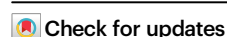


Cortical processing of discrete prosodic patterns in continuous speech

Received: 30 June 2023

Accepted: 29 January 2025

Published online: 03 March 2025



G. Nike Gnanateja^{1,10}, Kyle Rupp^{2,10}, Fernando Llanos³, Jasmine Hect², James S. German⁴, Tobias Teichert^{5,6}, Taylor J. Abel^{2,7,11}✉ & Bharath Chandrasekaran^{2,7,8,9,11}✉

Prosody has a vital function in speech, structuring a speaker's intended message for the listener. The superior temporal gyrus (STG) is considered a critical hub for prosody, but the role of earlier auditory regions like Heschl's gyrus (HG), associated with pitch processing, remains unclear. Using intracerebral recordings in humans and non-human primate models, we investigated prosody processing in narrative speech, focusing on pitch accents—abstract phonological units that signal word prominence and communicative intent. In humans, HG encoded pitch accents as abstract representations beyond spectrotemporal features, distinct from segmental speech processing, and outperforms STG in disambiguating pitch accents. Multivariate models confirm HG's unique representation of pitch accent categories. In the non-human primate, pitch accents were not abstractly encoded, despite robust spectrotemporal processing, highlighting the role of experience in shaping abstract representations. These findings emphasize a key role for the HG in early prosodic abstraction and advance our understanding of human speech processing.

Speech processing involves the high-fidelity encoding of continuous spectral and temporal information, as well as extracting and mapping relevant speaker-invariant dimensions onto linguistically relevant constructs such as phonemes, words, phrases, and sentences¹. Spoken words are generally viewed as providing the core information regarding a speaker's communicative intent. Words are composed of speech sounds or phonemes that are encoded in a spatially distributed manner along the superior temporal gyrus and sulcus (STG and STS)^{2,3}, with neural ensembles representing emergent properties of acoustically distinctive phonetic features (e.g., place and manner of articulation, or vowel height and backness). In addition to words, prosody in speech carries crucial cues regarding linguistic and affective content,

which ultimately serve to coherently structure the talker's intended message to the listener. Prosody is cued by multi-dimensional spectrotemporal features including pitch, duration, loudness, and gradient differences in vowel formant characteristics. In conversational speech, speakers often dynamically vary pitch in systematic ways to cue phonologically distinct intonational patterns that convey, for example, the type of speech act that is being intended (i.e., a statement vs. a question), that certain words or phrases are attentionally important, that specific information is either known or unknown to the listener, or an intent to change the topic of conversation^{4–6}. While much of the focus in the neuroscience of speech perception has been dedicated to the encoding of phonological patterns that contribute to spoken word

¹Speech Processing and Auditory Neuroscience Lab, Department of Communication Sciences and Disorder, University of Wisconsin-Madison, Madison, WI, USA. ²Pediatric Brain Electrophysiology Laboratory, Department of Neurological Surgery, University of Pittsburgh, Pittsburgh, PA, USA. ³UT Austin Neurolinguistics Lab, Department of Linguistics, The University of Texas at Austin, Austin, TX, USA. ⁴Aix-Marseille University, CNRS, LPL, Aix-en-Provence, France. ⁵Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. ⁶Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA. ⁷Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA. ⁸Roxelyn and Richard Pepper Department of Communication Sciences & Disorders, Northwestern University, Evanston, IL, USA. ⁹Knowles Hearing Center, Evanston, IL 60208, USA. ¹⁰These authors contributed equally: G. Nike Gnanateja, Kyle Rupp. ¹¹These authors jointly supervised this work: Taylor J. Abel, Bharath Chandrasekaran. ✉e-mail: abelt@pitt.edu; bchandra@northwestern.edu

processing^{1,3,7}, relatively little is known about the encoding of phonological patterns associated with intonation^{8,9}.

The STG and STS are considered important interfaces for the extraction of the high-level features that underlie spoken language processing^{2,3,10,11}. Sub-regions of the lateral STG independently encode distinctive features underlying phonological categories³. Prior work using intracranial recordings has argued for multiple, parallel pathways, including a direct pathway from the auditory thalamus to the STG that bypasses the primary auditory cortex (a sub-region within the Heschl's gyrus—HG) which encodes distinct speech features¹². Indeed, disruption of the HG due to lesions or via inhibitory cortical stimulation does not substantially impair speech perception¹². Tang et al.¹³ used controlled sentences to study the representation of sentence-level prosody in the STG via high-density surface intracerebral recordings. They found distinct representations of relative pitch trajectories associated with the different intonational patterns. Such representations were *independent* of the representation of segmental information. Prior studies^{12,13} argue that the lateral STG is the key region for processing segmental and *prosodic* information. However, there are challenges to this conclusion: these studies have rarely assessed prosodic encoding from both HG and STG; ablating the primary auditory cortex may not impact spoken word processing but is known to impair pitch processing¹⁴. Neuroimaging studies (invasive and non-invasive) have shown that the medial¹⁵ and anterolateral HG and planum temporale (PT)^{16–20} are key centers for processing behaviorally relevant pitch information. Further, pitch processing is crucial in narrative contexts which are rarely used in studies using intracerebral approaches.

Pitch accents are abstract phonological units of prosodic structure whose primary phonetic correlates are changes in the fundamental frequency contour (Fig. 1a, b) on and around specific syllables. In English, pitch accents are sparsely distributed, in that they do not occur on every word. When they do occur, they are typically localized to the lexically stressed syllable of a word. In the Autosegmental-Metrical (AM) theory of intonational phonology^{9,21}, pitch accents are composed of one or two discrete tonal units, high (H) or low (L), which determine the relative pitch levels as well as the temporal organization of the pitch movements associated with a given pitch accent. Only one tone may be associated directly with a pitch-accented syllable, thus if a pitch accent is composed of two tones (e.g., L + H* or L* + H), AM uses an asterisk (H* and L*, respectively) to identify which tonal target has a specific temporal alignment with the stressed syllable of the word that receives prominence. In other words, the entire pitch contour associated with a pitch accent may stretch over an interval that is longer

than a syllable, though substantially shorter than a sentence or utterance, within which they are highly localized. In standard varieties of American English, there are at least four discrete categories of pitch accents which are distinguished phonologically by both tonal composition and alignment, and include L*, H*, L + H*, and L* + H²². Pitch accents contribute additional prominence to stressed syllables not only in terms of pitch, but also duration, loudness, and vowel distinctiveness. Together with prosodic phrasal boundaries and their associated boundary tones, pitch accents form the scaffolding of sentential prosodic structure²³.

We leveraged the high spatio-temporal resolution of human intracerebral recordings to study the extent to which intonational categories extracted from continuous speech are encoded as discrete categories in the HG, relative to the STG^{13,24}. In contrast to previous intracerebral investigations of intonational prosody which used electrocorticography, we used stereoelectroencephalography (sEEG), which allows access to medial regions of the brain that are typically inaccessible to surface electrocorticography arrays, including the HG and the supratemporal plane (STP)^{25,26}. We scaffold our examination of phonologically distinct intonational categories within the AM framework, which argues for a tier of phonological information that is distinct from segmental information, but whose temporally organized sequences of high and low tones are aligned in specific ways with the latter. Importantly, the categories that make up the tonal tier carry information linked to the speaker's intentions⁸. For example, an H* pitch accent which is typically realized as shallow rising pitch movement on the stressed syllable indicates that the word carries salient (and new) information, while an L* pitch accent, realized as a low dipping pitch movement on the stressed syllable, marks that the word carries information already known by the listener (Fig. 1). These category differences in the shape and timing of the pitch trajectories are leveraged by listeners to access pragmatic information that goes beyond the lexical content conveyed by words. While the AM model is highly influential in shaping linguistic theory, the underlying premise that the proposed category distinctions have a neurobiological basis remains unknown. We propose that these intonational categories are robustly represented within the HG, that the phonological representations are an emergent abstract property, and that the neural organization for intonational categories is distinct from segmental feature representations within the HG and the STG. To further support this argument, we examine pitch accent encoding in a non-human primate model, shown to be an effective animal model for studying auditory processing similar to that of humans^{27–29}. We hypothesize that a primate without experience in understanding and using pitch

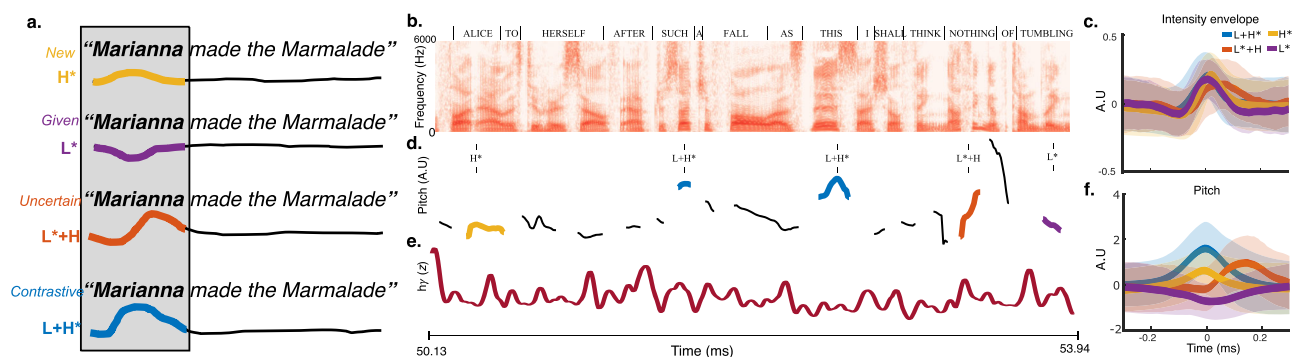


Fig. 1 | Illustration of the experiment and the acoustic correlates of pitch accents. **a** Examples of the four pitch accent categories and their associated meaning for four sentences with identical segmental content. **b** Subjects listened to the narrative 'Alice's Adventures in Wonderland,' with a representative spectrogram and annotated words shown here. **c** Mean stimulus envelope contours across all instances of each pitch accent. Shading shows the standard error of the mean. **d** Pitch contours of the narrative segment are shown in (b), with colors and

annotations marking individual pitch accent examples. These pitch accents are tightly tied to the stressed syllables in speech. **e** High gamma (H γ) responses from an example channel, time-locked to the narrative segment shown in (b) and (d). **f** Mean pitch contours across all instances of each pitch accent. Shading shows the standard error of the mean. A.U. indicates arbitrary units, as the figure here shows normalized pitch intensity and pitch contours for intensity and pitch.

