www.sciencemag.org/content/357/6353/797/suppl/DC1

Supplementary Materials for

Intonational speech prosody encoding in the human auditory cortex

C. Tang, L. S. Hamilton, E. F. Chang*

*Corresponding author. Email: edward.chang@ucsf.edu

Published 25 August 2017, *Science* **357**, 797 (2017)
DOI: 10.1126/science.aam8577

This PDF file includes:

Materials and Methods
Figs. S1 to S5
Table S1
Captions for Audio S1 to S96
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/357/6353/797/suppl/DC1)

Audio S1 to S96 (as a zipped archive)

Materials and Methods

The experimental protocol was approved by the Institutional Review Board at the University of California, San Francisco. Participants gave their written, informed consent before testing.

Participants

Ten individuals, with self-reported normal hearing, participated in this study (see Table S1 for demographic information). Each participant was a neurosurgical patient with intractable epilepsy who had a high-density electrode grid implanted subdurally for clinical monitoring of seizure activity. The placement of the grid was determined solely by clinical needs, typically covering the lateral cortical surface. 6 subjects had left-hemisphere grids and 4 subjects had right-hemisphere grids. Table S1 also includes the anatomical location of each participant's epilepsy focus. Electrodes were localized by aligning preimplantation MRI and post-implantation CT scans.

Data acquisition and neural signal processing

During experimental tasks, neural signals were recorded from the 256 channel (16 x 16, 4 mm spacing) ECoG grid (or from two 128 channel ECoG grids, 8 x 16, 4 mm spacing) using a multichannel amplifier optically connected to a digital signal processer. The local field potential at each electrode contact was amplified and sampled at 3052Hz. The raw waveform was visually examined, and channels containing continuous epileptiform activity or signal variation too low to be detectable from noise were removed. Time segments on remaining channels that contained electrical/movement-related artifacts or discrete epileptiform activity were manually marked and excluded. The signal was common average referenced and notch-filtered at 60 Hz, 120 Hz, and 180 Hz to remove line noise. Using the Hilbert transform, the analytic amplitude of eight Gaussian filters (center frequencies: 70-150 Hz) was computed. The high-gamma signal was taken as the average analytic amplitude across these eight bands. This signal was downsampled to 100 Hz and z-scored either to a silent baseline or across the entire recording block (Hγ). For each token, we analyzed the neural data in the window from 150 ms before stimulus onset to 650 ms after stimulus offset.

Stimulus design

Stimuli consisted of spoken sentences synthesized to have four linguistically distinct intonational contours. These contours, depicted in Fig. 1b, were Neutral, Emphasis 1, Emphasis 3, and Question. In the Neutral condition, the less variable pitch contour and falling pitch at the end of the sentence do not impart additional meaning beyond the declarative meaning of the words. In the Emphasis 1 condition, a pitch accent (rising followed by falling pitch) on the first word and low pitch throughout the rest of the sentence indicates the first word as the focus of the sentence, whereas the pitch accent on the third word in Emphasis 3 indicates that the third word was emphasized. Finally, rising pitch through the last word in the Question condition signals that the utterance was interrogative.

We applied these intonation contours to four declarative sentences ("**Humans** value **genuine** behavior", "**Movies** demand **minimal** energy", "**Lawyers** give a **relevant** opinion", and "**Reindeer** are a **visual** animal") constructed such that rising intonation at the end of the utterance would signal a question and a pitch accent could be added to either of the bolded words (referred to as the first word and third word even though in the latter two sentences the second bolded word is the fourth word). In order to precisely align the pitch contours across sentences, the sentences were designed to contain the same number of syllables.

Each sentence and intonation combination was recorded from a native female speaker of standard American English. To create base tokens, we duration-matched the syllables of the Neutral sentence recordings and equalized the average root-mean-square intensity across sentences. The total duration of each sentence was 2.2 seconds. To create the four intonation contours, the average pitch trace of recorded Neutral, Question, Emphasis 1, and Emphasis 3 sentences was taken and smoothed. We then applied each intonation contour to each base token using the pitch-synchronous-overlap-add (PSOLA) method (*52*).

Finally, we manipulated baseline pitch and formant values to create three speakers (two female, one male). In general, voices from different speakers are perceived on two main acoustic dimensions, fundamental frequency (f0) and formant frequencies, based on the length of the vocal folds and shape of the vocal tract, respectively (*53*, *54*). To model these two dimensions, the three speakers consisted of a low-pitch, low-formant male speaker (median pitch = 80 Hz, formants lowered by 15% of original recording), a high-pitch, low-formant female speaker (median pitch = 180 Hz, formants lowered by 15% of original recording), and a high-pitch, high-formant female speaker (median pitch = 180 Hz, formants lowered by 5% of original recording). The two female speakers had the same f0, but differing baseline formant frequencies, one of which matched the male speaker's baseline formant frequencies. Baseline pitch values were manipulated using PSOLA and baseline formants were manipulated by shifting the entire sound spectrum while maintaining duration and fundamental frequency. All speech synthesis was done using the linguistics software, PRAAT (*55*).

The resulting stimulus set consisted of 4 intonation contours x 4 sentences x 3 speakers = 48 tokens (Audio S1-48). The tokens were each played twice in one recording block in random order. The total experimental time for each block was about 5 minutes. We collected 2-4 blocks per participant (mean 2.7 blocks).

Data Analysis

All analyses were carried out using custom software written in MATLAB and Python. Open-source scientific Python packages used included numpy, scipy, pandas, scikit-learn, and statsmodels. Figures were created using matplotlib and seaborn. The code used to analyze the data and produce the figures is publicly available on Github at https://github.com/ChangLabUcsf/intonatang. The accompanying documentation can be found at https://changlabucsf.github.io/intonatang.

Raw data, experimental stimuli, and analysis code is accessible at https://doi.org/10.5281/zenodo.826950.

Single-electrode encoding analysis

Single-electrode encoding analyses were not restricted anatomically *a priori*. Using ordinary least-squares regression, we fit encoding models that predicted neural activity (Hγ) from stimulus conditions (e.g. Neutral, Question, Emphasis 1, and Emphasis 3 formed the set of intonation conditions). To determine how variance in the neural activity was explained by the intonation, sentence, and speaker conditions, we represented stimulus conditions and interactions with sets of dichotomous predictor variables. This regression analysis is mathematically equivalent to three-way, crossed ANOVA. To reduce timepoint by timepoint variability, we took neural activity as the average of Hγ in 60 ms windows, moving in 30 ms steps. For each model, the coefficient of determination, R^2, provides a measure of the proportion of variance in neural activity that is explained by stimulus conditions and interactions. The *p*-value associated with the omnibus *F*-statistic provides a measure of significance. We set the significance threshold at $\alpha = 0.05$ and corrected for multiple comparisons using the Bonferroni method, taking individual time points and electrodes as independent samples. The average number of significant electrodes for each participant was 17.7 (min: 5, max: 32).

