NEUROSCIENCE

Encoding of melody in the human auditory cortex

Narayan Sankaran[1], Matthew K. Leonard[1], Frederic Theunissen[2], Edward F. Chang[1]*

Melody is a core component of music in which discrete pitches are serially arranged to convey emotion and meaning. Perception varies along several pitch-based dimensions: (i) the absolute pitch of notes, (ii) the difference in pitch between successive notes, and (iii) the statistical expectation of each note given prior context. How the brain represents these dimensions and whether their encoding is specialized for music remains unknown. We recorded high-density neurophysiological activity directly from the human auditory cortex while participants listened to Western musical phrases. Pitch, pitch-change, and expectation were selectively encoded at different cortical sites, indicating a spatial map for representing distinct melodic dimensions. The same participants listened to spoken English, and we compared responses to music and speech. Cortical sites selective for music encoded expectation, while sites that encoded pitch and pitch-change in music used the same neural code to represent equivalent properties of speech. Findings reveal how the perception of melody recruits both music-specific and general-purpose sound representations.

INTRODUCTION

Musical communication is a hallmark of human behavior that requires listeners to extract multiple time-varying features from a dynamic acoustic signal. This process leverages core spectrotemporal organizing principles of the human auditory system (*1*–*3*) and may share evolutionary roots with language (*4*–*6*).

While music varies in its structure across cultures and genres, a defining feature across popular idioms is the serial arrangement of discrete pitch-units—or notes—to produce the emergent percept of melody (Fig. 1A). Considered in isolation, each constituent note within melody has a pitch, which is perceived along a low-to-high continuum according to its fundamental frequency (F_0). Considered within its melodic context, these notes are imbued with higher-order attributes, reflecting the integration of prior information at progressively longer time scales (*4*, *7*–*9*). For instance, the magnitude and direction of pitch-change between adjacent notes define the melodic interval and contour, respectively, linking the isolated sensory attribute of pitch with our perceptual experience of melody (*10*–*12*). Furthermore, listeners with prior exposure are familiar with the statistical structure of Western music, and use this knowledge to generate expectations about the likelihood of upcoming notes conditioned on the prior sequence. In Fig. 1A, for example, the third-to-last note is relatively unexpected, violating the pattern and tonality established earlier in the phrase. The continuum along which melody violates or fulfills our expectations plays a central role in our aesthetic experience of music (*13*–*19*), and composers will intentionally modulate expectations to systematically generate patterns of tension and resolution as we listen.

Although these melodic features—pitch, pitch-change, and expectation—convey distinct percepts, they derive from the same pitch-related information computed over progressively longer temporal windows. This raises fundamental questions about their representation in the brain, such as whether they are spatially dissociable. Specifically, are different features of melody selectively encoded by distinct neural populations (*20*–*21*), or do the same populations jointly encode melodic features by modulating afferent pitch representations (*22*, *23*)? This question has major implications for understanding how the brain represents information at different time scales within dynamic input streams (*24*).

In addition, the extent to which music is encoded by general auditory mechanisms versus ones specialized for music remains debated (*25*–*30*). Recent work has found subregions in the human superior temporal gyrus (STG) that selectively respond to music over other sounds such as speech (*31*–*33*). While this is thought to reflect a specialized neural pathway for "music," it remains unclear which musical properties drive this selectivity. Resolving this question is critical for understanding the nature and extent of specialization in the human brain for important domains of sound.

To address these questions, we used high-density electrocorticography (ECoG) to record neural activity on the human cortical surface while Western participants listened to a set of Western monophonic musical phrases (see Materials and methods for stimulus details). These direct high-density recordings are necessary to resolve fine-grained spatial tuning over millimeters of cortex to dynamic information changing over milliseconds. Within auditory cortex, we characterized the encoding of melodic pitch, pitch-change, and expectation and determined the extent to which these properties were encoded within separate or overlapping neural populations. The same participants also listened to natural speech, and we determined whether melodic feature encoding was specific to music or shared across domains.

RESULTS

To examine the neural encoding of melody, we created a naturalistic stimulus set consisting of 208 short musical phrases of varied instrumentation (audio S1). Stimuli were designed to vary along three fundamental pitch-related dimensions (Fig. 1A): (i) the absolute pitch (based on the fundamental frequency; F_0) of each note, (ii) the pitch-change between adjacent notes, and (iii) the expectation of each note conditioned on prior notes in the phrase. Expectation was calculated using a pretrained recurrent neural network [MelodyRNN; (*34*)] to estimate the surprisal of notes (negative $\log_2$ likelihood). Although these three measures are partially correlated, they contained sufficient independent variation for us to probe the extent to which they were independently encoded in the human auditory cortex

[1]Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA. [2]Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA.
*Corresponding author. Email: edward.chang@ucsf.edu

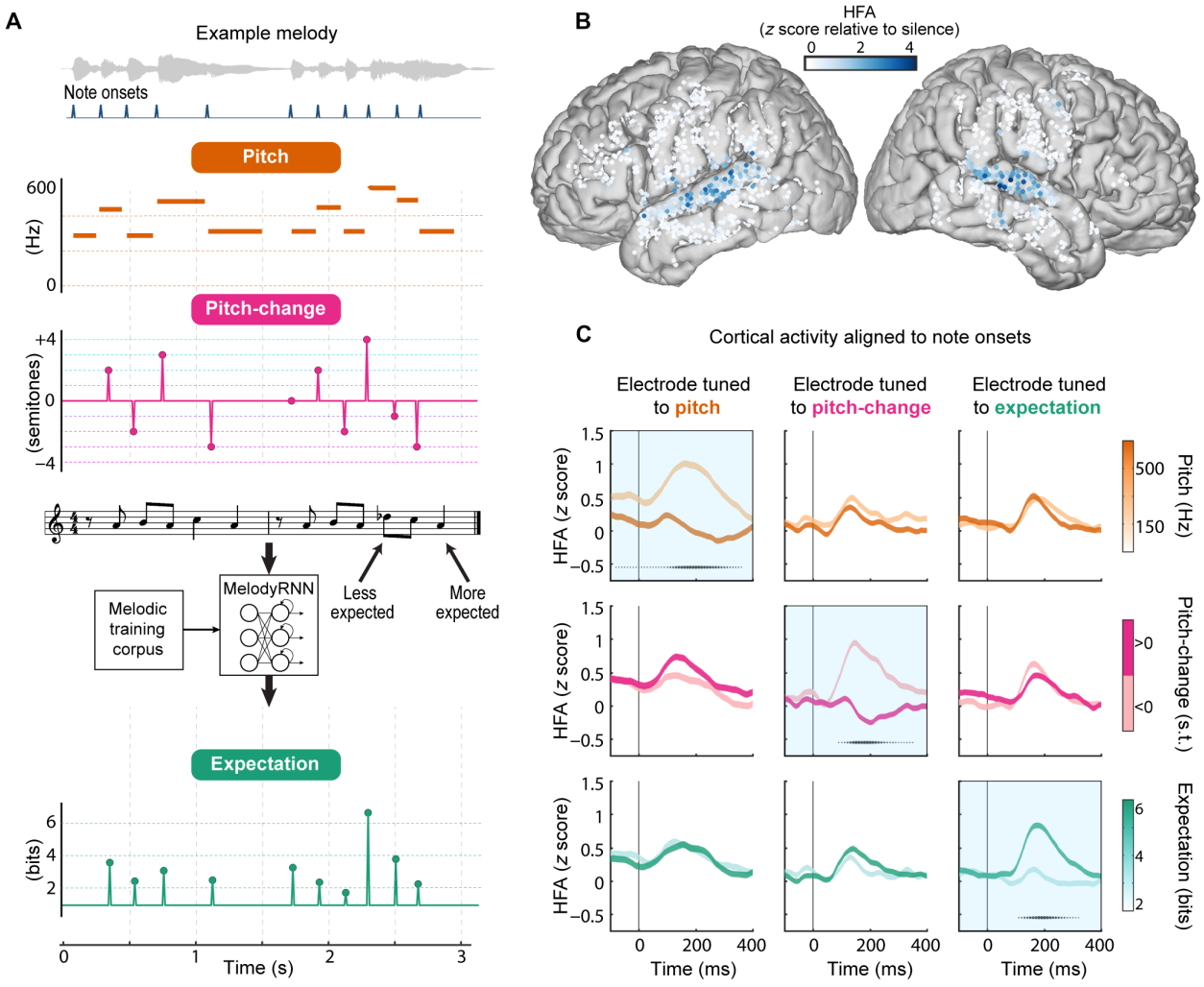

Fig. 1. Melodic pitch, pitch-change, and expectation modulate STG activity during music listening. (**A**) Three melodic features visualized for an example melody: (i) Absolute pitch (measured in Hz), defined by the fundamental frequency (F_0) of each discrete note. (ii) Pitch-change between adjacent notes. The magnitude of change (interval) is measured in semitones, while the direction of change (contour) is binary. (iii) Melodic expectation (measured in bits) indicates the surprisal of each note (negative $\log_2$ likelihood) conditioned on prior notes. Expectation was measured using a pretrained model of Western melody (MelodyRNN). (**B**) Electrodes across all participants ($N = 8$) plotted on a common brain. Color indicates the peak evoked high-frequency activity (HFA) averaged across all musical phrases. (**C**) Responses at three example electrodes demonstrating distinct tuning to pitch (left column), pitch-change (middle column), and expectation (right column). Each feature distribution is divided into two equal bins (median split). Traces indicate the mean ± SE of cortical responses within each bin. Black markers underneath traces indicate time points during which responses in the two bins significantly differed ($P < 0.001$; independent two-sample t test, Bonferroni corrected; marker size indicates t-statistic magnitude).

(fig. S1, Pearson's correlation: pitch versus pitch-change: $r = 0.14$; pitch versus surprisal: $r = -0.064$; pitch-change versus surprisal: $r = 0.025$).

Eight participants listened to musical stimuli while we recorded ECoG activity from high-density arrays placed over the lateral surface of the cortex. To identify music-responsive cortical sites, we extracted the average high-frequency activity (HFA; 70 to 150 Hz) at each electrode during presentation of musical phrases (Fig. 1B). We observed activity primarily in the bilateral STG, spanning posterior to mid-anterior subregions. Across all participants, we identified 224 electrodes with significant music-evoked responses relative to a silent baseline period ($P < 0.01$, Wilcoxon signed-rank tests, Bonferroni corrected).

Distinct neural populations encode pitch, pitch-change, and expectation in melody

Next, we examined the extent to which music-responsive neural populations encoded relevant melodic information. At three example electrodes, we aligned cortical activity to note onsets (Fig. 1C) and compared evoked responses to notes that had contrasting values of pitch (high versus low), pitch-change (ascending versus descending), or expectation (high versus low). At one electrode (Fig. 1C, left column), responses differentiated notes in distinct pitch ranges but did not differentiate notes with contrasting pitch-change or expectation values. At a different electrode (Fig. 1C, middle column), responses differentiated descending from ascending pitch-changes but not contrasts in either pitch or expectation. Last, at a third electrode

(Fig. 1C, right column), responses differentiated the expectation of a note but not its pitch or pitch-change. These response patterns demonstrate the sensitivity of local neural populations to distinct dimensions of melody (see fig. S2 for all electrodes).

To quantify the encoding of these melodic dimensions at each electrode, we used temporal receptive field (TRF) modeling, which predicts continuous neural activity from a set of stimulus features (*35*). In addition to the three melodic features of interest, we included the stimulus spectrogram and temporal landmarks indicating phrase and note onsets as predictors in TRF models. The inclusion of these additional predictor variables statistically controls for the cortical processing of features unrelated to—but potentially correlated with—pitch, pitch-change, and expectation (see fig. S3 for full TRF predictor matrix).

Across electrodes, melodic and acoustic features explained a substantial portion of the variance in neural activity [max R^2 (coefficient of determination) = 0.35, mean $R^2 = 0.12$], with the highest R^2 values in an example participant located at cortical sites in the mid-to-anterior STG (Fig. 2A). Performance was particularly high relative to the upper limit defined by each electrode's noise ceiling (*36*, *37*), with models, on average, predicting 70% of the explainable variance.