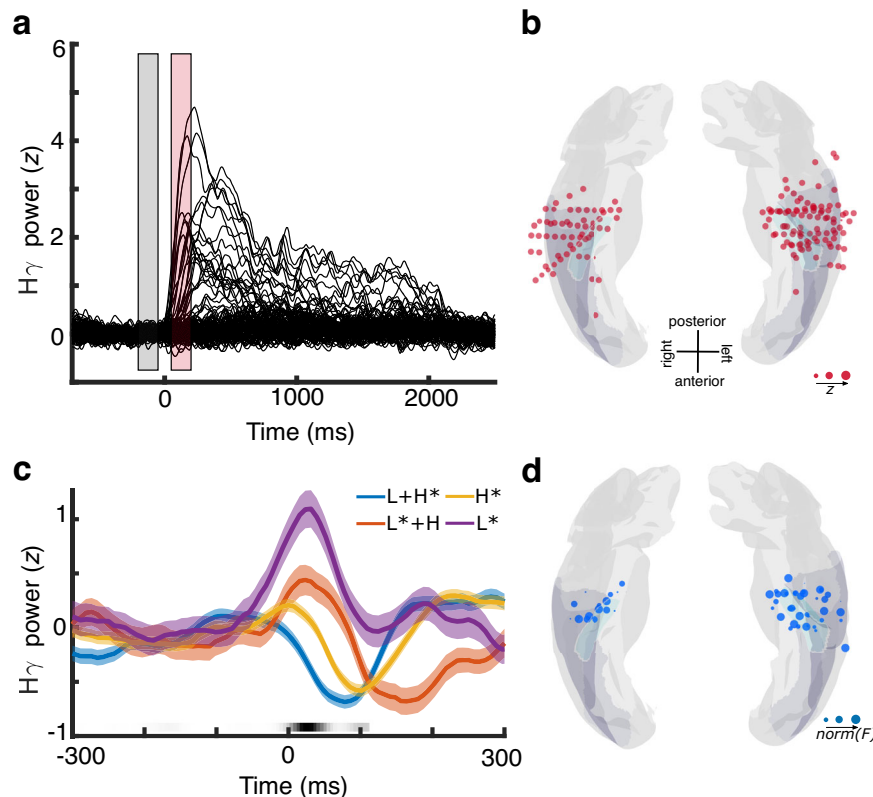


Fig. 2 | Speech responsiveness and pitch accent separability in the supratemporal plane. **a** High gamma (H γ) power event-related responses elicited using short speech tokens. The post-stimulus duration of the H γ response (shaded red) was compared with the pre-stimulus baseline (shaded gray) to select speech-responsive electrodes. **b** Speech-responsive electrodes (red dots) across all participants are plotted in a normalized space on the supratemporal plane of the MNI152 brain. Speech-responsive electrodes are primarily located in Heschl's gyrus (HG—green shading) and superior temporal gyrus (STG—blue shading). The size of the red dots indicates the magnitude of the Wilcoxon sign rank test statistic (z)

comparing the post-stimulus H γ power with the baseline as shown in (a). **c** Mean H γ responses from an electrode in HG, time-locked to the pitch accents. The shaded area shows the standard error of mean. The separability F -statistic across time is shown in grayscale at the bottom of the plot. **d** The electrodes that showed strong pitch accent separability are plotted as blue dots in a normalized space on the MNI152 brain. The size of the blue dots indicates the magnitude of pitch accent separability (cluster F statistic) normalized per individual. Pitch accent separable electrodes are primarily localized near the HG bilaterally.

accents—lacking both exposure to distinct pitch accent categories and their social and linguistic meanings—would not develop abstract representations of these units, while still robustly encoding the underlying acoustic cues.

Results

Speech-responsive electrodes

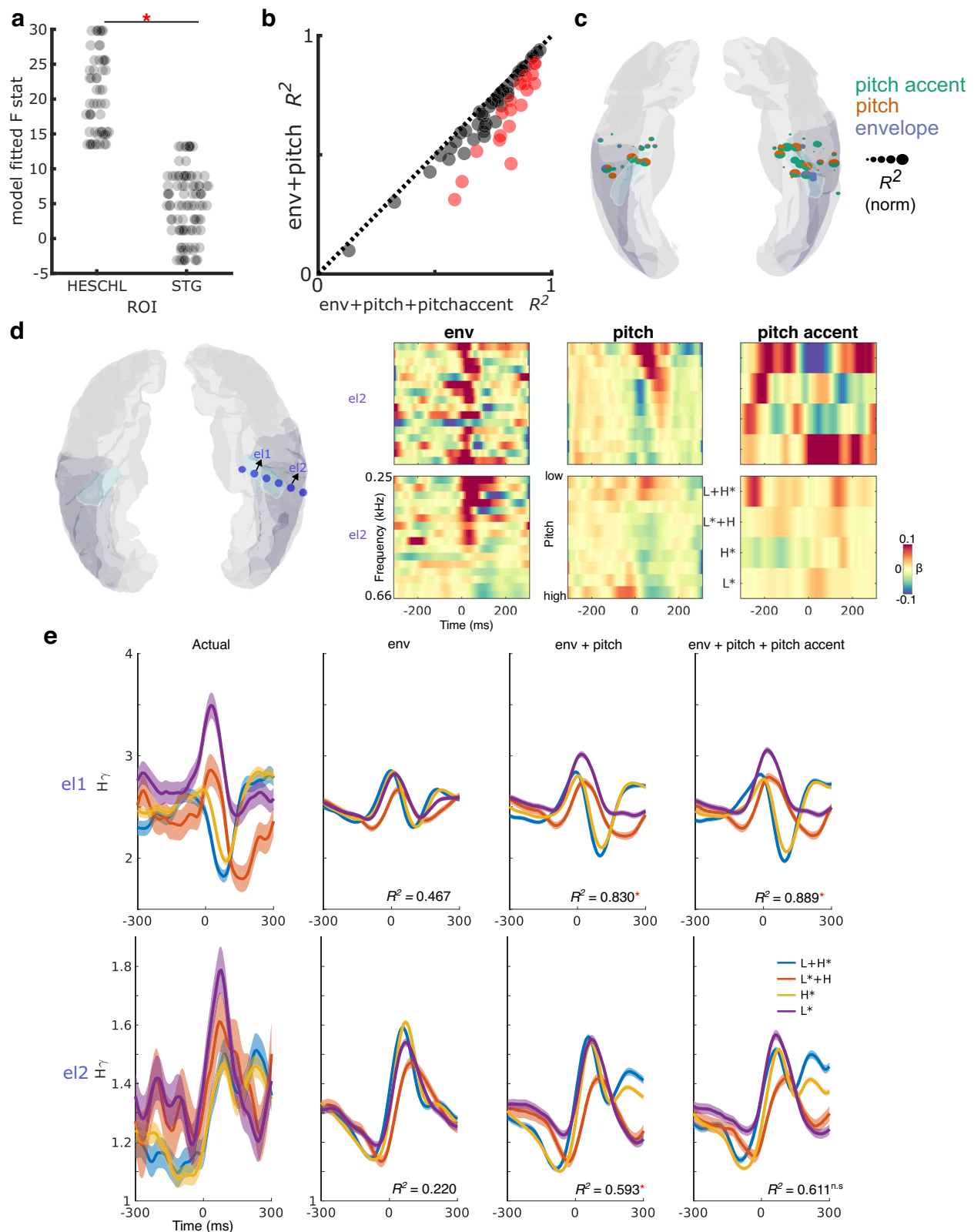
A total of 1561 electrode contacts were implanted across 11 participants undergoing sEEG monitoring for intractable epilepsy. The participants listened to an audio narrative of the book 'Alice's Adventures in Wonderland' by Lewis Carroll (produced by a male talker) during sEEG monitoring (Supplementary Fig. 1). No participants had any gross pathology in the STP. More details about the participants can be found in Supplementary Table 1 and Supplementary Fig. 1. Broadband high gamma (70–150 Hz) (H γ) was extracted across electrodes (Fig. 2a). In another session, the participants also listened to short tokens of natural speech. The H γ responses to these short tokens were used to identify speech-responsive electrodes (Fig. 2b). Among the electrodes implanted within the STP, 158 electrodes in the STP (Fig. 2b) were found to be speech responsive with significantly (false discovery rate-corrected $p < 0.01$) higher post-stimulus (50–200 ms) H γ responses than baseline (–200 to 50 ms). Further analyses were limited to only these electrodes.

Pitch accent separable electrodes in the supratemporal plane

H γ responses to the audiobook narrative were segmented based on the temporal loci (–300 to 300 ms relative to pitch accent location) of the

pitch accent (Fig. 2c). The time point 0 ms does not mark the onset of the pitch accent, but rather the location of the annotation landmark of the tonal target that is temporally aligned with the stressed syllable, which is typically the minimum or maximum fundamental frequency (f_0) in the stressed syllable depending on the pitch accent category (see 'Methods' section). Consequently, pitch accent cues emerge before the 0 ms mark, and thus may be reflected in H γ responses prior to this point as pitch accent categories. We measured the ratio of variance between and within-pitch accent categories to obtain F statistics. The F statistics were obtained for every time point between –100 and 300 ms, with a false discovery rate-corrected significance threshold of $q < 0.01$. A clustering approach was used to estimate significant clusters of time points that were temporally adjacent for at least four time bins. The sum of the F statistic within each cluster was used to obtain a cluster F statistic. The cluster statistic was used to infer the separability of pitch accent categories (henceforth, pitch accent separability).

Out of the 158 speech-responsive electrodes in the HG and STG, 63 electrodes showed significant pitch accent separability. The electrodes that showed maximum pitch accent separability were primarily localized to the HG in the STP bilaterally (Figs. 2d and 3a). A generalized linear mixed effects model was fit on the pitch accent separability across the 158 electrodes with fixed effects of the region of interest (ROI) (HG and STG) and participant as a random intercept, which showed that HG showed significantly more [$t(156) = -5.616$, $p = 8.749 \times 10^{-8}$] pitch accent separability than STG (Fig. 3a). Adding



the hemisphere as an additional fixed effect did not improve the model fit [*Likelihood ratio statistic*– $LRstat(2)=2.783$, $p=0.249$]. Lastly, the proportion of electrodes that showed significant separability did not differ between hemispheres for both HG [$\chi^2(1)=1.155$, $p=0.282$] and STG [$\chi^2(1)=1.155$, $p=0.282$].

Because the HG showed strong pitch accent separability, we further evaluated the extent to which pitch accent separability differed

along the 60 speech-responsive electrodes along the axis of the HG and the bounding sulci³⁰. This was performed by using the normalized MNI coordinates of the electrodes as fixed effects predicting the pitch accent separability using a generalized linear mixed effects model. The right and left hemisphere electrodes were combined in the analyses by multiplying the MNI medial-lateral coordinate by -1 . For easier interpretability of the model parameters, the coordinates of the electrodes

Fig. 3 | Pitch accent separability across the supratemporal plane. **a** Pitch accent separability in the Heschl's gyrus (HG) vs. the superior temporal gyrus (STG). The values shown on the y-axis are the generalized linear mixed effects model-fitted pitch accent separability (F stat) values with the fixed effects of region of interest (ROI), hemisphere (right vs. left) and their interaction effect with participants being the random intercept. Separability was significantly higher in the HG than the STG on a generalized linear mixed effects model ($p = 8.75 \times 10^{-8}$ two-tailed cluster corrected). **b** Variance explained (R^2) in the Hy responses by the acoustic-only and the acoustic + pitch accent category encoding models. Points on or near the diagonal represent electrodes where the pitch accent category model does not explain any variance beyond the acoustic model. Markers in red denote electrodes that showed significantly ($p < 0.01$) higher explained variance with the addition of the pitch accent feature (adjusted for additional features) to the acoustic model. **c** Supratemporal

plane view showing the unique variance explained by each feature. Each electrode is shown as a pie chart, with the total size denoting the full model R^2 normalized per participant, and the slices denoted the unique variance explained by each feature. **d** Temporal response functions (TRFs) for the different stimulus features in a representative subject from electrodes in HG (e11) and STG (e12). E11 shows strong encoding of all features, while e12 shows strong encoding of only env and pitch features. **e** Observed and TRF model-predicted Hy evoked responses at two electrodes shown in (d). Shaded regions show the standard error of mean. E11 shows distinct, highly separable responses across pitch accents, and the model-predicted Hy responses showed an increase in explained variance (* indicates a significant increase in R^2) with the addition of pitch accent features. E12 does not show strong differential responses across pitch accents. Model-predicted Hy responses at e12 closely match the observed Hy only when using the spectrotemporal acoustic features.