Variance partitioning and evaluation of interaction terms

The predictor variables were grouped into seven mutually exclusive sets. Three of the seven groups represented the main effects of intonation (In), sentence (Se), and speaker (Sp) condition. An additional three of the seven represented pairwise interactions, intonation × sentence (InSe), intonation × speaker (InSp), and sentence × speaker (SeSp). The last group of predictor variables represented the three-way interaction (InSeSp). For each token with intonation condition *i*, sentence condition *j*, and speaker condition *k*, the high-gamma was modeled as:

$$\mathrm{H}\gamma_{ijk}(\mathrm{t}) = \beta_0(\mathrm{t}) + \beta_{\mathrm{In}}(\mathrm{t}) \cdot \mathrm{In}_i + \beta_{\mathrm{Se}}(\mathrm{t}) \cdot \mathrm{Se}_j + \beta_{\mathrm{Sp}}(\mathrm{t}) \cdot \mathrm{Sp}_k + \beta_{\mathrm{InSe}}(\mathrm{t}) \cdot \mathrm{InSe}_{ij} + \beta_{\mathrm{InSp}}(\mathrm{t}) \cdot \mathrm{InSp}_{ik} + \beta_{\mathrm{SeSp}}(\mathrm{t}) \cdot \mathrm{SeSp}_{jk} + \beta_{\mathrm{InSeSp}}(\mathrm{t}) \cdot \mathrm{InSeSp}_{ijk}$$

The contribution of each group of predictor variables, including the groups for interaction terms, was evaluated by comparing the variance explained by the fully specified model with one that excluded the group. The proportion of variance uniquely explained by each group, R^2_G, was calculated as the difference in R^2 between those two models:

$$R^2_G = R^2_{\mathrm{full}} - R^2_{\mathrm{wo\_}G}$$

The significance of each group of predictors was evaluated using the *F*-test(m, $N{-}k{-}1$) with the following *F*-statistic:

$$F_G = \frac{R^2_G}{m} \Big/ \frac{1 - R^2_{\mathrm{full}}}{N - k - 1}$$

where *m* is the number of predictor variables coding for the group *G*, *k* is the number of predictor variables in the fully specified model, and *N* is the number of trials.

By applying a Bonferroni correction on the *p*-value, this method takes a conservative stance on finding significant values.

De-lexicalized non-speech control stimuli

To determine whether the cortical representation of intonational pitch contours is independent from the processing of phonetic information, we created a set of non-speech control stimuli that completely removed spectral information related to phonetic features and consisted of only a pitch contour (for 5 of 8 subjects who listened to non-speech control stimuli, these stimuli also had amplitude contours corresponding to the amplitude contours of each sentence condition (Audio S49 – S80); remaining 3 of 8 subjects heard (Audio S81 – S88).). These stimuli were created by summing a sinusoid and its second and third harmonics (*29*). The varying frequency of the sinusoid was matched to the pitch contour of each intonation condition.

To test whether neural responses to non-speech stimuli were similar in pattern to responses to the original speech stimuli, we used linear discriminant analysis (LDA) to ask whether the pattern of neural activity that differentiates intonation contours in the speech context generalizes to a non-speech context. To do this, we fit the model using neural responses to speech to predict the intonation condition from the neural activity time series from a single electrode (average Hy in 60 ms windows centered at -0.15 s to 2.85 s in 30 ms steps) and then tested the model on non-speech data. We then determined whether model performance on the non-speech data, measured as classification accuracy, was as good as performance for the speech data.

Specifically, we first computed a distribution of classification accuracies for speech. We trained an LDA classifier on a random 80% of the speech trials and then tested the model on a set with $N_{\text{non-speech}}$ trials bootstrapped from the remaining 20%. To prevent overfitting, we used a form of regularized LDA, diagonal LDA, which uses a diagonal covariance matrix that is shared between classes (i.e. intonation conditions). We performed this procedure 1000 times to arrive at the distribution of accuracies for speech data. We then trained an LDA classifier on 100% of the speech trials and used this model to test the non-speech trials and determined whether this accuracy fell within the 95% of values from the 2.5 to the 97.5 percentile for the speech trials.

Missing fundamental, non-speech control stimuli

To determine whether the neural activity we observed was a response to the psychoacoustic, perceptual attribute of pitch, rather than the physical, acoustic energy at the fundamental frequency (f0), we created a second type of non-speech control stimuli that did not contain energy at f0 (*30*, *31*) (Audio S89 – S96). These missing f0 stimuli contained the fourth, fifth, and sixth harmonics. To mask distortion products that may be introduced at the level of the cochlea (*56*, *57*), we also added pink noise from 0.25s before pitch contour onset through the duration of the stimulus. We presented these stimuli in random order to three participants, while we recorded their cortical activity. We then used the same analysis, described above, to test whether the pattern of neural activity to intonation contours was similar between the speech context and the missing f0 context. Briefly, we trained an LDA classifier to predict intonation condition from neural responses to speech stimuli. We then tested this classifier on neural responses to missing f0 and determined whether classification accuracy was as good for missing f0 data as it was for speech data (see section above for more details).

Absolute and relative pitch temporal receptive field analysis

To investigate how the cortical representation of intonation reflected the encoding of pitch values, we recorded neural activity as participants listened to a subset of the TIMIT continuous speech corpus (*28*). This stimulus set contained sentences spoken by hundreds of male and female speakers allowing for a statistical separation of absolute and relative pitch values (Fig. S3). Vocal pitch values were extracted using an automated autocorrelation method and corrected for halving and doubling errors. We defined absolute pitch as the natural logarithm of pitch values in Hz. There are two main methods used to compute relative pitch. One method normalizes values by interpolating each value between a speaker's minimum pitch and maximum pitch (*58*), while the other uses z-scoring (*59*). Here, we used to the z-score method to compute relative pitch values, and normalized absolute pitch values in ln Hz by the mean and standard deviation of each sentence as a proxy for speaker. We then discretized absolute and relative pitch values into 10 bins, equally spaced in pitch space from the 2.5 percentile to the 97.5 percentile value (Fig. S3H, I). The bottom and top 2.5% of the pitch values were placed into the first and last bins, respectively. By defining these percentile bounds, we prevent unstable estimates that can occur when the top and bottom bins contain too few data points.

To determine how absolute and relative pitch values in speech drive neural activity, we fit temporal receptive field models (*34*) that predicted neural activity from pitch values in the immediately preceding 400 ms window (sampled at 100 Hz) using L_2 regularized multiple linear regression. By including both absolute and relative pitch values as features, we could assess the unique contribution of absolute and relative pitch in predicting neural activity. We also included two additional temporal features using a continuous variable with intensity information and a binary variable when pitch values were present. This binary feature allows us to statistically control for the contribution of the presence of pitch (or voicing in the speech signal) when evaluating the contribution of absolute and relative pitch levels. To calculate the unique contribution of absolute and relative pitch, we calculated the R^2 gained when absolute or relative pitch features, respectively, were included in the model.

We used L_2 regularization (ridge regression) and cross-validation to prevent overfitting since the number of features in temporal receptive field models is typically large (>500). We evaluated the models using the correlation between actual and predicted values of neural activity on held on data. Specifically, we divided the data into three mutually exclusive sets containing 80%, 10%, and 10% of the total number of sentences. The first set of 80% was used as the training set. The second set was used to fit the L_2 regularization hyperparameter, and the final tenth was used as the test set. We performed this procedure 25 times and the performance of the model was taken as the mean of performance across all testing sets.

To calculate the significance of the R^2_{absolute} and R^2_{relative} values computed for each electrode, we used a permutation test. We shuffled the pitch and intensity contours between TIMIT sentences before using the same analysis pipeline to compute null values of R^2_{absolute} and R^2_{relative}. We ran this analysis 200 times to arrive at the null distribution. R^2_{absolute} and R^2_{relative} values above the 95th percentile were considered significant.