To determine the extent to which specific melodic features were encoded in single electrode activity, we computed the unique variance (ΔR^2) explained by pitch, pitch-change, and expectation within TRF models (see Materials and Methods). Within the STG of individual participants, we found electrodes that significantly encoded all three features ($P < 0.01$, permutation tests). Crucially, encoding at single electrodes tended to be dominated by a singular melodic feature, with little-to-no encoding of the other two features (Fig. 2B).

Across all participants, the three melodic features were encoded within largely nonoverlapping subsets of electrodes, with 80% of electrodes significantly tuned to a singular feature. Directly comparing the encoding of pitch, pitch-change, and expectation across electrodes (Fig. 2C), we found that the ΔR^2 attributed to any one feature was highly orthogonal to the ΔR^2 attributed to the other two features (linear mixed-effects, random effects grouped by participant, all β not significantly different from zero, all $t < 2.65$, all $P > 0.05$). Thus,

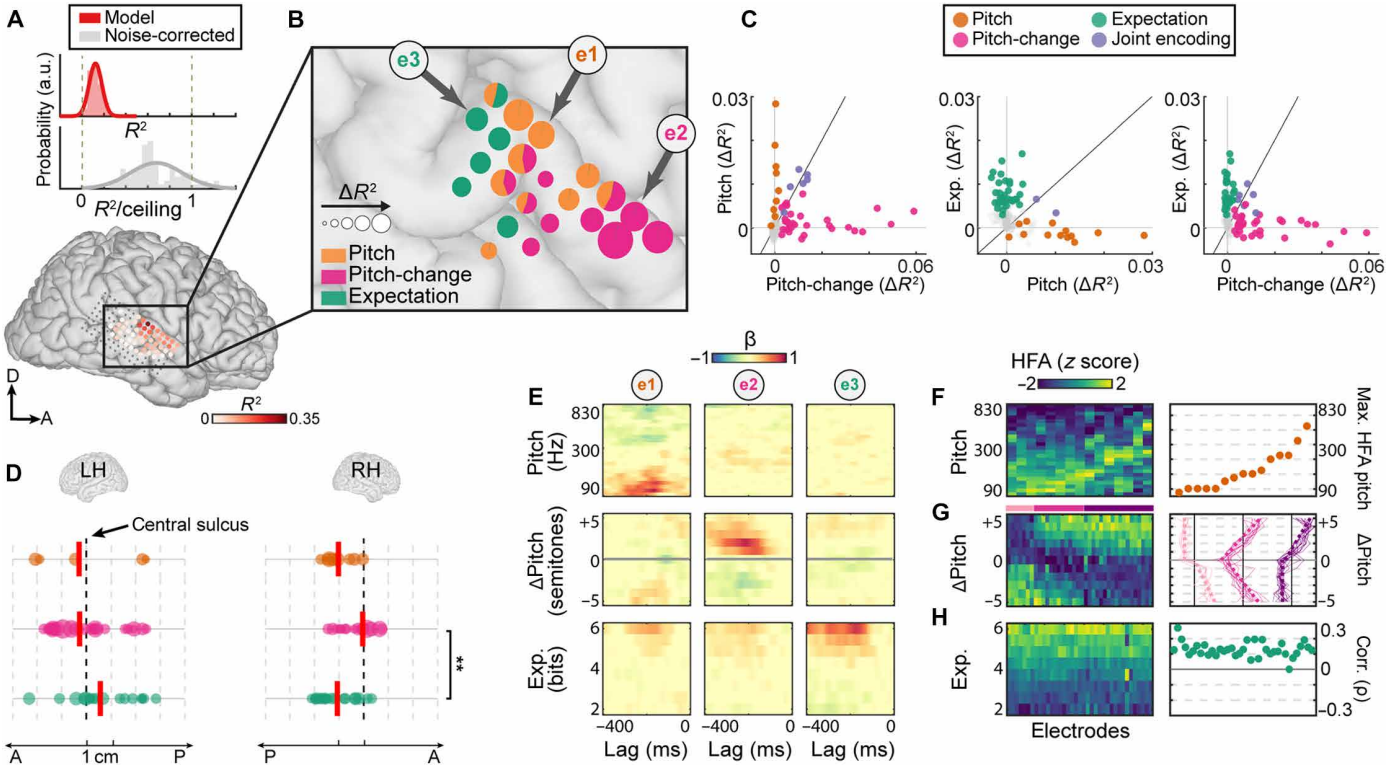

Fig. 2. Separate neural populations encode melodic pitch, pitch-change, and expectation. (**A**) Top: Distribution of TRF model and noise-corrected R^2 values across electrodes. Noise-corrected values are obtained by dividing model R^2 values by estimates of the noise ceiling. Bottom: Model R^2 values (indicated by darkness of markers) and corresponding neuroanatomical locations for an example participant. (**B**) Unique variance (ΔR^2) explained by pitch, pitch-change, and expectation. Size of pie charts indicate the magnitude of ΔR^2. (**C**) Scatter plots comparing ΔR^2 explained by two given features. Electrodes are represented by markers and pooled across all participants, with colors indicating features with significant ΔR^2 (permutation tests, $P < 0.01$). (**D**) Electrode locations along the posterior-anterior axis of each hemisphere, normalized across participants to a common anatomical landmark. Marker size indicates normalized ΔR^2. Vertical red lines indicate the weighted-mean location of feature encoding. Asterisks indicate significance level: $^{**}P < 0.01$. (**E**) TRF weights for three example electrodes whose locations are shown in (B). Weights indicate tuning to low pitch (e1), ascending pitch-change (e2), and unexpected notes (e3). (**F**) Tuning to pitch. Matrix columns correspond to electrodes with significant ΔR^2 explained by pitch, ordered by F_0 of peak response. Right: Orange markers indicate F_0 at which peak response tuning is observed. (**G**) Tuning to pitch-change. Electrodes are grouped by clusters (*k*-means, $k = 3$) with colored bars above matrix indicating cluster membership. Right: Tuning profiles for electrodes within each cluster. Thin solid lines indicate individual electrode tuning. Thick dashed lines indicate linear functions fit separately on ascending and descending ranges of pitch-change. (**H**) Tuning to expectation. Right: For each electrode, green markers indicate rank-order correlations between HFA and expectation across all notes. a.u., arbitrary unit. LH, left hemisphere; RH, right hemisphere.

melodic features were not jointly encoded within the same neural populations. Rather, each feature was selectively encoded within a distinct subpopulation in the STG.

We next asked whether the distinct subpopulations responsible for encoding pitch, pitch-change, and expectation were anatomically organized. We found that the three features were represented within highly overlapping regions along the posterior-to-anterior axis of the STG (Fig. 2D; see Fig. 2B and fig. S3B for organization within individual participants). In the right hemisphere, pitch-change was encoded anterior to expectation (linear mixed effects, randomized block design, Bonferroni corrected: $t_{31} = 3.50$, $P = 0.0056$). Despite this spatial difference, the representation of melodic features did not strongly segregate into anatomically distinct subregions. Rather, the independent encoding of pitch, pitch-change, and expectation occurred in an interdigitated subject-specific spatial map across the STG.

Beyond determining whether auditory cortical populations encode important dimensions of melody, we sought to specify the format in which this information is coded. To answer this, we first inspected model weights (Fig. 2E) for three example electrodes that had significant ΔR^2 explained by pitch (e1), pitch-change (e2), and expectation (e3). Consistent with their corresponding ΔR^2 values (Fig. 2C), weights were concentrated over a single feature at each electrode and revealed clear patterns of tuning to low pitch at e1, ascending pitch-changes at e2, and more unexpected notes at e3.

Next, we focused on the specific format in which pitch was coded. Across all electrodes with significant ΔR^2 explained by pitch, we characterized patterns of response tuning by visualizing the modulation in activity across the F_0 range spanned by the stimulus (Fig. 2F). This revealed a diversity of broad tuning profiles, with maximal responses tiling the F_0 range from low (90 Hz) to mid-high (300 to 700 Hz) values.

While we used F_0 as a proxy for pitch, pitch perception is influenced by several other acoustic properties that are correlated with F_0 (*38*, *39*), including the centroids of the spectral and spectral-modulation profiles ($r = 0.75$ and $r = -0.4$, respectively, across all notes in the current stimulus). We therefore asked whether these two alternate spectral features could explain pitch encoding above and beyond the F_0. First, we examined F_0 tuning in different stimulus subsets with distinct spectral profiles. We found highly consistent F_0 tuning across spectrally distinct stimuli (fig. S4). Next, we used a variance partitioning approach to determine whether spectral modulations could explain neural responses better than F_0. However, model R^2 values were significantly higher when using F_0 rather than modulations ($t_{14} = 4.6$; $P = 3.9 \times 10^{-4}$). In addition, F_0 continued to explain ΔR^2 even when TRF models controlled for spectral modulations (fig. S5). Thus, while pitch perception is influenced by multiple acoustic attributes, the pitch encoding we currently observe is best accounted for by F_0 and cannot be explained by the spectral profile or spectral modulations in the stimulus.

We next examined tuning across electrodes with significant ΔR^2 explained by pitch-change, visualizing the modulation in activity across the range of pitch-changes spanned by the stimulus (Fig. 2G and fig. S8A). Prior behavioral research suggests separability in the representation of a pitch-change's precise magnitude (interval) and general direction (contour) (*10*, *12*). In contrast, we found tuning patterns at local populations that represented both contour and interval information. Most electrodes (63%) had a rectified-linear tuning profile, with responses positively modulated by the magnitude of change within a given direction (ascending or descending) to which an electrode was selective (Fig. 2G, left and right clusters, *k*-means clustering). In addition, we found a smaller subset of electrodes that responded proportionally to interval magnitude regardless of direction (Fig. 2G, central cluster). Thus, STG populations tuned to pitch-change represented both melodic contour and interval information.

Psychophysical evidence suggests that listeners detect pitch-changes by tracking either the F_0 of adjacent notes or their individual harmonic components (*38*, *40*). We sought to determine which of these two possibilities drove the pitch-change tuning observed in ECoG activity. First, we divided the musical stimulus into two distinct subsets: one in which harmonics provided an unambiguous cue to the direction of pitch-change, and another in which they did not (leaving only the F_0 as a reliable cue; see Materials and Methods). Crucially, we found that tuning to pitch-change was consistent across these two subsets (fig. S6). To further dissociate harmonic cues from those based on the F_0, we created two sets of artificial melodic stimuli. One set was composed of harmonic complex tones, while another comprised tones in which higher-order components were jittered, rendering them inharmonic and lacking a reliable F_0. We identified an electrode at which activity was modulated by pitch-change in both original melodies and harmonic stimuli. Crucially, activity at this electrode was not modulated by pitch-change for inharmonic stimuli (fig. S7). Together, these findings suggest that the encoding of pitch-change in STG derives from the detection of changes in F_0 across successive notes.

Last, we examined the encoding of expectation, which reflects the degree to which successive notes in melody conformed to or departed from sequential patterns and learnt structural rules of Western tonal music (fig. S9). Across electrodes with significant ΔR^2 explained by expectation, we consistently found a monotonic relationship, whereby more unexpected notes evoked stronger responses (Fig. 2H). These results demonstrate that, in higher-order auditory cortex, perception of melody recruits multiple anatomically and functionally independent subpopulations, each selectively tuned to information along a different pitch-related dimension, spanning basic spectral to time-integrated and statistical structure.

Music-selective activity reflects encoding of melodic expectation

Beyond identifying how relevant information is encoded, a major goal of auditory neuroscience is to determine the nature and extent of specialization in the human brain for music compared with other acoustically complex and behaviorally relevant sounds such as speech. While recent work has found subregions in the STG that are selectively activated by music over other sounds (*31*–*33*), the underlying information to which these subregions are tuned remains unclear.

To address this question, we presented the same eight participants with naturally spoken English sentences (audio S2). We hypothesized that music selectivity reflects tuning to information that exclusively exists within music. Specifically, while pitch and pitch-change are acoustic properties that describe information that exists across different domains, melodic expectation describes the unique sequence structure of music. We therefore predicted that the degree to which populations selectively respond to music (over speech) directly reflects the extent to which they encode expectation.

To identify music selective electrodes, we first compared the relative magnitude of music and speech responses (Fig. 3A). From this, we derived a selectivity index (SI) ranging from −1 (speech selective)

to +1 (music selective), which quantifies the degree to which a given electrode preferentially responded to a given domain. While most electrodes responded to both music and speech, we nevertheless identified a substantial number of electrodes that were consistently and strongly selective for musical phrases over spoken utterances (Fig. 3B).

