Title

Vowel and formant representation in the human auditory speech cortex.

Journal

Neuron, 111(13)

Authors

Oganian, Yulia
Bhaya-Grossman, Ilina
Johnson, Keith
et al.

Publication Date

2023-07-05

DOI

10.1016/j.neuron.2023.04.004

Peer reviewed

Published in final edited form as:

Neuron. 2023 July 05; 111(13): 2105–2118.e4. doi:10.1016/j.neuron.2023.04.004.

Vowel and formant representation in human auditory speech cortex

Yulia Oganian[1,2,*], **Ilina Bhaya-Grossman**[1,3,*], **Keith Johnson**[4], **Edward F. Chang**[1,#]

[1]Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA

[2]Current address: Center for Integrative Neuroscience, University Medical Center Tuebingen, Ottfried-Mueller-Str. 25, 72076 Tuebingen, Germany

[3]University of California Berkeley, University of California, San Francisco Graduate Program in Bioengineering, Berkeley, CA 94720, USA

[4]Department of Linguistics, University of California, Berkeley

Summary:

Vowels, a fundamental component of human speech across all languages, are cued acoustically by formants, resonance frequencies of the vocal tract shape during speaking. An outstanding question in neurolinguistics is how formants are processed neurally during speech perception. To address this, we collected high-density intracranial recordings from the human speech cortex on the superior temporal gyrus (STG) while participants listened to continuous speech. We found that two-dimensional receptive fields based on the first two formants provided the best characterization of vowel sound representation. Neural activity at single sites was highly selective for zones in this formant space. Furthermore, formant tuning adjusted dynamically for speaker-specific spectral context. Yet, the entire population of formant-encoding sites was required to accurately decode single vowels. Overall, our results reveal that complex acoustic tuning in two-dimensional formant space underlies local vowel representations in STG. As a population code this gives rise to phonological vowel perception.

Graphical Abstract

Corresponding author: Edward F. Chang: edward.chang@ucsf.edu.

*These authors contributed equally.

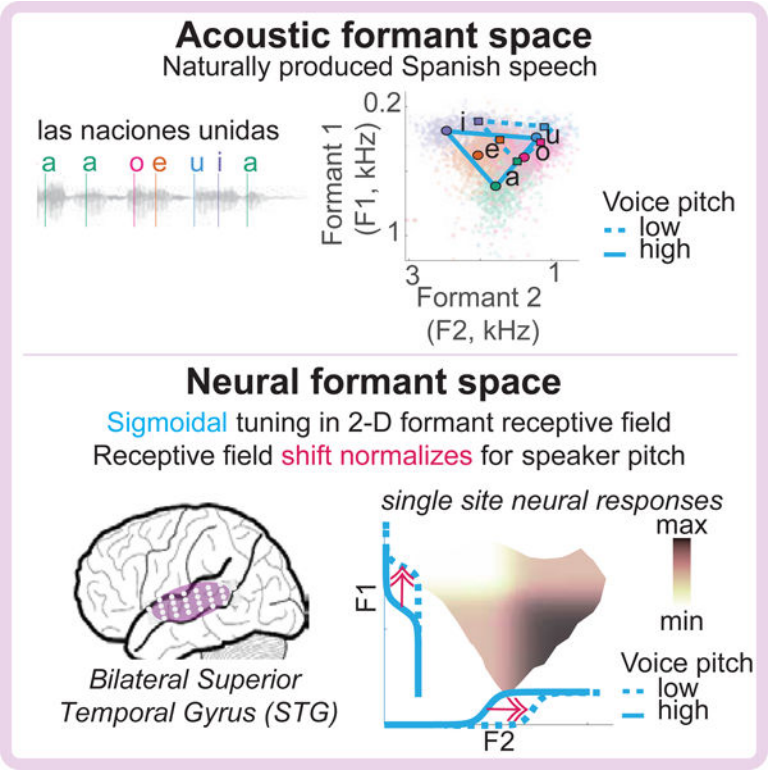
#Lead contact

Author Contributions

Conceptualization, YO, EFC; Methodology, YO, IBG, KJ; Software, YO, IBG; Formal Analysis, YO, IBG; Investigation, YO, IBG, KJ, EFC; Data Curation, YO, IBG; Writing - Original Draft, YO, IBG; Writing - Reviewing and Editing, YO, IBG, KJ, EFC; Visualization, YO, IBG, EFC; Supervision, YO, EFC; Project Administration, YO; Funding Acquisition, EFC.

Declaration of Interests

The authors declare no competing interests.



eTOC burb

Vowels, such as /a/ in had or /u/ in hood, are the centerpieces of speech across languages. Oganian et al. use direct electrophysiological recordings from the human speech cortex to reveal that local two-dimensional spectral receptive fields underlie a distributed representation of vowel categories.

Introduction:

Vowels are a significant component of all the world's languages, and they play a critical role in our ability to comprehend speech [1,2]. Vowel sounds are produced when vocal fold vibration is unobstructed, allowing a clear passage of air through the mouth, shaped by different positions of the jaw, tongue, and lips. For instance, the vowel sounds /i/ (as in "heed") is produced with the tongue close to the front of the mouth, whereas /a/ (as in "had") is produced with the tongue further back. These articulatory positions create distinct vowel sounds, distinguished acoustically by the value of the lowest two vocal tract resonance frequencies, which are known as the first and second formants (F1 and F2). Formants are considered the primary acoustic cues to vowel identity [3].

Sounds with different vowel identities can be very similar in absolute formant values when produced by different speakers. For example, /o/ (as in "hoed") produced by a speaker with a long vocal tract (and thus low voice) could have the same absolute formant values as /u/ (as in "who'd") produced by a speaker with a short vocal tract (and thus high voice). As a result, the accurate mapping of formant values to vowel identity requires a normalization operation that computes formant frequencies relative to the speaker's voice [4]. In sum, a rapid

and correct mapping of formants to vowel identity relies on both the precise identification of the formant frequencies as well as their interpretation given the speaker context.

In linguistics, the mental representation of formants, specifically the independence of F1 and F2 and how this representation supports vowel normalization, has been long debated. One fundamental theory suggests that they are independently extracted and represented, mapping individual vowel sound segments onto coordinates in two-dimensional F1-F2 space [5,6]. To account for speaker context, proponents of this theory have advocated for an explicit normalization on both F1 and F2 values using features such as speaker pitch or higher formants. Alternatively, it has been suggested that vowel identification relies on the *relationship* between formants, such as the one-dimensional distance between F1 and F2 [7]. This theory of vowel perception is bolstered by the fact that the relationship between F1 and F2 is more consistent across speakers than the independent absolute values, thus implicitly accounting for speaker context [8,9]. Examining neural responses to vowel sounds in the human speech cortex provides us with a unique opportunity to dissociate these theoretical models and elucidate the mechanisms that underlie speaker-normalized vowel identification.

Several studies have shown that neural activity in the human auditory cortex is sensitive to vowel sounds and that it discriminates vowel identity in highly controlled acoustic contexts [10–12]. In the primary auditory cortex (PAC), a tonotopically organized area [13] that is activated regardless of whether the presented stimulus is human speech (e.g. pure tones), this sensitivity is driven by narrow tuning to individual formant frequency bands [14,15]. In contrast, in the human speech cortex on the lateral superior temporal gyrus (STG), the majority of neural representations are spectrally complex and broadband [14]. Further, the STG shows stronger activity in response to speech and other complex natural sounds (e.g. music, [16,17] than to other sound stimuli, and this activity is more closely reflective of perceptual processing [18,19]. It thus remains an open question what neural representation in the STG underlies vowel perception in continuous, natural speech [20–22]. Specifically, it is not yet clear whether formants are represented separately or in combination and further, whether this representation is tuned to narrow-band frequencies, possibly centered on single vowel categories [23], or broad formant ranges.

To address this, we utilized high-density direct intracranial recordings of neural activity from the surface of the human STG. The highly resolved spatial scale afforded by this recording technique was critical, as neighboring cortical sites that are just a few millimeters apart can differ significantly in their spectral tuning [24,25]. The high temporal resolution of intracranial recordings allowed us to examine neural responses at the temporal scale of a single vowel sound. We used natural speech stimuli produced by a wide variety of speakers, which allowed us to record neural responses to a large set of vowel sounds, spanning the entire formant space.

With this approach we addressed four primary research questions. First, we asked how neural responses recorded at single electrodes in the STG were tuned to F1 and F2 in natural, continuous speech. We analyzed two-dimensional formant receptive fields of neural responses to determine whether neural tuning to F1 and F2 frequency ranges in human speech cortex is independent, and to characterize the mathematical properties of these tuning

functions [26,27]. Second, we asked how vowel information can be extracted from the neural representation of formants using a population decoding approach. Third, we asked how and to what extent formant receptive fields in the STG are normalized for speaker properties. Finally, we used a controlled set of artificial vowel-like sounds extending beyond the natural vowel formant space with experimentally decorrelated F1 and F2 to definitively test the independence of formant encodings.

We found that neural responses on most single electrode sites in the STG were jointly tuned to both F1 and F2, resulting in heightened sensitivity to a specific zone within the vowel formant space. This sensitivity was non-linear and sigmoidal along each of the separate formant dimensions (F1 and F2). However, the location of heightened sensitivity did not coincide with boundaries between single vowels, and we did not find single electrode sites with selectivity for a single vowel. Rather, single vowels could only be decoded at the population level, when information from differently tuned electrode sites was pooled together. Comparisons between neural responses produced by speakers with different vocal tract lengths showed that electrodes in the STG contain normalized, not absolute, formant representations, distinguishing it from the narrow-band frequency tuning in PAC. Though formant tuning on many active electrodes showed inverse tuning to the two formants (e.g., tuned to high F1 and low F2) when presented with natural speech, decorrelating the formants using a set of artificial vowels revealed a set of electrodes with the same direction of tuning to both formants (e.g., high F1 and high F2). This confirms that there exists a range of formant encoding types in the STG, and that F1 and F2 are neurally represented as coordinates in a two-dimensional formant space rather than as a ratio or distance.

Results:

Cortical activity at single electrodes over human STG is sensitive to vowel differences

In Experiment 1, Spanish monolingual patients (n=8, Table S1) listened to naturally produced Spanish sentences (Fig. 1A), while we recorded neural activity from the lateral surface of the STG using high-density ECoG electrode grids. The Spanish vowel system is well suited to study vowel representation for multiple reasons: Spanish has only 5 vowel sounds (as opposed to e.g., up to 20 in some English dialects, [28]), that span a large range of formant values. Thus, Spanish vowel categories clearly easily separated in the acoustic formant space: The median formant values for single vowel instances correspond strongly with their vowel label (Fig. 1B and C vowel clustering: median silhouette score = 0.099, permutation test with 500 repetitions; $p<0.002$). Moreover, while F1 and F2 tend to be inversely correlated across languages, this is less the case for Spanish than for English (Spanish $r = -0.01$, $p = 0.31$ in our stimulus set; English $r = -0.21$, $p<0.0001$ calculated from speech stimuli used in prior studies, e.g., [29].