Predicted neural activity from pitch temporal receptive fields

To determine whether the absolute or relative pitch temporal receptive field models captured stimulus features that were relevant for intonation-encoding electrodes, we used the pitch temporal receptive field (ptrf) models fit using TIMIT speech data to predict neural responses to the original intonation stimuli containing the four intonation conditions. To determine whether absolute or relative pitch better explained the neural responses, we compared the performance of the different ptrf models. We parameterized the original intonation stimuli using bins for absolute and relative pitch determined by the TIMIT data. We then used the absolute ptrf model and the relative ptrf model to predict the neural responses to the intonation stimuli. Since there are no features in the ptrf models associated with the spectral information that determine phonetic information, we averaged predictions and real neural responses over sentence conditions. We additionally averaged predictions and real neural responses over the two female speakers who did not differ in pitch. We then took the correlation (Pearson's *r*) between the predicted and actual neural responses for the absolute pitch only model ($r_{\text{abs\_pred}}$) and relative pitch only model ($r_{\text{rel\_pred}}$).

Fig. S1

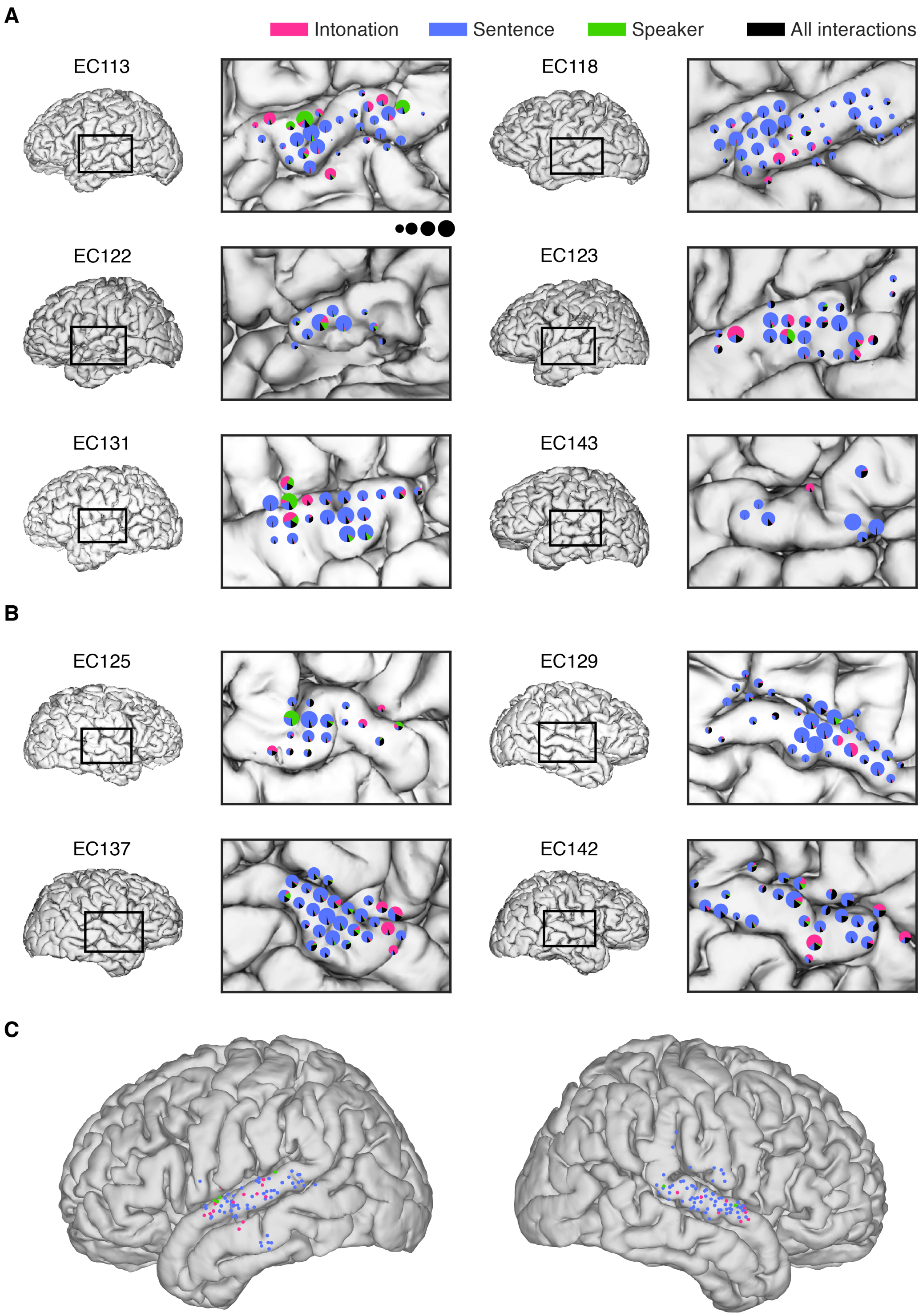

Fig. S1. Individual and group cortical maps of intonation, sentence, and speaker condition encoding.

Individual cortical maps in A and B show the proportion of variance explained by the main effects of intonation, sentence, and speaker condition, as well as the proportion explained by the sum of all pairwise and three-way interactions as pie charts. All significant electrodes (defined as those where the full encoding model was significant for at least 2 time points at $\alpha = 0.05$, Bonferroni corrected) are shown with the area of the pie chart proportional to the total R^2 of the fully specified model. The black circles represent the areas for 25%, 50%, 75%, and 100% of the maximum total R^2 across all significant electrodes for one subject. (A) Maps for participants with left hemisphere grids. (B) Maps for participants with right hemisphere grids. (C) Group cortical map showing Intonation, Sentence, and Speaker electrodes from all ten subjects warped to a common MNI brain.