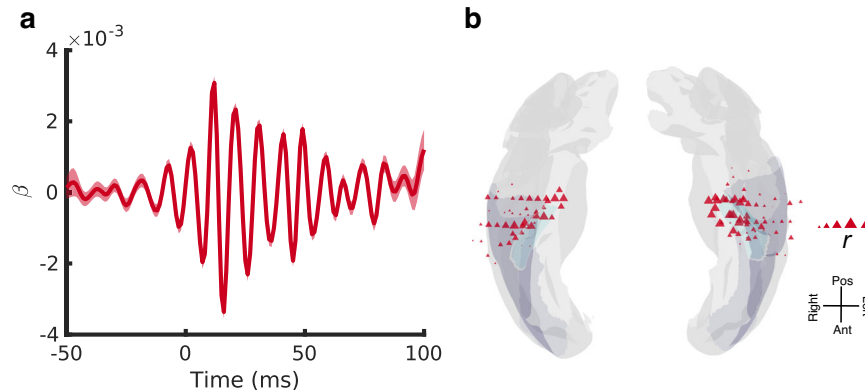


Fig. 4 | Periodicity encoding in the supratemporal plane. **a** Temporal response function for the periodicity from a representative electrode in the Heschl's gyrus. TRF was obtained on the local field potential. The shaded region shows the standard error of the mean. **b** Periodicity encoding in the supratemporal plane is

depicted by the prediction accuracy (marker size) of the encoding model of fundamental frequency onto the local field potential. The prediction accuracy was normalized per subject and projected on the MNI152 template brain.

were rotated such that the x coordinate was parallel to the HG axis, the y coordinate was perpendicular to the HG axis with an anteroposterior orientation, and the z coordinate was perpendicular to the HG with a dorsoventral orientation. The square of y and z coordinates were used in the model to assess the extent to which pitch accent separability decreased as the electrodes were farther from the center of the HG, and the participant identifier was used as the random intercept. The pitch accent separability did not show statistically significant (all p s > 0.01) difference across any of the three axes of HG [x : $t(52) = 2.426$, $p = 0.019$; y^2 : $t(52) = 2.553$, $p = 0.136$; z^2 : $t(52) = -2.601$, $p = 0.0121$], and none of the interaction effects were statistically significant [$x*y^2$: $t(52) = -2.217$, $p = 0.031$; y^2*z : $t(52) = 2.0728$, $p = 0.043$; $x*z^2$: $t(52) = -1.864$, $p = 0.043$; $x*y^2*z^2$: $t(52) = 1.262$, $p = 0.213$]. The pitch accent separability thus did not vary systematically along the HG. The addition of hemisphere to the model did not improve the model fit [$LRstat(1) = 0.174$, $p = 0.676$], suggesting that results did not differ between the two hemispheres. While the above analyses show that pitch accents are encoded differentially, they do not unequivocally suggest that the pitch accents are encoded as higher-order prosodic categories beyond acoustic representations.

Pitch accents are encoded as abstract representations beyond spectrotemporal acoustic cues

To investigate the extent to which pitch accent separability was driven by acoustic features or demonstrate evidence of abstract representations of the pitch accent category, we used multivariate encoding models. These encoding models resulted in linear kernels (temporal response functions: TRFs) that described the relationship between the stimulus features and the Hy responses. These linear kernels consist of regression coefficients in the time domain that describe the encoding strength of different stimulus features.

Multivariate encoding models with stimulus envelope and pitch features were used to predict the Hy responses to the pitch accents. Hy responses to pitch accents were averaged to obtain pitch accent event-related high-gamma responses (ER-Hy). Similarly, model-predicted Hy responses were also calculated for each stimulus feature using cross-validation, and ER-Hy traces were obtained using the TRF predictions (Fig. 3e). The addition of the pitch accent feature to the acoustic (env + pitch) model resulted in significantly higher (Fig. 3b) explained variance (adj. R^2 , p s < 0.01, see Variance Partitioning in Methods, an approach which accounts for a differing number of features between models) in the HG and STG electrodes, suggesting a higher-order abstract representation of pitch accents beyond encoding of the spectrotemporal acoustic cues. The ER-Hy responses for two electrodes in a representative subject are shown in Fig. 3e. The HG electrode showed distinct ER-Hy responses across the four pitch accent categories. These model-predicted ER-Hy showed that the cumulative addition of features significantly increased the explained variance in the responses in the HG electrode, while the same was not true in the STG electrode. The STG electrode only showed a significant increase in explained variance with envelope and pitch features and not the pitch accent feature. Variance partitioning of the multivariate encoding models was performed to extract unique R^2 explained by each feature¹³. Electrodes in which pitch accent features explained a significant unique R^2 were primarily localized near the HG (Fig. 3c and Supplementary Fig. 2), and were significantly higher in proportion than in the STG [$\chi^2(1) = 32.681$, $p = 1.085 \times 10^{-8}$]. It is worth noting that most of these electrodes in the HG and STG encoded both spectrotemporal acoustic features and higher-order categorical pitch accent features. This can also be seen in the weights of the TRF in an HG electrode and STG electrode in a representative participant (Fig. 3d). Further, the full

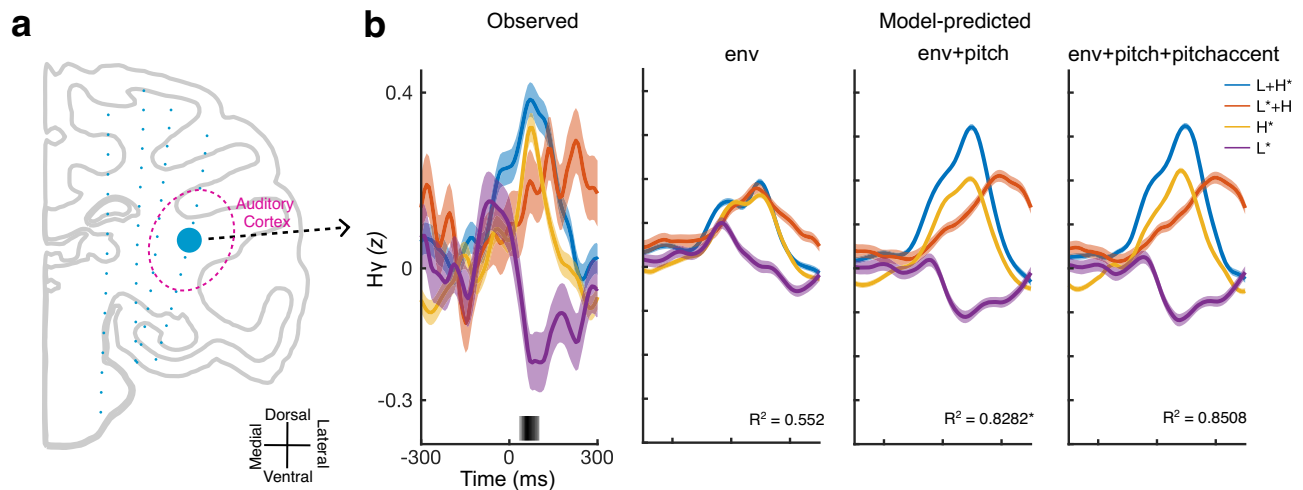


Fig. 5 | Pitch accent separability in the right hemisphere of a rhesus macaque. **a** Pitch accent separability (F -stat) shown as blue dots on a coronal brain slice. The size of blue dot indicates the magnitude of separability. Only one electrode in the auditory cortex showed pitch accent separability. **b** Pitch accent evoked high gamma responses from the auditory cortex electrode and model-predicted high

gamma responses. Shaded regions show the standard error of mean. The pitch accent separability was predicted by the acoustic (env and pitch) models, while the pitch accent feature did not explain significantly more variance than the acoustic features alone (* indicates significant increase in R^2 with $p < 0.01$).

spectrotemporal + discrete pitch accent features (actual) explained higher variance ($ps < 0.01$) than the spectrotemporal model + shuffled (100 times) pitch accent features. The variance explained by the full model beyond the mean of the shuffled model was significantly higher in the HG than the STG electrodes (Supplementary Fig. 3). This provides further converging evidence that the pitch accents are encoded as discrete features beyond spectrotemporal features, with higher encoding in the HG than the STG electrodes [$t(156) = -2.862$, $p = 0.005$].

Periodicity is a vital ingredient of intonation contours and pitch accent processing. We sought to assess the brain regions involved in encoding the low-level periodicity cues into more abstract pitch accent cues. To this end, the phase-locking of the local field potentials to periodicity (Fig. 4a) in the audiobook narrative was assessed in all the electrodes in the STP. We analyzed the local field potentials instead of the gamma information for periodicity encoding due to the partial overlap of the periodicity and high gamma frequencies. The phase-locking information was derived based on the magnitude of similarity between the periodicity of the stimulus and the periodicity in the local field potential at each electrode. This was done to evaluate if the neural regions that show higher pitch accent separability also show higher phase-locking to periodicity. First, a generalized linear mixed effects model was fit on the phase-locking with fixed effects of the ROI (HG and STG) and participant as a random intercept. The phase-locking was highly localized (Fig. 4b) in the HG compared to the STG [$t(156) = -10.796$, $p = 1.217 \times 10^{-20}$], and adding an interaction term of hemisphere did not improve model fit [$LRstat(2) = 3.429$, $p = 0.180$] suggesting similar phase-locking patterns between the hemispheres, which was the same as pitch accent separability. Then we assessed if the pitch accent separability was related to the phase-locking by fitting a generalized linear mixed effects model predicting pitch accent separability with fixed effects of periodicity encoding and interaction of periodicity and ROI (HG and STG). Pitch accent separability was strongly related to phase-locking [$t(155) = 9.534$, $p < 3.048 \times 10^{-17}$] and the interaction [$t(158) = -4.098$, $p = 6.686 \times 10^{-5}$] between phase-locking and ROI, suggesting that the electrodes in HG that showed pitch accent separability also showed higher phase-locking, while the same was not true in the STG. Further, a generalized linear mixed effects model predicting the unique R^2 of pitch accents was significantly related to the phase-locking magnitude [$t(156) = 5.323$, $p = 3.488 \times 10^{-7}$]. These results suggest that neural populations

displaying pitch accent separability also transform the spectrotemporal acoustic features into higher-order prosodic representations. Taken together, these results are consistent with a model in which the HG robustly encodes pitch and pitch accent information in a highly specialized manner much more robustly beyond spectrotemporal features compared to the STG.

Pitch accents are specific to language experience and not encoded in a non-human primate model of auditory perception

To bolster the premise that abstract pitch accent features are encoded in the human listeners beyond spectrotemporal representations and are tied to listeners' linguistic experience, and the learning of socially relevant linguistic categories, pitch accent processing was assessed using intracerebral recordings in a non-human primate model. We turned to the macaque monkey because (i) they do not use or understand pitch accents, but at the same time (ii) their auditory system processes low-level acoustic features much like humans. If our spectrotemporal model captures all relevant acoustic confounds, we would expect no encoding of the abstract pitch accent features in this species. Neural data were obtained from the right hemisphere of one monkey using an electrophysiological mesoscope consisting of three-dimensional penetrating electrode grid with 800 intracranial contacts distributed across 50 shafts arranged in 12 coronal slices. This provided a dense coverage of the brain relative to our human data, where the electrode coverage was more sparse. The ER-Hγ responses to pitch accents showed separability in only one electrode (Fig. 5a) located in the auditory cortex. Multivariate encoding models were fit in the same way as in the humans to evaluate if the pitch accent separability in the monkey was encoded as a higher-level abstract representation beyond the spectrotemporal acoustic representation. The pitch accent features did not explain variance in the ER-Hγ responses to pitch accents beyond the variance explained by the spectrotemporal acoustic features (Fig. 5b). Further assessment of phase-locking to the periodicity showed multiple electrodes in the auditory cortex and the brainstem that showed strong phase-locking (Fig. 6a, b). The pitch accent separable electrode also demonstrated strong phase-locking to the periodicity. This suggests that the single pitch accent separable electrode encoded multiple acoustic features, namely phase-locking information as well as the envelope and pitch contours. However, the pitch accents were not encoded as higher-order features beyond the low-level acoustic cues. A note worth mentioning is that, of the 11

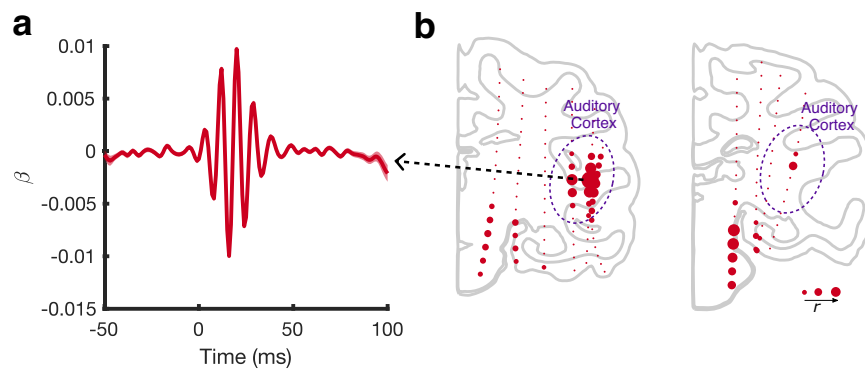


Fig. 6 | Periodicity encoding in the auditory cortex of the macaque. **a** Temporal response function (TRF) to speech periodicity recorded from an auditory cortex electrode; shaded regions show standard error of the mean across trials. **b** TRF prediction accuracy for periodicity was displayed on two coronal slices of the

macaque that showed the best phase-locking. Periodicity encoding was the strongest in auditory cortex and brainstem electrodes. The electrode that showed maximum periodicity encoding was not the one that showed maximum pitch accent separability.