Anatomically, music-selective electrodes (defined as electrodes with SI > 0.2; see Materials and Methods) were widely distributed and interdigitated with other sound-responsive populations in the STG (Fig. 3C). The music selectivity of electrodes was independent of location in the right hemisphere (linear mixed effects, randomized block design, Bonferroni corrected: $t_{123} = 1.97$, $P = 0.13$) and weakly biased toward posterior regions in the left hemisphere ($t_{130} = 2.45$, $P = 0.032$). Further, within-subject comparisons revealed only one participant in which the spatial distribution of music selective electrodes differed from that of nonselective electrodes (Wilcoxon rank sum tests, $Z = -3.03$, $P = 0.014$ in one participant, all other $P > 0.05$). Thus, we find little evidence for the clustering of music selectivity into a dedicated auditory subregion.

While the existence of music selective populations is consistent with prior studies (*31*), we next sought to evaluate our key hypothesis—that music selectivity specifically reflects the encoding of melodic expectation. We visualized the strength of expectation encoding (ΔR^2) as a function of domain selectivity (Fig. 3D). Consistent with our hypothesis, electrodes that encoded expectation were almost exclusively more responsive to music than speech. Furthermore, across all music-selective electrodes, the degree of expectation-encoding predicted the magnitude of music selectivity (Fig. 3E; partial correlation, $r = 0.31$, $P < 0.001$, permutation test). In contrast, there was no systematic relationship between selectivity and the encoding of pitch or pitch-change (pitch: $r = -0.02$, $P = 0.6$, contour: $r = 0.006$, $P = 0.47$. fig. S10).

As linguistic sequences can also be characterized by their statistical structure, and prior studies have shown robust encoding of phoneme-based expectation in the STG (*41*, *42*), we next sought to test the analogous hypothesis for populations that were speech selective. Using the same TRF modeling approach as before, we quantified the encoding of a set of acoustically and perceptually relevant speech

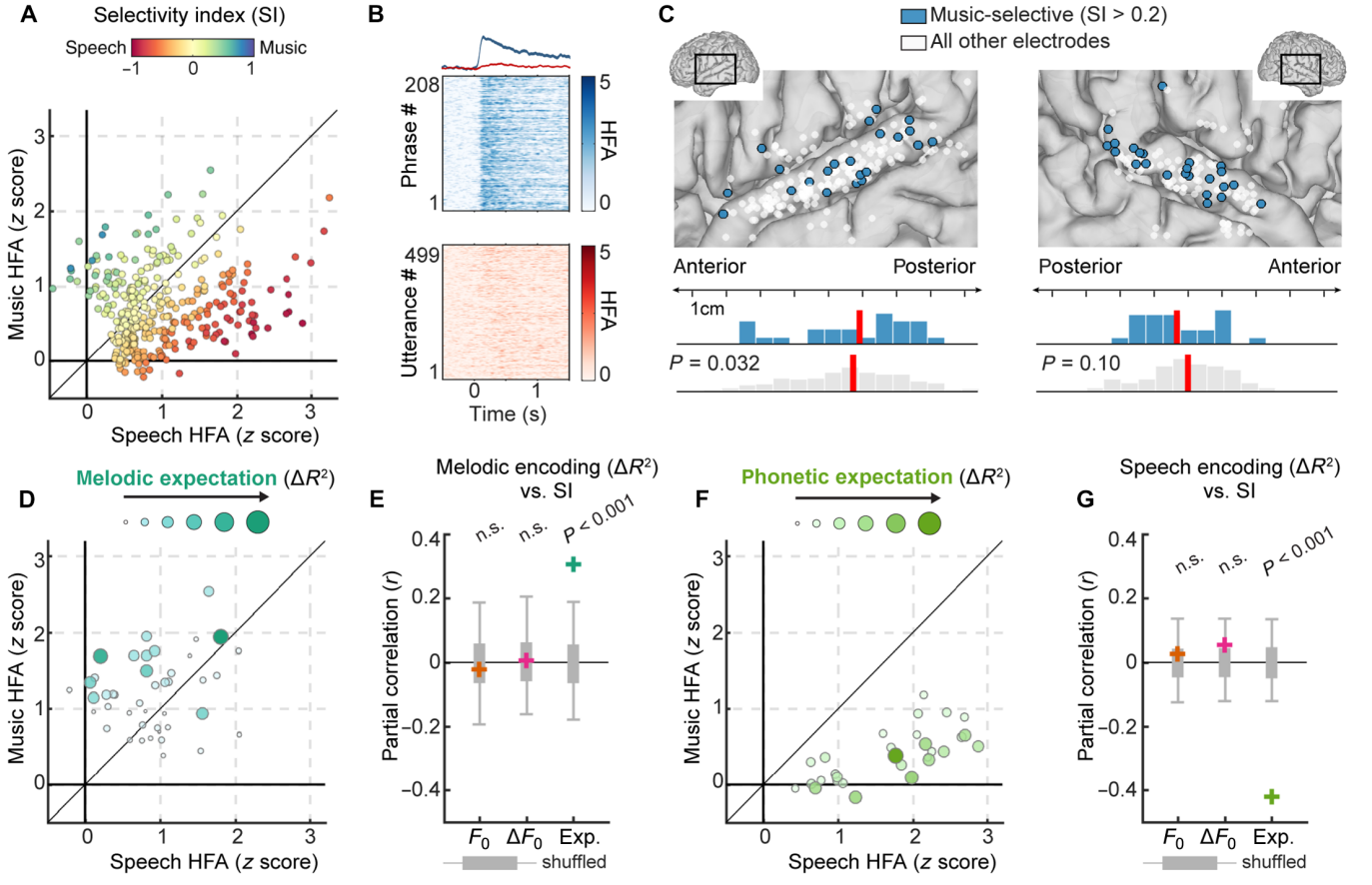

Fig. 3. Music selective activity reflects the encoding of melodic expectation. (**A**) Average music versus speech responses for all electrodes. Marker colors indicate SI. (**B**) Single-trial rasters for an example electrode that demonstrates selective responses to music (blue) over speech (red). (**C**) Anatomical location of music-selective electrodes (SI > 0.2; blue markers) indicating broad distribution throughout STG. Histograms indicate distribution of music-selective electrodes relative to all other electrodes along the posterior-anterior axis. Vertical red lines indicate median location of distributions. (**D**) Average music versus speech responses for electrodes that encode melodic expectation [axes are identical to (A)]. Marker size and color indicate ΔR^2 explained by expectation in TRF models. (**E**) Colored crosses indicate the partial correlation between SI and encoding of pitch (orange), pitch-change (magenta), and expectation (dark green) across all electrodes with SI ≥ 0. Gray error bars indicate 95th percentiles of permutation tests. (**F**) Average music versus speech responses for electrodes that encode phoneme-based expectation in speech. Marker size and color indicate ΔR^2 explained by phonetic expectation in TRF models of speech-evoked activity. (**G**) Partial correlation between SI and speech encoding across all electrodes with SI ≤ 0. n.s., not significant.

features in sentence-evoked activity (see Materials and Methods for details). Mirroring music, we found that phoneme-based expectation was encoded by speech selective electrodes (Fig. 3F), with the degree of encoding predicting the magnitude of speech selectivity (Fig. 3G; expectation: $r = -0.42$, $P < 0.001$, pitch: $r = -0.026$, $P = 0.65$, contour: $r = 0.054$, $P = 0.77$). Thus, rather than a domain-general mechanism for representing auditory sequence statistics, the relevant statistics of music and speech are encoded within two independent substrates of the STG.

So far, results suggest that selectivity for music reflects the encoding of its higher-order sequence structure, rather than its lower-order acoustic properties. However, music has an acoustic structure that is inherently different from that of speech, particularly in terms of the spectral and temporal modulation patterns to which STG populations are sensitive (*43*–*46*). We therefore sought to explicitly test whether low-level acoustic differences—rather than expectation—explain music selectivity. To do so, we created a third stimulus that was acoustically identical to speech except that it contained the pitch structure of melody (Fig. 4, A to C, and audio S2 and S3). Specifically, for every speech utterance, we created a "melodic speech" counterpart by warping the continuous pitch within each syllable onto the nearest discrete Western musical-scale tone. This manipulation left all other acoustic features, such as vowel formants and amplitude envelopes, unchanged (compare Fig. 4, A and B). As a result, the spectrograms of speech and melodic speech signals were highly similar, with correlations close to 1 along both the spectral ($r = 0.998$) and temporal ($r = 0.995$) axis. For a subset of original sentences (118 of 499), their melodic speech counterparts formed coherent Western diatonic melodies, which we determined using a musical key-finding algorithm (*47*, *48*).

To evaluate the perception of melodic speech, an independent cohort of listeners ($N = 11$) rated the extent to which they heard a melody within each token on a scale ranging from 0 (sounds such as regular speech) to 10 (sounds such as a song). We found that all tokens were perceived to contain melody, albeit with wide variation in the degree to which this was the case (Fig. 4D). Furthermore, ratings were consistent across listeners (average inter-rater reliability: $r = 0.35$, $P = 2.3 \times 10^{-27}$, permutation test), suggesting that they served as a reliable proxy for the degree to which ECoG participants experienced melodic speech as melody.

A subset of ECoG participants ($N = 2$) who previously heard music and speech stimuli were also presented with melodic speech. At music-selective electrodes, we first asked whether melodic speech elicited similar responses to music. If so, this would specifically implicate information imparted by the morphing of original speech into discrete musical tones—and not other spectrotemporal features—as the basis of music selectivity. For an example music-selective electrode, melodic speech elicited robust responses that were qualitatively similar in magnitude to those evoked by music and stronger than those evoked by speech (Fig. 4E). Such an enhanced response to melodic speech versus regular speech was primarily observed at music-selective electrodes (Fig. 4F; one-sample *t* tests; music-selective: $t_{11} = 3.99$, $P = 0.0011$; shared: $t_{46} = 1.12$, $P = 0.13$; speech-selective: $t_{35} = 0.27$, $P = 0.39$) and was positively correlated with electrodes' SI ($r = 0.66$; $P = 1.8 \times 10^{-6}$). Thus, inserting the pitch structure of melody into a stimulus that otherwise had the spectrotemporal characteristics of speech was sufficient to elicit music-like responses at music-selective electrodes.

We next asked whether enhanced responses to melodic (versus regular) speech specifically arose from the encoding of melodic expectation. Because listeners were unlikely to develop melodic expectations for stimuli that were not clearly perceived as melody, we first asked whether the magnitude of melodic-speech responses scaled with ratings of perceived melodiousness. At an example music-selective electrode, we observed a positive correlation between ratings and responses ($r = 0.24$; $P = 0.01$; Fig. 4G). This correlation was observed across all music-selective electrodes (one sample *t* tests; $t_{11} = 2.66$, $P = 0.011$; Fig. 4H) but not at shared ($t_{46} = 0.55$, $P = 0.29$) or speech-selective electrodes ($t_{35} = -1.96$, $P = 0.97$). These results imply that music-selective responses were driven by information only available to the extent that stimuli were perceived as melody.

Last, to explicitly examine how responses to melodic speech were modulated by features of melody, we extracted the pitch, pitch-change, and melodic expectation of each token (again using melodyRNN to extract expectation). We divided the distribution of these three features into two equal bins (median split) and examined the corresponding neural responses in each bin. At an example music-selective electrode, responses to melodic speech were significantly modulated by expectation (independent two-sample *t* tests; $P < 0.05$ from 90 to 210 ms after note onset). However, we only observed this modulatory effect for tokens that induced a relatively strong percept of melody (Fig. 4I, top versus bottom row). In contrast, responses were not modulated by pitch or pitch-change (independent two-sample *t* test, $P > 0.05$). Mirroring natural music, the degree to which expectation modulated melodic speech activity predicted electrode SI, while modulation due to pitch or pitch-change did not systematically vary with selectivity (partial correlations, significance evaluated via permutation tests; pitch: $r = -0.011$, $P = 0.55$; pitch-change: $r = -0.11$, $P = 0.79$; expectation: $r = 0.28$, $P = 0.016$; Fig. 4J). Together, the above results indicate that, when listening to both music and melodic speech alike, domain-selective activity specifically reflects the encoding of melodic expectation.

