Cortical substrates of perceptual confusion between pitch and timbre

Yongtian Ou[1,*], Emily J. Allen[1], Kendrick N. Kay[1,2], Andrew J. Oxenham[1]

[1] Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA.
[2] Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN 55455, USA.
[*] Correspondence: ou000036@umn.edu.

Abstract

Pitch and timbre, two fundamental perceptual attributes of sound, are commonly regarded as distinct features, but can be confused when varied simultaneously. Here we combine human behavior and fMRI to demonstrate a neural substrate to explain the well-known perceptual confusion effects. We identify orderly mappings of both pitch and timbral brightness within auditory cortex and reveal two independent lines of evidence for cortical confusion between them. First, voxels' preferred pitch decreases systematically as brightness increases, and vice versa, consistent with predictions based on perceptual confusion. Second, pitch and brightness mapping share a common high-low-high gradient across auditory cortex, implying a shared trajectory of cortical activation for changes in each dimension. The results provide a cortical substrate at both local and global scales for an established auditory perceptual phenomenon that is thought to reflect efficient coding of ubiquitous features in natural sound statistics.

Introduction

Pitch and timbre are commonly regarded as two distinct perceptual features of sound. In music, pitch defines a melody, whereas timbre describes the sound quality (e.g., of a musical instrument) (*1*). Pitch and timbre also have different acoustic correlates. The pitch of a sound is determined primarily by its repetition rate, or fundamental frequency (F0) (*2*). The timbre of a sound can be effectively manipulated by varying its spectral centroid (Fc) (*3*), which is a physical correlate of timbral brightness. Timbral brightness is generally regarded as one of the most salient dimensions of timbre and has been invariably identified as one of the primary dimensions of timbre in multidimensional scaling approaches (*3–5*).

Although treated as separate dimensions, pitch and timbral brightness are not completely independent, and numerous perceptual studies have found evidence for confusion between them. For example, when two tones are presented sequentially, a small increase in brightness from the first to the second can be misconstrued as an increase in pitch (*6*). Similarly, when pitch and brightness vary simultaneously, listeners' estimates of the pitch interval are larger for a given change in F0 when the brightness varies in the same direction (congruent) than when it moves in the opposite direction (incongruent) (*7*). Finally, listeners' judgments of the direction of pitch and brightness changes are faster and more accurate in congruent than in incongruent conditions (*8*, *9*). Taken together, these studies demonstrate that pitch and timbral brightness changes can be confused with each other. This confusion may be learned in early perceptual development (*10*) and may be a consequence of efficient neural encoding of dimensions that often co-vary in everyday stimuli, such as speech and music (*11*, *12*).

Previous studies have investigated the cortical representation of pitch and brightness independently. Variations in either can elicit distinguishable patterns of activation in regions of auditory cortex (*13*, *14*). Both pitch and brightness have shown an orderly mapping (or tonotopic pattern) across the cortex (*15*). Brightness mapping tends to be restricted mainly to regions of the primary auditory cortex, as expected from the tonotopy inherited from the peripheral auditory system and observed with pure tones (*15*). In contrast, regions sensitive to pitch tend to be more diffuse, extending to regions outside primary auditory cortex (*16–23*), where pitch mapping is most prominent (*15*).

While cortical sensitivity to both pitch and brightness has been established, it is unknown whether changes in one dimension affect the encoding of the other. This is an important gap in our knowledge, as such interactions could provide a neural substrate for the perceptual confusion effects observed in behavioral studies. Here, using stimuli that vary in both pitch and timbral brightness, we first confirm and quantify the perceptual confusion between pitch and brightness. We then report evidence for cortical interactions between the two dimensions within individual voxels, as well as overlapping gradients in both pitch and brightness mapping across auditory cortex. Taken together, the results provide a potential neural substrate for the well-known but poorly understood perceptual confusion between pitch and timbral brightness.

Results

Perceptual confusion between pitch and timbre

To confirm and quantify the perceptual confusion between pitch and timbral brightness in our stimuli, we measured the effect of pitch changes (manipulated by varying F0) on judgments of brightness changes (manipulated by varying Fc), and *vice versa*. The amount of change in the non-target dimension (brightness or pitch) varied from -2 to +2 difference limens (DLs) in steps of 1 DL, adjusted individually for each participant, where positive numbers imply congruent changes (both target and non-target dimensions change in the same direction). If the two dimensions were judged independently, changes in the non-target dimension should have no effect on performance. In contrast, we observed a significant positive relationship between performance and change in the non-target dimension, consistent with perceptual confusion between the two dimensions (Fig. 1A). Our paradigm enabled us to quantify the contribution, or perceptual weight, of the non-target dimension on the target discrimination, in terms of the slope of the linear functions shown in Fig. 1A: the steeper the slope, the more interference the non-target dimension produces on perceptual judgments (Fig. 1B; see *Data and statistical analysis* in Behavioral Experiment for details of the transformation). The perceptual weights assigned to the non-target dimension were significantly greater than zero for both pitch and brightness judgments (perceptual weight of brightness on pitch judgments = 0.51; two-tailed Wilcoxon signed-rank test: $N = 20$, $z = 3.73$, $p < 0.001$; perceptual weight of pitch on brightness judgments = 0.30; two-tailed Wilcoxon signed-rank test: $N = 20$, $z = 3.10$, $p = 0.0019$). The effect of irrelevant brightness changes on pitch judgments was somewhat greater than the effect of irrelevant pitch changes on brightness judgments (two-tailed Wilcoxon signed-rank test: $N = 20$, $z = 2.02$, $p = 0.044$). The positive weights indicate that participants' discrimination was facilitated when the two features changed in the same direction and was impaired when they changed in opposite directions. That fact that the weights were significantly greater than zero confirms that participants are unable to ignore simultaneous changes in one feature (brightness or pitch) when they are judging the direction of change in the other feature (pitch or brightness), despite the feedback provided after each trial throughout the task.