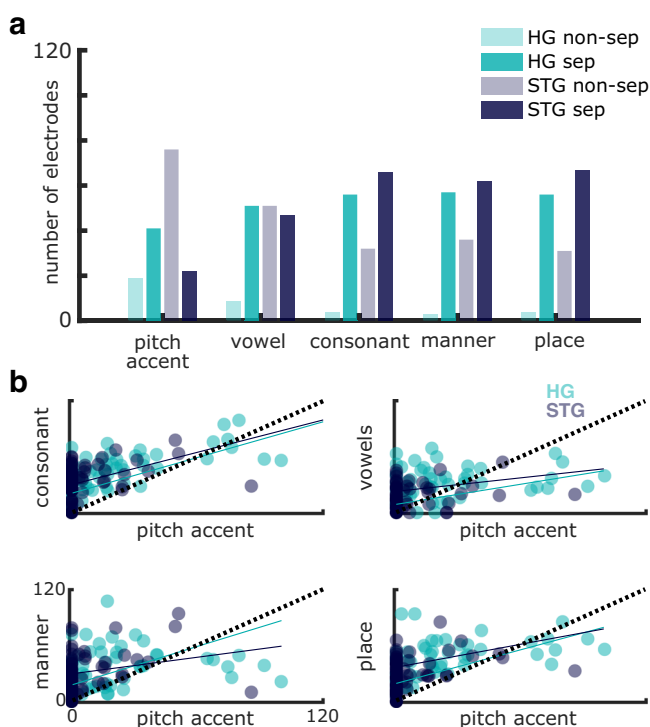


Fig. 7 | Pitch accent separability and segmental feature separability in the Heschl's gyrus (HG) and the superior temporal gyrus (STG). **a** Total number of feature separable electrodes across regions of interest (HG and STG). Significant association between separability and region of interest was seen for all features. The proportion of pitch accent separable to non-separable electrodes was higher in the HG than the STG, unlike the consonantal segmental features. **b** Scatter plot showing the relationship between the separability of pitch accent and segmental features (F_{clus}).

human participants, 4 had electrodes in right hemisphere ROIs, of which 1 participant showed no right hemisphere pitch accent separability.

Pitch accents and segmental speech features are differentially encoded in the supratemporal plane

Pitch accents are multi-dimensional short timescale inflections at the syllable level in the intonational contour of speech and are linked to stressed syllables, which often co-occur with changes in spectrotemporal acoustic cues. In contrast, the segmental/phonetic features

of speech are shorter timescale phoneme-level inflections primarily in the formant frequencies, slow temporal fluctuations, and spectral content that are produced as a consequence of the place, voicing, and manner of movement of the speech articulators in the vocal tract. However, both pitch accents and segmental features form discrete auditory unit-like representations that are relatively invariant across a range of absolute spectrotemporal cues. To determine whether pitch accents and segmental features in speech are encoded in similar brain regions, the pattern of electrode activation along the supratemporal plane was assessed for pitch accents, vowel features, and place and manner of articulation. Manner cues are primarily linked to longer timescale temporal fluctuations in speech, while the place of articulation is differentiated by short timescale spectrotemporal fluctuations, primarily associated with formant frequency transitions. Vowels also rely on formant frequency changes leading to the vowel formant shape. Thus, the separability of segmental features and their pattern of similarities with the pitch accent separability was assessed.

Separability was assessed across vowel categories, consonant categories, consonant place of articulation, and consonant manner of articulation (Figs. 7a and 8a) (see 'Methods' section for the full list of categories). In the HG, the proportion of pitch accent separable electrodes was significantly lower than the proportion of consonant ($z = -3.669$, $p = 2.437 \times 10^{-4}$), manner ($z = -4.021$, $p = 5.795 \times 10^{-5}$), and place [$z = -3.669$, $p = 2.437 \times 10^{-4}$] separable electrodes but did not differ significantly from the proportion of vowel separable electrodes ($z = -2.2015$, $p = 0.03$). In the STG, the proportion of pitch accent separable electrodes was significantly lower than the proportion of all segmental feature separable electrodes; consonant ($z = -7.081$, $p = 1.432 \times 10^{-12}$), manner ($z = -6.337$, $p = 2.335 \times 10^{-10}$), place ($z = -7.276$, $p = 3.449 \times 10^{-13}$), and vowels ($z = -3.880$, $p = 1.045 \times 10^{-4}$). Generalized linear mixed effects models were also fit for separate electrodes to assess whether the magnitude of pitch accent separability was linked to the magnitude of segmental feature separability; pitch accent separability was modeled using a fixed effect of segmental feature separability and interaction effect of ROI and segmental feature separability, with a random intercept of subject. This analysis was restricted only to the electrodes that showed significant pitch accent separability. The scatter plots in Fig. 7b show the relationship between pitch accent separability and segmental feature separability across all the speech-responsive electrodes in the HG and the STG. Pitch accent separability was significantly predicted by consonant separability [$t(60) = 5.335$, $p = 1.534 \times 10^{-6}$]. In contrast, pitch accent separability was not significantly predicted by vowel separability [$t(60) = 1.361$, $p = 0.179$], manner separability [$t(60) = 1.462$, $p = 0.149$], or place separability [$t(60) = 1.240$, $p = 0.220$]. None of the interaction effects were significant ($ps > 0.01$) indicating that the pattern did not differ across the HG and

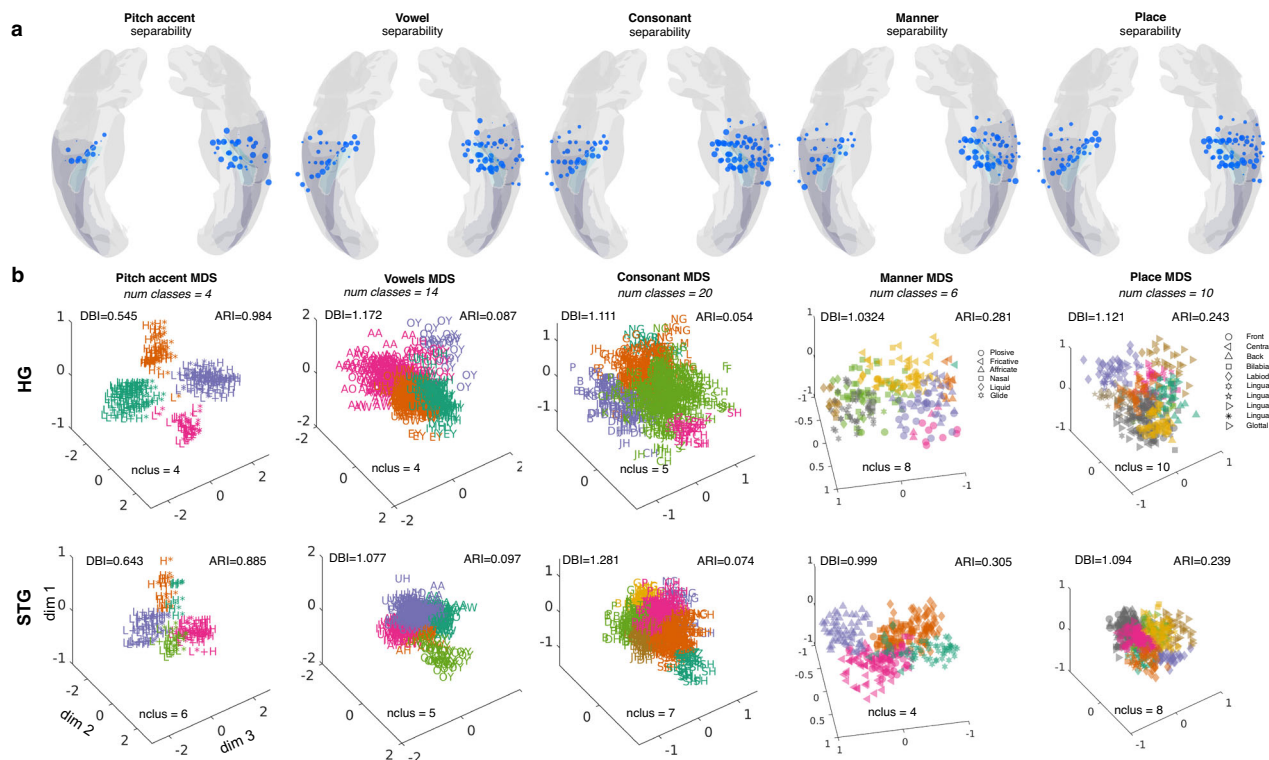


Fig. 8 | Comparison of electrode sites encoding prosodic and segmental features. **a** Electrode sites in the supratemporal plane (STP) showing separability of features (pitch accent, vowel identity, consonant identity, consonant manner, and consonant place of articulation). The size of the circles indicates the separability metric normalized per participant. The location of the electrode sites has been normalized to the MNI space and plotted on the MNI brain. Electrode sites showing pitch accent separability are primarily localized to posteromedial Heschl's gyrus

(pmHG), while electrodes showing vowel and consonant feature separability were localized to both posteromedial and anterolateral Heschl's gyrus, extending to superior temporal gyrus and planum temporale. **b** Multi-dimensional scaling (three dimensions) visualizing the distances between HG and STG high gamma responses to the pitch accents, vowels, and consonants across all participants. The colors of the labels show cluster group membership of the electrodes derived using hierarchical agglomerative clustering.

STG. Together, these results suggest that pitch accent categories were encoded by a subset of electrodes in the HG that encoded the consonant segmental features (Fig. 7a, b).