Encoding of pitch and pitch-change is shared across music and speech

Having established that the encoding of melodic expectation is functionally specialized for music, we next probed whether lower-level dimensions of melody—pitch and pitch-change—are represented by domain-general populations using a shared neural code across music and speech. First, we characterized speech along dimensions that are acoustically equivalent to melodic pitch and pitch-change. To do so, we extracted the pitch contour of each sentence and computed the suprasegmental changes between the median pitch of adjacent syllables (Fig. 5, A and B). Although not identical, distributions of pitch and pitch-change were highly overlapping across the two domains (Fig. 5C).

Next, for electrodes that significantly encoded pitch or pitch-change in music, we examined the extent to which they encoded information along the same dimension of speech. As with music, we computed the unique variance (ΔR^2) explained by each feature within TRF models of speech-evoked activity. We then directly compared ΔR^2 values across the two domains (Fig. 5D). We found strong correlations in the extent to which electrodes encoded either pitch ($r = 0.82$, $P = 2.3 \times 10^{-4}$) or pitch-change ($r = 0.79$, $P = 3.1 \times 10^{-8}$) across the two stimulus domains, indicating that neural populations strongly tuned to a given dimension of melody were generally tuned to the same dimension of speech.

While the above findings suggest that the same populations encode pitch and pitch-change across domains, we next probed the

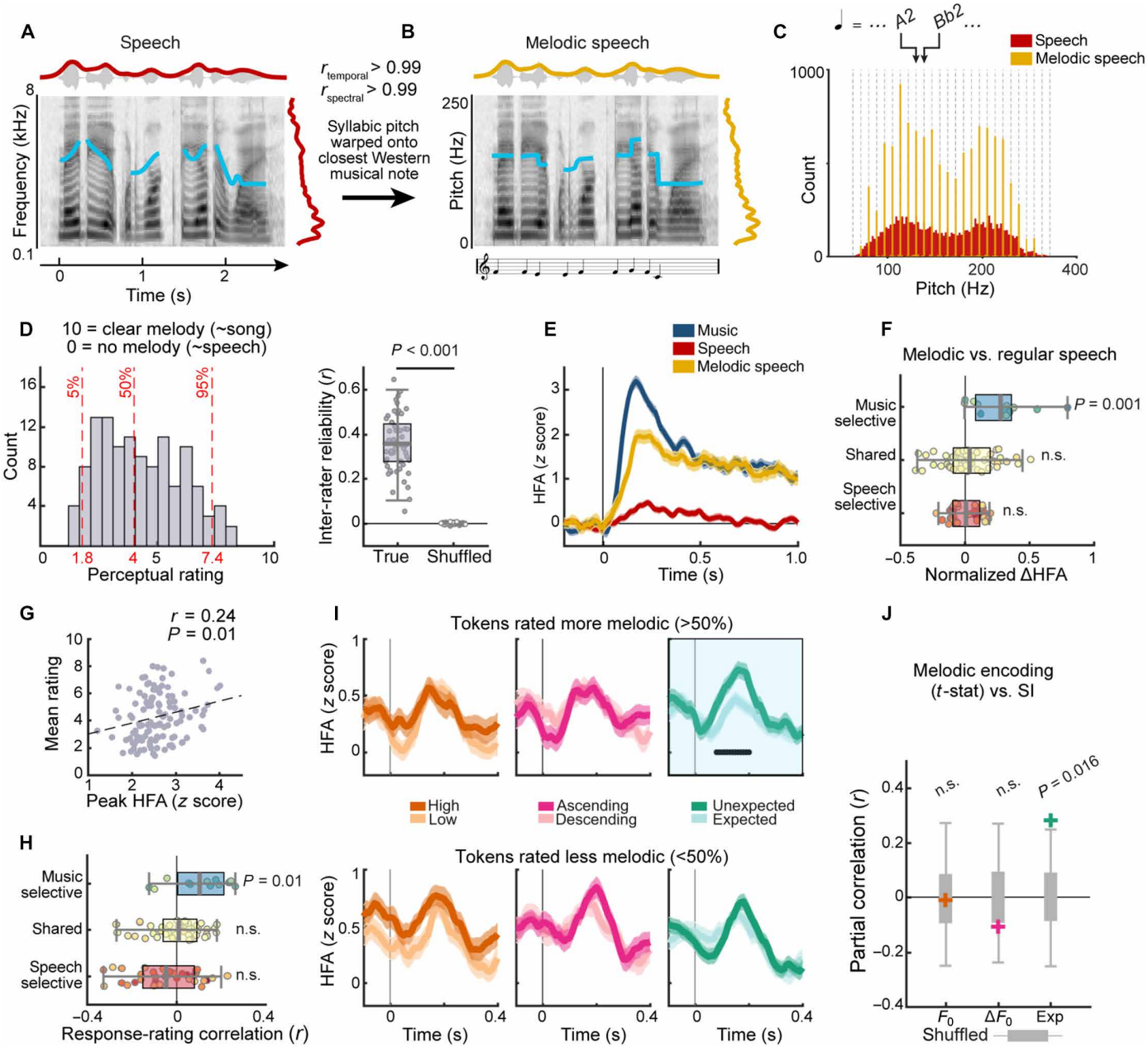

Fig. 4. Music selectivity is independent of low-level spectrotemporal properties. (**A**) Spectrogram of an example speech token. Overlaid blue lines indicate pitch contour. Dark red lines above and to the right indicate temporal and spectral envelopes respectively. (**B**) Melodic speech spectrogram for the same token as in (A), illustrating a similar spectrotemporal structure to speech. Musical notation underneath indicates the discrete musical pitch of each syllable. (**C**) Pitch distributions show that melodic speech discretizes syllabic pitch to the nearest Western musical note. (**D**) Left: Distribution of mean ratings indicating the extent to which independent listeners ($N = 11$) heard each melodic speech token as melody. Red dashed lines indicate 5th, 50th, and 95th percentiles of distribution. Right: Inter-rater reliability. (**E**) Mean evoked responses to music, speech, and melodic speech at a music-selective electrode. Responses are time-locked to the onset of tokens (music phrases or sentences) and averaged across all tokens. (**F**) Response difference indicating the extent to which melodic speech elicited larger responses than speech across electrodes split into speech-selective (SI < −0.2), shared (−0.2 < SI < 0.2), and music-selective (SI > 0.2) bins. Marker colors indicate electrodes' SI. (**G**) Scatter plot showing correlation between the peak response to melodic speech (*x* axis) and perceptual ratings (*y* axis) across all tokens for the same electrode as (E). (**H**) Response-rating correlations across all electrodes split into the same selectivity-dependent bins as in (F). (**I**) Tuning to melodic features within melodic speech for the same music-selective electrode as (E). Tuning is computed separately for tokens perceived more (top row) and less (bottom row) like melody. Responses are grouped based on the median split along each feature dimension. (**J**) Colored crosses indicate partial correlations between SI and the modulation of melodic speech responses by melodic features. Gray bars indicate the 95th percentiles of permutation tests.

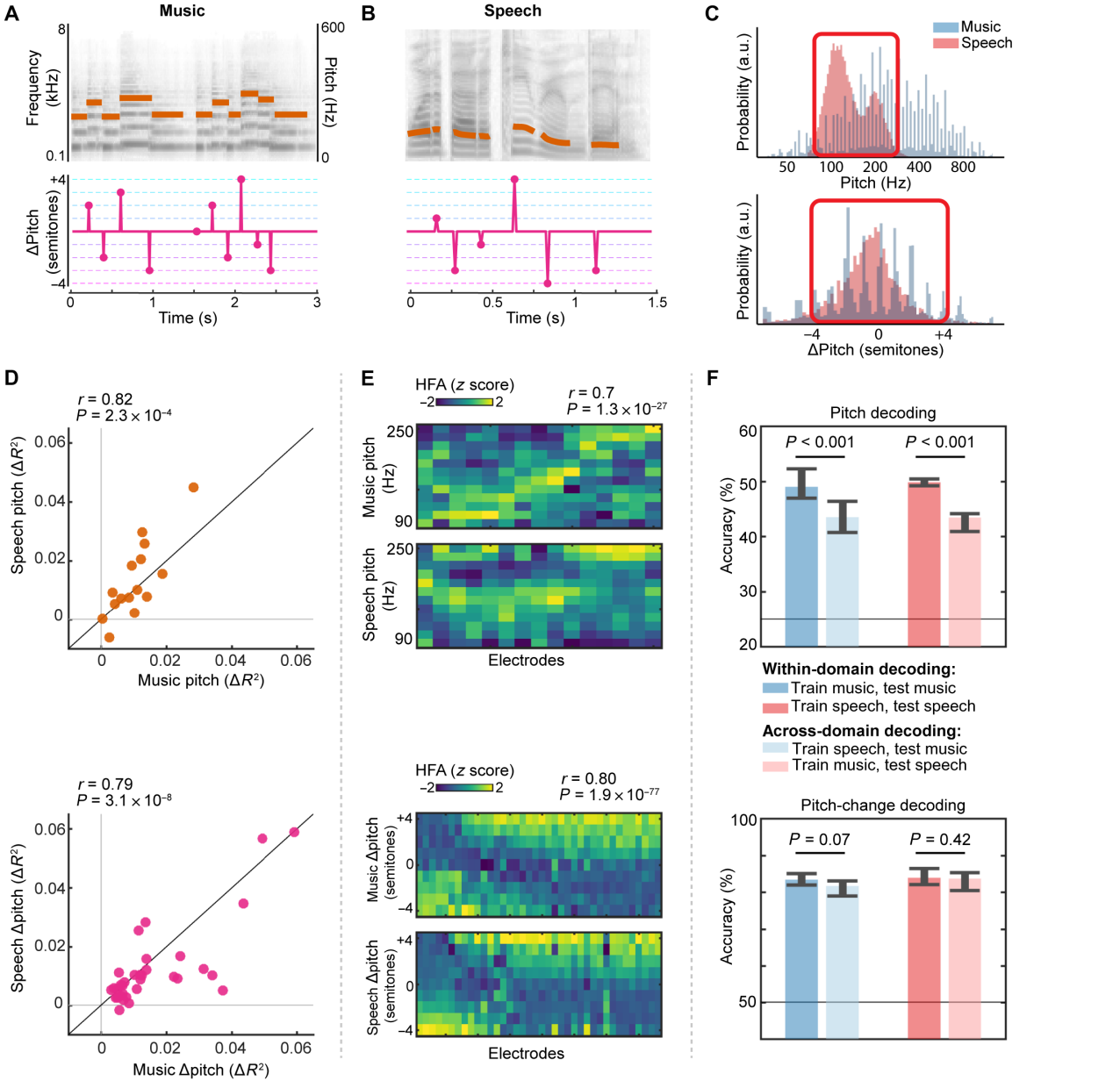

Fig. 5. Shared representations of pitch and pitch-change across music and speech. (**A** and **B**) Pitch and pitch-change representations for example music and speech tokens. Orange lines indicate pitch contours overlaid on stimulus spectrograms. Pitch-change in speech is based on differences in median syllabic pitch, specified at syllable onsets. (**C**) Partially overlapping pitch (top) and pitch-change (bottom) distributions in music and speech. Red boxes indicate overlapping regions of feature space. (**D**) ΔR^2 explained by pitch (top) and pitch-change (bottom) in TRF models of music (*x* axis) versus speech (*y* axis). (**E**) Tuning curves for pitch (top) and pitch-change (bottom) within music and speech. Tuning is characterized for every pitch or pitch-change encoding electrode (*x* axis) across the overlapping range of feature-space found in music and speech (*y* axis). For pitch, columns are ordered by the F_0 corresponding to peak HFA in music. For pitch-change, columns are ordered by increasing HFA difference between ascending and descending changes. (**F**) Linear classification accuracy when decoding pitch (top) or the direction of pitch-change (bottom) from corresponding neural activity across electrodes. Classifiers are trained and tested on neural activity either within (darker colored bars) or across domains (lighter colored bars). Errors indicate 95% distribution of bootstrap tests. Horizontal black lines indicate chance accuracy.

extent to which these populations represented information using a domain-general neural code. We computed tuning curves for the overlapping range of pitch and pitch-change across music and speech. This overlapping range spanned from 80 to 265 Hz for pitch, and −4 to +4 semitones for pitch-change (Fig. 5C). Both within and across electrodes, tuning profiles were highly correlated across domains for pitch ($r = 0.70$, $P = 1.6 \times 10^{-27}$; Fig. 5E, top) and pitch-change ($r = 0.82$, $P = 1.6 \times 10^{-74}$; Fig. 5E, bottom), indicating a highly conserved neural code across music and speech.