In our analyses we focused on evoked responses in the high gamma range (HGA). We found that evoked HGA responses on a subset of STG electrodes discriminated between vowel categories (n = 125 of 291 speech responsive, range: 2–26 per participant, one-way peak F-statistic across vowel categories > 5, Fig. 1E). Responses peaked at about 100–150 ms post vowel onset, with different response magnitudes corresponding to different vowels. For example, each of the three prototypical example electrodes E1-E3 (Fig. 1D) exhibited

strongest responses for a different vowel, with graded response magnitudes to all other vowels.

Notably, none of these electrode responses show a preference for a single vowel. That is, we did not observe instances where there was selectivity to a single vowel and no response to all other vowel sounds. All subsequent analyses included electrodes that discriminated between vowels (colored green in Fig. S1A).

Non-linear monotonic tuning to vowel formant frequencies in human STG

We first asked how vowel formants are represented by vowel-discriminating populations. Specifically, we evaluated three alternative hypotheses: Narrow-band nonlinear frequency tuning centered on vowel categories (Fig. 2A left), linear monotonic encoding of an entire vowel formant range (Fig. 2A, middle panel), or nonlinear monotonic encoding of formants within a limited dynamic range (Fig. 2A, right panel). In the case of narrow-band frequency tuning, we expect neural populations to preferentially respond to a narrow range of formant frequencies, as is typical for frequency-tuning in primary auditory cortices [14,30]. This model implies that the maximal neural response could be located in the center of the vowel's formant range. In contrast, in case of monotonic formant encoding, we expect neural responses to increase across the range of possible formant values, with the maximal response located at the edges of the formant range. While linear encoding implies equal sensitivity to formant differences across the entire range, nonlinear encoding would result in heightened sensitivity to a narrower range of formant frequencies, and little to no sensitivity to frequencies outside of this range.

We tested which of these models captured neural activity best separately for every single electrode. First, to distinguish between narrow non-monotonic formant tuning and monotonic encoding, we examined whether neural responses peaked at the edges or in the center of the formant frequency range (Fig. 2F). Then, to discriminate between linear and non-linear encoding, we compared linear and sigmoidal models of neural responses using cross-validated R^2. We estimated the neural response to vowel formants using feature temporal receptive field modeling, F-TRF [31,32]). Model features of interest were the spectro-temporal content of the speech signal in the formant frequency ranges. The model produced a regression weight time series (beta weights) for each frequency bin. For the main analysis, we extracted mean beta weights in a 50ms window around the peak in the beta weight time course (125 – 175 ms) and fit formant encoding models to these values.

For the representative electrodes E2 (Fig. 2B–C) and E3 (Fig. 2D–E), we found that responses were strongest towards one end of the vowel formant space, suggesting monotonic encoding of formant frequencies. On E2, response magnitudes and model beta weights increased with increases in F1 but decreased with increases in F2. In contrast, in electrode E3, beta weights were strongest for high F1 and low F2 values.

Across all vowel-discriminating electrodes, we found that neural responses peaked near the boundaries of the vowel formant space, reflected in the bimodal distribution of maximal beta values for both formants (Fig. 2F, mixed-effects F1: beta = −0.21, SE = 0.025, t(96) = −8.64, $p<0.001$; F2: beta = −0.99, SE = 0.082, t(96) = −12.1, $p<0.001$). We thus focused

our attention on the comparison between linear and sigmoidal encoding of vowel formants. Across electrodes, we found that cross-validated R^2 values were higher for the sigmoidal model than the linear model on 82.4% of electrodes for F1 and 81.5% of electrodes for F2. We found a mixture of preference for high and low formant values in both F1 and F2. This shows that STG neural populations have limited dynamic ranges: Each local population represents a subspace of the vowel formant space, such that a representation of the entire range emerges across the entire population.

Finally, we wanted to characterize the extent to which both formants are jointly encoded at a single electrode. Across electrodes, we found an inverse correlation between tuning to F1 and F2 (Fig. 2H, mixed-effects beta = −0.61, SE = 0.10, $t(103) = -5.90$, $p<0.001$, see Fig S2 for anatomical location of electrodes). That is, electrodes with preference for high F1 values also preferred low F2 values, and vice versa, as previously found for an English language dataset [29]. Here, we found that this trend was driven by two main patterns. First, most electrodes jointly encoded both formants (67 out of 80 electrodes), with a preference for negative tuning in F1 and positive tuning in F2 (41 out of 80 electrodes). In contrast, while a small subset of electrodes (n=10) had negative tuning to both formants, only a single electrode in our dataset had positive tuning to both formants. This raises the question: Does the lack of electrode sites with positive tuning to both formants reflect a specialization of STG for speech sound harmonics? Alternatively, it might be a confound of the limited formant frequency ranges in natural speech. We will address this question with Experiment 2 below.

Overall, these results show that neural activity at single electrode sites is only discriminative in a subspace of the overall vowel formant space. However, across electrodes, the range of formant tuning at the population level should be sufficient to represent the entire vowel space with a high fidelity. We directly test this in the next analysis step using population decoding.

Discrimination between vowel categories emerges at the population level

We evaluated whether single vowel categories are represented at the population level by comparing the accuracy of vowel decoding on different electrode subsets using a linear SVM approach. First, we compared classifier accuracies derived from single electrodes versus from the entire electrode population (Fig. 3A). We found that decoding from the entire electrode population was significantly more accurate when considering all pairwise comparisons (average improvement: 2.2 – 9.1% correct averaged across pairwise comparisons, t-test showed significant difference between all single electrodes versus the entire population accuracies with $p<0.0001$, Fig. 3A). Notably, the best single electrodes did not necessarily show selective responses to a single vowel (Fig. 3A right panel).

The improvement in classification accuracy from single electrode to the population could be due to an increase in the signal-to-noise ratio. However, this improvement may also reflect complementary vowel encoding. That is, single electrodes represent only select regions of the vowel formant range in high detail, and as a result, pooling information across electrodes with sensitivity to different select regions may be key to decoding vowel categories across all pairwise comparisons. To determine whether this was the case, we split electrodes into

the two dominant tuning subsets, namely (F1−/F2+) and (F1+/F2−) type-electrodes. Formant receptive fields averaged across electrodes with the same directionality of F1/F2 tuning (F1−/F2+: n = 36 electrodes; F1+/F2−: n = 15 electrodes, Fig. 3C) suggested that each set would be critical for a subset of comparisons. For instance, average receptive fields suggested that F1+/F2− populations would be able to discriminate between /a/, /e/, and /i/, whereas this would not be the case for the F1−/F2+ population.

Decoding performed separately on the two subsets of electrodes confirmed this prediction, as can be seen in Fig. 3D for five exemplary comparisons ($p<0.001$ for significant comparisons between decoding accuracies on different subsets, see Table S4 for details of pairwise comparisons). Moreover, this analysis clearly demonstrates that increased decoding accuracies at the population level reflect the addition of informational content rather than increases in signal-to-noise ratio. This is because accuracy of decoders based on the best electrode subset and both electrode subsets together did not significantly differ. Finally, Fig. 3D summarizes the accuracies of all pairwise comparisons, showing that the two subsets of electrodes discriminate different sets of vowel pairs. The F1−/F2+ electrodes show significantly higher accuracy for the /i/ − /o/ and /o/ − /u/ pairwise classification, while the F1+/F2− electrodes show higher accuracy for the /i/ − /a/, /i/ − /u/, and /e/ − /a/ classification. However, in conjunction, these two electrode sets contain the complementary information necessary to significantly discriminate between all vowel pairs.

Shifting dynamic ranges underlie normalization of vowel representation for speaker vocal tract length

It is well established that the mapping between vowel formants and vowel categories depends on a speaker's vocal tract length which is correlated with voice pitch [4,33]. We thus wanted to determine whether the formant tuning of neural populations in STG shifts with speaker voice quality during listening to continuous speech.

In line with prior work, we found that vowel formant frequencies increase with speaker pitch (Fig. 4A) and that normalizing for speaker pitch reduces the formant variance within vowel categories (Fig. 4B). We hypothesized that STG encoding of formants would reflect speaker-normalized rather than absolute formant frequencies. To test how speaker pitch affects the representation of vowel formants in human STG, we re-fit feature temporal receptive field models separately on subsets of data with low and high (>170 Hz) pitch levels, and estimate the sigmoidal fits to F1 and F2 model weights separately for each of the models. If STG representation of vowel formants reflects absolute formant frequencies, we expect to see no difference in formant tuning between models (Fig. 4C left). In contrast, if STG representations are normalized for speaker properties, we expect to see a shift in STG dynamic ranges, matching the shift in vowel formant space between single models (Fig. 4C right).

Fig. 4D–E shows tuning curves for low and high pitch speakers on a single example electrode. For this exemplary electrode, neural responses shifted their dynamic ranges to higher frequencies in response to vowels produced by speakers with a high pitch, in line with speaker normalization for both F1 and F2. To quantify the magnitude and extent of this shift on STG across all electrodes with significant formant frequency tuning (n = 105),

we focused on electrodes with robust frequency tuning curves for vowels produced by both high and low pitch speakers (F1 n = 25, F2 n = 45 (unique electrodes n = 68), 64.8% of electrodes). We extracted the sigmoidal fit inflection Points (IP) for each subset of speakers separately and assessed the difference in IP between models across electrode sites using linear mixed effects modeling (see methods for model details).

We found a systematic and robust shift of tuning curve inflection points (IP) on these electrodes towards higher formant ranges with increases in speaker pitch, with a most robust shift present on electrodes with overall good model fits for single formants (mixed-effects F1: pitch beta = 35.38, SE = 9.46, $t(111) = 3.7$, $p < 0.001$; F2: pitch beta = 95.03, SE = 31.7, $t(138) = 2.99$ $p = 0.003$; see Tables S2 and S3 for all fixed effects, Figure 4F, G). Remarkably, the average magnitude of the inflection point shift across electrodes mirrored the difference in average formant frequency between high and low pitch speakers (red lines in Fig. 4F, G). Finally, we also found that all electrodes with robust encoding of F1 and F2 also showed speaker normalization for both formants, Fig. 4H. Taken together, these analyses show that formant normalization for speaker voice characteristics is a general feature of formant encoding neural populations in the human STG.

Vowel formant encoding emerges from general complex frequency tuning on STG

Our analysis of the representation of vowel formants in STG raised two questions. First, are the opposite directions of preference for F1 and F2 due to the limited range and covariance between F1/F2 in natural speech? That is, when presented with complex harmonic sounds with a broader range of F1 and F2 values, will neural populations show sensitivity to formant values with the same direction of tuning? Second, does each formant value affect the neural responses independently (Fig. 5A top and middle row, joint independent encoding), or are co-encoding neural populations additionally integrating across F1 and F2, i.e. do responses to each formant depend on the value of the other formant (Fig. 5A bottom, joint interactive encoding)? Notably, the qualitative patterns differ between independent and interactive co-encoding: the latter shows a u-shaped tuning along one of the diagonals. Finally, we wanted to know whether STG encoding of the formant structure in sounds would continue outside the boundaries of the human vowel formant space. Alternatively, STG may contain separate neural populations encoding sound harmonic structure in speech and non-speech frequency ranges.