Fig. S2

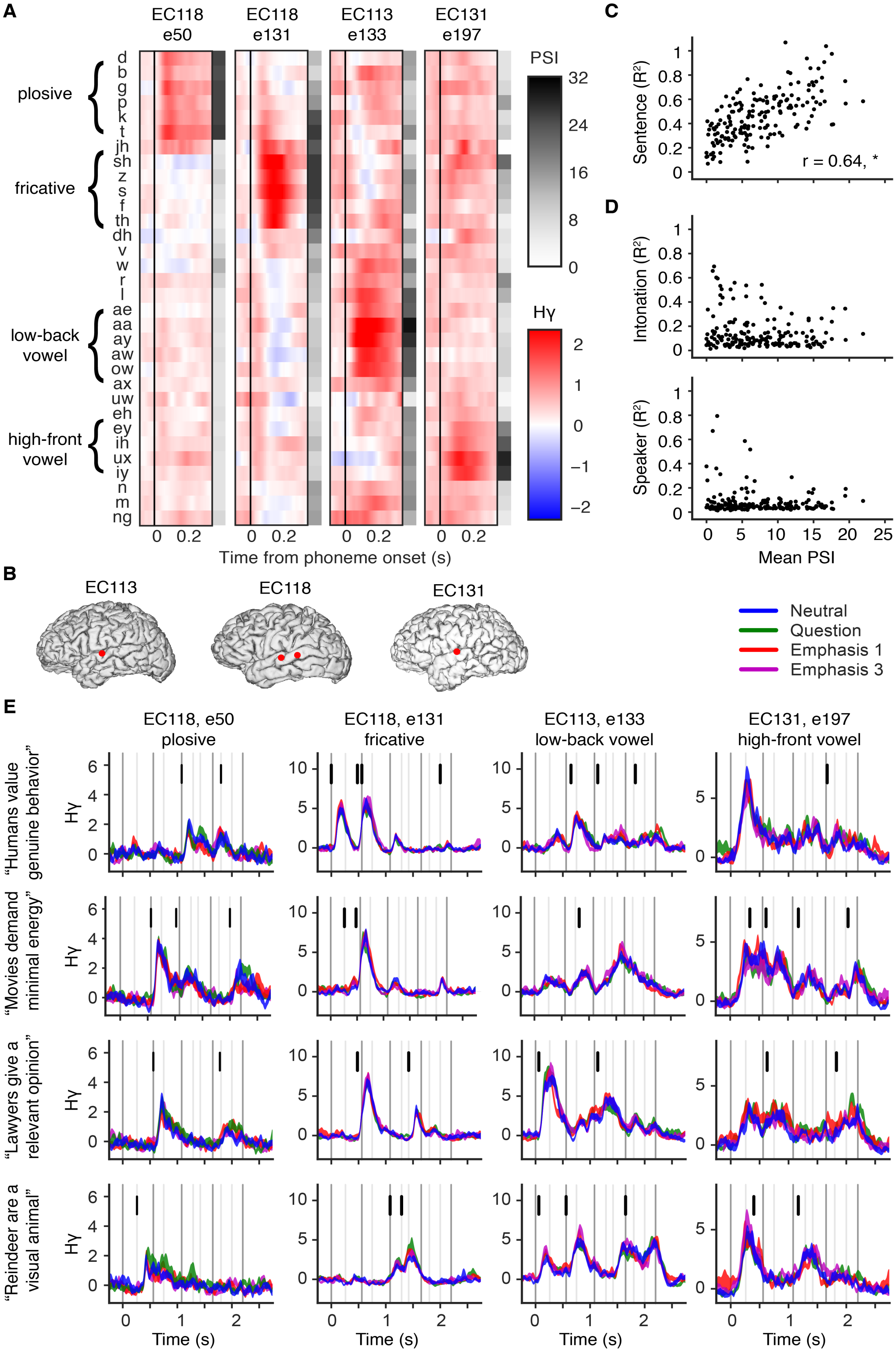
A
EC118 e50
EC118 e131
EC113 e133
EC131 e197
plosive
fricative
low-back vowel
high-front vowel
d b g p k t jh sh z s f th dh v w r l ae aa ay aw ow ax uw eh ey ih ux iy n m ng
0 0.2
Time from phoneme onset (s)
PSI
32 24 16 8 0
Hγ
2 1 0 −1 −2
C
Sentence (R²)
r = 0.64, *
D
Intonation (R²)
Speaker (R²)
1 0.8 0.6 0.4 0.2 0
0 5 10 15 20 25
Mean PSI
B
EC113
EC118
EC131
Neutral
Question
Emphasis 1
Emphasis 3
E
EC118, e50 plosive
EC118, e131 fricative
EC113, e133 low-back vowel
EC131, e197 high-front vowel
"Humans value genuine behavior"
"Movies demand minimal energy"
"Lawyers give a relevant opinion"
"Reindeer are a visual animal"
Hγ
Time (s)

Fig. S2. Activity that differentiates sentence conditions is driven by phonetic feature selectivity.

(A) Average neural responses time-locked to the onsets of individual phonemes from sentences in the TIMIT speech corpus. Each column shows the average response of an individual electrode. The phonetic selectivity index measures whether a response to a given phoneme can be discriminated from the response to all other phonemes. Grouping of phonemes into four phonetic categories is show to the left. (B) Anatomical location of electrodes shown in A. Each electrode is located in the STG. (C) Scatter plot showing each significant electrode's sentence condition encoding and average phonetic selectivity index ($r = 0.64$, p-value $< 1 \times 10^{-20}$). Data from 177 significant electrodes across 10 participants are shown. (D) Scatter plots of intonation encoding and average PSI on the top and of speaker encoding and average PSI on the bottom ($r = -0.18$, p-value < 0.05; $r = -0.15$, p-value > 0.05, respectively). (E) Average neural response of each example electrode in A to original stimulus set. Each row shows responses to a different sentence. For each column, tick marks indicate the onsets of phonemes which fall into the class written at the top of the column. The responses are colored by intonation condition. These phonetically-selective, Sentence electrodes are not sensitive to intonation and have a similar response regardless of what the intonation condition was.