To further quantify the extent of domain generalization, we first trained linear classifiers to decode pitch (Fig. 5F, top) or the direction of pitch-change (Fig. 5F, bottom) from electrode activity pooled across participants. We trained separate classifiers to decode information in music and speech activity. We then tested these classifiers

using activity from either the same domain (darker bars) or the opposite domain (lighter bars) to the one in which they were trained. As expected, within-domain decoding accuracy was well above chance for both features (bootstrap tests, all $P < 0.001$). Crucially, when decoders were tested across domains, accuracy remained well above chance for both pitch and pitch-change (bootstrap tests, all $P < 0.001$). In particular, for pitch-change, within and across domain decoding did not significantly differ (bootstrap difference tests: music generalization, $P = 0.07$; speech generalization, $P = 0.42$). Thus, in contrast to the encoding of melodic expectation, lower-level properties of melody are represented by general-purpose auditory populations using a neural code that is highly generalizable to speech.

DISCUSSION

While neuroimaging research has long implicated the STG in the perception of melody (*49*–*52*), the specific information represented in this region has remained unclear. Using high-density direct recordings from the human auditory cortex, we demonstrated the extraction of multiple perceptually critical features of melody. Furthermore, comparing the neural encoding of music with speech, we revealed how this process recruits both music-specialized and general-purpose mechanisms.

We showed that the STG contains a spatial map for representing different pitch-based properties of melody, consistent with a model in which distinct types of information are processed in dissociable pathways (*19*–*21*). The fact that pitch, pitch-change, and expectation are not jointly encoded is notable, given that each progressively higher-order feature derives from the temporal integration of lower-order features. Such a spatial code may arise via parallel projections from earlier cortical or subcortical regions (*53*). Future work should identify the inputs to each subpopulation to explicitly determine the network architecture supporting perception.

Current findings also clarify the nature and extent of specialization for music in the human brain. Previous research has found music-selective cortical activity (*31*–*33*, *46*), yet the stimulus information driving this activity has remained unclear. Here, we showed that music-selectivity is systematically driven by the encoding of melodic expectation. This encoding occurred in a format consistent with predictive coding theory, whereby more unexpected notes evoked larger responses (*18*, *54*–*56*). Our results thus demonstrate functional specialization in the human brain for encoding the statistical structure of a behaviorally relevant domain of sound. Future work should determine the degree to which this encoding is bottom-up, reflecting regularities within recent stimulus history, versus top-down, potentially requiring feedback from higher-order cortical regions (*57*).

Last, we revealed the extent to which auditory representations are shared across music and speech. Decades of neurophysiological work in nonhuman animals has characterized the encoding of pitch agnostic to sound domain (*58, 59*). More recently, human electrophysiology has examined the encoding of pitch within the domain of speech (*19*, *60*). Despite these advances, a cross-domain comparison to evaluate the extent to which music and speech recruit shared auditory representations has been outstanding. Evidence from behavioral and lesion studies has been conflicting, with research providing evidence for both shared (*27, 61*–*62*) and domain-specific (*63*) mechanisms. Leveraging the spatiotemporal resolution of ECoG, we provided direct evidence that STG representations of pitch and pitch-change are largely shared across domains. These domain-general representations localized to anterior regions of the STG, consistent with prior reports of pitch-sensitive cortical regions (*64*).

Perception of melody requires the successful extraction of multiple features from a dynamic acoustic signal. Correspondingly, we have shown that melody is not processed by a single monolithic region. Rather, leveraging both general-auditory and music-specific mechanisms, different neural subpopulations across a spatial map encode distinct melodic features, spanning basic spectrotemporal to time-integrated and statistical structure.

MATERIALS AND METHODS

Participants

Eight patients undergoing treatment for intractable epilepsy participated in the study (see table S1 for demographic and clinical information). One additional patient participated in a control experiment involving presentation of harmonic or inharmonic complex tones (see below) and was also presented with a partial dataset of music (three of the six blocks). Participants were implanted with 4-mm-spaced subdural electrode grids unilaterally over peri-Sylvian regions for clinical monitoring of seizure-related activity. Grid placement was determined solely by clinical considerations. One patient reported a history of tinnitus while all others reported having normal hearing. All patients were nonmusicians with the extent of prior formal musical training ranging from 0 to 8 years (table S1). All participants provided written informed consent before experimental testing. The study was approved by the University of California, San Francisco Committee on Human Research.

Software

Analyses were carried out using custom-written scripts in MATLAB 2016b (MathWorks; www.mathworks.com) and Python. Cortical surface reconstruction was performed using Freesurfer and electrodes were localized using a Python package (img-pipe). Melodic expectation was extracted using publicly available Python toolboxes (https://github.com/magenta/) (*34*). All other melodic features were extracted in MATLAB using the music information retrieval toolbox (*65*) and custom-written scripts. Melodic speech stimuli were created by manipulating the pitch of speech stimuli in Praat (*66*). Syllable onsets and offsets in speech were detected using a forced aligner (https://babel.ling.upenn.edu/phonetics).

Stimuli and procedure

All participants passively listened to natural music and speech stimuli, while we recorded ECoG activity. Two of these participants additionally listened to a control stimulus, which we refer to as melodic speech. All stimuli were delivered through free-field speakers from a Windows laptop at a sample rate of 44.1 kHz (melody) or 16 kHz (speech and melodic speech).

Music

A stimulus set comprising 214 distinct monophonic musical phrases (total duration = 23.4 min, mean phrase duration = 6.57 s) was compiled by sampling directly from natural solo instrumental recordings. Six purely percussive phrases were excluded from analysis. Remaining phrases comprised 4578 discrete notes, with a mean density of 3.4 notes/s. Phrases collectively featured 18 different Western instruments in genres broadly categorized as classical, jazz, or folk (see table S2 for example of source material). Phrases were presented in pseudorandom order and separated by silent intertrial

intervals ranging from 0.7 to 1.5 s. Participants heard each phrase once, with data collected across five separate listening blocks that each lasted approximately 4 min. An additional block contained 10 repetitions of 10 phrases. Musical phrases were chosen to avoid well-known melodies, and participants reported being unfamiliar with most phrases.

Speech

Speech stimuli comprised a selection of 499 English sentences from the TIMIT corpus (*67*), spoken by a variety of male and female speakers with regional North American accents. Speech stimuli were presented to participants in a similar fashion to music—in pseudorandom order across four separate listening blocks with an additional block containing 10 repetitions of 10 sentences. Sentences had a mean duration of 2.05 s (SD = 0.4).

Melodic speech

Melodic speech control stimuli were created by altering the pitch of TIMIT tokens while leaving all other spectrotemporal features of speech intact. To create melodic speech, we first calculated the median pitch of each syllable and identified the closest Western musical pitch. We then warped each syllable's continuous pitch onto its nearest discrete musical value, thereby transforming each syllable into a discrete musical pitch event (Fig. 4, A to C). To verify that this process did not significantly change the spectrotemporal structure of speech, we compared the spectral and temporal profiles of melodic speech and speech. Specifically, to compare temporal structure, we averaged power across all frequency bands of the spectrogram to extract the temporal envelope for every token. We concatenated all tokens into two vectors for speech and melodic speech, respectively, and computed their linear correlation. As expected, temporal envelopes were nearly identical ($r = 0.995$). To compare spectral structure, we concatenated power across frequency bands at every time point in the spectrograms into two vectors for speech and melodic speech, respectively. Spectral envelopes were strongly correlated ($r = 0.998$). To evaluate whether melodic speech generated coherent melodies, we first applied an automatic musical key-finding algorithm to each token (*47, 48*). This algorithm returns a series of 24 correlation coefficients, indicating the extent to which the sequence of pitches aligns with the canonical pitch distributions of the 24 Western diatonic keys (12 major and 12 minor). We retained tokens with a maximum correlation coefficient greater than 0.6, resulting in 118 melodic speech tokens. For comparison, the mean correlation coefficient when applying key finding to the natural musical stimuli was 0.77 (SD = 0.11). To encourage listeners to process melodic speech in a musical manner, and thus generate melodic expectations, we presented melodic-speech tokens grouped by their tonality and primed the first token within each tonality with a diatonic triad chord indicating the tonality. The duration of the entire melodic speech stimulus was 3 min.

Harmonic and inharmonic tones

For a subset of the original music stimulus (47 of 214 phrases), we used the F_0 and note durations of melodies to generate synthesized tone sequences. This subset was pseudo-randomly chosen, such that it contained pitch and pitch-change ranges representative of the entire musical stimulus. We synthesized two versions of each melody, one consisting of harmonic tones and the other consisting of inharmonic tones. Each tone consisted of six harmonic components including the F_0 with the relative power of each component preserved from original melodies. We applied a tapered cosine window to each tone with 10-ms onset and offset ramps. To create inharmonic tones, we followed the identical procedure to that used in (*38*, *40*). Specifically, we jittered the frequency of each harmonic, excluding the fundamental, by a random amount chosen from the uniform distribution $U(-0.5, 0.5)$. This value was multiplied by the F_0 and added to the frequency of the respective harmonic. To reduce beating, jitter values were further constrained, such that all frequency components were separated by at least 30 Hz. For a given melody, the same profile of jitter values was applied to every note.

Melodic feature extraction

Note onset times and their corresponding absolute pitch values were extracted using the Music Information Retrieval toolbox (*65*) and verified manually. Pitch-change was then calculated by subtracting the pitch value of the previous note from that of the current note and expressed in semitones. To extract melodic expectations, we used a recurrent neural network model (MelodyRNN; https://github.com/magenta/) that applies natural language modeling approaches to model melodic sequence structure. We used an off-the-shelf implementation in which the model was pretrained on a large corpus of approximately 45,000 Western popular melodies pulled from the Lakh MIDI dataset (https://colinraffel.com/projects/lmd/). The model was trained with the following parameters: learning rate = 0.001, batch size = 128, number of layers = 2 × 512 nodes, attention length = 40, and dropout rate = 0.5. Internal model weights were optimized during training to maximize the probability mass assigned to the note occurring at t_{n+1} given the pitch and duration of previous notes from $t_{1:n}$.We used the "Attention" configuration of melodyRNN [see (*34*, *68*) for a detailed explanation], which enables the model to learn long-term dependencies characteristic of music that traditional Markov-based approaches fail to capture (*69*). For all phrases in the current stimulus set, MelodyRNN was used to calculate the surprisal of each note, defined as the negative log probability of the note e that occurred at position i, given the MIDI pitch and duration (quantized to the nearest 16th note) of preceding notes in the melody

$$\text{Surprisal}(e_i) = -\log_2 p(e_i|e_1, \ldots, e_{i-1})$$

We also computed the uncertainty of the melody at each time step, which is defined as the entropy of the probability distribution over an alphabet of $K = 120$ possible notes

$$\text{Uncertainty}(e_i) = -\sum_{e \in K} p(e_i|e_1, \ldots, e_{i-1}) \log_2 p(e_i|e_1, \ldots, e_{i-1})$$

Surprisal and uncertainty are complementary measures in that the former indicates the extent to which an event deviated from preexisting expectations, while the latter conveys the specificity of those expectations in anticipation of the event. While these measures are highly correlated ($r = 0.32$, $P < 0.001$; fig. S1) and produced similar patterns of neural tuning (fig. S3C), we found that surprise modulated evoked responses to a greater degree than uncertainty (fig. S2). While MelodyRNN has been found to generate realistic melodies (*34*, *68*), we sought to validate its ability to accurately model listeners' expectation. To do so, we also computed expectation using a traditional Markov approach (*70*). Values obtained using the two different modeling approaches (MelodyRNN versus Markov-based) were correlated (surprise: $r = 0.59$, $P < 0.001$; uncertainty: $r = 0.38$, $P < 0.001$); however, we found that MelodyRNN explained a greater amount of variance in neural activity than Markov-based estimates

($t = 4.48$; $P = 1.01 \times 10^{-5}$), which may reflect its ability to model long-distance dependencies as noted above.