To test these questions, we presented a group of ECoG patients (n=8) with a set of isolated artificial vowel-like sounds with F1 and F2 values in and outside the natural vowel space (Black outline in Fig 5B, see Fig S3 for perceptual rating of stimuli). We report all results across English and Spanish native speakers; but note that no systematic differences between native speakers of the two languages were found. Notably, this task was language independent, as vowels across languages fall within the same space due to physical constraints on vowel production. We chose to use isolated harmonic sounds rather than consonant-vowel combinations to reduce any effects of coarticulation and to ensure that tokens outside the vowel formant range could be perceived as non-linguistic. We found robust evoked HGA responses to single stimulus tokens on a subset of STG electrodes, which stereotypically peaked 150–200 ms after stimulus onset, in line with responses to

natural speech. We first tested whether electrodes responded differently to tokens within and outside the natural formant range (black vs gray in Fig. 5B). Then, to model the encoding of F1 and F2, we extracted peak HGA for each stimulus token and evaluated the effects of F1, F2, and their linear interaction (IA, F1*F2) on peak HGA magnitude. Across all 8 patients, analyses were focused on 160 electrodes (10–33 per patient), for which the best linear regression model explained at least 10% of total response variance.

Fig. 5C–E shows patterns of neural responses for three exemplary electrodes. Electrode E1 responded equally to tokens inside and outside the natural formant range (Fig. 5C left), with the strongest responses to high F2 values (Fig. 5D left), and no effect of F1 or the interaction term (Fig. 5D bottom). In contrast, responses on electrode E2 were highest for the combination of low F1 and high F2 values, as supported by the significant interaction term on this electrode (Fig. 5D middle). Finally, electrode E3 responded stronger to tokens outside the vowel formant range (Fig. 5C right), which was due to a significant positive interaction effect (Fig. 5D right). On all three electrodes the pattern of responses is generally smooth around the vowel space boundaries, suggesting that any difference between responses inside and outside this space are due to spectral and not speech tuning. Crucially, we found a high overlap between formant receptive fields derived from the synthetic vowel tokens and from natural speech stimuli, suggesting that both stimuli drive neural responses on STG to the same degree and in a comparable manner (Fig. 5E).

Comparison between formant encoding in synthetic vowel sounds and in natural speech

Across all electrodes, we found that the main effects of F1 and F2 explained the most unique variance on single electrodes (F1: R^2 median=0.07, max=0.71; F2: R^2 median=0.07, max=0.72), with a minor but significant contribution of interaction terms (F1*F2: R^2 median=0.03, max=0.32; Fig. 6A). Notably, effect magnitudes for F1 and F2 were not correlated ($r = -0.1$, $p=0.2$, mixed-effects model $t(158) = -1.48$,n.s., Fig. 6B), suggesting that each contributes independently to neural responses. In contrast, main and interaction effect magnitudes were negatively correlated ($r=-0.52$, $p<0.001$, mixed-effects model $t(158) = -6.4$, $p<.001$, Fig. 6C). That is, electrodes with large interaction effects had little independent contribution of F1 and F2 main effects and low R^2 overall. This suggests that encoding of F1 and F2 and integration across formants are implemented by distinct STG populations with independent joint encoding of F1 and F2 dominating neural response patterns.

In a second step, we asked whether STG representation of vowel formants is tailored to formant ranges found in natural speech. Unlike in natural speech, in our synthetic vowel stimuli we found only a weak negative correlation between F1 and F2 model weights ($r = -0.2$, $p=0.01$, mixed-effects model: $t(158) = -3.31$, $p=.001$, Fig. 6D). Importantly, the range of tuning to different combinations of F1 and F2 values (N: +/+ 36, −/− 44, +/− 28, −/+ 52 tuning direction) in our data speaks in favor of independent co-encoding of F1 and F2, rather than differential joint encoding.

We hypothesized that this discrepancy was due to the naturally limited range of F1 and F2 values in natural speech. To test this, we reran our analyses on a subset of stimuli with formants falling within the natural formant frequency range (as marked in Fig. 5B). Figure

6F shows the model R^2 for full and subset models by electrode formant tuning. We found that in addition to the expected overall lower model R^2 with a subset of stimuli (main effect of model: b = −0.21, SE = 0.04, t(316) = −5.94, p<.001) and marginally higher R^2 values in electrodes with opposite tuning to F1 and F2 (main effect of tuning directions: b = 0.16, SE = 0.08, t(316) = 2.02, p=0.04), this reduction was more pronounced for electrodes with the same direction of tuning for both formants (interaction of model and tuning direction: b = −0.19, SE = 0.05, t(316) = −3.67, p<.001, Fig. 6G). Overall, this shows that populations with the same directionality of tuning for both formants are not as strongly activated by vowels but rather tuned to other harmonic sounds.

Discussion

This study provides a comprehensive account of the representation of vowels in the human speech cortex on lateral superior temporal gyrus (STG). We found that vowel responses at local electrode sites in the STG were best characterized by complex two-dimensional receptive fields, defined by the first two formant frequencies, and normalized for speaker voice characteristics. Along each formant range (or receptive field axis), electrode sites showed nonlinear, monotonic frequency tuning, with high sensitivity to a specific zone in the natural formant space. Because the spectral location of this zone often did not correspond to any single vowel, discrimination between vowel categories at single electrode sites was unreliable. However, when neural population response patterns were aggregated across electrode sites, vowel categories could be decoded with high accuracy. Finally, neural responses to artificial vowel-like sounds with experimentally de-correlated F1 and F2 values showed that the complex tuning to formants in the STG independently represents F1 and F2, and extends beyond the natural vowel formant range.

The primary objective of this study was to use neural data to adjudicate linguistically informed theoretical models of vowel representation. Two fundamental models of vowel representation have been proposed, one in which the mental representation of vowels is described by a one-dimensions relationship between F1 and F2 (F1-F2) [7] and another in which it is described by a two-dimensional coordinate in (F1, F2) space [5,6]. While the neural responses to vowel sounds in this and earlier studies [29,34,35] seem to support the former theory, the results from our synthesized vowel experiment suggest that this interpretation is misleading due to the limited range of formants in natural speech. By presenting subjects with formant combinations that extended beyond the range found in natural speech, we show that though a majority of electrode sites inversely encode F1 and F2 (high F1 values and low F2 values, or vice versa), there exist electrode sites with other joint encoding patterns (e.g., high F1 values and high F2 values) or encoding to only a single formant. The fact that the encoding of formants on individual electrode sites spans all possible tuning combinations suggests that neural populations on the human speech cortex represent F1 and F2 as two distinct dimensions of the vowel space, rather than a relationship between them [7,36].

Our second objective was to determine how the neural encoding of formants gives rise to representation of vowel information and speculate as to why the encoding of formants should organize in this way [20,21]. Although prior studies found categorical vowel representations in auditory cortex [12,37], our decoding analysis revealed no evidence for the

categorical representation of vowels at local electrode sites on the STG. That is, electrode sites responded strongly to subareas of the two-dimensional formant space, but these zones were not centered on single vowel categories. However, population-level decoding, where neural responses at single electrode sites were aggregated, allowed for the robust decoding of vowel categories. This is in line with an accumulating number of ECoG studies on the neural representation of consonants that also reported that specific discrete phonemes can only be reliably decoded at the population level [24,25,29].

Our decoding results add to the existing evidence that a distributed representation of vowel sounds on the lateral STG is organized as a heterogeneous spatial code [10,11,18,19,38]. Electrode sites that jointly encoded F1 and F2 belonged to two spatially interspersed encoding types: those with the strongest neural responses to high F1 and low F2 and vice versa. Together, the two encoding types represented the complementary formant information necessary to decode vowel categories. Since a majority of electrode sites were nonlinearly tuned to F1 and F2 (Fig. 2) resulting in maximal neural responses at the edges of the formant range, formant encoding on the STG is also likely the basis for nonlinear, categorical vowel perception and perceptual magnet effects [39–41].

The third objective of this study was to determine the degree to which speaker normalization takes place on the lateral STG. Building on prior work that used active, isolated word-level tasks [33,42], we found that local electrode sites on the STG represented vowel formant frequencies normalized for the speaker even when subjects passively listened to natural speech. These results strongly support the notion that neural representations of speech on the lateral STG are context-sensitive and contradict recent findings suggesting that these representations reflect absolute frequency content of harmonic sounds [43]. This discrepancy may be in part due to the different spatial resolutions of ECoG and EEG scalp recordings [44].

To perform speaker normalization in natural speech, the auditory system can draw on several distinct and often co-occurring acoustic cues [45], such as the distributions of F1 and F2 for the speaker [42], the ratio between F1 and F3 [36], and the average F0 of the speaker (e.g. [46]). We did not experimentally isolate these cues and thus are unable to make a definitive argument regarding the cue used by neural populations on the STG to initiate speaker normalization. However, our results did show that the degree of normalization on single neural populations matched the distance between the average formants of speakers with low and high pitch, differing from previous ECoG studies that report only partial normalization [42]. It is possible that the larger magnitude normalization effects observed in natural speech are due to the presence of co-occurring acoustic cues versus only speaker F1 values that were manipulated in the prior study [42].

We can briefly speculate about the mechanisms that may account for the observed normalization effect [4]. First, these effects may be explained by the general auditory mechanisms that give rise to adaptation and contrast-enhancing sensory representations in both speech and non-speech contexts [47,48]; examples of such mechanisms include stimulus specific adaptation and gain control [4,49,50] or critical band behavior sensitive to the density of harmonics [51]. It is also possible that speaker normalization reflects an integration of spatially interspersed but functionally distinct cortical areas encoding talker identity and

might involve other parts of the temporal cortex, such as the superior temporal sulcus [10,11,52,53]. We note that the possibilities listed here as mechanisms for normalization processes are not mutually exclusive and may both play a role in causing the effects we observe.

Finally, we were interested in understanding to what extent STG representation of complex harmonic sounds is specialized for human vowels. While it is well established that vowel representation in PAC relies on narrowband frequency tuning of tonotopically organized neural populations that are not specifically tuned to human speech [54–56], the vowel representation in non-primary auditory areas has not been fully explored. Here, we show that vowel-discriminating neural populations in the STG encode F1 and F2, with nonlinear, sigmoidal tuning along each separate formant dimension. Notably, our artificial vowel data showed that such representations also support the encoding of other harmonic sounds in the environment, namely, formant combinations outside those found in naturally produced vowels. The specialization for different harmonic ranges may support discrimination of non-speech complex harmonic sounds and possibly underlies recent findings of distinct neural populations for speech and music in this area [16,57].