Fig. S3

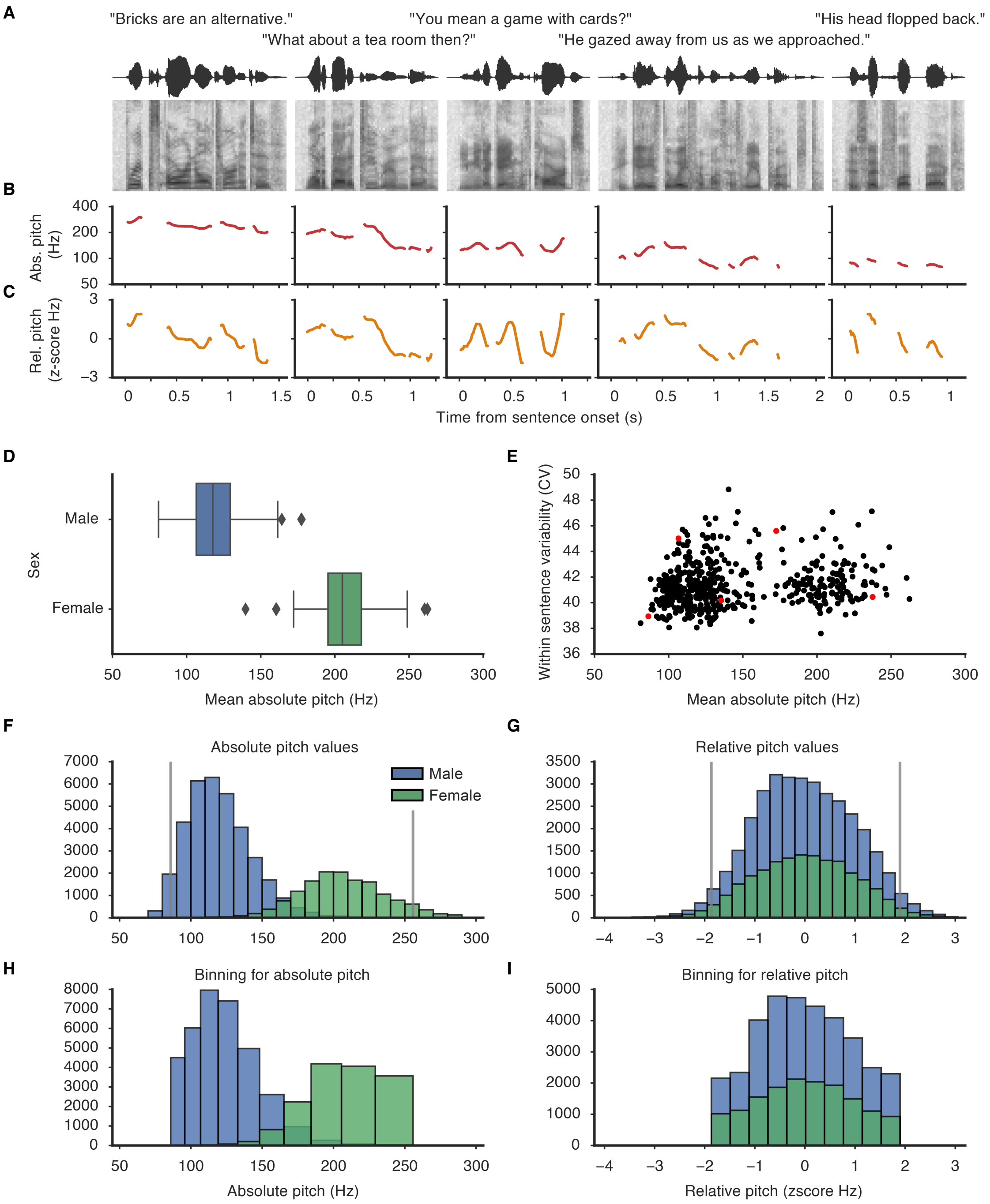

Fig. S3. Absolute and relative pitch contours and variability in TIMIT speech corpus.

The TIMIT corpus contains continuous speech recorded from hundreds of male and female speakers. Having many speakers who differ in their baseline absolute pitch allows for the statistical separation of absolute and relative pitch. (A) Five example tokens from the TIMIT dataset with their acoustic amplitude signal and spectrograms. The first two were spoken by female speakers while the last three were spoken by male speakers. (B) Absolute pitch contours for the five tokens in (A). The fundamental frequency was extracted from the speech signal using an autocorrelation method. (C) Relative pitch contours for the five tokens in (A). The relative pitch was calculated as the z-score of absolute pitch values (in ln Hz) for each token. (D) Distribution of mean pitch values for the speakers in the subset of TIMIT used in this study. (E) Scatterplot of mean pitch values and pitch variability (coefficient of variation expressed as percentage of the mean). Each dot represents one token and red dots indicate the five tokens from (A). (F, G) Histogram of all the absolute pitch (F) and relative pitch (G) values calculated from TIMIT tokens, with values from male and female speakers shown separately. Gray lines indicate the 2.5 and 97.5 percentile. (H, I) Histograms showing the binning used to parameterize absolute (H) and relative (I) pitch values for the pitch temporal receptive field models. Ten equally spaced bins (for absolute pitch, bins are equally spaced on a logarithmic scale) were created between the 2.5 and 97.5 percentile of all pitch values.

Fig. S4

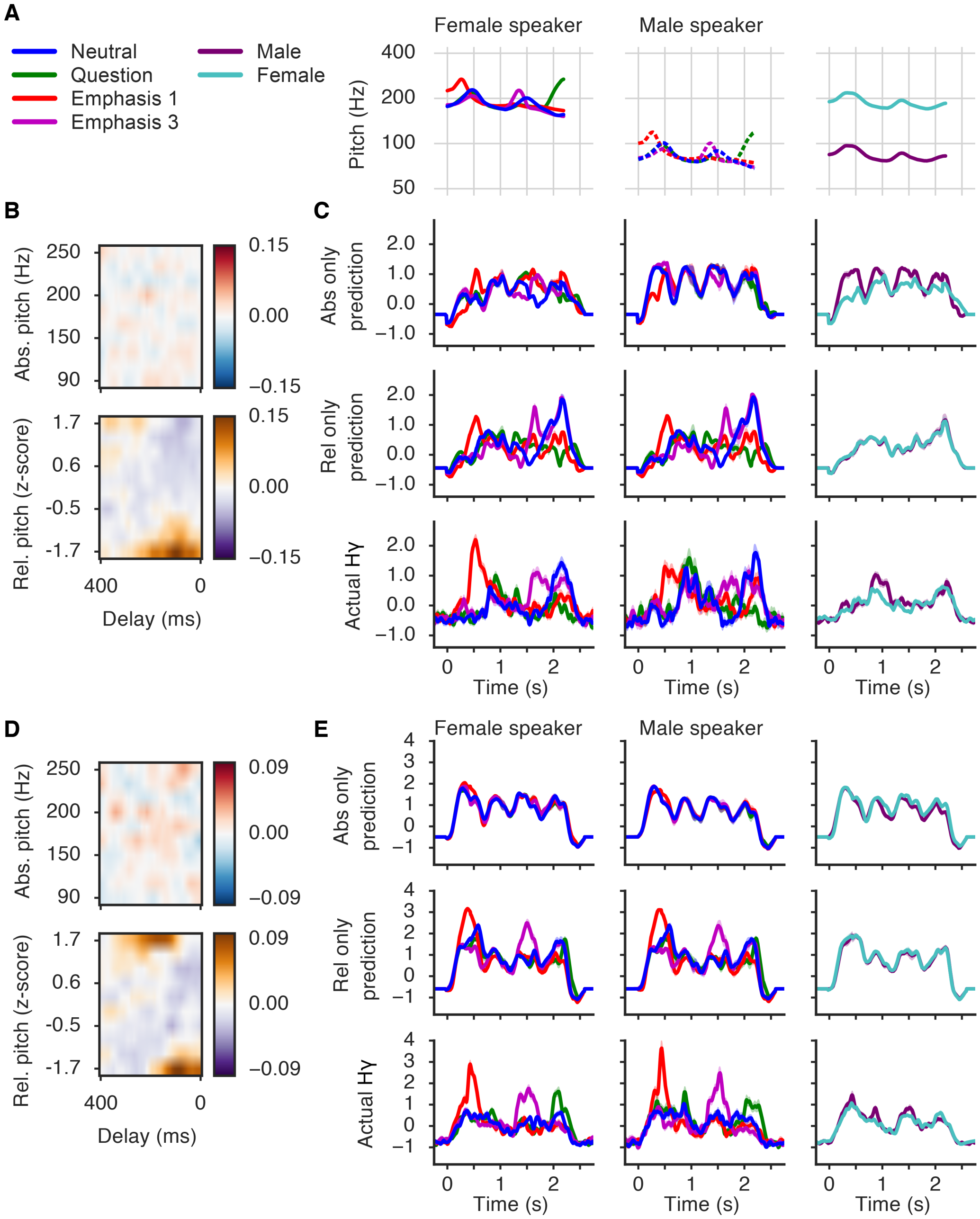

Fig. S4. Example relative pitch encoding electrodes tuned for low relative pitch and high-to-low relative pitch.