Although some studies suggest that sensitivity to pitch changes may differ between tone language speakers and non-tone language speakers (*24*, *25*), our data provided no evidence for differences in F0DL, FcDL, or the perceptual interference for either pitch or brightness judgments between the two groups (Mann-Whitney U-tests: $N_{\text{tone}} = 9$, $N_{\text{nontone}} = 11$; $U_{\text{F0DL}} = 100$, $p = 0.71$; $U_{\text{FcDL}} = 90$, $p = 0.77$; $U_{\text{pitch judgment}} = 101$, $p = 0.66$; $U_{\text{brightness judgment}} = 98$, $p = 0.82$; Fig. S1 and Fig. 1C). Musical training is known to facilitate pitch discrimination (*26*, *27*), and our data are consistent with these earlier findings (Spearman's correlation: $N = 20$; $\rho_{\text{F0DL}} = -0.744$, $p < 0.001$; Fig. S2). However, we found no effect of years of musical training on the perceptual interference on either pitch or brightness judgments (Spearman's correlation: $N = 20$, $\rho_{\text{pitch judgment}} = 0.184$, $p = 0.46$; $\rho_{\text{brightness judgment}} = 0.176$, $p = 0.33$; Fig. 1D), consistent with the results from only prior study to compare such interference between musicians and non-musicians (*8*). The fact that neither language background nor musical training appears to affect the perceptual interference between pitch and brightness justifies the inclusion of native tone- and non-tone-language speakers and participants with a wide range of musical experience for the fMRI portion of the study.

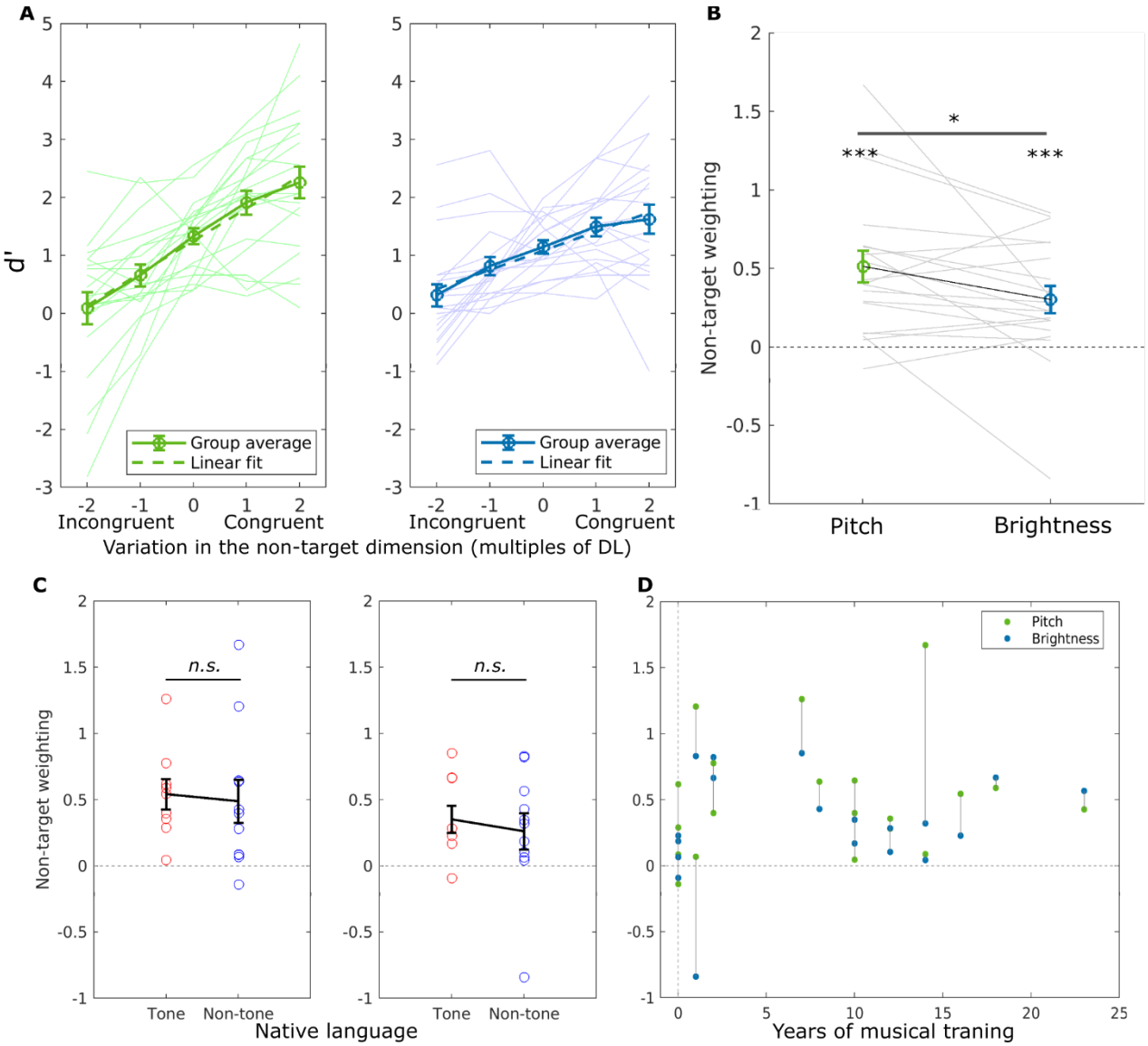

Fig. 1 | Behavioral confusion between pitch and timbral brightness. (A) Discrimination of the pitch (left) and brightness (right) as a function of the variation in the non-target dimension. A positive slope implies that increases in pitch or brightness are confused for increases in the other dimension. Thick solid line: group average; thick dashed line: linear fit of the group average; thin solid lines: individual data. **(B)** Perceptual weight of the non-target dimension on judgments of the target dimension, obtained from individual participants' linear fits in (A). Black line connects group average perceptual weights for pitch and timbre discrimination. Grey lines connect perceptual weights of individual participants. **(C)** Comparison of behavioral confusion between tone (red) and non-tone (blue) language speakers for pitch judgments (left panel) and brightness judgments (right panel). **(D)** Relationship between years of musical training and perceptual interference of brightness on pitch judgments (green) and pitch on brightness judgments (blue). Vertical gray lines connect results obtained from the same participant. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; *n. s.*: not significant. Error bars: ±1 standard error of the mean (SEM) across participants.

Cortical tuning model incorporating interactions provides best description of data

We measured fMRI responses (3T, 2mm^3 resolution) from a subset of ten participants from the behavioral experiment while they listened to tones that varied across five F0 values and five Fc values

with each trial sampling a specific pair of F0 and Fc values (see Fig. 2A and 2B and *Stimuli* in MRI Experiment). The conditions, consisting of 25 F0-Fc pairs, were presented in a pseudorandom order with each condition repeated 48 times per participant. We resampled the fMRI data onto the mid-grey cortical surface (*28*) (see Fig. 2C) and computed percent BOLD change ("beta") for each trial using GLMsingle (*29*). To characterize the tuning properties for each vertex (the smallest unit that was imaged on the cortical surface; subsequently referred to as "voxel" for simplicity) within the regions of auditory cortex, we averaged beta values across repetitions of each condition and fitted five candidate Gaussian-based encoding models (see Fig. 3A and *Fitting and evaluating encoding models*).