Unsupervised hierarchical clustering was performed on the multi-dimensional scaled distances of ER-Hγ responses to assess whether the pitch accents were encoded as discrete representations in the HG and STG are different from the representation of segmental feature of vowel and consonants. Clustering was performed separately at the HG electrodes and STG electrodes (Fig. 8b). Two indices of clustering were derived: the Davies–Bouldin index (DBI) which is a combined index of cluster tightness and cluster separability, and the Adjusted Rand Index (ARI) which measured the accuracy of clustering the true categories. Lower DBI indicates tighter and well-separated clusters, while higher ARI indicates better clustering accuracy. A bootstrapped independent *t*-test was performed on the DBI and ARI to evaluate the accuracy of clustering in the HG vs. STG (Supplementary Fig. 4). The pitch accents clustered more tightly [$t(998) = -87.140$, $p = 0$, $d = -5.511$] and with higher accuracy [$t(998) = 104.815$, $p = 0$, $d = 6.629$] in the HG than the STG. For the vowel and consonant labels, multiple labels clustered together suggesting that there were no phoneme-specific clusters in the HG and STG, rather the clusters formed based on common patterns across vowel and consonant labels. For the vowels, the clusters were less tight [$t(998) = 45.281$, $p = 3.191 \times 10^{-244}$, $d = 2.864$] and showed poorer accuracy [$t(998) = -17.938$, $p = 1.399 \times 10^{-62}$, $d = -1.134$] in the HG than STG. Similarly, for the consonants, clusters were less tight [$t(998) = -82.186$, $p = 0$, $d = -5.198$] and showed poorer accuracy [$t(998) = -46.677$, $p = 3.623 \times 10^{-253}$, $d = -2.952$] in the HG than STG. Each consonant label is cued by multiple segmental features such as manner (affricate, liquid, etc.) and place (front, back, etc.) of articulation. Both the HG and STG show that segmental features of manner and place are clustered more

accurately in the brain than the consonant labels themselves. Manner clusters in the STG were tighter [$t(998) = 18.9346$, $p = 1.498 \times 10^{-68}$, $d = 1.197$] but similar in accuracy [$t(998) = -2.592$, $p = 0.01$, $d = -0.164$] when compared to the HG. Place features showed less tighter clustering [$t(998) = 18.935$, $p = 1.498 \times 10^{-68}$, $d = 1.197$] but similar accuracy [$t(998) = 3.160$, $p = 0.002$, $d = 0.200$] in the STG as in the HG. Taken together, the pitch accents showed more meaningful clusters in the HG than the STG, while the place and manner features showed more meaningful clustering in the STG than the HG, These results suggest that the pitch accents are strongly encoded in the HG, and with a pattern that is different from the encoding of consonant segmental features.

Furthermore, we tested for *hemispheric asymmetry* in the encoding of pitch accent and segmental speech features, motivated by the differences in temporal scales across which the pitch accent and segmental features of vowels and consonants unfold. This was done by comparing the proportion of electrodes encoding each feature within the HG and STG across the two hemispheres (Fig. 9). Chi-square tests revealed no significant associations between hemisphere and electrode proportion for pitch accent feature in both the HG [$\chi^2(1) = 0.96$, $p = 0.327$] and STG [$\chi^2(1) = 0.578$, $p = 0.447$]. Similarly, all other segmental features showed no hemispheric asymmetry ($p > 0.01$) except manner features in STG [$\chi^2(1) = 9.168$, $p = 0.002$].

Discussion

The human auditory system parses multi-dimensional cues in natural speech to facilitate seamless speech perception across a variety of listening situations. Natural speech is rich in prosodic cues that add a critical layer of meaning beyond the segmental features of speech. We characterized the processing of pitch accents, which are discrete phonological units of prosodic structure associated with the melodic

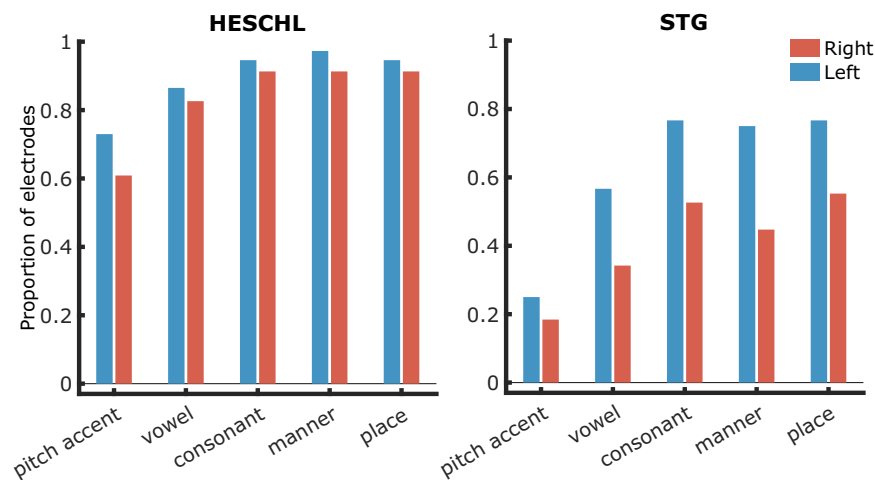


Fig. 9 | Hemispheric differences in the encoding of speech features. Proportion of speech-responsive electrode sites in the HG and STG that show feature separability (within-class variance > across-class variance). The proportion of electrode sites is separately shown for left and right HG and STG.

and rhythmic profile of speech and which are substantially shorter in timescale than an entire sentence or utterance. Specifically, we traced the encoding of pitch accents along the STP spanning the HG, and STG, in contrast to a previous study¹³ which investigated the processing of distinct intonational patterns over longer sentential duration in the STG but with sparse STP sampling. However, the encoding of intonational patterns in the STG may be inherited from the HG in the STP, which has yet to be evaluated. Using high-density sampling of the STP, we examined the encoding of short-timescale prosodic features (i.e., pitch accent categories) that form the building blocks of sentential prosody. We evaluated the extent to which pitch accents are encoded in the STP as discrete representations and traced their transformation along the medial-lateral axis of the STP. Primarily, we focused on how pitch accent representations transform along both the HG, which contains the earliest stages of cortical auditory processing, and the STG. Furthermore, we investigated the extent of dissociation between the processing of pitch accents and segmental speech cues, to characterize specialized areas for processing the multi-dimensional cues in natural speech.

Our results demonstrate that the HG is the cortical locus for processing pitch accents as discrete representations, expanding its role in speech processing beyond the low-level spectrotemporal representations suggested previously¹². We show discrete representations of pitch accent categories in the HG, beyond the associated acoustic representations of pitch and/or intensity responses. Studies that have used controlled non-speech stimuli have suggested that diffuse regions in the medial and anterolateral HG and the PT^{16,18,19,31} comprise the pitch center in the brain. Further, studies showed that intonation contours of sentences are encoded in specific electrode sites along the STG, while absolute pitch is encoded in the HG or the anterolateral border of the core auditory cortex^{12,13}. In contrast, we show that pitch accents are robustly encoded in the HG. The HG sites that encoded pitch accents are also strongly tuned to periodicity cues, relative pitch, and amplitude envelope cues. However, the pitch accents showed maximum separability only in the HG and not the STG. The pitch accent separability was distributed across the HG and not specifically limited to the anterolateral or the posteromedial portion of the HG. This suggests that HG specializes in extracting discrete pitch accent information from natural speech by integrating multi-dimensional cues. It also performs higher-order computations that go beyond spectrotemporal processing or simple phase-locking to periodicity pitch. Similarly, pitch processing across multiple time-scales and complexities has been shown to be differentially encoded along the HG^{15,32}, suggesting that the centers for processing pitch and

pitch accent information differ based on relevance and hierarchical complexity.

We sought to evaluate the extent to which cortical sites encoding the pitch accents are similar to those encoding segmental features of vowels or consonants. Pitch accent identification and categorization are known to be sensitive to the pitch contour's alignment with changes in intensity and periodic energy, as well as durational variability, associated with features of the underlying segments^{33–37}. As with pitch and pitch accents, the acoustic cues for vowels unfold over time and must be integrated dynamically with other types of spectrotemporal variation including, most notably, the rising-falling nature of the speech envelope of the vowel itself. The primary cues to vowel category are the first and second spectral formants. For most vowel categories, these are not *stable* properties in continuous speech but are realized as transitions stretched over the temporal span of the vowel. In contrast to pitch accents and vowels, the primary manner cues to consonants such as plosives, fricatives, and nasals occur as a sequentially organized series of relatively stable states. The acoustic cues for consonantal manner of articulation comprise the entirety of the speech signal with a combination of duration and spectral energy cueing the manner of articulation. In contrast, place cues, while relatively short in duration, interact in complex ways with the quality of the following vowel and typically involve transitions over and above those required for cueing the category of the vowel itself. It is an open question, therefore, whether place features for plosives are expected to pattern with vowels, and by extension, pitch accents, in terms of the need for dynamic integration with overlapping spectrotemporal cues. The vowel features, unlike consonant segmental features, were primarily encoded in the HG, and the encoding of vowels was not related to pitch accent processing. Similarly, pitch accent processing was not related to place feature representation; however, it was related to the manner feature representation. This dissociation between representations of segmental features and pitch accents is consistent with the fact that pitch accents unfold over time and must be dynamically integrated with other spectrotemporal features for perceptual identification and category discrimination, while consonant features are temporally more stable and constitute bundles of acoustic features that are perceptually independent. The difference observed in vowel representation in contrast to pitch accent representation is potentially driven by the importance of formant spectral representation for vowel perception in contrast to pitch contours cueing pitch accent representation.

The anatomical distribution of pitch accent representations differed from segmental speech features. Pitch accents and vowel representations were primarily localized to the HG, with pitch accents

exhibiting the smallest spatial spread. Representations of place and manner information were seen both in the HG and the STG, with place and manner features clustering better in the STG than in HG. While most explorations of phonetic feature representation have been demonstrated in the STG using intracerebral recordings^{2,12}, the current study and another recent study³⁸ demonstrate phonetic feature representation in the HG and STG. The specialization of the HG for preferentially processing intonational/prosodic features and distributed segmental feature processing in the HG and STG may suggest a parallel organization for speech processing.

The encoding of pitch accents upstream in the processing hierarchy suggests a role in dynamically shifting focus³⁹ to the stressed syllable, which could aid in the extraction of segmental information. While we clearly show neural representations of pitch accents and their dissociation from segmental speech features, it remains an open question how these speech cues interact and aid in a coherent speech percept. Pitch accent processing compared to segmental feature processing showed clear categorical representations in the HG. While segmental features showed meaningful clustering, multiple segmental features were clustered together as shared representations. Future research focused on a unified model of prosody and segmental information will help clarify the functional role of pitch accents and their neurophysiological basis in speech perception.

The non-human primate model bolsters evidence that pitch accent processing is a specialized function related to human speech processing, as evidenced by the absence of higher-order pitch accent representations beyond spectrotemporal representations in the primate auditory cortex. Interestingly, macaques are a social species that use vocalizations to communicate, and their vocal repertoire includes category-relevant variations along the primary acoustic dimensions of pitch accents, namely intensity and pitch contour shape. Despite the behavioral relevance of these acoustic dimensions for macaques, they do not use or understand pitch accents. Our results showing acoustic but not category encoding of pitch accents in primate auditory cortex are consistent with this argument, providing a critical piece of evidence for empirical questions surrounding the relevance of pitch accent encoding to human speech processing.

In addition to exploring pitch accent representations, we also sought to study whether the brain regions that support the processing of these higher-order features also encode an important feature of pitch processing, i.e., phase-locking to the periodicity of speech. We found that neural populations that encoded pitch accents also phase-locked to the periodicity. These results hint at the existence of populations of neural ensembles that transform periodicity features into higher-order pitch accent features in the HG. Previous explorations of pitch processing argue for a pitch center in the anterolateral HG and the STG^{18,40–42}; in contrast, we show that encoding of higher-order pitch features in speech is distributed across the HG. Similar results have also been reported in recent multiunit explorations of pitch-relevant information in the HG, which found a pitch-sensitive region distributed across the HG^{15,43,44}. While we demonstrate that extraction of higher-order linguistically relevant pitch accent information is distributed in the HG, it should not be generalized to the existence of a universal pitch center. The focus of the current study was to evaluate pitch accent processing for speech, and we did not explore the various dimensions of pitch perception using controlled pitch magnitudes as in previous studies^{15,17,19,45}. Our results shed light on the perception of linguistically relevant, naturally produced pitch accents in natural speech containing rich harmonic information, and do not necessarily explore a pitch center in the brain common to processing of pitch across speech and non-speech stimuli.

Limitations and future directions

The current study used naturally produced pitch accent stimuli to study their neural representation, but we stopped short of

comprehensively exploring their interaction with segmental feature representations. This would require using experimental stimuli with controlled manipulation of pitch accents to understand how they interact with different speech cues to generate a coherent speech percept, conveying both the speaker's message and intent. Furthermore, we did not explore connectivity between neural populations that represent pitch accents and those that represent segmental features, which would elucidate how pitch accents influence representations of the accompanying segmental features. Future research exploring the above will inform how upstream and downstream information for speech processing are integrated in the brain to perceive speech.

