the dominant distinctive features are manner-of-articulation that contribute most for phoneme classification, providing support to auditory theories of speech perception (Stevens 1972, 1989, 2002) over the motor theory one (Liberman et al., 1967; Liberman and Mattingly 1985; Galantucci et al., 2006), and are consistent with an hierarchical organization of features (Clements, 1985; Keyser and Stevens, 1994).

Furthermore, we found that most of the sonorant and obstruent phonemes cluster separately and that strident fricatives form a subcluster of the obstruent one. Our findings point to a functional organization based on acoustic cues. First, sonorants are highly resonant and have identifiable formant structure compared to obstruents. Second, stridents have a clear acoustic footprint, characterized by high intensity and high-frequency energy. These findings are in agreement with

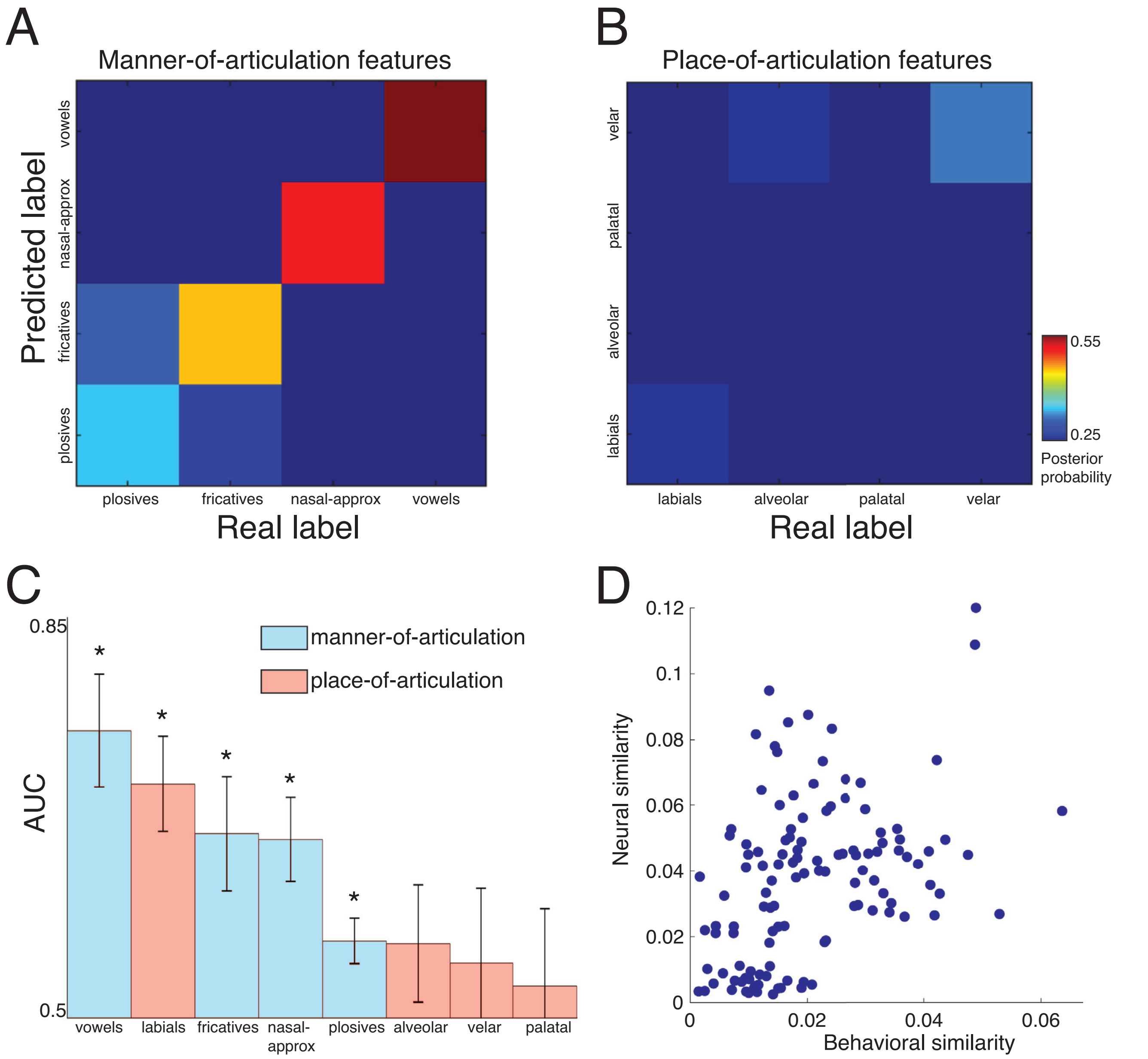

Fig. 5. (A) Confusion matrix among manner-of-articulation features of consonants: (1) plosives; (2) fricatives; (3) nasals and approximants; and (4) vowels. **(B)** Confusion among place-of-articulation features of consonants: labial, alveolar, palatal, and velar(chance level = 0.25). **(C)** AUC values for each binary feature, e.g., [+nasal] vs. [-nasal], [+labial] vs. [-labial]. AUC values were determined from the posterior probabilities of the Naïve-Bayes model and phoneme identities of the test samples; Error-bars are calculated across test sets. **(D)** A comparison between neural and behavioral similarity. Each dot represents a pair of phonemes, X-axis values represent perceptual phoneme similarity, estimated based on confusion rates among phonemes stimuli, which were collected in a behavioral experiment with healthy participants [24]. Yaxis values represent neural similarity from patient data (see Materials and Methods). The Spearman correlation between the behavioral and neural similarities is $\rho = 0.45$ (p<0.001).

Pasley et al. (2012), who showed that speech waveforms can be reconstructed from LFPs in the lateral STG, suggesting that encoded information in this region is mainly acoustic.

To further quantify the representations of different phonemes, we trained a probabilistic classifier, which mimics the generation process of spikes, as recorded by the units. We then compared model predictions when grouping phonemes according to various phonological features. Findings show that the confusion matrix of manner features is more diagonal compared to place-of-articulation features. Similarly, area under the curve for binary classification for each feature resulted with significant prediction for all four manner features whereas only one place-of-articulation feature was above chance. Although our findings are based on recordings from a relatively small number of neurons, and categorical conclusions are therefore limited, taken together, all analyses suggest that spiking activity of few cells encodes phonemes according to manner-of-articulation features, which have acoustic correlates.