Importantly, this study relied on two passive listening paradigms, one in which subjects listened to natural speech and one in which subjects listened to synthetic sound stimuli. We thus cannot exclude the possibility that task-dependent effects modulated the recorded neural responses. It is known that when subjects attend to and comprehend natural, meaningful speech, speech-relevant receptive fields are enhanced [58] and further, that task demands and complexity critically alter connectivity patterns across speech and language cortical networks [59,60]. Task-specific enhancement effects may reflect top-down inputs from prefrontal areas [61] via task-dependent oscillatory phase alignment [62] or selective enhancement of receptive fields for task-relevance (e.g. [63,64]). However, prior work comparing data from active and passive listening paradigms found qualitatively similar fMRI responses across task conditions in PAC and STG [65] and comparable receptive fields in ECoG [24], suggesting that the effect of an active versus a passive task paradigm is more likely to fine tune existing representations rather than alter them completely.

The current study on the neural representation of vowel formants in the human speech cortex leaves several important questions open. First, we focused analysis on discrete vowel categories and static formant values, and as a result, did not explore the effects of vowel duration or formant temporal dynamics on neural activity in the STG. [66–70]. We believe that this study lays the groundwork for the future exploration of the neural encoding of such complex formant dynamics. Second, we did not explicitly address the extent to which language experience affects neural vowel representation in the current work. Language experience influences vowel recognition [71–73] and a targeted paradigm is needed to address this. Finally, unlike past findings of functional asymmetries in speech processing across hemispheres [74–77], but in line with prior ECoG studies [29,78], we did not observe hemispheric differences with respect to any of our findings. This indicates at least some level of bilateral involvement in vowel processing. However, alternative neuroimaging modalities with bilateral coverage in single subjects are better suited to make claims about hemispheric asymmetries.

In conclusion, the results of this study demonstrate the broad and complex tuning of local electrode sites on the lateral STG to the formants in natural vowel sounds. This tuning is best described by nonlinear, two-dimensional formant receptive fields that adapt to speaker voice. While most local electrode sites jointly encode the first two formants, without a strong preference for a single vowel, vowel categories can be extracted from the neural responses if aggregated across the population. Overall, we provide a comprehensive account of the representation of vowels in non-tonotopic areas of the auditory parabelt instrumental in the sensory processing of speech.

STAR Methods

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Edward F. Chang (edward.chang@ucsf.edu).

Materials availability—This study did not generate new unique reagents.

Data and code availability

- The data that support the findings of this study are available on request from the lead contact. The data are not publicly available because they could compromise research participant privacy and consent.
- All original code and summary data for figure replication has been deposited on zenodo.org and will be publicly available as of the date of publication. Details are listed in the Key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The study was approved by the University of California, San Francisco Committee on Human Research and all participants gave informed written consent before experimental testing. Fifteen (7 female) patients were implanted with 256-channel, 4-mm electrode distance, subdural ECoG grids as part of their treatment for intractable epilepsy. Electrode grids were placed over the peri-Sylvian region of one of the patients' hemispheres, as determined by clinical assessment. Eight Spanish-native speakers (5 male, 4 LH) with little to no knowledge of English participated in the DIMEx corpus speech experiment. Eight participants (2 Spanish-native; 6 English-native, 7LH) listened to the synthesized vowel tokens of Experiment 2 (SOM Table S1). Participants in the synthesized vowel experiment also listened to the DIMEx corpus. Two Spanish-native participants took part in both experiments.

All participants had normal hearing and left-dominant language functions.

Participant demographics and further implantation and resection details can be found in Table S1.

METHOD DETAILS

Data acquisition methods closely follow those reported in our previous work [14,80].

Neural data acquisition—ECoG signals were recorded with a multichannel PZ2 amplifier, connected to an RZ2 digital signal acquisition system [TuckerDavis Technologies (TDT), Alachua, FL, USA], with a sampling rate of 3052 Hz. The audio stimulus was split from the output of the presentation computer and recorded in the TDT circuit, time-aligned with the ECoG signal. In addition, the audio stimulus was recorded with a microphone and also input to the RZ2. Data were online referenced in the amplifier without any further re-referencing.

Data preprocessing—All data analyses were based on the analytic amplitude of neural responses in the high gamma range (HGA; 70 to 150 Hz), which is closely related to local neuronal firing and tracks neural activity at the temporal scales of natural speech [44,81]. Offline preprocessing of the data included (in this order) downsampling to 400 Hz, notch-filtering of line noise at 60, 120, and 180 Hz, extraction of the analytic amplitude in the high-gamma frequency range (70 to 150 Hz, HGA), exclusion of bad channels, and exclusion of bad time intervals.

HGA was extracted using eight band-pass filters [Gaussian filters, logarithmically increasing center frequencies (70 to 150 Hz) with semi-logarithmically increasing bandwidths] with the Hilbert transform. The high-gamma amplitude was calculated as the first principal component of the signal in each electrode across all eight high-gamma bands, using principal components analysis. Bad channels were defined by visual inspection as channels with excessive noise. Bad time points were defined as time points with noise activity in HG band, which typically stemmed from movement artifacts, interictal spiking, or non-physiological noise.

Last, the HGA was downsampled to 100 Hz, and z-scored relative to the mean and SD of the data within each experimental block. All further analyses were based on the resulting time series.

Electrode localization—For anatomical localization, electrode positions were extracted from postimplantation computer tomography scans, coregistered to the patients' structural magnetic resonance imaging and superimposed on three-dimensional reconstructions of the patients' cortical surfaces using a custom-written imaging pipeline [82]. Freesurfer was used to create a 3d model of the individual subjects' pial surfaces, run automatic parcellation to get individual anatomical labels, and warp the individual subject surfaces into the cvs_avg35_inMNI152 average template.

Experiment 1: Continuous speech (DIMEx)

Stimuli and procedure

Participants passively listened to a selection of 500 Spanish sentences from the DIMEx corpus [79,83], spoken by a variety of native Mexican-Spanish speakers. Data in this task were recorded in five blocks of approximately 7-min duration each. Four blocks contained distinct

sentences, and one block contained 10 repetitions of 10 sentences. Sentences were 2.5 to 8.03 s long and presented with an intertrial interval of 800 ms. The repeated block was used for validation of temporal receptive field models (TRF; see details below).

All stimuli were presented at a comfortable ambient loudness (~70 dB) through free-field speakers (Logitech) placed approximately 80 cm in front of the patients' head using custom-written MATLAB R2016b (MathWorks, www.mathworks.com) scripts. Speech stimuli were sampled at 16000 Hz for presentation in the experiment. Participants were asked to listen to the stimuli attentively and were free to keep their eyes open or closed during the stimulus presentation.

Stimulus spectrograms were calculated using the NSL toolbox (http://nsl.isr.umd.edu/downloads.html) for Matlab. Continuous formant values were extracted using the praat software [84], https://www.fon.hum.uva.nl/praat/,). We found that median formant values discriminate between vowel categories with a high accuracy. Thus all depictions of vowel tokens in two dimensional formant space reflect the token's median formant values.

Electrode selection

Analyses included electrodes located on the STG, for which the per electrode peak HGA response (over a contiguous window of 50 ms) after vowel onset significantly discriminated between vowel categories (using a one-way F-test and a non-corrected threshold of $p < 0.001$). Final analyses included 122 electrodes, 2 to 26 within single patients (median = 12). Selected electrodes were equally distributed across hemispheres, with no hemispheric differences in vowel discriminability or electrode location along the anterior-posterior axis of the STG (Fig. S2). For each of the below analysis, a subset of electrodes from this set were selected based on relevant criteria.

Feature temporal receptive field analysis (F-TRF)

We fit neural data with a linear temporal receptive field (F-TRF) model with different sets of speech features as predictors. In this model, the neural response at each time point [HGA(t)] is modeled as a weighted linear combination of features (f) of the acoustic stimulus (X) in a window of 600 ms before that time point, resulting in a set of model coefficients, $b_{1...,d}$ for each feature f, with d = 60 for a sampling frequency of 100 Hz and inclusion of features from a 600 ms window (See previous work, [29].

$$HGA(t) = \sum_{k=1}^{d} \sum_{f=1}^{F} b(k, f) X(f, t-k)$$

The models were estimated separately for each electrode, using five-fold cross-validation (80% train, 20% test). The regularization parameter was estimated using a 10-way bootstrap procedure on the training dataset for each electrode separately. Then, a final value was chosen as the average of optimal values across all electrodes for each patient. For all models, predictors and dependent variables were z-scored and scaled to between −1 and 1 before entering the model. This approach ensured that all estimated beta values were scale free and

could be directly compared across predictors, with beta magnitude interpreted as an index for the contribution of a predictor to model performance.

Predictors in the model included spectral ranges that spanned the 5–95th percentile of frequency values for the first two formants for all vowels (F1: 250 – 800 Hz, F2: 1000 – 2500 Hz), as well as sentence onsets, vowel onsets, and predictors for timing and magnitude of peakRate, a marker of rising envelope edge dynamics (for peakRate feature structure see Oganian & Chang, 2019). Vowel and sentence onset predictors were timed to onsets of the respective phonemes in the speech.

Linear and sigmoidal model comparisons on F-TRF betas

Linear and sigmoidal curves were fit to the mean beta weights around the beta peak in a 600 ms window (125–175ms), as estimated by the F-TRF model described above. Curve fitting for all model types was leave-one-out cross-validated (16 and 26 times for F1 and F2 bins respectively). This allowed us to directly compare models with different numbers of free parameters. Corresponding model R^2 values were calculated based on the average error derived from the cross-validated predictions. R^2 values were calculated as $1 - RSS/SST$, where RSS is defined as the residual sum of squares and SST is defined as the total sum of squares. Note, this coefficient of determination measure (R^2) could be negative if the regression predictions are further from the true value than a model that predicts the sample average. Bayesian Information Criterion (BIC) was used to validate the R^2 values. According to this metric, the sigmoidal model outperformed the linear model on 31.4 % of electrodes for F1 and 25.7% of electrodes for F2. To test for bimodality, effects of tuning direction on the frequency value of maximal beta weight were assessed using linear mixed effects modeling with fixed effects of tuning direction, and random intercepts and slopes for subject and electrode (Maximal Beta Position ~Tuning Direction + (1 | Subject) + (1 | Electrode:Subject)).

Vowel decoding

Binary SVM linear classifiers (using pre-built function from Matlab Statistics and Machine Learning Toolboxes) were trained on neural data to distinguish pairs of vowel (all possible pairs of the five Spanish vowel categories, /a/, /e/, /i/, /o/, /u/) from a fixed window of 50 ms around the time point of peak discriminability between vowels (centered at approximately 150 ms post-vowel onset). Four types of classifiers were constructed: population level classifiers, two sets of population-subset classifiers, and single electrode classifiers. Population level classifiers were trained and tested on the output of principal component analysis (PCA) applied to neural activity from a population of electrodes (n = 54) spanning 4 subjects with sufficiently overlapping stimulus sets. To ensure that the pairwise decoding accuracy was comparable across pairs, data were subset to contain equal numbers of samples per vowel, equal to the number of samples for the least frequent vowel (/u/) resulting in approximately 110 samples per vowel. Population-subset classifiers were also trained and tested on the output of a PCA and included electrodes with either F1+/F2− (n = 40) or F1−/F2+ (n = 14) individual tuning profiles derived from previous analysis. Single electrode classifiers were trained and tested on the neural activity recorded at single electrodes (n=105). Each classifier was 5-fold cross-validated and reported accuracy

measures were averaged across each of the cross-validation sets. Significance testing was based on classification accuracies for data with permuted vowel labels, based on 50 permutations.