Methods

Humans

Participants. A total of 11 participants (2 female) participated in the study. The age of the participants ranged from 9 to 24 years, and 2 of these participants were left handed. Additional details about the participants are provided in Supplementary Table 1. The participants were undergoing intracranial electroencephalography monitoring for clinical evaluation of epilepsy surgery using sEEG. All participants were native speakers of English. The participants had no other relevant medical history, and electrode placement was based purely on clinical necessity. The families/participants gave written informed consent to participate in the study. All research protocols were approved by the Institutional Review Board of the University of Pittsburgh.

Stimuli and annotation. An audio narrative of the book 'Alice's Adventures in Wonderland' by Lewis Carroll was used as the stimulus (Chapters 1–7, <http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-5>). The book was narrated in American English by a male speaker (F0: $M = 152.10$ Hz, $SD = 40.82$ Hz). Silences in the narrative >300 ms were trimmed. The narrative was broken down into 45 contiguous audio tracks each ~65 s long. The narrative audio and written version of the book were then aligned in time using the Montreal Forced Aligner to locate the onsets and offsets of phonemes and syllables in the narrative. The forced alignment was visually inspected by three experienced linguists, and the annotation was manually corrected following the guidelines for phoneme boundary placement⁴⁶. A full prosodic annotation was then carried out following the Tones and Break Indices (ToBI) guidelines^{22,47}, which is a set of annotation conventions designed for American English and motivated by the AM theoretical model^{9,21}. These annotations were performed by an expert annotator who is not one of the authors. This annotation included pitch accents, phrase accents, boundary tones, and break indices. Only four pitch accent categories ($L + H^*$: 1188 occurrences, $L^* + H$: 322, H^* : 1610, and L^* : 318) were considered for analysis in the study due to their comparatively higher frequency of occurrence in the recorded narrative. The * in the labels indicates which tone in a pitch accent (either L or H) is phonologically associated with the stressed syllable, which for bitonal pitch accents ($L + H^*$, $L^* + H$) is realized phonetically as a difference in how the low and high f0 'targets' align relative to the stressed syllable (Fig. 1 and Supplementary Fig. 5). The stimuli were presented through ER-3C headphones driven by an audio interface through psychtoolbox in MATLAB. The stimulus intensity was calibrated for each patient prior to the start of the experiment by adjusting the volume to a level deemed loud but comfortable. Participants listened to at least 30 of the 45 audio tracks.

Stereo-electroencephalography. sEEG electrodes were inserted into the brain using robot-assisted implantation^{48,49}. Electrode trajectories were inserted along different brain regions to test seizure localization hypotheses based on non-invasive evaluations⁵⁰. Each electrode had between 8 and 18 cylindrical contacts with a length of 2 mm and a diameter of 0.8 mm. The center-to-center distance between each

electrode contact was 3.5 mm. The choice of electrode sampling (spatial resolution) across the trajectories was made based on clinical necessity. The anatomical locations of the electrode sites were estimated using high-resolution computed tomography (CT) and structural MRI. For each patient, a post-operative CT scan and a pre-operative structural MRI were jointly used to identify the precise anatomical location of each electrode. A cortical surface reconstruction was generated from the MRI using *Freesurfer*⁵¹, and electrodes were localized using coregistration between the CT and MRI and then manually marking electrode locations in the CT using *Brainstorm*, a MATLAB-based third-party application⁵².

The sEEG data were recorded with a Grapevine Nomad processor (Ripple) using the Trellis recording software. The sEEG was recorded at a sampling rate of 1000 Hz, and an online notch filter was applied at 60/120/180 Hz to reduce electrical line interference. The audio signal was synchronously recorded by the Grapevine system at a sampling rate of 30,000 Hz and was used to temporally synchronize the audio and the sEEG. The participants listened to the audio narrative *Alice's Adventures in Wonderland*, with every ~65-s segment followed by two multiple-choice comprehension questions to encourage participants' attention.

Data pre-processing. A regression-based algorithm *cleanline* implemented in EEGLAB 14.1.2b was used to clean any residual electrical line interference in the EEG data at frequencies 60, 120, 180, and 240 Hz. A Laplacian-style re-referencing approach⁵³ was applied to cleaned sEEG data, where every electrode contact in an electrode trajectory was re-referenced to its immediately neighboring contacts. Laplacian re-referencing reduces the volume conduction spread of neural activity and aids in increasing the spatial resolution. High gamma power was extracted from the re-referenced data to obtain population-level activity from neural ensembles proximal to the electrode contact. High gamma (Hy) amplitude was extracted between the range of 70–150 Hz using the *naplib* toolbox⁵⁴. A 1/7th octave filterbank was used to filter the EEG data to extract the EEG in multiple frequency bands. A Hilbert transform was then applied in each band, and the magnitude of the analytic signal in each band was extracted. The resultant signal was averaged across all the bands and resampled at 128 Hz. This signal was then z-scored based on the mean and variance in a baseline period before the onset of each audio track and was referred to as Hy activity.

Speech-responsive electrodes. Speech stimuli from the Natural Sounds stimulus set used in earlier studies^{55,56} were used to identify speech-responsive electrodes. ER-Hy responses time-locked to the onset of the speech tokens were extracted with an epoch ranging from –1000 to 2000 ms (ref: stimulus onset). The post-stimulus ER-Hy epochs were z-scored using the pre-stimulus baseline from –200 ms to –50 ms. The ER-Hy epochs were then averaged to obtain the averaged ER-Hy responses. The ER-Hy responses in the post-stimulus epoch (50–200 ms) were then statistically compared with the baseline using Wilcoxon Sign rank tests with false discovery rate correction ($q = 0.01$) to obtain electrodes that were speech responsive. All data presented in the manuscript are only from speech-responsive electrodes obtained in this step.

Pitch accent separability. High gamma responses were temporally aligned with the stimulus based on the auxiliary audio input channel. Annotated markers for pitch accents were used to segment high gamma responses into 1000 ms long epochs (–500 to 500 ms relative to the locus of the pitch accent) for each pitch accent. In contrast with phoneme onsets or word onsets where cues start at 0 ms^{2,57}, pitch accents are tightly tied to stressed syllables and are marked at their maximum realizations, which is typically an f0 peak for the categories H* and L + H*, and an f0 minimum for L* and L* + H. Since time points prior to 0 ms contain relevant information, we considered both

negative and positive epoch limits for data analysis. The epochs were further baseline corrected using a –300 to –150 ms baseline to enhance the responses elicited by the pitch accent; the function *rmbase.m* in EEGLAB 14.1.2b was used. These baseline-corrected epoch data were trimmed to –300 to 300 ms and considered the ER-Hy.

The trial-specific ER-Hy was then used to calculate pitch separability. The ratio of the within-pitch accent category variance to the variance across categories (F statistic) was calculated using the *naplib* toolbox and served as our metric of pitch accent separability. F -statistic significance was estimated using numerator degrees of freedom $k-1$ ($k = 4$, the number of pitch accents), and denominator degrees of freedom $n-k$ (n = number of ER-Hy epochs across the four pitch accents). F -statistics were calculated for every time point from –100 to 300 ms; this time range was chosen because it contained the most deviations from zero in both the mean ER-Hy and stimulus pitch contours. As the F -statistic was calculated at 58 time points, the significance level (p value) was adjusted based on the false-discovery rate (FDR) correction procedure to control the inflation of the family-wise error rate due to multiple comparisons⁵⁸ at a q value of 0.01. A conservative FDR-adjusted p value of 0.01 was used to consider each individual F -statistic as significant. Furthermore, only time points with significant p values that were contiguous (i.e., clustered) in time for at least ~31 ms (four time bins) were considered significant. This step was used to ensure that only neurophysiologically plausible data that shows temporal quasi-stationarity⁵⁹ were considered significant. The F -statistic was summed across all time points within each cluster to get a cluster statistic (F_{clus}), and maximum F_{clus} was considered as the metric of pitch accent separability. The pitch accent separability was estimated at all speech-responsive electrode contacts.

Vowel and consonant separability. Separability metrics were also estimated for vowels (vowel identity), consonants (consonant identity), and segmental features (manner and place) in the same way as the pitch accent separability metrics with degrees of freedom adjusted by the number of categories in each feature. Consonant and vowel identity refers to the individual phoneme labels that are cued by multi-dimensional features, while manner and place are features that collectively contribute to the identity of each phoneme (consonant or vowel). A total of 14 vowels were assessed (/AA/, /AE/, /AH/, /AO/, /AW/, /AY/, /EH/, /EY/, /IH/, /IY/, /OW/, /OY/, /UH/, /UW/). In total, 21 consonants were analyzed (B/, /CH/, /D/, /DH/, /ER/, /F/, /G/, /JH/, /K/, /L/, /M/, /N/, /NG/, /P/, /R/, /S/, /SH/, /T/, /TH/, /V/, /Z/). The labels were based on the ARPAbet. Ten place features were assessed (Front, Central, Back, Bilabial, Labiodental, Linguadental, Lingualveolar, Linguapalatal, Linguavelar, Glottal). Six manner features were analyzed (Plosive, Fricative, Affricate, Nasal, Liquid, Glide). Consonant manner features were defined as the degree and type of constriction or obstruction of airflow as it passes through the vocal tract when producing the consonant. Place of articulation was defined as the location in the vocal tract where a speech sound is produced based on the position of the speech articulators.

Multivariate encoding models. Multivariate forward encoding models were fit to high gamma responses across the entire narrative. Stimulus features used for the models were the *envelope*, *pitch*, and *pitch accents* (Fig. 3).

Envelope. A multiband amplitude envelope was derived from the stimulus waveform. The stimulus waveform was first filtered into 16 equally spaced bands on an ERB scale^{60–63}. A Hilbert transform was applied to the output of each filter band and the magnitude of the analytic signal was extracted to obtain the envelope at each band. The amplitude of the multiband envelope was compressed using a factor of 0.6 to mimic the compression in the inner ear. This multiband envelope was then resampled to 128 Hz to match the sampling rate of the Hy.

Pitch. Pitch accents are closely linked to the shape of the pitch contour. Thus, a pitch model was fit to the neural data to assess if the neural responses to pitch accents merely represent the pitch contours or if they represent an object-related invariant response that is informed by the pitch contours. The stimulus pitch was estimated with the autocorrelation method in Praat⁶⁴ using a pitch floor of 75 Hz, a time frame of 40 ms, and time steps of 10 ms. We used the MATLAB implementation of praat, mPraat⁶⁵, to extract the pitch contours. The pitch contours in each audio track were z-scored to get a normalized estimate of the pitch. These pitch contours were epoched in the same way as the neural data; Fig. 1d demonstrates the difference in pitch contours across the four pitch accents. However, the pitch contours were not directly used in encoding models. Generally speaking, linear models estimate unit changes in the high gamma magnitude given unit changes in an input stimulus feature. This is straightforward for the stimulus envelope, as increases in stimulus amplitudes change the magnitude of neural activity. We would not expect a similar relationship with stimulus pitch, as a reduction in stimulus pitch would not necessarily reduce the neural response. To circumvent this problem, pitch contours were discretized into ten bins, similar to the approach used by Tang et al.¹³. The discretization was performed by creating a binary vector of 0s and 1s, which signified instances when the pitch contour was within a bin's edges.

Pitch accent model. To assess whether pitch accents form discrete object-like representations beyond the spectrotemporal acoustic cues of envelope and pitch, separate models were built using pitch accent categories. Binary vectors of 0s and 1s were created with a 1 denoting the locus of each pitch accent. Four pitch accent categories (L + H*, L* + H, H*, and L*) were used.