Remarkably, spiking activity from this set of neurons reflected perceived similarities derived from behavioral results, based on phoneme-confusion experiments using the same set of stimuli. The distinct neural representation of nasal and approximant features with respect to other feature classes, corresponded to their relatively distinct perceptual saliency. These results suggest that the perceptual representation of phonemes can be observed at the level of single neurons.

The Superior Temporal Gyrus (STG) is one of the central regions for speech perception and language processing in the brain (Geschwind, 1970; Wernicke, 1874). In this region, it is believed that neural processes transform continuous acoustic signals into discrete linguistic code such as phonemes, syllables, words and phrases (Poeppel et al., 2008; Yi et al., 2019). A complete theory of speech perception should thus strive to provide links between the neural mechanism level and the linguistic code. Our study goes one step in this direction. First, it suggests that neural representations derived from microelectrode recordings of single cells reflect a functional organization observed at the *meso* scale in EcoG and EEG studies. This suggests that these different levels both follow an organization based on linguistic features (Jakobson et al., 1951; Mesgarani et al., 2014). Second, the structure of the representations at the neural level reflects those derived from behavioral measurements, suggesting a remarkable consistency across levels of description.

Finally, our results suggest phonological feature distinctions, based on manner-of-articulation, as the appropriate kinds of distinctions at which treatment should be targeted for aphasic patients with deficits in various phonological levels; including the phonological lexicon (ascribed to the STG, Bles and Jansma 2008; de Zubicaray et al. 2002; Indefrey 2011; Graves et al., 2007; Levelt et al., 1998; Wilson et al., 2009), and possibly also deficits in the phonological buffers (Gvion and Friedmann 2012; Friedmann et al., 2013).

Data and code availability statement

The data that support the findings of this study are available on request from the corresponding author YL or co-senior author IF. The data are not publicly available because the data include information that could compromise the privacy of research participants. Codes are available on request from the corresponding author YL or co-first author OO.

Author statement

Conceptualization: YL, OO, NF, RM, IF; Methodology: YL, OO, NF, RM, IF; Data curation: YL, OO, IF; Formal Analysis: YL, OO; Writing – Review and Editing: YL, OO, NF, RM, IF; Funding Acquisition: NF, RM, IF.