Neural recordings and preprocessing

ECoG activity was acquired at a sampling rate of 3051.8 Hz using either a 256-channel PZ2 amplifier or 512-channel PZ5 amplifier connected to an RZ2 digital acquisition system (Tucker-Davis Technologies, Alachua, FL, USA). We recorded the local field potential of each electrode and removed line noise using notch filters at 60, 120, and 180 Hz. Bad channels with variance indistinguishable from noise or continuous epileptiform activity were removed, and time segments on remaining channels that contained electrical or motor artifacts were marked and excluded. We used the log-analytic amplitude of the Hilbert transform to filter the signal into eight log-spaced bands in the high-gamma range from 70 to 150 Hz and took the first principal component across these bands to extract stimulus-related activity (*71*, *72*). The resulting HFA was downsampled to 100 Hz. When fitting TRF models, we normalized activity by subtracting the mean and dividing by unit variance (i.e., the *z* score) using activity across entire recording blocks. In general, for naturalistic stimuli such as music, we have found that this normalization approach yields higher TRF performance and more stable weights than normalizing to a local baseline. However, for analyses involving the comparison of activity across stimulus domains (for instance, when deriving the SI), signals were renormalized relative to the mean and SD of a 500-ms silent period preceding each individual token. This was done to account for the unequal stimulus-to-silence ratio of music and speech blocks, ensuring that activity was normalized to an unbiased baseline across different listening domains. We verified that this approach produced equal prestimulus baseline HFA values across music, speech, and melodic speech (e.g., Fig. 4E).

Electrode localization

Electrodes were localized on each participant's brain by coregistering the preoperative structural T1 magnetic resonance imaging (MRI) with postoperative tomography scans. Locations were superimposed onto a three-dimensional reconstruction of each patient's cortical surface using a custom-written imaging pipeline (*73*). For localizing electrodes on a common atlas across patients, we used a nonlinear alignment procedure described in (*73*) to warp electrode locations from the patient's native space to the cvs_avg36_inMNI152 template (*74*).

Electrode selection

We first aligned continuous ECoG activity to the onset of musical phrases and viewed the peak of each electrode's evoked response, averaged across all phrases, on a common brain map (Fig. 1B). To screen for music-responsive electrodes, we applied a statistical test comparing the activity at each time point during sound presentation to activity during a silent period preceding each phrase (signed-rank $P < 0.05$; Bonferroni correction for multiple time points and electrodes). Electrodes with a significantly higher magnitude of activity during sound presentation than baseline for a continuous window of at least 200 ms were considered responsive and included in subsequent analyses ($n = 224$ music-responsive electrodes across eight participants). The same statistical procedure was applied for evaluating speech-responsive sites, and we included the union of music and speech-responsive electrodes ($n = 342$) in all cross-domain analyses (Figs. 3, 4, and 5). For an additional participant presented with synthesized harmonic/inharmonic tones, we identified sound-responsive electrodes using the same procedure as above ($n = 7$ electrodes).

Single-electrode sensitivity to contrasts in melodic features

To directly visualize whether responses at individual electrodes were sensitive to the pitch, pitch-change, or expectation of notes, we divided each feature's distribution into two equal bins (median split) and examined the corresponding cortical activity within each bin during the epoch from 100 ms before to 400 ms after the onset of notes (Fig. 1C). To limit intrinsic correlations that exist between pitch and pitch-change ($r = 0.14$, $P < 0.001$; fig. S1), we excluded notes from the upper and lower quartile of the pitch distribution from this analysis only (we later use TRF modeling to overcome the issue of correlated features). Binned responses at each time point were compared using independent two-sample *t* tests ($P < 0.001$, Bonferroni corrected for time points, temporal threshold > 50 ms). The above procedure was applied for every electrode to understand how activity was modulated by each melodic feature (fig. S2).

TRF modeling of music evoked activity

To further quantify the extent to which different sources of melodic information were encoded in continuous activity at each electrode, we fit linear TRF models. We discretized the pitch, pitch-change, and expectation of each phrase into *N* bins (see below), with each bin forming a unique row in the [features × time] stimulus matrix. We used binary predictors that were sparsely coded at note onsets to specify a given feature's value (see fig. S3A). For pitch, we discretized values into 24 equally spaced (in log-hertz space) bins between the 5th and 95th percentile of the pitch distribution. Pitch values in the lowest and highest 5% of the distribution were placed into the first and last bins, respectively. By defining these percentile bounds, we prevent unstable estimates that can occur when extreme bins contain too few data points. The number of bins was chosen such that the lowest 12 pitch bins overlapped with the pitch distribution in speech (see below). We used the same approach to discretize both surprisal and uncertainty into eight bins. For pitch-change, the distribution in music is naturally discretized into semitone bins. To avoid bins with little data points, we created distinct bins for each pitch-change between [−5, +5] semitones and placed pitch-changes lower or higher than −5 and +5 semitones, respectively, in two additional bins. In addition to modeling neural responses to the three melodic features, the full TRF model also included the auditory spectrogram (a peripheral stimulus representation), which we extracted using the NSL toolbox (*75*) and downsampled to 25 logarithmically spaced bands spanning six octaves between 0.08 and 8 kHz. We also included two binary predictors specifying the onset location of phrases and notes, respectively. These temporal landmark features allow us to statistically control for the contribution of onset-from-silence responses (*76*) and the presence or absence of a note when evaluating the contribution of specific melodic information. Last, we included a sparse predictor at note onsets that specified the number of times a note with the same pitch had consecutively occurred previously within the phrase. This "note repeat" predictor controls for the effects of stimulus specific adaptation that may be confounded with low expectation contexts. Before fitting, predictors were scaled between 0 and 1 by dividing them by the magnitude of the maximum value in each bin. This ensured that all estimated beta values were scale free and comparable across predictors, with beta magnitude being an index for the contribution of a predictor to model performance.

Neural activity at each time point HFA(t) was modeled as a weighted linear combination of features of the stimulus X within a window spanning $t - 400$ ms to $t + 100$ ms. For each feature f, this resulted in a set of weights, $b_{1...,d}$, with $d = 50$ for a sampling frequency of 100 Hz across the 500-ms window (Fig. 2B and fig. S3).

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

Models were estimated separately for each electrode using L_2 regularization (ridge regression) and fivefold cross-validation. For each cross-validation fold, we trained models on 80% of the data and evaluated them on the held out 20%. The regularization parameter was estimated using a 10-way bootstrap procedure within each training fold before a final value was chosen as the average optimal value across folds. Model performance was evaluated as the Pearson's correlation between actual and predicted brain responses. These correlations were squared to obtain the R^2, a measure of the portion of variance in neural activity explained by the model (Fig. 2A and fig. S3).

Noise ceiling

To evaluate how well TRF models performed relative to an upper limit on explainable variability, we computed the noise ceiling for each electrode using the adjusted split-half correlation approach (*36*, *37*) applied to a subset of musical phrases that were repeated 11 times to participants. First, we randomly split trials into two groups and computed the R^2 value between the averages of the two groups. We repeated this process many times to obtain an average of the distribution of values, which we inserted into the equation

$$\left(\frac{1}{R^2_{\mathrm{upper}}}\right) - 1 = \frac{1}{2}\left(-M + M\sqrt{\frac{1}{R^2_{\mathrm{split}}}}\right)$$

where M represents the number of stimulus repetitions. We divided true TRF R^2 values by these noise-ceiling estimates to obtain the distribution of noise-corrected R^2 values (Fig. 2A).

Variance partitioning of TRF models

To estimate the contribution of a specific feature to the full TRF models (Fig. 2, B and C), we computed its unique variance explained (ΔR^2). For a given feature of interest G, we fit a reduced TRF model that predicted neural activity using all features except G. We then evaluated the unique variance explained by G as the difference in R^2 between the full and the reduced models

$$\Delta R^2_G = R^2_{\mathrm{full}} - R^2_{\mathrm{without}\,G}$$

The unique variance explained by expectation was estimated as the combined contribution of both surprisal and uncertainty. The significance of each feature's unique variance contribution was calculated using a permutation test. Specifically, we shuffled the rows of the predictor matrix corresponding to a feature of interest, leaving all other rows (corresponding to other features that were not being tested) intact. We then fit a TRF model to the permuted predictor matrix. Repeating this procedure 200 times, we arrived at a null distribution. True ΔR^2 values that fell above the 95th percentile of these permuted ΔR^2 values were considered significant.

Independent encoding of melodic features

To determine whether the three melodic features were encoded by overlapping or independent neural populations, we directly compared encoding using linear mixed-effects models. Specifically, for each of the three electrode subsets that significantly encoded pitch, pitch-change, and expectation, respectively, we modeled the response variable as the encoding (ΔR^2) of a given feature. We included fixed-effects terms for the encoding of the other two features and random effects for the intercept and slope grouped by participant. Inclusion of these random effects terms ensured that results were not driven by unequal representation of a given participant in pooled data.

To examine whether the encoding of pitch, pitch-change, and expectation was spatially organized along the STG (Fig. 2D), we extracted electrode locations along the posterior-anterior axis of the brain and normalized each participant's locations to the point at which the central sulcus meets the sylvian fissure. Normalizing to a common anatomical landmark avoids warping individuals' electrode locations to a common brain, better preserving the relative spatial differences in the encoding of each feature. Within each hemisphere, we assessed whether the distribution of electrodes encoding different features differed in their posterior-to-anterior location using linear mixed-effects models with a randomized block design in which blocks correspond to participants.

Melodic feature tuning

For electrodes that significantly encoded a melodic feature (based on TRF ΔR^2), we sought to further characterize their tuning patterns—that is, the format by which information was encoded. In addition to inspecting corresponding TRF weights (Fig. 2E), we examined tuning patterns in the raw HFA using the same bin edges as that used in constructing TRF stimulus matrices. To produce time-collapsed tuning matrices (Fig. 2, F to H), we estimated each electrode's peak encoding latency K relative to the onset times of notes in the stimulus. This estimate was based on the lag at which the magnitude of TRF weights was maximal. For every musical note onset at time t, we then binned the HFA at time $t + K$. Responses in each bin were then averaged, and tuning across bins was normalized by removing the mean and dividing by unit variance.

Tuning to pitch

To further quantify tuning across pitch-encoding electrodes, we calculated the F_0 corresponding to the peak-HFA in tuning curves for each electrode (Fig. 2F, right). While we used F_0 as a proxy for pitch, to examine whether pitch tuning could be explained by other acoustic variables that are correlated with F_0, we examined how tuning was impacted by the spectral profile (fig. S4) or the spectral modulation profile (fig. S5) of notes. To dissociate between the spectral profile and F_0, we examined pairs of note clusters that spanned comparable pitch ranges but differed in their spectral profiles (fig. S4, B and C). This was achieved by grouping notes by instrument or by manually selecting clusters of notes with anticorrelated spectral profiles across fixed F_0 ranges. For every electrode, we characterized F_0 tuning separately within each cluster, and correlated the two resulting F_0-tuning curves (fig. S4E). To test whether F_0 tuning could be explained by tuning to rates of spectral modulation, we first extracted the spectral modulation profile of notes using the nsltoolbox (*75*) using 16 bins between 0.25 and 4 octaves per cycle. We then fit a TRF model that included these 16 spectral modulation bins as predictor variables along with other melodic and acoustic predictors. We assessed whether replacing F_0 with spectral modulation predictors could

yield comparable model R^2 values (fig. S5C). We also assessed whether F_0 was able to explain unique variance in the models above and beyond what spectral modulation variables could explain (fig. S5D).