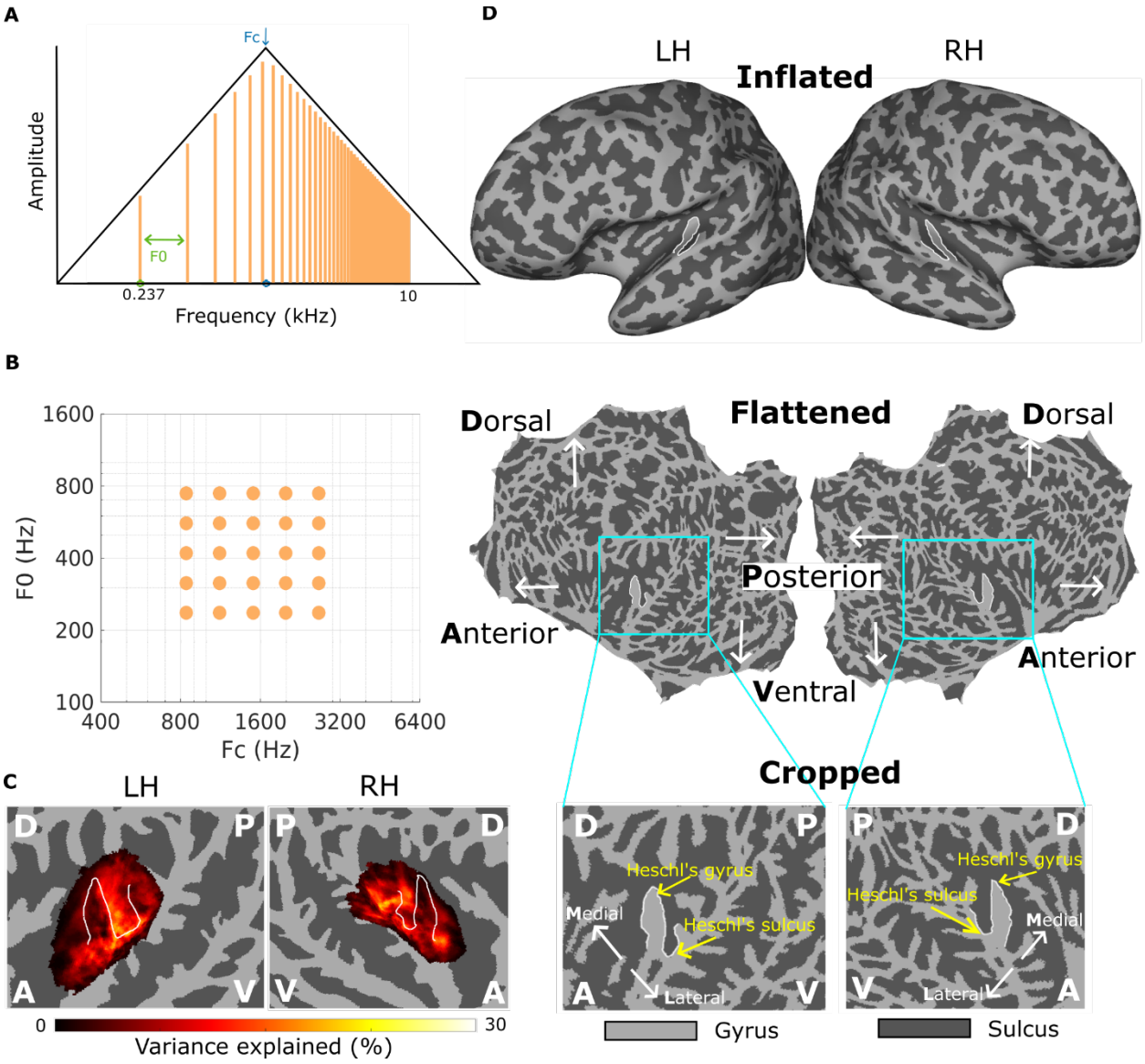

Fig. 2 | Schematic diagrams of the stimuli and visualization of the auditory cortex. (A) Spectrum of a harmonic complex tone used in the study plotted on log-log axes. The spacing between frequency components corresponds to the F0 (perceived as pitch). The peak of the spectrum indicates Fc (perceived as brightness). **(B)** The 25 dots represent the 25 pairs of Fc and F0 values used for the stimuli of the study. **(C)** Steps to prepare individual brain maps for group averaging. The inflated cortical surface (top row) is constructed from the individual participants' anatomical T1- and T2-weighted images. The inflated brain is then flattened and mapped onto a standard fsaverage space (middle row). Two rectangular areas, one for each hemisphere (bottom row), are then specified on this standard space. **(D)** Group-averaged variance explained by all sound stimuli against baseline silence within the group-level ROIs (regions of interest, regions that cover at least 50% of the individual ROIs, see *ROI definition*). Gyri are indicated in light grey color and sulci are indicated in dark grey color. White contours delineate the boundaries of the Heschl's gyrus (HG) and the Heschl's sulcus (HS). LH = left hemisphere; RH = right hemisphere; D = dorsal; V = ventral; A = anterior; P = posterior. Group average maps in the following figures all use the same plotting schemes.

Figure 3A shows the responses from five voxels, each considered representative of one of the five candidate encoding models. The first and the second columns of Fig. 3A show the actual response and the predicted response by their model fit, respectively. The response pattern of the first voxel illustrates one that is well captured by the *interaction* model: it exhibits tuning to both F0 and Fc (i.e., its responses

vary in both dimensions) and its preferred frequency (i.e., the peak of the fitted response) in one dimension varies as the other dimension varies; in this case the preferred Fc of the voxel decreases as the F0 increases. The pattern of the second voxel illustrates one that is captured by the *independent* model: it also exhibits tuning to both F0 and Fc, but its preferred frequency in one dimension does not depend on the other dimension. The third and fourth voxels illustrate the responses of voxels that appear to be tuned only to Fc or F0 (*Fc-tuned* and *F0-tuned*, respectively), and the fifth voxel provides an example of one that responds to sound, but appears to be tuned to neither pitch nor brightness (*untuned*). Overall, the two models tuned to both dimensions performed best (Fig. 3B), with the *interaction* model providing a small but significant advantage over the *independent* model, as evaluated by N-fold leave-one-run-out cross-validation (see *Fitting and evaluating encoding models*) (Friedman test across five models: $\chi^2(4) = 38$, $p < .001$; post-hoc pair-wise two-tailed Wilcoxon sign-rank test between the *interaction* model and the *independent* model: $N = 10$, $W = 54$, $p = .004$). There were also significant differences between all other pairs of conditions ($p < 0.01$ in all cases).