Declaration of Competing Interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Acknowledgments

The authors thank the patients for participating in the study. We also thank Gal Chechik for helpful suggestions and comments, and Aviad Albert and Evan-Gary Cohen for help with stimulus generation. We thank M. Tran and G. Kalendar for administrative help and E. Ho, T. Fields, and E. Behnke for technical assistance; This study was supported by the NINDS GRANT RO1 NS084017 to IF, the I-CORE Program of the Planning and Budgeting Committee (Grant No. 51/11), The Israel Science Foundation (Grants No. 1771/13 and 2043/13; R.M.), and The Human Frontiers Science Program (RGP0057/201, Friedmann). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117499.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. 98 (23), 13367–13372.

Arsenault, J.S., Buchsbaum, B.R., 2015. Distributed neural representations of phonological features during speech perception. J. Neurosci. 35 (2), 634–642.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T, 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.

Bles, M., Jansma, B.M., 2008. Phonological processing of ignored distractor pictures, an fMRI investigation. BMC Neurosci. 9, 20.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F, 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327.

Chan, A.M., Dykstra, A.R., Jayaram, V., Leonard, M.K., Travis, K.E., Gygi, B., Baker, J.M., Eskandar, E., Hochberg, L.R., Halgren, E, 2013. Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex 24, 2679–2693.

Cheung, C., Hamilton, L.S., Johnson, K., Chang, E.F, 2016. The auditory representation of speech sounds in human motor cortex. Elife 5, e12577.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York.

Clements, G.N., 1985. The geometry of phonological features. Phonology yearbook 2, 225–252.

Creutzfeldt, O., Ojemann, G., Lettich, E, 1989. Neuronal activity in the human lateral temporal lobe. Exp. Brain Res. 77, 451–475.

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

Desai, R., Liebenthal, E., Waldron, E., Binder, J.R, 2008. Left posterior temporal regions are sensitive to auditory categorization. J. Cogn. Neurosci. 20, 1174–1188.

DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proc. Natl. Acad. Sci. 109, E505–E514.

de Zubicaray, G.I., McMahon, K.L., Eastburn, M.M., Wilson, S.J., 2002. Orthographic/phonological facilitation of naming responses in the picture-word task: an event-related fMRI study using overt vocal responding. Neuroimage 16, 1084–1093.

Donchin, O., Gribova, A., Steinberg, O., Bergman, H., de Oliveira, C.S., Vaadia, E., 2001. Local field potentials related to bimanual movements in the primary and supplementary motor cortices. Exp. Brain Res. 140 (1), 46–55.

Formisano, E., De Martino, F., Bonte, M., Goebel, R, 2008. "Who" is saying" what"? Brain-based decoding of human voice and speech. Science 322, 970–973.

Fried, I., Wilson, C.L., Maidment, N.T., Engel Jr, J., Behnke, E., Fields, T.A., Macdonald, K.A., Morrow, J.W., Ackerson, L, 1999. Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. J. Neurosurg. 91, 697–705.

Friedmann, N., Biran, M., Dotan, D., 2013. Lexical retrieval and breakdown in aphasia and developmental language impairment.. In C. Boeckx & K. K. Grohmann (Eds.), The Cambridge Handbook of Biolinguistics 350–374.

Galantucci, B., Fowler, C.A., Turvey, M.T., 2006. The motor theory of speech perception reviewed. Psychon. Bull. Rev. 13 (3), 361–377.

Geschwind, N., 1970. The organization of language and the brain. Science 170 (3961), 940–944.

Graves, W.W., Grabowski, T.J., Mehta, S., Gordon, J.K., 2007. A neural signature of phonological access: distinguishing the effects of word frequency from familiarity and length in overt picture naming. J. Cogn. Neurosci. 19, 617–631.

Grodzinsky, Y., Nelken, I., 2014. The neural code that makes us human. Science 343, 978–979.

Gvion, A., Friedmann, N., 2012. Phonological short term memory in conduction aphasia. Aphasiology 26 (3–4), 579–614. doi:10.1080/02687038.2011.643759.

Indefrey, P., 2011. The spatial and temporal signatures of word production components: a critical update. Front. Psychol. 2, 1–16.