Speaker normalization

To determine the extent to which cortical responses to formants depend on speaker physiology, separate F-TRF models were were fit to neural data using subsets of speakers with an average fundamental frequency across the sentence of either less than or greater than 170 Hz, using the same predictors as the model described above. A total of 233 sentences (4840 vowel instances) were used in the low pitch speaker model and 267 sentences (5750 vowel instances) were used in the model corresponding to high pitch speakers. Curve fitting was performed on the speaker subset F-TRF model beta weights in the same way as described above. Inflection point shifts were calculated using the parameters derived from each of the F-TRF model types.

Effects of speaker subset on inflection point shift was determined using linear mixed effects modeling (using the Matlab Statistics and Machine Learning Toolbox) with fixed effects of model R^2, speaker subset, and their interaction, and random intercepts and slopes for speaker within electrode (Inflection Point Position ~ Model R^2 * SpeakerSubset + (1 | Subject) + (1 | Electrode:Subject)).

Experiment 2: Synthetic vowels

Stimuli and procedure

Participants passively listened to a set of synthesized vowel tokens. Stimuli were synthesized using an online version of the Klatt vowel synthesizer (www.source-code.biz/klattSyn/), with fixed pitch (250 Hz), F3 (3.1 kHz) and F4 (3.3 kHz). F1 and F2 were varied orthogonally to cover the entire vowel formant range as well as values outside those occurring in natural speech. For F1, we selected 10 values between 200 and 100 Hz (200, 255, 310, 420, 530, 640, 750, 860, 915, 970); for F2 we selected 10 values between 500 and 3000 Hz (500, 650,850,1200,1550, 1900, 2250, 2600, 2800, 2950). Stimuli covered all F1/F2 combinations with F2 larger than F1. Data in this task were recorded in five blocks of approximately 4-min duration each, resulting in 8 to 10 repetitions per token in each patient. Tokens were 300 ms long and were presented with an average intertrial interval of 800 ms, randomly sampled from a uniform distribution between 700 and 900 ms.

All stimuli were presented at a comfortable ambient loudness (~70 dB) through free-field speakers (Logitech) placed approximately 80 cm in front of the patients' head using custom-written MATLAB R2016b (MathWorks, www.mathworks.com) scripts and psychtoolbox [85–87]. Stimuli were sampled at 16 kHz for presentation in the experiment. Participants were asked to listen to the stimuli attentively and were free to keep their eyes open or closed during the stimulus presentation.

Analysis

Electrode selection: Analyses included electrodes located on the STG, which showed robust evoked responses to vowel stimuli, defined as electrodes for which the best linear model, either with only main effects or with main effects and interaction terms, explained at least 10 % of the variance. Analyses contained 160 electrodes, 10 to 33 within single patients.

Analysis of variance: As neural responses had a stereotypical evoked response peaking between 100 and 400 ms after stimulus onset, we focused our analyses on the mean HGA in that time window. Single time point analyses of the entire HGA time course produced qualitatively the same results. We modeled the main effects of F1, F2 and their linear interaction (F1*F2) onto evoked HGA responses to synthetic vowel onset using linear models. To assess the unique variance for main effects, we compared a main effect model (HGA ~ F1 + F2) to models containing only one formant (e.g., HGA ~ F1). To assess the unique variance explained by the interaction term, a full model (HGA ~ F1 + F2 + F1*F2) was compared to the model containing main effects only (HGA ~ F1 + F2). For comparability across electrodes, predictors and HGA were z-scored prior to model fitting. For comparison to natural speech data, S-TRF models were fitted to natural speech response data on the same electrodes, using the same procedures as described above. To assess the effect of the formant range onto electrode response properties, models were fitted twice: First using all stimulus tokens, and second using only the subset of stimuli with formant values within the DIMEx vowel formant range.

Analyses across electrodes: For the correlations between beta weights across electrodes the model was beta(F2) ~ beta(F1) + (1|subj) + (1|el:subj).

For the comparison between models on full data and reduced stimulus sets the model was: Rsq ~model*tuningDirection + (1|subj) + (1|el:mod).

QUANTIFICATION AND STATISTICAL ANALYSIS

We used R^2 as a metric of model fit for model comparisons between linear and sigmoidal models.

Analysis across electrodes—Analysis across electrodes were conducted using Pearson's correlations across all electrodes as well as with mixed-effects modeling with random intercepts and slopes for subjects and electrodes. Mixed effects models fit using the matlab function fitlme. Specific models are listed in the corresponding sections above.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Will Schuermann, and Matthew K. Leonard for discussions of the manuscript. We thank Ben Speidel for help with anatomical localization of electrodes. We thank all members of the Chang Lab for feedback and support throughout the project.

This work was supported by grants from the NIH (R01-DC012379 and U01-NS117765 to E.F.C.). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2034836 to IBG. This research was also supported by Bill and Susan Oberndorf, the Joan and Sandy Weill Foundation, and the William K. Bowes Foundation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Fogerty D, and Humes LE (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. J. Acoust. Soc. Am 131, 1490–1501. [PubMed: 22352519]
2. Fogerty D, and Kewley-Port D (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. J. Acoust. Soc. Am 126, 847–857. [PubMed: 19640049]
3. Ladefoged P, and Johnson K (2014). A course in phonetics (Nelson Education)
4. Johnson K, and Sjerps MJ (2021). Speaker Normalization in Speech Perception. The Handbook of Speech Perception, 145–176. 10.1002/9781119184096.ch6.
5. Strange W (1989). Evolving theories of vowel perception. J. Acoust. Soc. Am 85, 2081–2087. [PubMed: 2659637]
6. Nearey TM (1989). Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am 85, 2088–2113. [PubMed: 2659638]
7. Syrdal AK, and Gopal HS (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. The Journal of the Acoustical Society of America 79, 1086–1100. 10.1121/1.393381. [PubMed: 3700864]
8. Miller JD (1989). Auditory-perceptual interpretation of the vowel. The Journal of the Acoustical Society of America 85(5), 2114–2134. [PubMed: 2659639]
9. Peterson GE (1961). Parameters of vowel quality. J. Speech Hear. Res 4, 10–29. [PubMed: 13734834]
10. Formisano E, De Martino F, Bonte M, and Goebel R (2008). "Who" is saying" what"? Brain-based decoding of human voice and speech. Science 322, 970–973. [PubMed: 18988858]
11. Bonte M, Hausfeld L, Scharke W, Valente G, and Formisano E (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J. Neurosci 34, 4548–4557. [PubMed: 24672000]
12. Levy DF, and Wilson SM (2020). Categorical Encoding of Vowels in Primary Auditory Cortex. Cereb. Cortex 30, 618–627. [PubMed: 31241149]
13. Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, and Goebel R (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40, 859–869. [PubMed: 14622588]
14. Hamilton LS, Oganian Y, Hall J, and Chang EF (2021). Parallel and distributed encoding of speech across human auditory cortex. Cell 184, 4626–4639.e13. [PubMed: 34411517]
15. Khalighinejad B, Patel P, Herrero JL, Bickel S, Mehta AD, and Mesgarani N (2021). Functional characterization of human Heschl's gyrus in response to natural speech. Neuroimage 235, 118003. [PubMed: 33789135]
16. Norman-Haignere S, Kanwisher NG, and McDermott JH (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron 88, 1281–1296. 10.1016/j.neuron.2015.11.035. [PubMed: 26687225]
17. Leaver AM, and Rauschecker JP (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci 30, 7604–7612. [PubMed: 20519535]
18. Yi HG, Leonard MK, and Chang EF (2019). The Encoding of Speech Sounds in the Superior Temporal Gyrus. Neuron 102, 1096–1110. [PubMed: 31220442]
19. Bhaya-Grossman I, and Chang EF (2022). Speech Computations of the Human Superior Temporal Gyrus. Annu. Rev. Psychol 73, 79–102. [PubMed: 34672685]
20. Obleser J, and Eisner F (2009). Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn. Sci 13, 14–19. [PubMed: 19070534]

21. Kohonen T, and Hari R (1999). Where the abstract feature maps of the brain might come from. Trends in Neurosciences 22, 135–139. 10.1016/s0166-2236(98)01342-3. [PubMed: 10199639]
22. Rauschecker JP, and Scott SK (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci 12, 718–724. [PubMed: 19471271]
23. Shestakova A, Brattico E, Soloviev A, Klucharev V, and Huotilainen M (2004). Orderly cortical representation of vowel categories presented by multiple exemplars. Brain Res. Cogn. Brain Res 21, 342–350. [PubMed: 15511650]
24. Fox NP, Leonard M, Sjerps MJ, and Chang EF (2020). Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. Elife 9. 10.7554/eLife.53051.
25. Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, and Knight RT (2010). Categorical speech representation in human superior temporal gyrus. Nat. Neurosci 13, 1428–1432. [PubMed: 20890293]
26. Bertalmío M, Gomez-Villa A, Martín A, Vazquez-Corral J, Kane D, and Malo J (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. Sci. Rep 10, 16277. [PubMed: 33004868]
27. Fischer BJ, Anderson CH, and Peña JL (2009). Multiplicative auditory spatial receptive fields created by a hierarchy of population codes. PLoS One 4, e8015. [PubMed: 19956693]
28. Hagiwara RE (2004). An Acoustic Analysis of Vowel Variation in New World English (review). Language 80, 903–903. 10.1353/lan.2004.0191.
29. Mesgarani N, Cheung C, Johnson K, and Chang EF (2014). Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010. [PubMed: 24482117]
30. Bitterman Y, Mukamel R, Malach R, Fried I, and Nelken I (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature 451, 197–201. [PubMed: 18185589]
31. Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, and Gallant JL (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 12, 289–316. [PubMed: 11563531]
32. Gill P, Zhang J, Woolley SMN, Fremouw T, and Theunissen FE (2006). Sound representation methods for spectro-temporal receptive field estimation. J. Comput. Neurosci 21, 5–20. [PubMed: 16633939]
33. Johnson K (2020). The ΔF method of vocal tract length normalization for vowels. Lab. Phonol 11. 10.5334/labphon.196.
34. Diesch E, and Luce T (1997). Magnetic fields elicited by tones and vowel formants reveal tonotopy and nonlinear summation of cortical activation. Psychophysiology 34, 501–510. [PubMed: 9299904]
35. Obleser J, Elbert T, Lahiri A, and Eulitz C (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. Brain Res. Cogn. Brain Res 15, 207–213. [PubMed: 12527095]
36. Monahan PJ, and Idsardi WJ (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. Lang. Cogn. Process 25, 808–839. [PubMed: 20606713]
37. Scharinger M, Idsardi WJ, and Poe S (2011). A Comprehensive Three-dimensional Cortical Map of Vowel Space. Journal of Cognitive Neuroscience 23, 3972–3982. 10.1162/jocn_a_00056. [PubMed: 21568638]
38. Obleser J, Leaver AM, Vanmeter J, and Rauschecker JP (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Front. Psychol 1, 232. [PubMed: 21738513]
39. Iverson P, and Kuhl PK (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. J. Acoust. Soc. Am 97, 553–562. [PubMed: 7860832]
40. Kuhl PK (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Percept. Psychophys 50, 93–107. [PubMed: 1945741]
41. Feldman NH, and Griffiths TL (2007). A rational account of the perceptual magnet effect. In Proceedings of the annual meeting of the cognitive science society Vol. 29