To examine the spatial distribution of voxels with different tuning properties, we plotted the best model (the one that produced the best cross-validated prediction among the five models) of each voxel within the region of interest (ROI, see *ROI definition*) on the cortical surface (Fig. 3C). Voxels that favored the *interaction* and the *independent* models spanned the core areas (i.e., HG and HS), whereas voxels that favored the *Fc-tuned* or the *F0-tuned* models tended to flank the core areas and were relatively sparse. Comparing between the two models tuned to just one dimension, the *Fc-tuned* model had better performance (Fig. 3B), and was favored by more voxels, than the *F0-tuned* model (Fig. 3C; for maps of all individual participants see Fig. S3), suggesting more robust and more widespread tuning to Fc than F0 within auditory cortex, consistent with earlier findings of robust tonotopy across auditory cortex (*15*). Taken together, these results demonstrate that the model incorporating the interaction between tuning to pitch and timbre best captures the tuning properties of the auditory cortex. Given the consistent advantage of the *interaction* model and the fact that the *interaction* model subsumes all other models (see *Fitting and evaluating encoding models*), we used the fits from that model to explore local and global tuning patterns in the following sections.

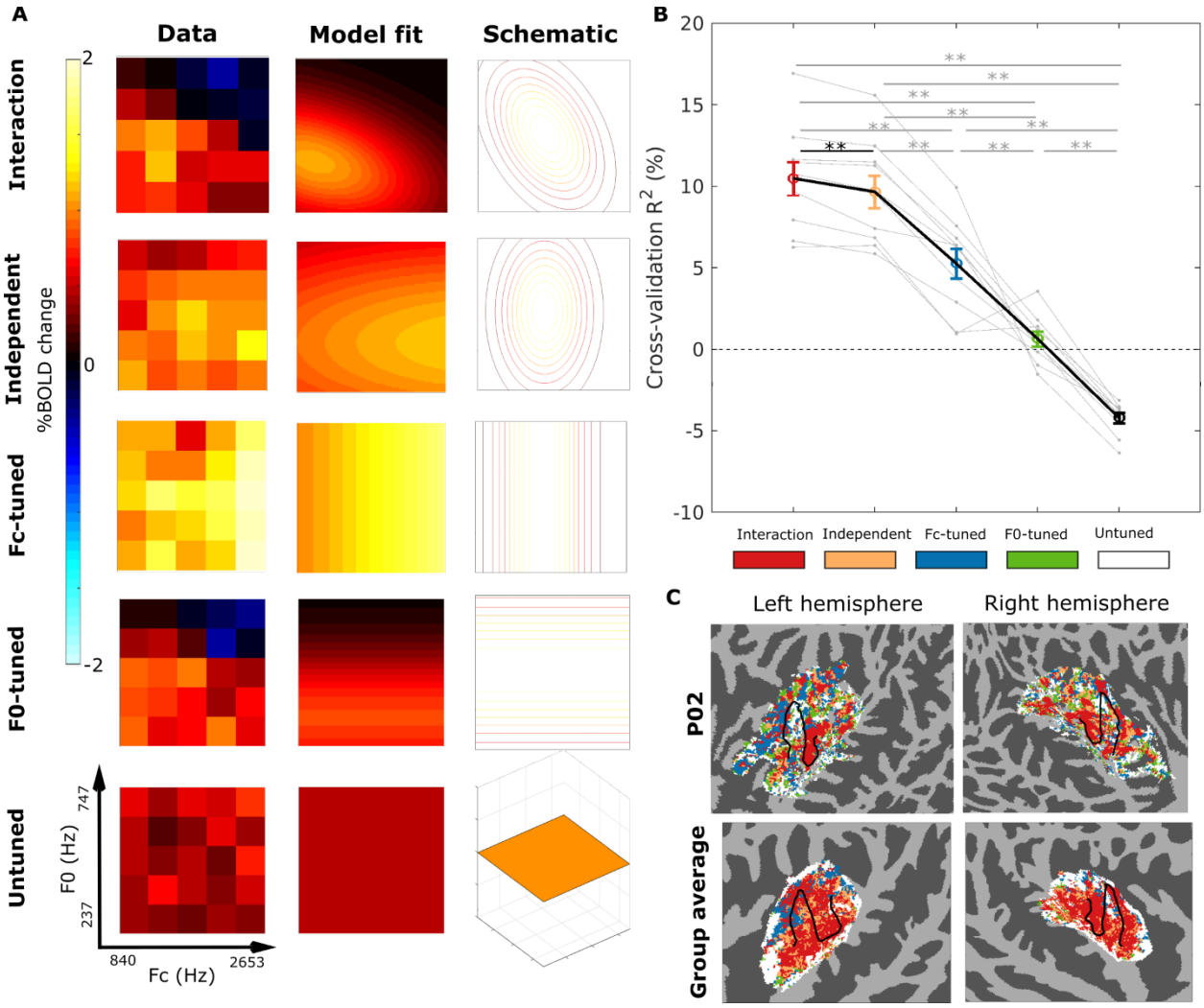

Fig. 3 | Interaction model best captures voxel-wise tuning to F0 and Fc. (A) BOLD responses to all 25 conditions (left column) and their model fits (middle column) of five representative voxels, one for each of the five candidate encoding models. The right column provides schematic diagrams (not reflecting real data) of the five models. The color of the matrices reflects the averaged beta value across trials. **(B)** Cross-validated R^2 for each of the five encoding models averaged across voxels that have a positive cross-validated R^2 for any of the encoding models except for the *untuned* model. Grey dots are individual participants' averaged R^2 values connected with grey lines. Circles are group means connected with black solid lines. Error bars represent ±1 SEM across participants. **(C)** Spatial distribution of voxels with different tuning properties ("best model map"). Color of the voxels within the ROI indicates the best-performing model for this voxel. Black contours delineate the boundaries of HG and HS. For the individual participant P02 (top row), only sound-responsive vertices within the ROI were shown. For the group average map (bottom row), the best model was defined as the model with the highest cross-validated R^2, averaged across participants. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Within-voxel substrates of perceptual confusion