Tuning to pitch-change

For pitch-change encoding electrodes, we used *k*-means clustering to group electrodes with similar tuning profiles (Fig. 2G, right). We chose $k = 3$ as this produced distinct and interpretable clusters. We sought to determine whether the encoding of pitch-change relied on tracking F_0-changes versus changes in individual frequency components. We took two complementary approaches to address this question. First, we asked whether tuning to pitch-change was dependent on whether high-order frequency components (above the F_0) provided a reliable cue to pitch-change direction (fig. S6). We classified every pitch-change in the music stimulus based on whether components (excluding F_0) provided an unambiguous versus ambiguous cue to the direction of pitch-change. To classify a pitch-change as either ambiguous or unambiguous, for every harmonic component of a given order, we determined whether the most proximate component (in log-hertz space) within the next note had the same or different order. Pitch-changes in which at least 50% of all nearest components were of a different order were classified as ambiguous. Consistent with prior research, we found ambiguity primarily at interval magnitudes greater than three semitones (*38*). In a second approach, we analyzed neural data from one participant presented with synthesized melodies composed of either harmonic or inharmonic tones in addition to music (fig. S7). Because of limited electrode coverage over auditory regions, we found relatively few sound-responsive electrodes ($n = 7$; see electrode selection procedure above). We compared evoked responses to ascending versus descending changes across all three conditions (music, harmonic, and inharmonic; independent two-sample *t* tests, $P < 0.05$, temporal threshold > 50 ms). In addition to pitch-change, we identified one electrode at which responses were significantly modulated by F_0 in music and harmonic tone conditions (Pearson's correlation between F_0 and HFA $r > 0.2$ for both conditions, $P < 0.001$) but not the inharmonic condition ($r = 0.04$, $P > 0.05$). No electrodes were significantly modulated by expectation across all three conditions (independent two-sample *t* test on median-split responses, all $P > 0.05$).

Tuning to expectation

For expectation encoding electrodes, we tested for monotonic encoding by computing the rank order correlation between the response to each note and its corresponding expectation value (Fig. 2H, right). To further visualize tuning, we divided the distribution of expectation into four equal-spaced bins and plotted the corresponding neural responses across time in each bin (fig. S9A).

Domain selectivity

To characterize the extent to which activity at each electrode was stronger for music or speech (Fig. 3A), we derived a domain SI ranging from −1 (speech selective) to +1 (music selective). For each electrode, we concatenated activity evoked during the first second of music and speech tokens, respectively, into two vectors, which we then compared using an independent two-sample *t* test. We used the resulting test statistic as a proxy for domain selectivity, normalizing values so that the largest magnitude across all electrodes was equal to 1. To determine whether music selectivity was anatomically clustered (Fig. 3C), we extracted electrode locations along the posterior-anterior axis of the brain and normalized each participant's locations to the point at which the central sulcus meets the sylvian fissure. Within each hemisphere, we then compared the location of music-selective electrodes (SI > 0.2) with that of all other electrodes (SI < 0.2) using linear mixed-effects models with a randomized block (where blocks correspond to participants). We further conducted within-participant comparisons of the location of music-selective versus nonselective electrodes using Bonferroni-corrected rank sum tests. To ensure results were not an artifact of the SI threshold of 0.2, we repeated these analyses using a cutoff value of 0.33 and verified that results remained unchanged. Specifically, linear mixed-effects analyses confirmed that music-selective electrodes did not anatomically differ from other electrodes in both left ($t = 1.15$, $P = 0.25$) and right ($t = 0.23$, $P = 0.82$) hemispheres, and all within-subject comparisons between the two distributions were insignificant ($P > 0.1$, ranksum tests). Last, we conducted a power analysis to examine whether the absence of evidence for a music-selective region was due to limitations in sample size. On the basis of prior work showing the existence of an anterior music-selective patch of STG, we estimated a 20-mm difference in the mean location of music-selective versus nonselective electrode distributions along the posterior-anterior axis. From initial pilot data collected from two participants, we estimated a 1:7 ratio of music-selective to nonselective electrodes (which closely approximated the eventual ratio of 47 of the 342 electrodes). We used an alpha = 0.05 and statistical power = 80%. Power analyses revealed that nine music-selective and 63 nonselective electrodes was sufficient to achieve statistical power. Because our empirical sample size exceeded this in both hemispheres (left hemisphere = 23 music selective, 194 nonselective; right hemisphere = 24 music selective, 101 nonselective), we concluded that the current sample sizes were sufficient.

Relationship between melodic feature encoding and selectivity

To determine whether music-selectivity was explained by the encoding of pitch, pitch-change, or expectation, we computed the partial correlation between the ΔR^2 explained by each feature and the SI across all electrodes with SI > 0 (Fig. 3E and fig. S10). For a given feature, partial correlations controlled for the ΔR^2 explained by the other two features. Partial correlations also controlled for the magnitude of the mean music response on each electrode (electrodes with stronger music responses were more likely to be music-selective and have greater ΔR^2 values by default, leading to spurious correlations). Significance was evaluated by randomly permuting the SI and ΔR^2 values before correlating permuted variables. This procedure was repeated 10,000 times to define a null distribution. Cases in which the true correlation exceeded 95% of the null distribution were considered significant. Similarly, to determine whether speech selectivity was explained by the encoding of relevant features of speech (fig. 3G, see below), we applied the same procedure as above, evaluating partial correlations between ΔR^2 and SI across all electrodes with SI < 0.

TRF modeling of speech

To model speech-evoked activity at each electrode, we applied the same TRF pipeline that was previously used to model music-evoked activity. We modeled features of speech that were analogous to musical pitch, pitch-change, and expectation. To code pitch in the stimulus matrix, we discretized the continuous pitch contour of sentences into 12 equally spaced bins. To extract pitch-change, we computed the difference between the median pitch of the current and prior syllable, discretized values into semitone-spaced bins, and coded each interval as a separate row in the stimulus matrix at the onset of

syllable nuclei. To consider sequence statistics in speech that are comparable to melodic expectation, we extracted and modeled both phoneme surprisal and cohort entropy, as defined by (*42*). These values are mathematically equivalent to those calculated earlier for music, with phonemes in place of notes. Specifically, phoneme surprisal is the inverse of the conditional probability of each phoneme give the preceding phonemes in a word

$$\text{Surprisal}(e_i) = -\log_2\left[\frac{\text{freq}(\text{cohort}_i)}{\text{freq}(\text{cohort}_{i-1})}\right]$$

where cohort_i is the set of all possible words at position i and $\text{freq}(c)$ is the summed frequency of all words in cohort c. Cohort entropy is the Shannon entropy of the cohort at each phoneme, given by

$$\text{Entropy}(e_i) = -\sum_{\text{word}}^{\text{cohort}_i} p_{\text{word}} \log_2 p_{\text{word}}$$

where p_{word} is the relative frequency count of the given word within a language corpus [see (*42*) for details]. We specified these values in the stimulus matrix at the onset of each phoneme and discretized both variables into eight bins. To control for the contribution of extraneous speech features, we included the spectrogram and two binary predictors specifying sentence and syllable-nuclei onset locations. All other aspects of the TRF fitting procedure and estimation of unique variances were identical to the modeling approach used for music.

Analysis of melodic speech

We evaluated the extent to which melodic speech tokens elicited a melodic percept by conducting a behavioral test on an independent set of listeners ($N = 11$; Fig. 4D). After hearing each token, subjects rated the extent to which they heard a melody on a continuous scale of 0 to 10. Subjects were explicitly instructed to use regular speech and song as perceptual anchors corresponding to ratings of 0 and 10, respectively.

To assess the extent to which electrodes responded more strongly to melodic speech over regular speech (Fig. 4F), we computed the normalized difference in evoked activity at each electrode

$$\Delta\,\text{HFA} = \frac{(\text{HFA}_{\text{mel.sp.}} - \text{HFA}_{\text{sp}})}{(\text{HFA}_{\text{mel.sp.}} + \text{HFA}_{\text{sp}})}$$

We evaluated the ΔHFA for electrodes within three different groups: speech selective (SI < −0.2), shared (−0.2 < SI < 0.2), and music selective (SI > 0.2). Within each bin, we used a one-sample t test to assess whether the mean ΔHFA was significantly greater than zero, which would indicate enhanced responses to melodic versus regular speech. We verified that enhanced responses to melodic versus regular speech remained even when the tokens preceded by chords were excluded from the analysis. To evaluate whether melodic speech responses scaled with the degree to which they were perceived as melody, we correlated the average ratings across listeners with the peak HFA elicited by corresponding tokens (Fig. 4G). We computed this correlation for all electrodes and grouped values into the same three SI-based bins as above (Fig. 4H). Within each bin, we used one-sample t tests to evaluate whether mean correlations were significantly different from zero.

We examined tuning to features of melody in melodic speech by first extracting pitch, pitch-change, and expectation from every token. We divided each feature distribution into two bins (using the median split) and inspected corresponding neural responses in each bin. Because melodic speech tokens with female speakers were rated as more melodic than those with male speakers (Wilcoxon rank-sum: $P = 6.65 \times 10^{-5}$, $z = -4$), to avoid speaker identity–driven effects, we binned neural responses separately for male and female subsets of the data before averaging across gender. We did this separately for tokens that were perceived as more versus less melodic (using the median split of perceptual ratings). For a given feature, we compared electrode responses in the two bins using independent-sample t tests and used the resulting test statistic as a proxy for the strength of tuning to that feature. We used partial correlations to predict SI from tuning to melodic features, evaluating significance using permutation-based null distributions (Fig. 4J).

Cross-domain comparison of pitch and pitch-change encoding

To compare encoding of pitch and pitch-change across domains, we correlated the ΔR^2 values explained by a given feature across TRF models of music and speech (Fig. 5D). To compare music and speech tuning curves (Fig. 5E), we confined our analysis to the range of overlapping pitch and pitch-change values across the two domains (Fig. 5C). For pitch, we focused on the F_0 range from [86 Hz to 252 Hz]. This corresponded to the 5th to 49th percentiles and the 2.5th to 97.5th percentiles of the pitch distributions in music and speech, respectively. For pitch-change, we focused on the set of nine intervals from [−4, +4] semitones, corresponding to the 9th to 91st percentiles and the 3.5th to 98th percentiles of the pitch-change distributions in music and speech, respectively. As when previously computing tuning curves (Fig. 2, F to H), we only used HFA at the time point corresponding to peak encoding (estimated based on the time point of maximal TRF weights). To characterize pitch tuning, we grouped neural responses to notes (for music) or to voiced portions of an utterance (for speech) into 12 log-hertz spaced bins. We quantified the similarity in neural codes across music and speech by computing the linear correlation between tuning curves across all electrodes.

Cross-domain decoding of pitch and pitch-change

We used a cross-domain decoding approach to quantify the extent to which neural codes for pitch and pitch-change generalized across domains (Fig. 5f). Specifically, we trained linear discriminant classifiers (*77*, *78*) to decode either the pitch or the direction of pitch-change from cortical responses pooled across electrodes. For pitch, we discretized the distribution into four classes (chance accuracy = 25%). For pitch-change, we classified activity for descending versus ascending changes (chance accuracy = 50%). For the purposes of evaluating generalization of pitch-change, we chose to classify contour (ascending versus descending) rather than specific interval information, as intervals are not well defined in speech. To evaluate decoding accuracy within domain, we used 10-fold cross validation to maintain independent test and train datasets. For decoding across domain (e.g., train on music, test on speech), we used all the data in each domain to train and test models. Before decoding, we equalized the number of observations of each class (observations per class for pitch: $N = 355$ for music, $N = 6598$ for speech; observations per class for pitch-change: $N = 1243$ for music, $N = 689$ for speech) and centered the data in each domain to have mean = 0. Statistical inference was performed by bootstrapping data and re-applying the entire decoding analysis pipeline on each run (N runs = 200).