42. Sjerps MJ, Fox NP, Johnson K, and Chang EF (2019). Speaker-normalized sound representations in the human auditory cortex. Nat. Commun 10, 2465. [PubMed: 31165733]
43. Daube C, Ince RAA, and Gross J (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. Curr. Biol 29(12), 1924–1937.e9. [PubMed: 31130454]
44. Mukamel R, and Fried I (2012). Human intracranial recordings and cognitive neuroscience. Annu. Rev. Psychol 63, 511–537. [PubMed: 21943170]
45. Fujisaki H, and Kawashima T (1968). The roles of pitch and higher formants in the perception of vowels. IEEE Trans. Audio Electroacoust 16, 73–77.
46. Johnson K (1988). Processes of speaker normalization in vowel perception
47. Laing EJC, Liu R, Lotto AJ, and Holt LL (2012). Tuned with a Tune: Talker Normalization via General Auditory Processes. Front. Psychol 3, 203. [PubMed: 22737140]
48. Huang J, and Holt LL (2012). Listening for the norm: adaptive coding in speech categorization. Front. Psychol 3, 10. [PubMed: 22347198]
49. Rabinowitz NC, Willmore BDB, Schnupp JWH, and King AJ (2011). Contrast gain control in auditory cortex. Neuron 70, 1178–1191. [PubMed: 21689603]
50. Pérez-González D, and Malmierca MS (2014). Adaptation in the auditory system: an overview. Front. Integr. Neurosci 8, 19. [PubMed: 24600361]
51. Fishman YI, Reser DH, Arezzo JC, and Steinschneider M (2000). Complex tone processing in primary auditory cortex of the awake monkey. II. Pitch versus critical band representation. J. Acoust. Soc. Am 108, 247–262. [PubMed: 10923889]
52. von Kriegstein K, Eger E, Kleinschmidt A, and Giraud AL (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res. Cogn. Brain Res 17, 48–55. [PubMed: 12763191]
53. Okada K, Rong F, Venezia J, Matchin W, Hsieh I-H, Saberi K, Serences JT, and Hickok G (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb. Cortex 20, 2486–2495. [PubMed: 20100898]
54. Fisher JM, Dick FK, Levy DF, and Wilson SM (2018). Neural representation of vowel formants in tonotopic auditory cortex. Neuroimage 178, 574–582. [PubMed: 29860083]
55. Obleser J, Boecker H, Drzezga A, Haslinger B, Hennenlotter A, Roettinger M, Eulitz C, and Rauschecker JP (2006). Vowel sound extraction in anterior superior temporal cortex. Hum. Brain Mapp 27, 562–571. [PubMed: 16281283]
56. Ohl FW, and Scheich H (1997). Orderly cortical representation of vowels based on formant interaction. Proc. Natl. Acad. Sci. U. S. A 94, 9440–9444. [PubMed: 9256501]
57. Boebinger D, Norman-Haignere SV, McDermott JH, and Kanwisher N (2021). Music-selective neural populations arise without musical training. J. Neurophysiol 125, 2237–2263. [PubMed: 33596723]
58. Mesgarani N, and Chang EF (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236. [PubMed: 22522927]
59. Saarinen T, Jalava A, Kujala J, Stevenson C, and Salmelin R (2015). Task-sensitive reconfiguration of corticocortical 6–20 Hz oscillatory coherence in naturalistic human performance. Hum. Brain Mapp 36, 2455–2469. [PubMed: 25760689]
60. Liljeström M, Kujala J, Stevenson C, and Salmelin R (2015). Dynamic reconfiguration of the language network preceding onset of speech in picture naming. Hum. Brain Mapp 36, 1202–1216. [PubMed: 25413681]
61. Fritz JB, David S, and Shamma S (2013). Attention and Dynamic, Task-Related Receptive Field Plasticity in Adult Auditory Cortex. In Neural Correlates of Auditory Cognition, Cohen YE, Popper AN, and Fay RR, eds. (Springer New York), pp. 251–291.
62. Bonte M, Valente G, and Formisano E (2009). Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. J. Neurosci 29, 1699–1706. [PubMed: 19211877]
63. Atiani S, David SV, Elgueda D, Locastro M, Radtke-Schuller S, Shamma SA, and Fritz JB (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. Neuron 82, 486–499. [PubMed: 24742467]

64. Fritz JB, David SV, Radtke-Schuller S, Yin P, and Shamma SA (2010). Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. Nat. Neurosci 13, 1011–1019. [PubMed: 20622871]
65. Vannest JJ, Karunanayaka PR, Altaye M, Schmithorst VJ, Plante EM, Eaton KJ, Rasmussen JM, and Holland SK (2009). Comparison of fMRI data from passive listening and active-response story processing tasks in children. Journal of Magnetic Resonance Imaging 29, 971–976. 10.1002/jmri.21694. [PubMed: 19306445]
66. Strange W, Verbrugge RR, Shankweiler DP, and Edman TR (1976). Consonant environment specifies vowel identity. J. Acoust. Soc. Am 60, 213–224. [PubMed: 956528]
67. Strange W (1989). Dynamic specification of coarticulated vowels spoken in sentence context. The Journal of the Acoustical Society of America 85, 2135–2153. 10.1121/1.397863. [PubMed: 2732388]
68. Jenkins JJ, Strange W, and Edman TR (1983). Identification of vowels in “vowelless” syllables. Percept. Psychophys 34, 441–450. [PubMed: 6657448]
69. Kuwabara H (1985). An approach to normalization of coarticulation effects for vowels in connected speech. J. Acoust. Soc. Am 77, 686–694. [PubMed: 3973240]
70. Hillenbrand J, Getty LA, Clark MJ, and Wheeler K (1995). Acoustic characteristics of American English vowels. J. Acoust. Soc. Am 97, 3099–3111. [PubMed: 7759650]
71. Näätänen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 432–434. [PubMed: 9009189]
72. Dufour S, Brunellière A, and Nguyen N (2013). To what extent do we hear phonemic contrasts in a non-native regional variety? Tracking the dynamics of perceptual processing with EEG. J. Psycholinguist. Res 42, 161–173. [PubMed: 22460687]
73. Fox RA, Flege JE, and Munro MJ (1995). The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. J. Acoust. Soc. Am 97, 2540–2551. [PubMed: 7714272]
74. Boemio A, Fromm S, Braun A, and Poeppel D (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat. Neurosci 8, 389–395. [PubMed: 15723061]
75. Schönwiesner M, Rübsamen R, and Von Cramon DY (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. European Journal of Neuroscience 22, 1521–1528. 10.1111/j.1460-9568.2005.04315.x. [PubMed: 16190905]
76. Hickok G, and Poeppel D (2007). The cortical organization of speech processing. Nat. Rev. Neurosci 8, 393–402. [PubMed: 17431404]
77. Zatorre RJ (2001). Spectral and Temporal Processing in Human Auditory Cortex. Cerebral Cortex 11, 946–953. 10.1093/cercor/11.10.946. [PubMed: 11549617]
78. Towle VL, Yoon H-A, Castelle M, Edgar JC, Biassou NM, Frim DM, Spire J-P, and Kohrman MH (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. Brain 131, 2013–2027. [PubMed: 18669510]
79. Pineda LA, Castellanos H, Cuétara J, Galescu L, Juárez J, Llisterri J, Pérez P, and Villaseñor L (2010). The Corpus DIMEx100: transcription and evaluation. Lang Resources & Evaluation 44, 347–370.
80. Oganian Y, and Chang EF (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. Sci Adv 5, eaay6279. [PubMed: 31976369]
81. Ray S, and Maunsell JHR (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol 9, e1000610. [PubMed: 21532743]
82. Hamilton LS, Chang DL, Lee MB, and Chang EF (2017). Semi-automated Anatomical Labeling and Inter-subject Warping of High-Density Intracranial Recording Electrodes in Electrocorticography. Front. Neuroinform 11, 62. [PubMed: 29163118]
83. Pineda LA, Pineda LV, Cuétara J, Castellanos H, and López I (2004). DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. In Advances in Artificial Intelligence – IBERAMIA 2004 (Springer Berlin Heidelberg), pp. 974–983.
84. Boersma P, and Weenink D Praat: doing phonetics by computer. [Computer Program] https://www.fon.hum.uva.nl/praat/.

85. Kleiner M, Brainard D, and Pelli D (2007). What's new in Psychtoolbox-3?

86. Brainard DH (1997). The Psychophysics Toolbox. Spat. Vis 10, 433–436. [PubMed: 9176952]

87. Pelli DG (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis 10, 437–442. [PubMed: 9176953]

Highlights

- Human speech cortex represents vowels based on local 2D formant tuning maps
- Formant tuning is sigmoidal along the speaker pitch-normalized formant range
- Vowel category representation emerges at the population-level
- Formant tuning reflects spectral content of harmonic sounds, not just speech