Linear kernels (models) were estimated, which when convolved with the stimulus features resulted in predicted high gamma responses (1), and called the TRFs. Each of these models was fit separately using linear ridge regression implemented using the mTRF toolbox⁶⁶. Leave-one-out cross-validation was used for hyperparameter tuning (λ , ridge parameter) and model fit (r , Pearson's correlation between observed and model-predicted high gamma responses). The leave-one-out step trained the models on all tracks except the k th track and predicted the response to the k th track. This step was repeated until predicted responses were generated for all tracks.

$$Hy(t) = \sum w(t) \times s(t - \tau) + \varepsilon(t) \quad (1)$$

The high gamma response $Hy(t)$ is modeled as the linear convolution of time-lagged versions of the stimulus features $s(t - \tau)$ and the TRF $w(t)$, where t denotes time and τ denotes the time lag. Time lags of -300 to 300 ms were used to estimate the TRFs. TRFs were practically estimated using Eq. 2, where λ is the ridge parameter and M is a matrix that quadratically penalizes the neighboring terms in w and smoothens the TRF to avoid overfitting and produce better generalization (2).

$$w(t) = (s^t \times s + \lambda M)^{-1} \times (s^t \times r) \quad (2)$$

The observed ER-Hy was compared to the model-predicted ER-Hy, which was generated by averaging across predicted Hy responses within each pitch accent category (Fig. 3e). Three TRF models (env, env + pitch, and env + pitch + pitch accent) were fit separately to predict the Hy responses, and the ER-Hy were extracted from the predicted Hy responses. No baseline correction was performed while comparing the observed ER-Hy and the model-predicted ER-Hy to facilitate direct comparisons.

Phase-locking to periodicity. Phase-locking to the periodicity of speech stimuli was assessed by fitting the forward encoding model to map the fundamental frequency waveform of the narrative onto the

local field potential. The fundamental frequency waveform was extracted using empirical mode decomposition⁶⁷. Unlike the above sections where encoding models were fitted on the Hy responses, Laplacian referenced local field potential waveforms that were band-pass filtered from 70 to 300 Hz were used to analyze the phase-locking to the periodicity. The periodicity waveform and the neural data were segmented into successive 3000 ms segments for estimating the TRFs. This resulted in ~600 to ~900 segments. The TRFs were fitted using linear ridge regression with a lambda value of 10,000, which was empirically found to be appropriate. The time lags used for estimating the TRFs were -50 to 100 ms. TRFs were fitted for each of the segment. Leave-one-out cross-validation was performed to obtain the goodness of fit of the periodicity. This was obtained by predicting the LFP of the left-out trial using the average of TRFs across all segments except one, and this procedure was cross-validated. The goodness of fit was obtained using Pearson's correlation (r value) between the TRF-predicted local field potential and the fundamental frequency waveform to obtain phase-locking magnitude. The phase-locking magnitude was averaged across all the 3000 ms segments per electrode.

Anatomy. The cortical surfaces of each participant were reconstructed from a pre-operative MRI using Freesurfer⁵¹. The MRI was then co-registered with a post-operative CT scan, and the electrode locations were localized in Brainstorm⁵². For visualization and data analyses, the MRIs were normalized using Brainstorm's implementation of SPM12 non-linear warping. This MNI deformation field was then used to warp the Julich volumetric atlas into patient space and each channel was localized to a ROI by finding the label of the closest voxel. The electrodes that corresponded to STG, HG, ID3, and ID4 insula ROIs were used for further analysis. The electrodes in the ID3 and ID4 insula were combined with the HG for data analyses, as these showed auditory responses that potentially bled from the HG. Further, ROI labels of each electrode were manually corrected wherever appropriate.

Macaques

Electrophysiological data recording. Mesoscopic electrophysiological recordings were performed in an adult male macaque monkey (*Macaca mulatta*). The treatment of the monkey was in accordance with the guidelines set by the US Department of Health and Human Services (National Institutes of Health) for the care and use of laboratory animals. All methods were approved by the Institutional Animal Care and Use Committee at the University of Pittsburgh. The animal was 10 years old and weighed 15.7 kg at the time of the experiments.

The mesoscopic electrophysiology recording system consisted of a three-dimensional grid of 800 intracranial electrode contacts distributed across 50 multi-electrode electrode shafts arranged on a grid on the horizontal plane to cover the volume of the right hemisphere. The electrodes traversed the brain in the dorsoventral direction and penetrated the cerebrum, cerebellum, diencephalon, and parts of the mesencephalon. The inter-electrode distance varied between 0.4 and 2.7 mm between consecutive electrode contacts, depending on the length of each shaft. The locations of the electrode shafts and contacts were assessed using a CT scan with all the implanted electrode shafts in place. We then registered the D99 atlas to the T1 image of the animal. This allowed us to assign anatomical labels to each electrode contact.

The recordings were performed in a small (4 feet wide × 4 feet deep × 8 feet high) sound-attenuating and electrically insulated recording booth (Eckel Noise Control Technologies). The monkey was positioned and head-fixed in custom-made primate chairs (Scientific Design). Neural signals were recorded with a SpikeGadgets recording system at a sampling rate of 30 kHz. Stimuli were presented in the same paradigm as in humans through a custom MATLAB package *monkeylogic*. The sound files were presented through a PCI audio interface (model M-192, M-Audiophile) at a rate of 96 kHz. The analog

output of the audio interface was amplified by a 330 W amplifier (QSC GX3) and converted into sound waves using a single element 4-inch full range speaker (model W4-1879, Tang Band) located 8 inches in front of the monkey and presented at an intensity of 78 dB SPL. To synchronize the electrophysiological data with the audio, the analog audio signal from another channel of the audio interface was routed to one of the unused analog inputs of the data recording system and was used later in the offline processing.

Data processing. The electrophysiological data were downsampled to 1000 Hz, and data at each electrode were Laplacian referenced using the two closest electrodes in 3D space to enhance spatial resolution. The data processing beyond this step was the same as in the human sEEG data. During analysis, unlike the human sEEG data, the responses were not masked based on speech responsivity, as we did not have the same dataset as was used to estimate speech responsiveness in humans (responses to the Natural Sounds stimulus set). However, this was not an important factor, as only a single electrode in the monkey showed significant pitch accent separability and robust ER-Hy responses to the pitch accents.

Statistical analysis

All statistical analyses were performed in Matlab R2022a. Generalized linear mixed analysis was performed for statistical comparisons to predict the pitch accent separability based on the ROI, hemispheres, and different separability phase-locking metrics. The proportion of electrodes that encoded each feature were compared using chi-square tests of associations. All statistical comparisons considered a level of significance of 0.01. Wherever appropriate, false discovery rate corrections were applied for multiple comparisons. All tests performed were two-sided.

Variance partitioning. To evaluate the unique variance explained (R^2) by a given feature, a partial model was built in which that feature was excluded. The R^2 for this partial model was then subtracted from the R^2 for the full model (env + pitch + pitch accent)¹³ (3–5).

$$R_{pitch\ accent}^2 = R_{env + pitch + pitch\ accent}^2 - R_{env + pitch}^2 \quad (3)$$

$$R_{pitch}^2 = R_{env + pitch + pitch\ accent}^2 - R_{env + pitch\ accent}^2 \quad (4)$$

$$R_{env}^2 = R_{env + pitch + pitch\ accent}^2 - R_{pitch + pitch\ accent}^2 \quad (5)$$

The significance of the unique variance explained was calculated using an F -test ($m, N - k - 1$) with the following statistic, where k is the number of predictor variables in the full model and N is the total number of time points in the ER-Hy responses. This approach accounts for the difference in the number of predictors between the full model and the underspecified model¹³. The R^2 in each electrode in each participant was normalized by the maximum R^2 per participant for plotting (6).

$$F_{feature} = \frac{R_{feature}^2}{m} \bigg/ \frac{1 - R_{full}^2}{N - k - 1} \quad (6)$$

The same approach was used to evaluate if the addition of additional predictors statistically increased explained variance in Fig. 3e.

Multi-dimensional scaling (MDS) analysis and clustering. MDS was performed separately on the averaged ER-Hy responses to five different feature sets, namely pitch accents, vowel identity, consonant identity, manner, and place of articulation. All electrodes in each ROI (HG and STG) were used to assess how the neural activity differed between categories within a feature set in latent three-dimensional space. The epoch used for paired distances was −100 to 300 ms for the pitch accents and 0–300 ms for the segmental features. The MDS

space in all electrodes was Procrustes transformed (without scaling) to the mean of Euclidean distance averaged across all electrodes. These transformed MDS spaces were clustered using agglomerative hierarchical clustering. The number of clusters was optimized by minimizing the DBI⁶⁸. The range of number of clusters evaluated was 4–10. Statistics were performed on the cluster outputs to obtain metrics of clustering tightness, separability, and accuracy. The DBI was used as a combination metric for cluster separability and compactness.

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \{D_{i,j}\} \quad (7)$$

$$D_{i,j} = \frac{(\bar{d}_i + \bar{d}_j)}{d_{i,j}} \quad (8)$$

where \bar{d}_i is the average distance between each point in the i th cluster and the centroid of the i th cluster. \bar{d}_j is the average distance between each point in the j th cluster and the centroid of the j th cluster. $d_{i,j}$ is the Euclidean distance between the centroids of the i th and j th clusters. The maximum value of $D_{i,j}$ represents the worst-case within-to-between cluster ratio for cluster i . The optimal clustering solution has the smallest DBI value.

ARI^{69,70} was used as a metric for clustering accuracy by comparing the cluster membership with ground truth.

$$RI = \frac{TP + TN}{FP + FN + TP + TN} \quad (9)$$

where TP (True Positives) are pairs of samples that are in the same cluster in both the true and predicted partitions, TN (True Negatives) are pairs of samples that are in different clusters in both the true and predicted partitions, FP (False Positives) are pairs of samples that are in the same cluster in the predicted partition but in different clusters in the true partition, and FN (False Negatives) are pairs of samples that are in different clusters in the predicted partition but in the same cluster in the true partition. The ARI was then calculated which adjusts the Rand Index by considering the expected similarity by chance, providing a correction for random agreement. The ARI was more suitable for our application as it assesses the similarity between two clustering solutions, particularly when dealing with different numbers of clusters or when random agreement needs to be taken into account.

$$ARI = \frac{RI - \text{Expected } RI}{\max(\text{Total } RI) - \text{Expected } RI} \quad (10)$$

Bootstrapped statistics were obtained to compare the clustering metrics between ROIs. Bootstrapping was performed using 500 draws of the data with permutation (with replacement). Clustering metrics were obtained on the bootstrapped samples and independent t -tests were used to compare the metrics between ROIs. Cohen's d was estimated for effect size.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data in this study are available on request from the lead contacts. The data are not publicly available because they could compromise research participant privacy and consent.

Code availability

The code used to support the findings of this study is available on request from the lead contacts.