When comparing tuning to Fc across low and high stimulus F0s, we observed a "warmer" map of preferred Fc values as the stimulus F0 increased from low to medium to high (Fig. 4A), indicating that the preferred Fc of a given voxel tends to decrease with increasing stimulus F0. This trend is illustrated in the *difference* map of preferred Fc between the highest F0 and the lowest F0 (High-Low; bottom row in Fig. 4A), which exhibits cool colors overall, suggesting a downward shift of preferred Fc as the F0 increases. Individual voxel tuning to F0 follows a similar pattern (Fig. 4B), with preferred F0 decreasing

with increasing stimulus Fc. Quantifying this effect, we found that as the stimulus F0 increased from lowest to highest, significantly more voxels decreased their preferred Fc than increased it (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 49$, $p = 0.027$); similarly, as the stimulus Fc increased from lowest to highest, significantly more voxels decreased than increased their preferred F0 (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 54$, $p = 0.004$). This effect is highly robust and can be seen in individual participants (Fig. 4D and Table S1). Additionally, the difference between the number of voxels shifting downwards and upwards in preferred F0 was significantly greater than that in preferred Fc (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 55$, $p = 0.002$). In other words, brightness changes had a greater effect on the preferred pitch than pitch changes had on the preferred brightness.

The observed pattern of tuning dependence between dimensions can explain the pattern of perceptual confusion observed here and in previous perceptual studies. For this explanation, we assume that the pitch and brightness of a stimulus are indicated by the voxels with the strongest activation. We observed that an increase in stimulus Fc generally causes the voxels' preferred F0s to decrease. Fig. 4C illustrates an array of voxels, tuned to a range of F0s. If a given F0 (green vertical line) leads to maximal stimulation of one voxel (4 in Fig. 4C) at a low Fc, then the same F0 will lead to maximal stimulation of a different voxel (5 in Fig. 4C) at a higher Fc, resulting in the perception of a higher pitch. Similarly, if a given Fc leads to maximal stimulation of one voxel at a low F0, the same Fc will lead to maximal stimulation of a different voxel that was previously tuned to a higher Fc, resulting in the perception of a brighter timbre.

In summary, the within-voxel tuning has shown two lines of evidence for the neural correlates of the perceptual confusion between pitch and brightness: (i) the tuning dependence between pitch and brightness observed at both the individual and group levels and (ii) the stronger dependence of preferred pitch on brightness than the dependence of preferred brightness on pitch.

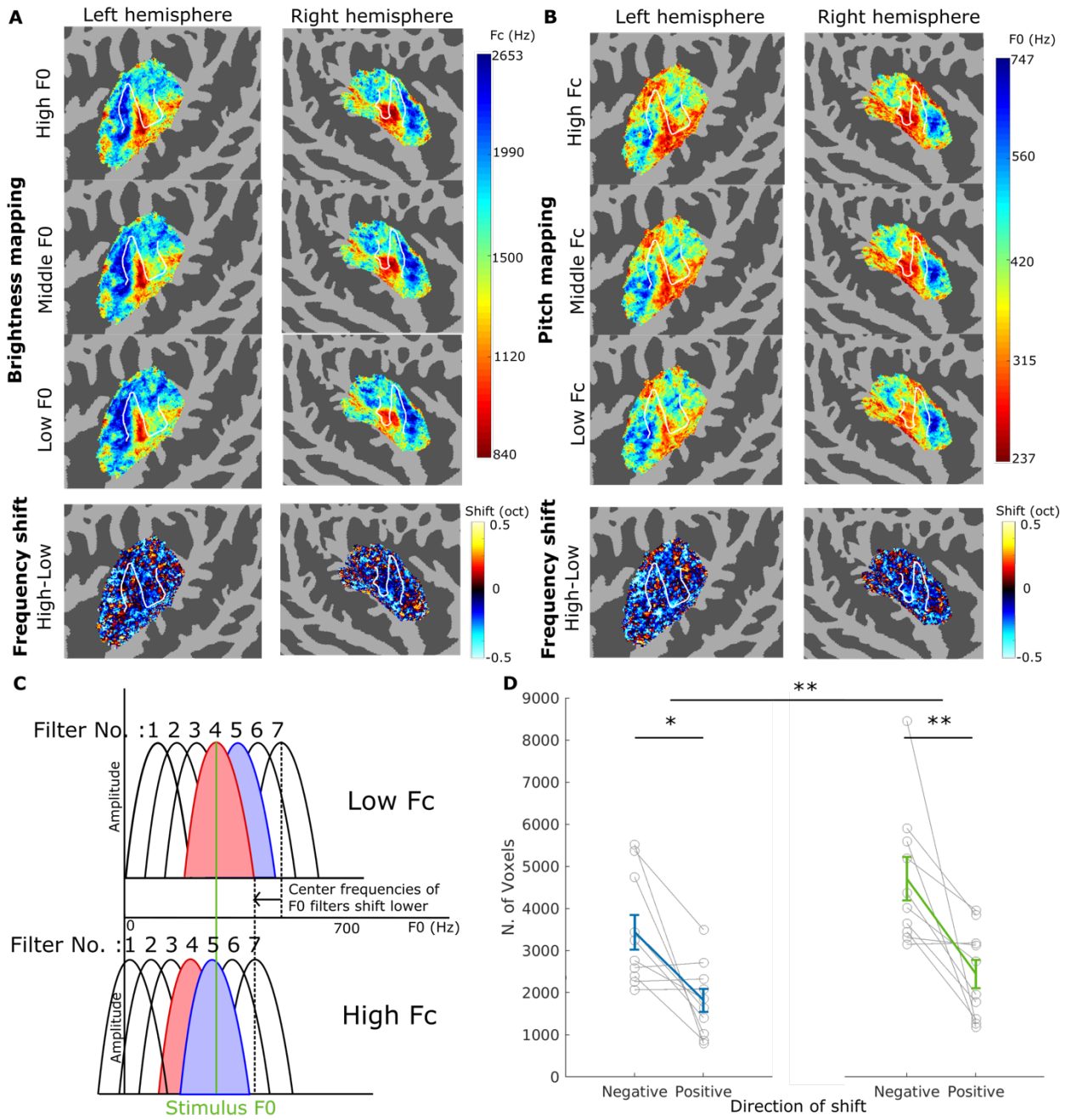

Fig. 4 | Within-voxel interaction between F0 and Fc as a potential substrate of perceptual confusion. (A) Top three rows: group average maps of preferred Fc values at the lowest (237 Hz), middle (420 Hz), and the highest (747 Hz) F0s. Bottom row: difference in preferred Fc values between the highest and the lowest F0s. **(B)** Top three rows: group average maps of preferred F0s at the lowest (840 Hz), middle (1500 Hz), and the highest (2653 Hz) Fc values. Bottom row: difference in preferred F0s between the highest and the lowest Fc values. Group averaging and subtraction of preferred frequencies were computed on a log scale. **(C)** Schematic diagram explaining how within-voxel interaction between F0 and Fc may underlie perceptual confusion. Top row: relationship between the stimulus F0 and F0 filters at a low Fc. Bottom row: relationship between the stimulus F0 and F0 filters at a high Fc. Black curves are theoretical F0 filters. The filter number indicates perceived pitch. The example stimulus indicated by the green line excites filter 4 the most at low Fc but excites filter 5 the most at high Fc, implying a higher perceived pitch. (D) Left: number of voxels with a positive shift or a negative shift in preferred Fc from a low stimulus F0 to a high stimulus F0. Right: number of voxels with a positive shift or a negative shift in preferred F0 from a low stimulus Fc to a high stimulus Fc.