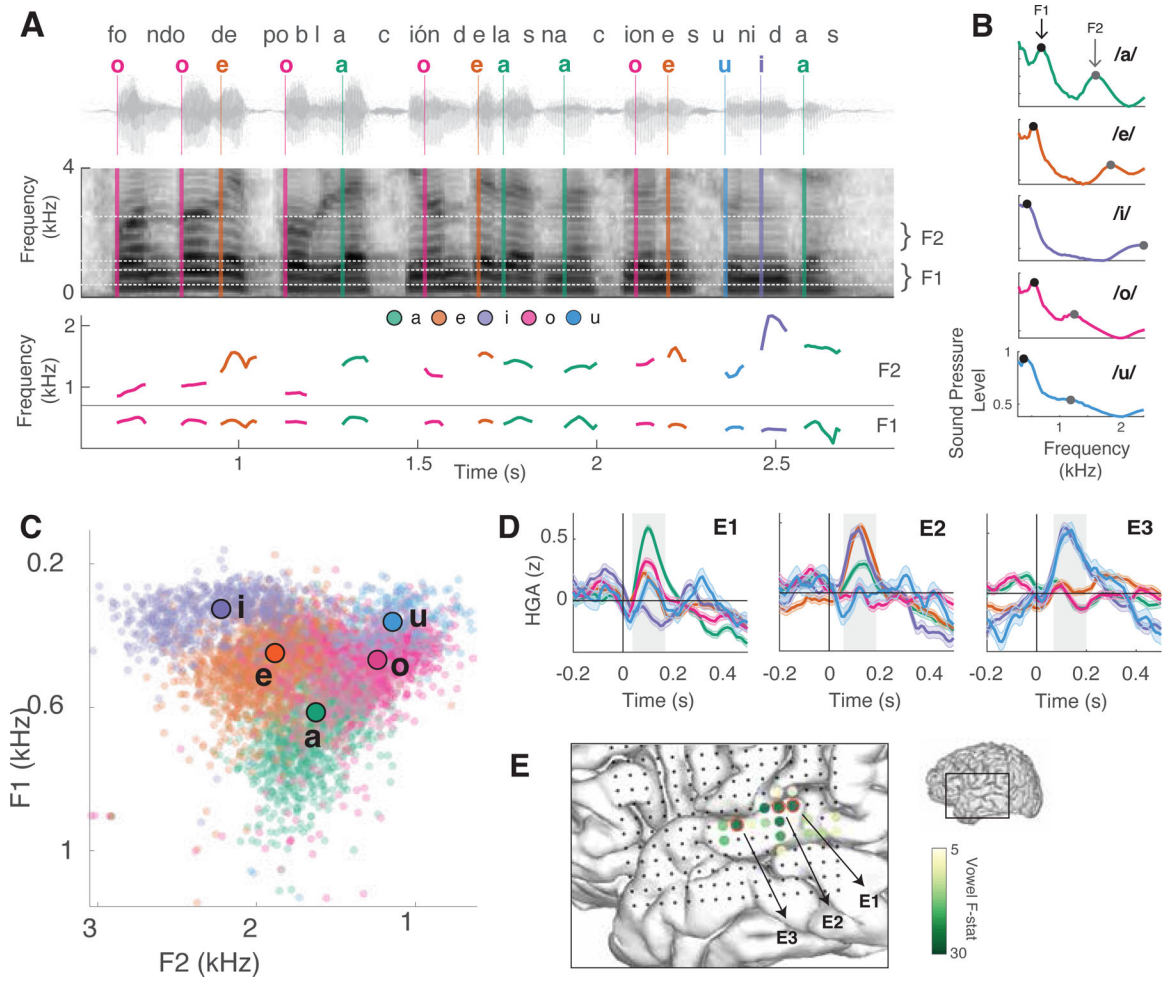


Figure 1. Neural activity on single electrodes in bilateral human STG is sensitive to vowels.

A. Example Spanish stimulus sentence waveform (top), spectrogram (middle) and extracted formant trajectories (bottom). Vertical lines mark vowel onsets, colored by vowel identity. **B.** Median frequency spectrum for a single speaker per vowel. Formant spectral peaks are marked by filled circles **C**. Median F1 and F2 values across all vowel instances in our stimulus set. **D.** Differential average HGA responses to single vowel categories on three example STG electrodes. Error bars indicate standard error of the mean, gray shaded area marks time window of averaging for electrode selection. **E**. Vowel discriminative electrodes for a single participant. Gray circles demarcate all grid electrodes, example electrodes from panel D are marked in red. See Figure S1B and C for anatomical information across all participants.

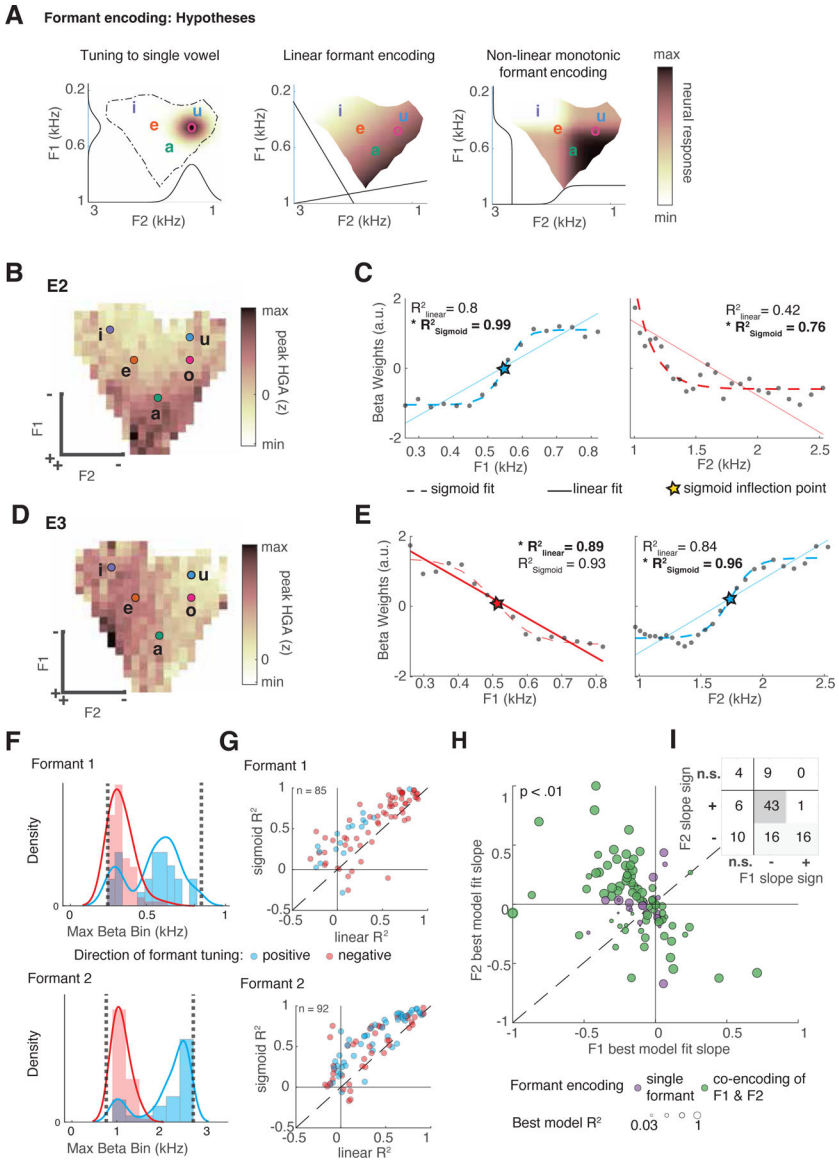


Figure 2. Non-linear monotonic encoding of vowel formant frequencies in human STG.

A. Three alternative hypotheses for encoding of vowel formants on single electrodes. **B, D.** Example electrodes' formant receptive fields. **C, E**. Formant tuning curves of the electrodes in B and D. Red: Decrease in beta weight as formant values increase; blue: increase in beta weight as formant values increase. **F.** Frequency of maximal beta weights in F1 and F2 ranges. **G**. Comparison between linear and sigmoid model fits for F1 (top) and F2 (bottom). **H**. Distribution of slopes for F1 and F2 across vowel-responsive electrodes- see Figure S2 for anatomical maps. **I.** Inset showing the total number of electrodes that encode F1 and F2 and the direction of their slopes, corresponding to the number of electrodes in each quadrant in **H.**

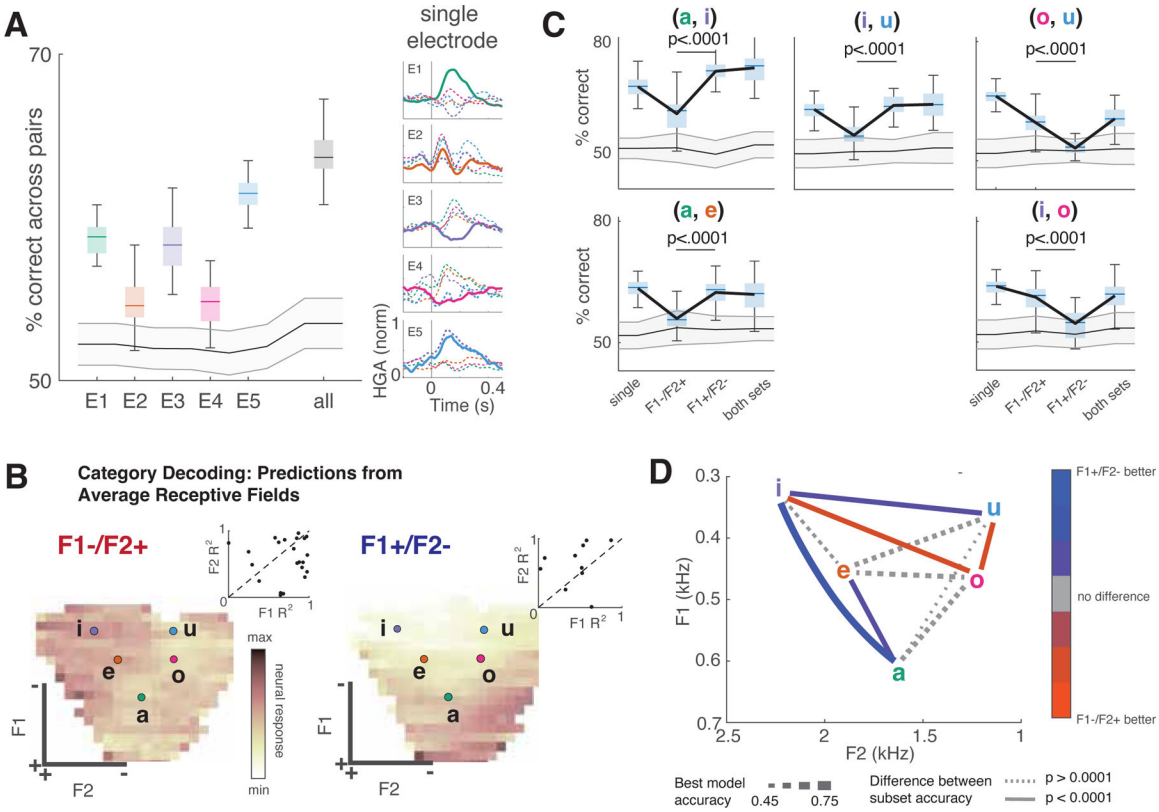


Figure 3: Emergent vowel representation at the population level.

A. Vowel decoding accuracy from the best single electrode per vowel (five unique electrodes across four subjects) as compared to the decoding accuracy using all electrodes. All accuracies averaged across ten pairwise comparisons. Right: Average HG response to each vowel in five electrodes used in single electrode decoding in the leftmost panel. Solid line corresponds to the vowel for which the selected electrode shows best average accuracy, dashed lines correspond to all other categories. Gray shaded bar indicates empirical chance performance over repetitions. **B.** Average formant receptive fields for electrodes with different tuning types and corresponding F1/F2 sigmoidal model R^2. **C.** Decoding accuracy from five vowel pairs separated by electrode sub-group (best single electrode, each tuning type, both tuning types). **D.** Summary of decoding accuracy across vowel pairs (all pairwise statistics in Table S4).