The pitch temporal receptive field (ptrf) fit using neural responses to the TIMIT speech corpus and the prediction of this encoding model to the original set of stimuli are shown for two example relative pitch encoding electrodes. The ptrf indicates how different values of absolute and relative pitch at different time delays affects neural activity. (A) Pitch contours of original set of stimuli. The left panel shows the pitch contours for each intonation condition for the female speakers. The middle panel shows the pitch contours for each intonation condition for the male speaker. The right panel shows the average pitch contour for the male versus the female speakers. (B) Electrode that encoded relative pitch ($R^2_{\text{relative}} = 0.07$, significant by permutation test; $R^2_{\text{absolute}} = -0.01$, not significant) and was tuned to low relative pitch. (C) The top two rows show the predicted neural responses from the absolute pitch only and relative pitch only models. The bottom row shows the actual neural responses. The actual response of this electrode to the original stimulus set was better predicted by the relative pitch only model ($r_{\text{rel\_pred}} = 0.76$; $r_{\text{abs\_pred}} = 0.55$). (D) Electrode that encoded relative pitch ($R^2_{\text{relative}} = 0.02$, significant by permutation test; $R^2_{\text{absolute}} = -0.01$, not significant) and was tuned to high relative pitch at a delay of ~180 ms and low relative pitch at a delay of ~100 ms. (E) The activity on this electrode was better predicted by the relative pitch only model than the absolute pitch only one ($r_{\text{rel\_pred}} = 0.85$; $r_{\text{abs\_pred}} = 0.74$).

Fig. S5

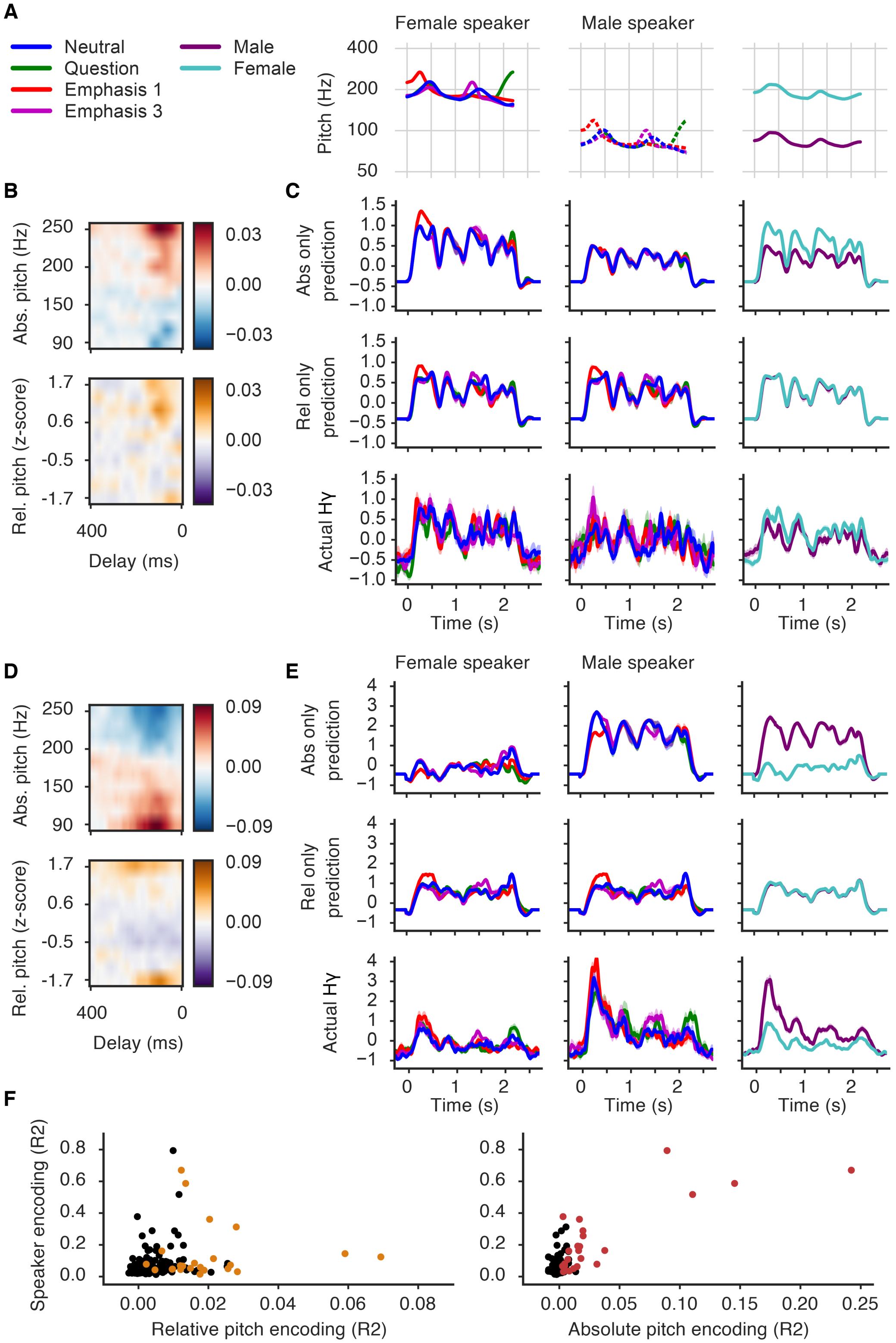
A
Neutral
Question
Emphasis 1
Emphasis 3
Male
Female
Female speaker
Male speaker
Pitch (Hz)
400
200
100
50
B
Abs. pitch (Hz)
250
200
150
90
0.03
0.00
−0.03
Rel. pitch (z-score)
1.7
0.6
-0.5
-1.7
400
0
Delay (ms)
C
Abs only prediction
Rel only prediction
Actual Hγ
1.5
1.0
0.5
0.0
−0.5
−1.0
Time (s)
0
1
2
D
0.09
0.00
−0.09
E
4
3
2
1
0
−1
F
Speaker encoding (R2)
0.8
0.6
0.4
0.2
0.0
0.00
0.02
0.04
0.06
0.08
Relative pitch encoding (R2)
0.00
0.05
0.10
0.15
0.20
0.25
Absolute pitch encoding (R2)

Fig. S5. Example absolute pitch encoding electrodes tuned for high and low absolute pitch.

The pitch temporal receptive field, predicted responses of the ptrf model to the original set of stimuli, and the actual responses are shown for two example absolute pitch encoding electrodes. The ptrf indicates how different values of absolute and relative pitch at different time delays affects neural activity. (A) Pitch contours of original set of stimuli. (B) Electrode that encoded absolute pitch ($R^2_{\text{relative}} = 0.00$, not significant by permutation test; $R^2_{\text{absolute}} = 0.02$, significant) and was tuned to absolute pitch values greater than 180 Hz. (C) The top two rows show the predicted neural responses from the absolute pitch only and relative pitch only models. The bottom row shows the actual neural responses. This electrode had a greater response to the female speakers than the male speaker. The actual response of this electrode was better predicted by the absolute pitch only model ($r_{\text{rel\_pred}} = 0.75$; $r_{\text{abs\_pred}} = 0.78$). (D, E) Electrode that encoded absolute pitch and was tuned to low absolute pitch ($R^2_{\text{relative}} = 0.01$, significant; $R^2_{\text{absolute}} = 0.15$, significant; $r_{\text{rel\_pred}} = 0.59$; $r_{\text{abs\_pred}} = 0.68$). (F) Scatterplot between relative and absolute pitch encoding with neural discriminability of speaker conditions ($r_{\text{relative\_speaker}} = 0.21$, p-value < 0.05; $r_{\text{absolute\_speaker}} = 0.79$, p-value < 1×10^{-38}). Colored markers indicate electrodes with significant (permutation test; $R^2 >$ 95th percentile of null distribution) relative pitch and absolute pitch encoding for the left and right panels, respectively.