Global cortical substrates of perceptual confusion

One of the most established tuning properties throughout the auditory pathway, including the auditory cortex, is the orderly frequency-to-place mapping, or tonotopy (*30*, *31*). Such orderly mapping has also been found for Fc and, to some extent, F0 in the auditory cortex (*15*). To examine if such orderly mapping of Fc and F0 exists in our data, we show the preferred Fc value at the middle F0 (420 Hz) for each voxel (top row, Fig. 5) and the preferred F0 value at the middle Fc (1500 Hz) for each voxel (bottom row, Fig. 5), both based on the fitted responses from the *interaction* model. Surprisingly, the two maps show many similarities, with both exhibiting high-low-high gradients that are spatially overlapped. Specifically, low-frequency-preferred regions for both mappings were in the anterolateral HG, with both also showing two low-high gradients stemming from the same low-frequency-preferred region and extending out in the anteromedial and posterior directions. Such a pattern is evident in individual participants (Fig. S4) and is also reflected in the significant positive correlation between the preferred Fc and preferred F0 across the auditory cortex (Spearman's correlation, $\rho > 0.22$, $p < 0.001$, for every participant). The overlapping gradients imply that increases (or decreases) in Fc lead to changes in activation patterns that are similar to those resulting from increases (or decreases) in F0. For example, increases in Fc alone and in F0 alone would both lead to a decrease of response in the anterolateral HG (low-frequency-preferred region) and an increase of response outwards (high-frequency-preferred regions) (magenta arrows in Fig. 5). Such similarity provides a global correlate of confusion between pitch and brightness. Note that this global correlate is independent from the previous result showing correlates of perceptual confusion at the single-voxel level: the previous correlate of confusion is derived from the tuning dependency between the two features *within* individual voxels, whereas the current correlate is derived from the spatial relationship *across* voxels with different preferred F0 and Fc values. Nevertheless, they are both consistent with the behavioral evidence, showing that a change in one feature (pitch or brightness) can be confused for a change in the same direction of the other.

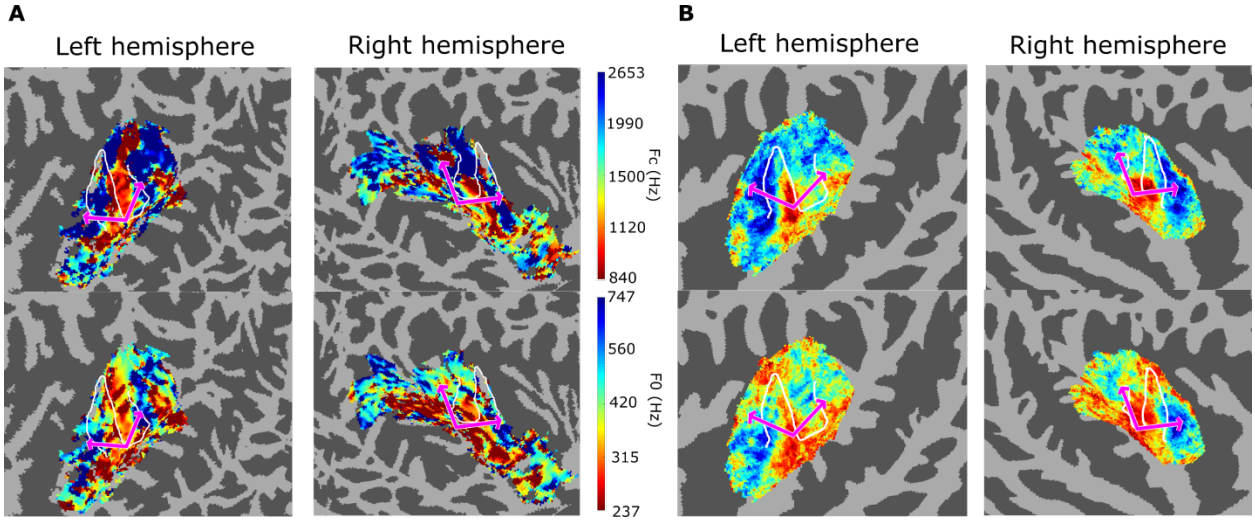

Fig. 5 | Pitch and brightness mappings have overlapping gradients. (A) Pitch and brightness mappings for an individual participant (P07). **(B)** Pitch and brightness mappings averaged across 10 participants. Top row: preferred Fc at the middle F0 (420 Hz). Bottom row: preferred F0 at the middle Fc (1500 Hz). Magenta arrows indicate the trajectories going from a low-frequency-preferred region to a high-frequency-preferred region for both F0 and Fc at the same locations on the cortex.

Combining within- and across-voxel patterns to predict perceptual confusion

If behavioral confusion is reflected in activation patterns of auditory cortex, we would expect congruent changes in pitch and brightness to produce more distinct (less similar) brain activation patterns than incongruent changes, mirroring the better behavioral discrimination of congruent than incongruent changes observed in our behavioral experiment (see *Perceptual confusion between pitch and timbre*). To test this hypothesis, we computed the cosine similarity between all pairs of conditions in their activation pattern across the auditory cortex and generated a similarity matrix according to their difference in F0 and Fc. To be consistent with perceptual confusion, we should observe pairs of conditions with congruent changes in pitch and brightness (both up or both down) having a lower similarity than the pairs with incongruent changes (one up, the other down). Therefore, the first and the third quadrants of the matrix (containing similarity between pairs of conditions with congruent changes) should have lower values than the second and the fourth quadrants (containing pairs with incongruent changes). Indeed, both the group average (Fig. 6) and all ten individual matrices (Fig. S5) exhibited this pattern, thereby providing compelling evidence that congruent changes in F0 and Fc induce less similar, and hence more discriminable, activation patterns across the auditory cortex than incongruent changes. Thus, the similarity matrices at the group and individual level provide a potential cortical substrate of perceptual confusion between pitch and brightness. Note that both the within-voxel changes in tuning and the more global mapping properties across voxels, described in the previous two sections, can contribute to the evidence for confusion provided by these similarity matrices (see simulation results in Fig. S6), despite these results being independent of each other.