References

- Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
- Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
- Yi, H. G., Leonard, M. K. & Chang, E. F. The encoding of speech sounds in the superior temporal gyrus. *Neuron* **102**, 1096–1110 (2019).
- Collier, R. & Hart, J. T. The role of intonation in speech perception. in *Structure and Process in Speech Perception* (eds. Cohen, A. & Nooteboom, S. G.) 107–123 (Springer, Berlin, Heidelberg, 1975). https://doi.org/10.1007/978-3-642-81000-8_7.
- Trott, S., Reed, S., Kaliblotzky, D., Ferreira, V. & Bergen, B. The role of prosody in disambiguating English indirect requests. *Lang. Speech* **66**, 118–142 (2023).
- Shattuck-Hufnagel, S. & Turk, A. E. A prosody tutorial for investigators of auditory sentence processing. *J. Psycholinguist. Res.* **25**, 193–247 (1996).
- Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* **28**, 3976–3983.e5 (2018).
- Pierrehumbert, J. & Hirschberg, J. B. The meaning of intonational contours in the interpretation of discourse. in *Intentions in Communication* (eds. Cohen, P. R. et al.) 271–311 (MIT Press, Cambridge, MA, 1990).
- Pierrehumbert, J. B. *The Phonology and Phonetics of English Intonation* (Massachusetts Institute of Technology, 1980).
- Bhaya-Grossman, I. & Chang, E. F. Speech computations of the human superior temporal gyrus. *Annu. Rev. Psychol.* **73**, 79–102 (2022).
- Peelle, J. E., Gross, J. & Davis, M. H. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* **23**, 1378–1387 (2013).
- Hamilton, L. S., Oganian, Y., Hall, J. & Chang, E. F. Parallel and distributed encoding of speech across human auditory cortex. *Cell* **184**, 4626–4639.e13 (2021).
- Tang, C., Hamilton, L. S. & Chang, E. F. Intonational speech prosody encoding in the human auditory cortex. *Science* **357**, 797–801 (2017).
- Russell, S. M. & Golfinos, J. G. Amusia following resection of a Heschl gyrus glioma: case report. *J. Neurosurg.* **98**, 1109–1112 (2003).
- Griffiths, T. D. et al. Direct recordings of pitch responses from human auditory cortex. *Curr. Biol.* **20**, 1128–1132 (2010).
- Griffiths, T. D., Büchel, C., Frackowiak, R. S. & Patterson, R. D. Analysis of temporal structure in sound by the human brain. *Nat. Neurosci.* **1**, 422–427 (1998).
- Bendor, D. & Wang, X. Neural coding of periodicity in marmoset auditory cortex. *J. Neurophysiol.* **103**, 1809–1822 (2010).
- Hall, D. A. & Plack, C. J. Pitch processing sites in the human auditory brain. *Cereb. Cortex* **19**, 576–585 (2009).
- Bendor, D., Osmanski, M. S. & Wang, X. Dual-pitch processing mechanisms in primate auditory cortex. *J. Neurosci.* **32**, 16149–16161 (2012).
- Bendor, D. & Wang, X. The neuronal representation of pitch in primate auditory cortex. *Nature* **436**, 1161–1165 (2005).
- Beckman, M. E. & Pierrehumbert, J. B. Intonational structure in Japanese and English. *Phonol. Yearb.* **3**, 255–309 (1986).
- Beckman, M. E., Hirschberg, J. B. & Shattuck-Hufnagel, S. Chapter 2: *The Original ToBI System and the Evolution of the ToBI Framework* 9–54 <https://doi.org/10.7916/D87P97T5> (2004).
- Llanos, F., German, J. S., Gnanateja, G. N. & Chandrasekaran, B. The neural processing of pitch accents in continuous speech. *Neuropsychologia* **158**, 107883 (2021).
- Li, Y., Tang, C., Lu, J., Wu, J. & Chang, E. F. Human cortical encoding of pitch in tonal and non-tonal languages. *Nat. Commun.* **12**, 1161 (2021).
- Nourski, K. V. Auditory processing in the human cortex: an intracranial electrophysiology perspective. *Laryngoscope Investig. Otolaryngol.* <https://doi.org/10.1002/lio.2.73> (2017).
- Parvizi, J. & Kastner, S. Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.* **21**, 474–483 (2018).
- Gnanateja, G. N. et al. Frequency-following responses to speech sounds are highly conserved across species and contain cortical contributions. *eNeuro* **8**, <https://doi.org/10.1523/ENEURO.0451-21.2021> (2021).
- Bodin, C. et al. Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques. *Curr. Biol.* **31**, 4839–4844.e4 (2021).
- Erb, J., Henry, M. J., Eisner, F. & Obleser, J. The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *J. Neurosci.* **33**, 10688–10697 (2013).
- Nourski, K. V. et al. Functional organization of human auditory cortex: investigation of response latencies through direct recordings. *NeuroImage* **101**, 598–609 (2014).
- Plack, C. J., Barker, D. & Hall, D. A. Pitch coding and pitch processing in the human brain. *Hear. Res.* **307**, 53–64 (2014).
- Gander, P. E. et al. Direct electrophysiological mapping of human pitch-related processing in auditory cortex. *NeuroImage* **202**, 116076 (2019).
- Albert, A., Cangemi, F. & Grice, M. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Proceedings Speech Prosody*. Vol. 9, 13–16, (2018).
- Barnes, J., Brugos, A., Veilleux, N. & Shattuck-Hufnagel, S. On (and off) ramps in intonational phonology: rises, falls, and the Tonal Center of Gravity. *J. Phon.* **85**, 101020 (2021).
- Barnes, J., Veilleux, N., Brugos, A. & Shattuck-Hufnagel, S. Tonal Center of Gravity: a global approach to tonal implementation in a level-based intonational phonology. *Lab. Phonol.* **3**, 337–383 (2012).
- Cangemi, F., Albert, A. & Grice, M. Modelling intonation: beyond segments and tonal targets. In *Proc. 19th International Congress of Phonetic Sciences*, 572–576 (Australasian Speech Science & Technology Association Inc., Melbourne, Australia, 2019).
- D’Imperio, M. *The Role of Perception in Defining Tonal Targets and Their Alignment* (The Ohio State University, 2000).
- Khalighinejad, B. et al. Functional characterization of human Heschl’s gyrus in response to natural speech. *NeuroImage* **235**, 118003 (2021).
- Schafer, A., Carlson, K., Clifton, H. & Frazier, L. Focus and the interpretation of pitch accent: disambiguating embedded questions. *Lang. Speech* **43**, 75–105 (2000).
- Patterson, R. D., Uppenkamp, S., Johnsrude, I. S. & Griffiths, T. D. The processing of temporal pitch and melody information in auditory cortex. *Neuron* **36**, 767–776 (2002).
- Penagos, H., Melcher, J. R. & Oxenham, A. J. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *J. Neurosci.* **24**, 6810–6815 (2004).
- Schönwiesner, M. & Zatorre, R. J. Depth electrode recordings show double dissociation between pitch processing in lateral Heschl’s gyrus and sound onset processing in medial Heschl’s gyrus. *Exp. Brain Res.* **187**, 97–105 (2008).
- Berger, J. I. et al. Distribution of multiunit pitch responses recorded intracranially from human auditory cortex. *Cereb. Cortex* **33**, 9105–9116 (2023).
- Kumar, S. & Schönwiesner, M. Mapping human pitch representation in a distributed system using depth-electrode recordings and modeling. *J. Neurosci.* **32**, 13348–13351 (2012).
- Langner, G., Dinse, H. R. & Godde, B. A map of periodicity orthogonal to frequency representation in the cat auditory cortex. *Front.*

- Integr. Neurosci.* **3**, <https://doi.org/10.3389/neuro.07.027.2009> (2009).
46. Turk, A., Nakai, S. & Sugahara, M. Acoustic segment durations in prosodic research: a practical guide. In *Methods in Empirical Prosody Research*. 1–28, <https://doi.org/10.1515/9783110914641.1> (De Gruyter, 2006).
 47. Beckman, M. E. & Elam, G. A. *Guidelines for ToBI Labelling, Version 3*, Ohio State University. https://www.ling.ohio-state.edu/research/phonetics/E_ToBI/ (1997).
 48. Abel, T. J. et al. Frameless robot-assisted stereoelectroencephalography in children: technical aspects and comparison with Talairach frame technique. *J. Neurosurg. Pediatr.* **22**, 37–46 (2018).
 49. Faraji, A. H., Remick, M. & Abel, T. J. Contributions of robotics to the safety and efficacy of invasive monitoring with stereoelectroencephalography. *Front. Neurol.* **11**, 570010 (2020).
 50. Chabardes, S., Abel, T. J., Cardinale, F. & Kahane, P. Commentary: understanding stereoelectroencephalography: What's next? *Neurosurgery* **82**, E15–E16 (2018).
 51. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
 52. Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. Brainstorm: a user-friendly application for MEG/EEG Analysis. <https://doi.org/10.1155/2011/879716> (2011).
 53. Li, G. et al. Optimal referencing for stereo-electroencephalographic (SEEG) recordings. *NeuroImage* **183**, 327–335 (2018).
 54. Khalighinejad, B., Nagamine, T., Mehta, A. & Mesgarani, N. NAPLib: An open source toolbox for real-time and offline Neural Acoustic Processing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 846–850 (IEEE, 2017).
 55. Norman-Haignere, S., Kanwisher, N. G. & McDermott, J. H. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88**, 1281–1296 (2015).
 56. Rupp, K. et al. Neural responses in human superior temporal cortex support coding of voice representations. *PLoS Biol.* **20**, e3001675 (2022).
 57. Di Liberto, G. M. & Lalor, E. C. Indexing cortical entrainment to natural speech at the phonemic level: methodological considerations for applied research. *Hear. Res.* **348**, 70–77 (2017).
 58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
 59. Brunet, D., Murray, M. M. & Michel, C. M. Spatiotemporal analysis of multichannel EEG: CARTOOL. *Comput. Intell. Neurosci.* **2011**, 1–15 (2011).
 60. McHaney, J. R., Gnanateja, G. N., Smayda, K. E., Zinszer, B. D. & Chandrasekaran, B. Cortical tracking of speech in delta band relates to individual differences in speech in noise comprehension in older adults. *Ear Hear.* **42**, 343–354 (2021).
 61. Dial, H. R. et al. Cortical tracking of the speech envelope in logopenic variant primary progressive aphasia. *Front. Hum. Neurosci.* **14**, 597694 (2021).
 62. Reetzke, R., Gnanateja, G. N. & Chandrasekaran, B. Neural tracking of the speech envelope is differentially modulated by attention and language experience. *Brain Lang.* **213**, 104891 (2020).
 63. Quique, Y. M., Gnanateja, G. N., Dickey, M. W., Evans, W. S. & Chandrasekaran, B. Examining cortical tracking of the speech envelope in post-stroke aphasia. *Front. Hum. Neurosci.* **17**, 1122480 (2023).
 64. Boersma, P. Praat, a system for doing phonetics by computer. *Glottol. Int.* **5**, 341–345 (2002).
 65. Bořil, T. & Skarnitzl, R. Tools rPraat and mPraat. in *Text, Speech, and Dialogue* (eds. Sojka, P., et al.) 367–374 (Springer International Publishing, Cham, 2016). https://doi.org/10.1007/978-3-319-45510-5_42.
 66. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016).
 67. Forte, A. E., Etard, O. & Reichenbach, T. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife* **6**, e27203 (2017).
 68. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).
 69. qqfssxx. *Rand and Adjusted Rand Index Calculator for Cluster Analysis - File Exchange - MATLAB Central* (Matab Central File Exchange, 2024).
 70. Warrens, M. J. & van der Hoef, H. Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *J. Classif.* **39**, 487–509 (2022).

Acknowledgements

We thank the participants, families, and epilepsy monitoring unit staff. This investigation was supported by the National Institutes of Health Grants; R01-DC-013315 to B.C., T.J.A., and T.T., and R21DC019217-01A1 awarded to T.J.A., the Vice-Chancellors Research and Graduate Education, and College of Letters and Sciences UW Madison Funds to G.N.G. This work, partially carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). This work was also supported in part by the resources provided through the University of Pittsburgh Center for Research Computing. Preliminary findings from this investigation were presented at the Association of Research in Otolaryngology Mid-Winter Meeting 2022.

Author contributions

G.N.G., K.R., T.J.A., and B.C. designed research. K.R. and T.J.A. performed data collection in humans. T.T. performed data collection in Macaque. G.N.G. designed analytic tools for iEEG data processing. K.R. designed analytic tools for structural MRI processing and visualization. F.L., J.G., and B.C. contributed to stimulus annotation. G.N.G. and B.C. analyzed the data. G.N.G., K.R., J.H., F.L., J.G., T.T., T.J.A., and B.C. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-56779-w>.

Correspondence and requests for materials should be addressed to Taylor J. Abel or Bharath Chandrasekaran.

Peer review information *Nature Communications* thanks Daniela Sammler, who co-reviewed with Seung-Cheol Baek, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025