Subject	Hem	Age	Sex	Handedness	Language dominance	Epilepsy focus
EC113	L	22	M	R	L	Left hippocampus and anterior temporal lobe
EC118	L	31	F	R	L	Left insula and temporal lobe
EC122	L	28	M	R	L	Left anterior temporal lobe
EC123	L	33	F	R	L	Left temporal lobe
EC125	R	35	M	R	L	Right anterior medial temporal lobe
EC129	R		F	R	L	Right superior frontal gyrus
EC131	L	28	M	R	L	Left anterior-mesial temporal lobe
EC137	R	20	M	R	L	Right hippocampus
EC142	R	20	M	R	L	Right supramarginal gyrus
EC143	L	21	M	R	L	Left superior temporal gyrus

Table S1. Clinical and demographic details for subjects.
Hem = hemisphere of implantation, R = right, L = left

Audio S1 – S48:

Synthesized set of speech stimuli that independently varies intonation contour, phonetic content, and speaker. Each individual .wav file is named sn*X*_st*Y*_sp*Z*.wav, where *X* is 1-4, *Y* is 1-4, and *Z* is 1-3. *X* indicates the sentence condition (sentence number or sn). *Y* indicates the intonation contour (sentence type or st). *Z* indicates the speaker (speaker or sp).

Audio S49 – S80:

Set of non-speech stimuli that preserves intonational pitch contour but removes spectral content related to phonetic features. These stimuli also have varying amplitude contours corresponding to the sentence condition from the original set of speech stimuli. Each wav file is named purr_*Z*_st*Y*_sn*X*, where *Z* is either "female" or "male", *Y* indicates the intonation contour (1: Neutral, 2: Question, 3: Emphasis 1, 4: Emphasis 3), and *X* indicates which sentence condition the amplitude contour came from (1-4). These stimuli were played to 5/10 total participants.

Audio S81 – S88:

Set of non-speech stimuli that preserves intonational pitch contour, but removes spectral content related to phonetic features. These stimuli have flat amplitude contours. Each wav file is named purr_stretch_0_*Z*_st*Y*, where *Z* is either "female" or "male" and *Y* indicates the intonation contour (1: Neutral, 2: Question, 3: Emphasis 1, 4: Emphasis 3). These stimuli were played to 3/10 total participants.

Audio S89 – S96:

Missing fundamental stimuli that preserves intonational pitch contour. These stimuli are the combination of the fourth, fifth, and sixth harmonics of the fundamental frequency contour with pink noise added 0.25 before pitch contour onset to mask energy at the fundamental frequency that may be introduced at the level of the cochlea. Each wav file is named purr_missing_f0_noise_first_stretch_0_*Z*_st*Y*, where *Z* is either "female" or "male" and *Y* indicates the intonation contour (1: Neutral, 2: Question, 3: Emphasis 1, 4: Emphasis 3). These stimuli were played to 3/10 total participants.

References and Notes

1. A. Cutler, D. Dahan, W. van Donselaar, Prosody in the comprehension of spoken language: A literature review. *Lang. Speech* **40**, 141–201 (1997). doi:10.1177/002383099704000203 Medline

2. D. R. Ladd, *Intonational Phonology* (Cambridge Univ. Press, 2008).

3. S. Shattuck-Hufnagel, A. E. Turk, A prosody tutorial for investigators of auditory sentence processing. *J. Psycholinguist. Res.* **25**, 193–247 (1996). doi:10.1007/BF01708572 Medline

4. I. R. Titze, Physiologic and acoustic differences between male and female voices. *J. Acoust. Soc. Am.* **85**, 1699–1707 (1989). doi:10.1121/1.397959 Medline

5. E. D. Ross, The aprosodias. Functional-anatomic organization of the affective components of language in the right hemisphere. *Arch. Neurol.* **38**, 561–569 (1981). doi:10.1001/archneur.1981.00510090055006 Medline

6. K. M. Heilman, D. Bowers, L. Speedie, H. B. Coslett, Comprehension of affective and nonaffective prosody. *Neurology* **34**, 917–921 (1984). doi:10.1212/WNL.34.7.917 Medline

7. M. D. Pell, S. R. Baum, The ability to perceive and comprehend intonation in linguistic and affective contexts by brain-damaged adults. *Brain Lang.* **57**, 80–99 (1997). doi:10.1006/brln.1997.1638 Medline

8. J. Witteman, M. H. van Ijzendoorn, D. van de Velde, V. J. J. P. van Heuven, N. O. Schiller, The nature of hemispheric specialization for linguistic and emotional prosodic perception: A meta-analysis of the lesion literature. *Neuropsychologia* **49**, 3722–3738 (2011). doi:10.1016/j.neuropsychologia.2011.09.028 Medline

9. E. Plante, M. Creusere, C. Sabin, Dissociating sentential prosody from sentence processing: Activation interacts with task demands. *Neuroimage* **17**, 401–410 (2002). doi:10.1006/nimg.2002.1182 Medline

10. M. Meyer, K. Alter, A. D. Friederici, G. Lohmann, D. Y. von Cramon, FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Hum. Brain Mapp.* **17**, 73–88 (2002). doi:10.1002/hbm.10042 Medline

11. M. Meyer, K. Steinhauer, K. Alter, A. D. Friederici, D. Y. von Cramon, Brain activity varies with modulation of dynamic pitch variance in sentence melody. *Brain Lang.* **89**, 277–289 (2004). doi:10.1016/S0093-934X(03)00350-X Medline

12. J. Gandour, Y. Tong, D. Wong, T. Talavage, M. Dzemidzic, Y. Xu, X. Li, M. Lowe, Hemispheric roles in the perception of speech prosody. *Neuroimage* **23**, 344–357 (2004). doi:10.1016/j.neuroimage.2004.06.004 Medline

13. C. P. Doherty, W. C. West, L. C. Dilley, S. Shattuck-Hufnagel, D. Caplan, Question/statement judgments: An fMRI study of intonation processing. *Hum. Brain Mapp.* **23**, 85–98 (2004). doi:10.1002/hbm.20042 Medline

14. A. D. Friederici, K. Alter, Lateralization of auditory language functions: A dynamic dual pathway model. *Brain Lang.* **89**, 267–276 (2004). doi:10.1016/S0093-934X(03)00351-1 Medline

15. Y. Tong, J. Gandour, T. Talavage, D. Wong, M. Dzemidzic, Y. Xu, X. Li, M. Lowe, Neural circuitry underlying sentence-level linguistic prosody. *Neuroimage* **28**, 417–428 (2005). doi:10.1016/j.neuroimage.2005.06.002 Medline

16. J. Kreitewolf, A. D. Friederici, K. von Kriegstein, Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *Neuroimage* **102**, 332–344 (2014). doi:10.1016/j.neuroimage.2014.07.038 Medline

17. T. D. Griffiths, C. Büchel, R. S. Frackowiak, R. D. Patterson, Analysis of temporal structure in sound by the human brain. *Nat. Neurosci.* **1**, 422–427 (1998). doi:10.1038/1637 Medline

18. 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). doi:10.1016/S0896-6273(02)01060-7 Medline

19. H. Penagos, J. R. Melcher, A. J. Oxenham, A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *J. Neurosci.* **24**, 6810–6815 (2004). doi:10.1523/JNEUROSCI.0383-04.2004 Medline