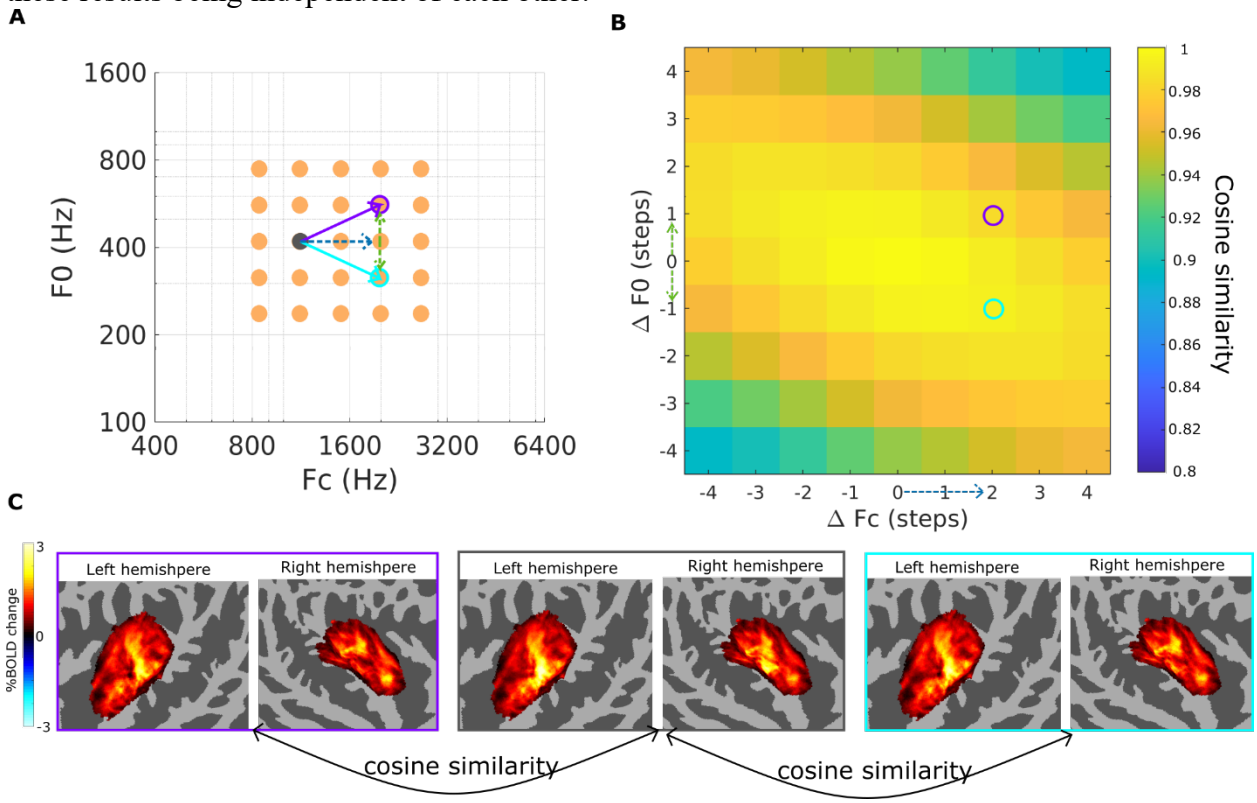

Fig. 6 | Cortical activation patterns are less similar for congruent changes than for incongruent changes between pitch and timbral brightness. (A) Stimulus grid showing two example pairs of conditions. The congruent condition (purple arrow and circle) relative to the reference condition (grey dot) is two steps higher in Fc (blue dashed arrow) and one step higher in F0 (green dashed arrow). The incongruent condition (cyan arrow and circle) is two steps higher in Fc (blue dashed arrow) and one step lower in F0 (green dashed arrow). **(B)** Similarity matrix showing cosine similarity between pairs of conditions. Similarity of the congruent pair is placed at position [2, 1] (purple circle) and similarity of the incongruent pair is placed at

position [2, -1] (cyan circle). Cosine similarities were averaged across all pairs of conditions populating the same location of the matrix. **(C)** Cortical activation patterns of the three conditions. Purple = congruent; grey: reference; cyan: incongruent.

Hemispheric asymmetry in the tuning to pitch and brightness

When we compared the overall tuning to *both* F0 and Fc, reflected in the performance of the *interaction* model, the number of voxels favoring the *interaction* model did not differ significantly between hemispheres (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 19$, $p = 0.432$); however, the overall performance of the *interaction* model was somewhat better for the right hemisphere than for the left (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 5$, $p = 0.020$).

We also compared hemispheres in their *exclusive* tuning to F0 or Fc separately by examining the performance of the *F0-tuned* and the *Fc-tuned* models. The numbers of voxels favoring the *F0-tuned* model did not differ significantly between hemispheres (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 18.5$, $p = 0.375$); the same was the case for the *Fc-tuned* model (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 28$, $p = 1$). However, in terms of model performance, the right hemisphere showed better performance than the left for the *Fc-tuned* model (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 5$, $p = 0.020$), whereas the performance of the *F0-tuned* model did not differ significantly between hemispheres (two-tailed Wilcoxon signed-rank test: $N = 10$, $W = 18$, $p = 0.375$). In sum, we did not observe a hemispheric difference in terms of the *size* of the cortical regions tuned to these two features, but we did observe a small advantage of the right over the left hemisphere in terms of *accuracy* of joint tuning to both pitch and brightness as well as exclusive tuning to brightness.