Supplementary Materials

This PDF file includes:
Figs. S1 to S10
Tables S1 and S2
Legends for audio S1 to S3

Other Supplementary Material for this manuscript includes the following:
Audio S1 to S3

REFERENCES AND NOTES

1. R. Santoro, M. Moerel, F. De Martino, G. Valente, K. Ugurbil, E. Yacoub, E. Formisano, Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4799–4804 (2017).
2. A. A. Brewer, B. Barton, Maps of the Auditory Cortex. *Annu. Rev. Neurosci.* **39**, 385–407 (2016).
3. A. M. Leaver, J. P. Rauschecker, Functional Topography of Human Auditory Cortex. *J. Neurosci.* **36**, 1416–1428 (2016).
4. A. D. Patel, *Music, Language, and the Brain* (Oxford Univ. Press, 2010); https://play.google.com/store/books/details?id=qekVDAAAQBAJ).
5. J. McDermott, The evolution of music. *Nature* **453**, 287–288 (2008).
6. W. T. Fitch, The biology and evolution of music: A comparative perspective. *Cognition* **100**, 173–215 (2006).
7. J. H. McDermott, A. J. Oxenham, Music perception, pitch, and the auditory system. *Curr. Opin. Neurobiol.* **18**, 452–463 (2008).
8. B. Tillmann, B. Poulin-Charronnat, Auditory expectations for newly acquired structures. *Q J Exp Psychol.* **63**, 1646–1664 (2010).
9. E. Bigand, B. Tillmann, B. Poulin-Charronnat, A module for syntactic processing in music? *Trends in Cognitive Sciences* **10**, 195–196 (2006).
10. F. Attneave, R. K. Olson, Pitch as a Medium: A new approach to psychophysical scaling. *The American Journal of Psychology* **84**, 147 (1971).
11. J. Plantinga, L. J. Trainor, Memory for melody: infants use a relative pitch code. *Cognition* **98**, 1–11 (2005).
12. W. J. Dowling, D. S. Fujitani, Contour, interval, pitch recognition in memory melodies. *J. Acoust. Soc. Am.* **49**, 524–531 (1971).
13. D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation* (MIT Press, 2008; https://play.google.com/store/books/details?id=sgr-DwAAQBAJ).
14. L. B. Meyer, *Emotion and Meaning in Music* (Univ. of Chicago Press, 1956; https://play.google.com/store/books/details?id=HuWCVGKhwy0C).
15. S. Koelsch, Brain correlates of music-evoked emotions. *Nat. Rev. Neurosci.* **15**, 170–180 (2014).
16. Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, *31*(5), 559–575.
17. P. Vuust, O. A. Heggli, K. J. Friston, M. L. Kringelbach, Music in the brain. *Nat. Rev. Neurosci.* **23**, 287–305 (2022).
18. S. Koelsch, P. Vuust, K. Friston, Predictive processes and the peculiar case of music. *Trends Cogn. Sci.* **23**, 63–77 (2019).
19. G. M. Di Liberto, C. Pelofi, R. Bianco, P. Patel, A. D. Mehta, J. L. Herrero, A. de Cheveigné, N. Mesgarani, Cortical encoding of melodic expectations in human temporal cortex. *eLife* **9**, e51784 (2020).
20. C. Tang, L. S. Hamilton, E. F. Chang, Intonational speech prosody encoding in the human auditory cortex. *Science* **357**, 797–801 (2017).
21. P. Belin, S. Fecteau, C. Bédard, Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* **8**, 129–135 (2004).
22. P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, N. Mesgarani, Joint representation of spatial and phonetic features in the human core auditory cortex. *Cell Rep.* **24**, 2051–2062.e2 (2018).
23. D. Amaro, D. N. Ferreiro, B. Grothe, M. Pecka, Source identity shapes spatial preference in primary auditory cortex during active navigation. *Curr. Biol.* **31**, 3875–3883.e5 (2021).
24. C. H. C. Chang, S. A. Nastase, U. Hasson, Information flow across the cortical timescale hierarchy during narrative construction. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209307119 (2022).
25. I. Peretz, D. Vuvan, M.-É. Lagrois, J. L. Armony, Neural overlap in processing music and speech. *Philos. Trans. R. Soc., Lond. B Biol. Sci.* **370**, 20140090 (2015).
26. R. J. Zatorre, P. Belin, V. B. Penhune, Structure and function of auditory cortex: music and speech. *Trends Cogn. Sci.* **6**, 37–46 (2002).
27. A. D. Patel, J. R. Iversen, J. C. Rosenberg, Comparing the rhythm and melody of speech and music: the case of British English and French. *J. Acoust. Soc. Am.* **119**, 3034–3047 (2006).
28. N. J. Zuk, E. S. Teoh, E. C. Lalor, EEG-based classification of natural sounds reveals specialized responses to speech and music. *Neuroimage* **210**, 116558 (2020).
29. E. Fedorenko, A. Patel, D. Casasanto, J. Winawer, E. Gibson, Structural integration in language and music: Evidence for a shared system. *Mem. Cognit.* **37**, 1–9 (2009).
30. M. G. Jantzen, E. W. Large, C. Magne, *Overlap of Neural Systems for Processing Language and Music* (Frontiers in Psychology, 2016), pp. 115; www.frontiersin.org/articles/10.3389/fpsyg.2016.00876/full).
31. S. Norman-Haignere, N. G. Kanwisher, J. H. McDermott, Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88**, 1281–1296 (2015).
32. S. V. Norman-Haignere, J. Feather, D. Boebinger, P. Brunner, A. Ritaccio, J. H. McDermott, G. Schalk, N. Kanwisher, A neural population selective for song in human auditory cortex. *Curr. Biol.* **32**, 1454–1455 (2022).
33. D. Boebinger, S. V. Norman-Haignere, J. H. McDermott, N. Kanwisher, Music-selective neural populations arise without musical training. *J. Neurophysiol.* **125**, 2237–2263 (2021).
34. E. Waite, D. Eck, A. Roberts, D. Abolafia, *Project magenta: Generating long-term structure in songs and stories* (*Github*, 2016); https://github.com/magenta/magenta/tree/main/magenta/models/melody_rnn.
35. F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, J. L. Gallant, Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network* **12**, 289–316 (2001).
36. A. Hsu, A. Borst, F. E. Theunissen, Quantifying variability in neural responses and its application for the validation of model predictions. *Network* **15**, 91–109 (2004).
37. C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, F. E. Theunissen, Encoding and decoding models in cognitive electrophysiology. *Front. Syst. Neurosci.* **11**, 61 (2017).
38. M. J. McPherson, J. H. McDermott, Relative pitch representations and invariance to timbre. *Cognition* **232**, 105327 (2023).
39. E. J. Allen, A. J. Oxenham, Symmetric interactions and interference between pitch and timbre. *J. Acoust. Soc. Am.* **135**, 1371–1379 (2014).
40. M. J. McPherson, J. H. McDermott, Diversity in pitch perception revealed by task dependence. *Nat Hum Behav.* **2**, 52–66 (2018).
41. M. K. Leonard, K. E. Bouchard, C. Tang, E. F. Chang, Dynamic encoding of speech sequence probability in human temporal cortex. *J. Neurosci.* **35**, 7203–7214 (2015).
42. C. Brodbeck, L. E. Hong, J. Z. Simon, Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* **28**, 3976–3983.e5 (2018).
43. T. M. Elliott, F. E. Theunissen, The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* **5**, e1000302 (2009).
44. N. C. Singh, F. E. Theunissen, Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **114**, 3394–3411 (2003).
45. P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, E. F. Chang, Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci.* **36**, 2014–2026 (2016).
46. S. V. Norman-Haignere, J. H. McDermott, Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biol.* **16**, e2005127 (2018).
47. C. L. Krumhansl, E. J. Kessler, Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychol. Rev.* **89**, 334–368 (1982).
48. M. A. Schmuckler, R. Tomovski, Perceptual tests of an algorithm for musical key-finding. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 1124–1149 (2005).
49. R. D. Patterson, S. Uppenkamp, I. S. Johnsrude, T. D. Griffiths, The processing of temporal pitch and melody information in auditory cortex. *Neuron* **36**, 767–776 (2002).
50. V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, E. Brattico, Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *Neuroimage* **59**, 3677–3689 (2012).
51. I. Burunat, P. Toiviainen, V. Alluri, B. Bogert, T. Ristaniemi, M. Sams, E. Brattico, The reliability of continuous brain responses during naturalistic listening to music. *Neuroimage* **124**, 224–231 (2016).
52. A. Angulo-Perkins, W. Aubé, I. Peretz, F. A. Barrios, J. L. Armony, L. Concha, Music listening engages specific cortical regions within the temporal lobes: differences between musicians and non-musicians. *Cortex* **59**, 126–137 (2014).
53. L. S. Hamilton, Y. Oganian, J. Hall, E. F. Chang, Parallel and distributed encoding of speech across human auditory cortex. *Cell* **184**, 4626–4639.e13 (2021).
54. K. Friston, The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
55. K. Friston, S. Kiebel, Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 1211–1221 (2009).
56. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
57. P. Kern, M. Heilbron, F. P. de Lange, E. Spaak, Cortical activity during naturalistic music listening reflects short-range predictions based on long-term experience. *elife* **11**, e80935 (2022).
58. D. Bendor, X. Wang, The neuronal representation of pitch in primate auditory cortex. *Nature* **436**, 1161–1165 (2005).
59. M. J. Tramo, P. A. Cariani, C. K. Koh, N. Makris, L. D. Braida, Neurophysiology and neuroanatomy of pitch perception: auditory cortex. *Ann. N. Y. Acad. Sci.* **1060**, 148–174 (2005).

60. Y. Li, C. Tang, J. Lu, J. Wu, E. F. Chang, Human cortical encoding of pitch in tonal and non-tonal languages. *Nat. Commun.* **12**, 1161 (2021).
61. M. Scharinger, C. A. Knoop, V. Wagner, W. Menninghaus, Neural processing of poems and songs is based on melodic properties. *Neuroimage* **257**, 119310 (2022).
62. C. Semal, L. Demany, K. Ueda, P. A. Hallé, Speech versus nonspeech in pitch memory. *J. Acoust. Soc. Am.* **100**, 1132–1140 (1996).
63. R. J. Zatorre, S. R. Baum, Musical melody and speech intonation: singing a different tune. *PLoS Biol.* **10**, e1001372 (2012).
64. S. Norman-Haignere, N. Kanwisher, J. H. McDermott, Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.* **33**, 19451–19469 (2013).
65. O. Lartillot, P. Toiviainen, T. Eerola, A Matlab Toolbox for music information retrieval in *Data Analysis, Machine Learning and Applications* (Springer, 2008), pp. 261–268; http://dx.doi.org/10.1007/978-3-540-78246-9_31.
66. P. Boersma, D. Weenink, Praat: Doing phonetics by computer (version 6.0.14) (retrieved from last access: 29 April 2018).
67. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM (NIST Speech Disc 1-1.1, 1993), p. 27403.
68. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 [cs.CL] (2014).
69. N. J. Verosky, E. Morgan, Pitches that wire together fire together: scale degree associations across time predict melodic expectations. *Cognit. Sci.* **45**, e13037 (2021).
70. M. T. Pearce, thesis, City University London (2005).
71. S. Ray, J. H. R. Maunsell, Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9**, e1000610 (2011).
72. E. Edwards, M. Soltani, W. Kim, S. S. Dalal, S. S. Nagarajan, M. S. Berger, R. T. Knight, Comparison of time–frequency responses and the event-related potential to auditory speech stimuli in human cortex. *J. Neurophysiol.* **102,** 377–386 (2009).
73. L. S. Hamilton, D. L. Chang, M. B. Lee, E. F. Chang, Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. *Front. Neuroinform.* **11**, 62 (2017).
74. B. Fischl, M. I. Sereno, R. B. Tootell, A. M. Dale, High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
75. T. Chi, P. Ru, S. A. Shamma, Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **118**, 887–906 (2005).
76. L. S. Hamilton, E. Edwards, E. F. Chang, A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Curr. Biol.* **28**, 1860–1871.e4 (2018).
77. N. Sankaran, T. A. Carlson, W. F. Thompson, The rapid emergence of musical pitch structure in human cortex. *J. Neurosci.* **40**, 2108–2118 (2020).
78. N. Sankaran, W. F. Thompson, S. Carlile, T. A. Carlson, Decoding the dynamic representation of musical pitch from human brain activity. *Sci. Rep.* **8**, 839 (2018).

Acknowledgments: We thank all members of the Chang Laboratory for useful feedback and discussion. We thank B. Speidel for the help with anatomical localization of electrodes. Special thanks to M. Tilson Thomas and J. Robison. **Funding:** This work was supported by grants to EFC from the National Institutes of Health (R01-DC012379), Bill and Susan Oberndorf Foundation, Bowes Foundation, and Somesh Dash. **Author contributions:** Conceptualization: N.S., M.K.L., and E.F.C. Methodology: N.S., M.K.L., F.T. Software: N.S. Formal analysis: N.S. Investigation: N.S., M.K.L., E.F.C. Data curation: N.S. Writing—original draft: N.S. Writing—review and editing: N.S., M.K.L., F.T., and E.F.C. Visualization: N.S. Supervision: E.F.C., M.K.L., F.T. Funding acquisition: E.F.C. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Code and summary data needed to replicate figures in the paper are publicly available (https://doi.org/10.5281/zenodo.8381672) as of the date of publication.

Submitted 27 July 2023
Accepted 17 January 2024
Published 16 February 2024
10.1126/sciadv.adk0010