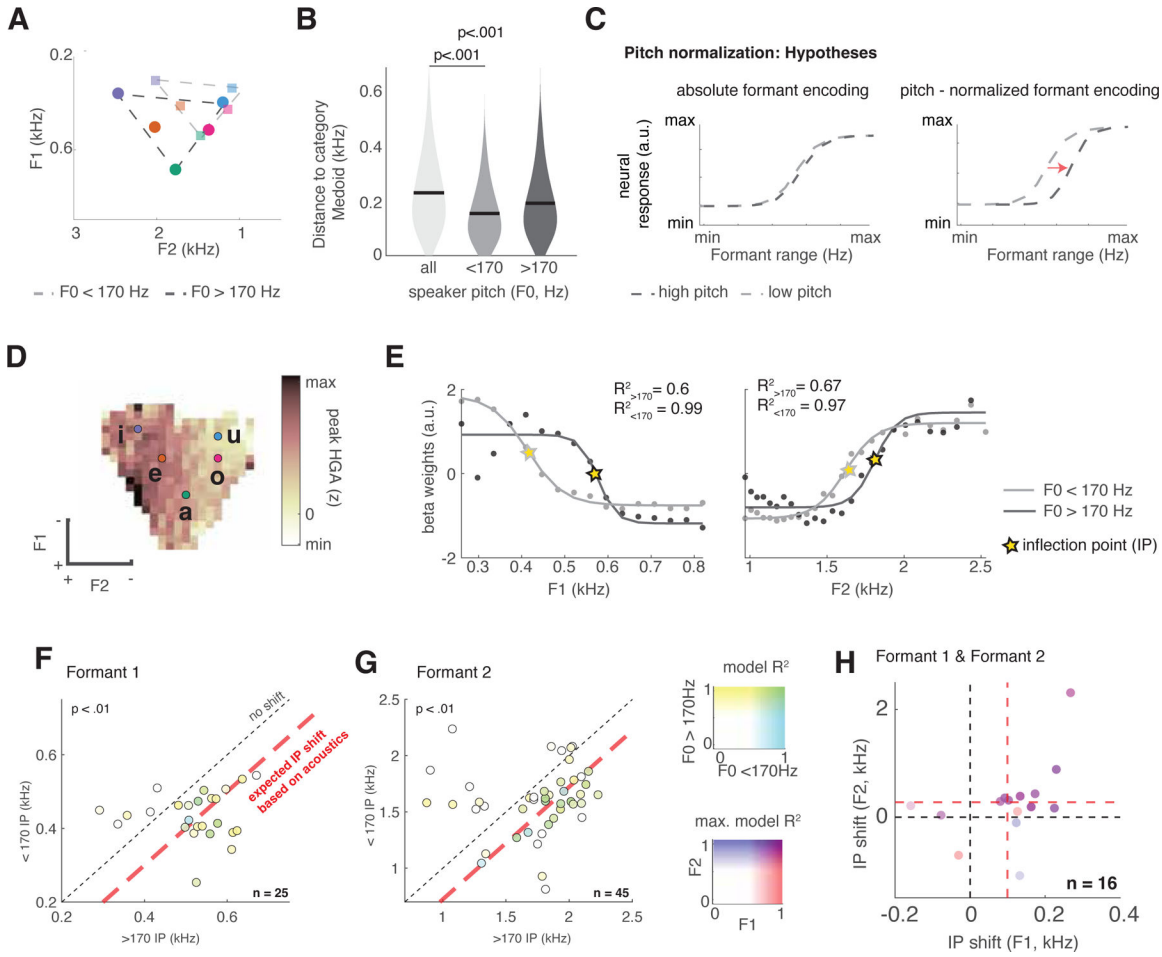


Figure 4. Shifting dynamic ranges underlie normalization of vowel representation for speaker pitch

A. Vowel formants shift between speakers with short and long vocal tracts (and thus high and low pitch). **B**. Formant normalization in two groups by high/low pitch reduces formant variability within vowel categories. **C.** Hypotheses for absolute (left) and pitch-normalized (right) encoding of vowel formants. **D.** Formant receptive field for an exemplary electrode. **E.** Formant frequency tuning for F1 (left) and F2 (right) on the same exemplary electrode, split by speaker pitch. **F-G.** Tuning curve inflection points (IP), calculated separately for sentences with high and low speaker pitch, across all F1 (**F**) and F2 (**G**) encoding electrodes. **H**. IP shift for F1 and F2 on electrodes encoding both formants.

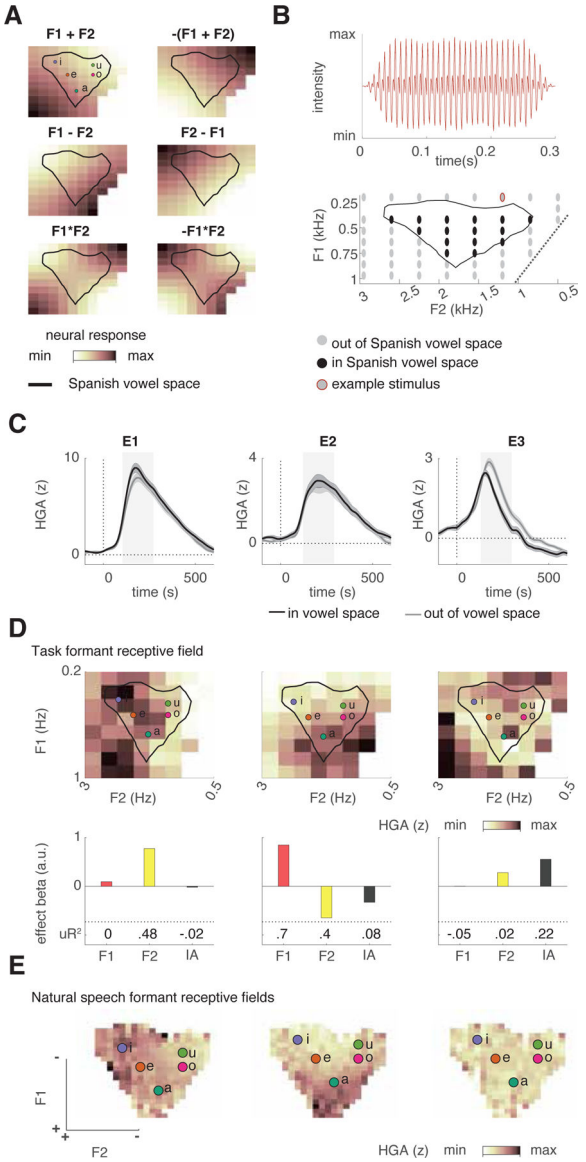


Figure 5. Vowel formant tuning emerges from complex frequency tuning on STG.

A. Schematic of independent (top two rows) and interactive encoding of F1 and F2 (bottom row). **B.** Stimulus design for synthetic vowel task; Top: example token waveform. Bottom: F1 and F2 values task. see Figure S3 for perceptual ratings of all stimuli. **C.** Mean responses (+-SEM) to stimulus tokens that fall within and outside the Spanish vowel space on three example STG electrodes. Dark background marks peak area for averaging and further analyses. **D**. Formant receptive fields (top) and linear model effect R^2 (bottom) for the same example electrodes. Black outline in formant receptive fields shows the vowel formant range in continuous Spanish sentences in the natural speech task. **E.** Formant receptive fields derived from the DIMEx corpus of Mexican Spanish [79] for the same example electrodes.

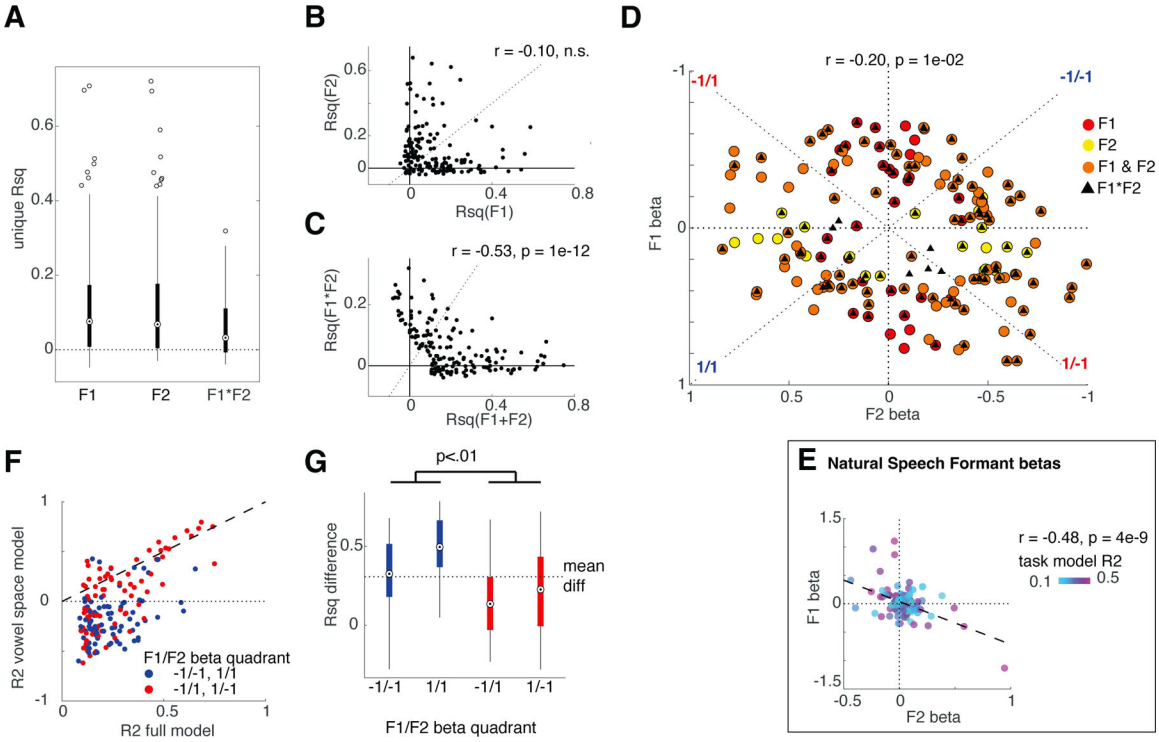


Figure 6.

Comparison between formant encoding in synthetic vowel sounds and in natural speech. **A.** Effect R^2 distribution across electrodes. **B-C.** Correlation between effect R^2 for main effects of F1 and F2 (**B**) and between main effects and the linear interaction effect (**C**). **D-E.** Across all electrodes, direction of effects for F1 and F2 is less correlated in the synthetic vowel task (**D**) than in the natural speech corpus (**E**). **F.** Model R^2 for model fit on the entire vowel task stimulus set vs only on stimuli with formants located within the natural formant space. **G.** R^2 values for models fit on the full and reduced stimulus sets.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Matlab 2021b	Mathworks.com	
Custom code and data	This paper	Zenodo DOI 10.5281/zenodo.7620900
Imaging pipeline for coregistration of electrodes to CT and MRI scans	Hamilton et al. 2019	10.3389/fninf.2017.00062
Other		
Human patient participants recruited from neurosurgical patients at UCSF (see Table S1).	This paper	N/A