20. M. Schönwiesner, R. J. Zatorre, Depth electrode recordings show double dissociation between pitch processing in lateral Heschl's gyrus and sound onset processing in medial Heschl's gyrus. *Exp. Brain Res.* **187**, 97–105 (2008). doi:10.1007/s00221-008-1286-z Medline

21. D. Bendor, X. Wang, The neuronal representation of pitch in primate auditory cortex. *Nature* **436**, 1161–1165 (2005). doi:10.1038/nature03867 Medline

22. W. A. van Dommelen, Acoustic parameters in human speaker recognition. *Lang. Speech* **33**, 259–272 (1990). doi:10.1177/002383099003300302 Medline

23. M. Steinschneider, Y. I. Fishman, J. C. Arezzo, Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cereb. Cortex* **18**, 610–625 (2008). doi:10.1093/cercor/bhm094 Medline

24. S. Ray, J. H. R. Maunsell, Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLOS Biol.* **9**, e1000610 (2011). doi:10.1371/journal.pbio.1000610 Medline

25. 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). doi:10.1152/jn.90954.2008 Medline

26. N. E. Crone, D. Boatman, B. Gordon, L. Hao, Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* **112**, 565–582 (2001). doi:10.1016/S1388-2457(00)00545-9 Medline

27. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014). doi:10.1126/science.1245994 Medline

28. J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, V. Zue, “TIMIT acoustic-phonetic continuous speech corpus,” Linguistic Data Consortium (1993); https://catalog.ldc.upenn.edu/ldc93s1.

29. G. P. Sonntag, T. Portele, PURR—A method for prosody evaluation and investigation. *Comput. Speech Lang.* **12**, 437–451 (1998). doi:10.1006/csla.1998.0107

30. J. C. R. Licklider, “Periodicity” pitch and “place” pitch. *J. Acoust. Soc. Am.* **26**, 945 (1954). doi:10.1121/1.1928005

31. C. J. Plack, A. J. Oxenham, R. R. Fay, *Pitch: Neural Coding and Perception* (Springer, 2006), vol. 24.

32. P. C. M. Wong, R. L. Diehl, Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *J. Speech Lang. Hear. Res.* **46**, 413–421 (2003). doi:10.1044/1092-4388(2003/034) Medline

33. C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, J. Terken, The perceptual prominence of fundamental frequency peaks. *J. Acoust. Soc. Am.* **102**, 3009–3022 (1997). doi:10.1121/1.420355 Medline

34. 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. *Netw. Comput. Neural Syst.* **12**, 289–316 (2001). doi:10.1080/net.12.3.289.316 Medline

35. P. Belin, S. Fecteau, C. Bédard, Thinking the voice: Neural correlates of voice perception. *Trends Cogn. Sci.* **8**, 129–135 (2004). doi:10.1016/j.tics.2004.01.008 Medline

36. P. Belin, P. E. G. Bestelmeyer, M. Latinus, R. Watson, Understanding voice perception. *Br. J. Psychol.* **102**, 711–725 (2011). doi:10.1111/j.2044-8295.2011.02041.x Medline

37. R. Jakobson, G. Fant, M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates* (MIT Press, 1951).

38. J. H. McDermott, A. J. Lehr, A. J. Oxenham, Is relative pitch specific to pitch? *Psychol. Sci.* **19**, 1263–1271 (2008). doi:10.1111/j.1467-9280.2008.02235.x Medline

39. K. R. Kluender, J. A. Coady, M. Kiefte, Sensitivity to change in perception of speech. *Speech Commun.* **41**, 59–69 (2003). doi:10.1016/S0167-6393(02)00093-6 Medline

40. J. D. Warren, S. Uppenkamp, R. D. Patterson, T. D. Griffiths, Separating pitch chroma and pitch height in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 10038–10042 (2003). doi:10.1073/pnas.1730682100 Medline

41. E. J. Allen, P. C. Burton, C. A. Olman, A. J. Oxenham, Representations of pitch and timbre variation in human auditory cortex. *J. Neurosci.* **37**, 1284–1293 (2017). Medline

42. R. J. Zatorre, K. Delhommeau, J. M. Zarate, Modulation of auditory cortex response to pitch variation following training with microtonal melodies. *Front. Psychol.* **3**, 544 (2012). doi:10.3389/fpsyg.2012.00544 Medline

43. E. Formisano, F. De Martino, M. Bonte, R. Goebel, "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008). doi:10.1126/science.1164318 Medline

44. P. A. Cariani, B. Delgutte, Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J. Neurophysiol.* **76**, 1698–1716 (1996). Medline

45. P. A. Cariani, B. Delgutte, Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. *J. Neurophysiol.* **76**, 1717–1734 (1996). Medline

46. Y. I. Fishman, C. Micheyl, M. Steinschneider, Neural representation of harmonic complex tones in primary auditory cortex of the awake monkey. *J. Neurosci.* **33**, 10312–10323 (2013). doi:10.1523/JNEUROSCI.0020-13.2013 Medline

47. J. K. Bizley, K. M. M. Walker, B. W. Silverman, A. J. King, J. W. H. Schnupp, Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *J. Neurosci.* **29**, 2064–2075 (2009). doi:10.1523/JNEUROSCI.4755-08.2009 Medline

48. J. K. Bizley, K. M. M. Walker, A. J. King, J. W. H. Schnupp, Neural ensemble codes for stimulus periodicity in auditory cortex. *J. Neurosci.* **30**, 5078–5091 (2010). doi:10.1523/JNEUROSCI.5475-09.2010 Medline

49. J. K. Bizley, K. M. M. Walker, F. R. Nodal, A. J. King, J. W. H. Schnupp, Auditory cortex represents both pitch judgments and the corresponding acoustic cues. *Curr. Biol.* **23**, 620–625 (2013). doi:10.1016/j.cub.2013.03.003 Medline

50. J. Pierrehumbert, The perception of fundamental frequency declination. *J. Acoust. Soc. Am.* **66**, 363–369 (1979). doi:10.1121/1.383670 Medline

51. C.-Y. Lee, Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *J. Acoust. Soc. Am.* **125**, 1125–1137 (2009). doi:10.1121/1.3050322 Medline

52. E. Moulines, F. Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* **9**, 453–467 (1990). doi:10.1016/0167-6393(90)90021-Z

53. G. Fant, *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations* (De Gruyter, 1971), vol. 2.

54. O. Baumann, P. Belin, Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychol. Res.* **74**, 110–120 (2010). doi:10.1007/s00426-008-0185-z Medline

55. P. Boersma, Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2002).

56. A. J. Oxenham, C. J. Plack, A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing. *J. Acoust. Soc. Am.* **101**, 3666–3675 (1997). doi:10.1121/1.418327 Medline

57. D. Pressnitzer, R. Patterson, "Distortion products and the pitch of harmonic complex tones," in *Physiological and Psychological Bases of Auditory Function*, D. J. Breebaart, A. J. M.

Houtsma, A. Kohlrausch, V. F. Prijs, R. Schoonhoven, Eds. (Maastricht Shaker, 2001), pp. 97–104.

58. M. A. Earle, *An Acoustic Phonetic Study of Northern Vietnamese Tones* (Speech Communications Research Laboratory, 1975).

59. P. Rose, Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Commun.* **6**, 343–352 (1987). doi:10.1016/0167-6393(87)90009-4