Jakobson, R., 1968. Child Language, Aphasia and Phonological Universals. Walter de Gruyter Mouton, Oxford, England.

Jakobson R., Fant C.G., Halle M. 1951. Preliminaries to speech analysis: the distinctive features and their correlates.

Keyser J., S., Stevens N., K., 1994. Feature geometry and the vocal tract.. Phonology 11 (2), 207–236.

Khalighinejad, B., da Silva, G.C., Mesgarani, N, 2017. Dynamic encoding of acoustic features in neural responses to continuous speech. J. Neurosci. 2383 -2316.

Lakretz Y., Chechik G., Cohen E.-.G., Treves A., Friedmann N. 2018. Metric learning for phoneme perception. arXiv preprint arXiv:180907824.

Levelt, W.J.M., Praamstra, P., Meyer, A.S., Helenius, P., Salmelin, R., 1998. An MEG study of picture naming. J. Cogn. Neurosci. 10, 553–567.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M, 1967. Perception of the speech code. Psychol. Rev. 74, 431–461.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.

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

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A, 2005. Neural substrates of phonemic perception. Cereb. Cortex 15, 1621–1631.

Liebenthal, E., Desai, R., Ellingson, M.M., Ramachandran, B., Desai, A., Binder, J.R, 2010. Specialization along the left superior temporal sulcus for auditory categorization. Cereb. Cortex 20, 2958–2970.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F, 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338–352.

Möttönen, R., Calvert, G.A., Jääskeläinen, I.P., Matthews, P.M., Thesen, T., Tuomainen, J., Sams, M, 2006. Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. Neuroimage 30, 563–569.

Mukamel, R., Fried, I., 2012. Human intracranial recordings and cognitive neuroscience. Annu. Rev. Psychol. 63, 511–537.

Mukamel, R., Nir, Y., Harel, M., Arieli, A., Malach, R., Fried, I, 2011. Invariance of firing rate and field potential dynamics to stimulus modulation rate in human auditory cortex. Hum. Brain Mapp. 32 (8), 1181–1193.

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A., Brugge, J.F, 2009. Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. 29 (49), 15564–15574.

Ossmy, O., Fried, I., Mukamel, R, 2015. Decoding speech perception from single cell activity in humans. Neuroimage 117, 151–159.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F, 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10, e1001251.

Poeppel, D., Idsardi, W.J., Van Wassenhove, V., 2008. Speech perception at the interface of neurobiology and linguistics. Philos. Trans. R. Soc. B: Biol. Sci. 363 (1493), 1071–1086.

Quiroga, R.Q., Nadasdy, Z., Ben-Shaul, Y, 2004. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. 16, 1661–1687.

Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I, 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102.

Rolls, E.T., Treves, A., 2011. The neuronal encoding of information in the brain. Prog. Neurobiol. 95 (3), 448–490.

Sankaran, N., Swaminathan, J., Micheyl, C., Kalluri, S., Carlile, S., 2018. Tracking the dynamic representation of consonants from auditory periphery to cortex. J. Acoust. Soc. Am. 144 (4), 2462–2472.

Shepard, R.N., 1987. Toward a universal law of generalization for psychological science. Science 237, 1317–1323.

Stevens, K.N., 1972. The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data. Human Communication: A Unified View, New York.

Stevens, K.N., 1989. On the quantal nature of speech. J. Phonet. 17, 3–45.

Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. 111, 1872–1891.

Tversky, A., 1977. Features of similarity. Psychol. Rev. 84, 327.

Venezia, J.H., Thurman, S.M., Richards, V.M., Hickok, G., 2019. Hierarchy of speech–driven spectrotemporal receptive fields in human auditory cortex. NeuroImage: 186, 647–666.

Wernicke, C., 1874. Der Aphasische Symptomencomplex: Eine Psychologische Studie auf Anatomischer Basis. Cohn & Weigert.

Wilson, S.M., Isenberg, A.L., Hickok, G., 2009. Neural correlates of word production stages delineated by parametric modulation of psycholinguistic variables. Hum. Brain Mapp. 30, 3596–3608.

Yi, H.G., Leonard, M.K., Chang, E.F., 2019. The encoding of speech sounds in the superior temporal gyrus. Neuron 102 (6), 1096–1110.