Spatial distribution of exclusive tuning and selectivity to pitch and brightness

Previous studies have explored cortical responses to each dimension separately, using both simple and complex, natural and artificial stimuli (*15*, *18*, *32*, *33*). In contrast, this study investigated cortical responses when two dimensions varied simultaneously. To determine the extent to which our results match those of previous studies, we compared the spatial distribution of voxels strongly tuned to either Fc or F0. For exclusive tuning to Fc, we selected voxels that favored the *Fc-tuned* model in each participant and showed their preferred Fc values for the *interaction* model at the middle F0 (420 Hz) (top row of Fig. 7A). The spatial distribution of preferred Fc forms orderly gradients that follow a high-low-high pattern *across* HG. Conversely, when we followed the same procedure with voxels that favored the *F0-tuned* model, plotting their preferred F0 values for the *interaction* model at the middle Fc (1500 Hz), we observed orderly gradients of preferred F0s forming a high-low-high gradient *surrounding* HG (middle row of Fig. 7A), with regions preferring low F0s located both as a band along STG, as well as a smaller region at the posteromedial end of HG (see Fig. 7B for a schematic depiction). This pattern bears a strong resemblance to what was found in an earlier study of just Fc or F0 tuning (*15*) (bottom row of Fig. 7A). In addition, pitch mapping can be seen as having a low-high-low gradient *along* the long axis of HG (Fig. 7B). The partial orthogonality of pitch and timbre gradients is in line with previous fMRI and MEG studies (*32*, *34*) that reported indications of orthogonal representations of periodicity (acoustic correlate of pitch) and spectral content (acoustic correlate of timbral brightness) in the human auditory cortex. Note that the distinct patterns in the *exclusive* tuning to pitch or timbre observed here is not in conflict with the similarity between the two mappings observed in Fig. 5, as the same gradients were observed in both results (see magenta arrows in Fig. 5 and Fig. 7B). The two gradients formed a V-shaped high-low-high tonotopic pattern that resembled the classic tonotopy

obtained with pure tones (*35*), especially in core regions that exhibited joint tuning to both pitch and timbre (see Fig. 3C).

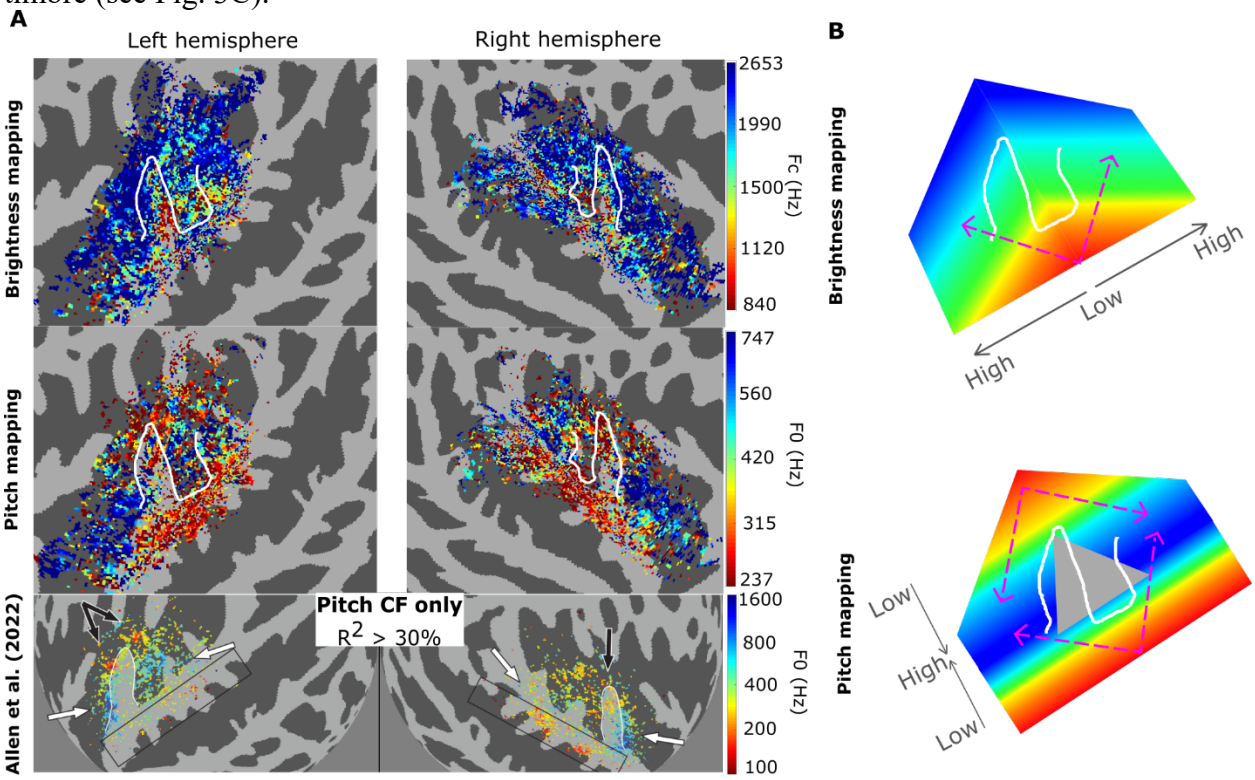

Fig. 7 | Distinct tonotopic patterns in the exclusive tuning to timbre and pitch. (A) Top row: Preferred Fc at the middle F0 (420 Hz) for voxels that were exclusively tuned to Fc but not to F0, pooled across all participants. Middle row: Preferred F0 at the middle Fc (1500 Hz) for voxels that were exclusively tuned to F0 but not to Fc, pooled across all participants. Bottom row: Preferred F0 for voxels that were tuned to F0 at an Fc of 2400 Hz pooled across all participants, data and figure from ref. (*15*). Note that top and middle rows are flattened surface maps, whereas the bottom row shows the cortical surface in a spherical space; also, the color ranges were set differently according to the frequency range used in these studies in order to better visualize the gradient transitions. These differences in visualization should be taken into account when comparing within and across studies. **(B)** Schematic depiction of the tonotopic gradients across the auditory cortex observed from this study and ref. (*15*). Grey and magenta arrows point from low-frequency-preferred regions to high-frequency-preferred regions. The grey triangle indicates the region that is not exclusively tuned to pitch. White contours delineate the boundaries of HG and HS. Only the left hemisphere is shown.

To compare the extent to which the response of a voxel depends on pitch and timbre, we quantified the selectivity of a feature as the coefficient of variation in this dimension. For example, for Fc selectivity, we averaged Fc tuning curves (fitted response at a given F0) across the five values of F0 and computed the standard deviation of the averaged curve, divided by the mean of the curve. For F0 selectivity, we averaged F0 curves across values of Fc and performed the same computation. Generally, Fc (brightness) selectivity was stronger than F0 (pitch) selectivity, especially in the core auditory regions, whereas tuning to F0 was somewhat stronger in the surrounding regions (see group average in Fig. 8 and individual maps in Fig. S7). Although the metrics used to compute voxel selectivity were different for this study and ref. (*15*), the pattern of results are similar across the two studies.

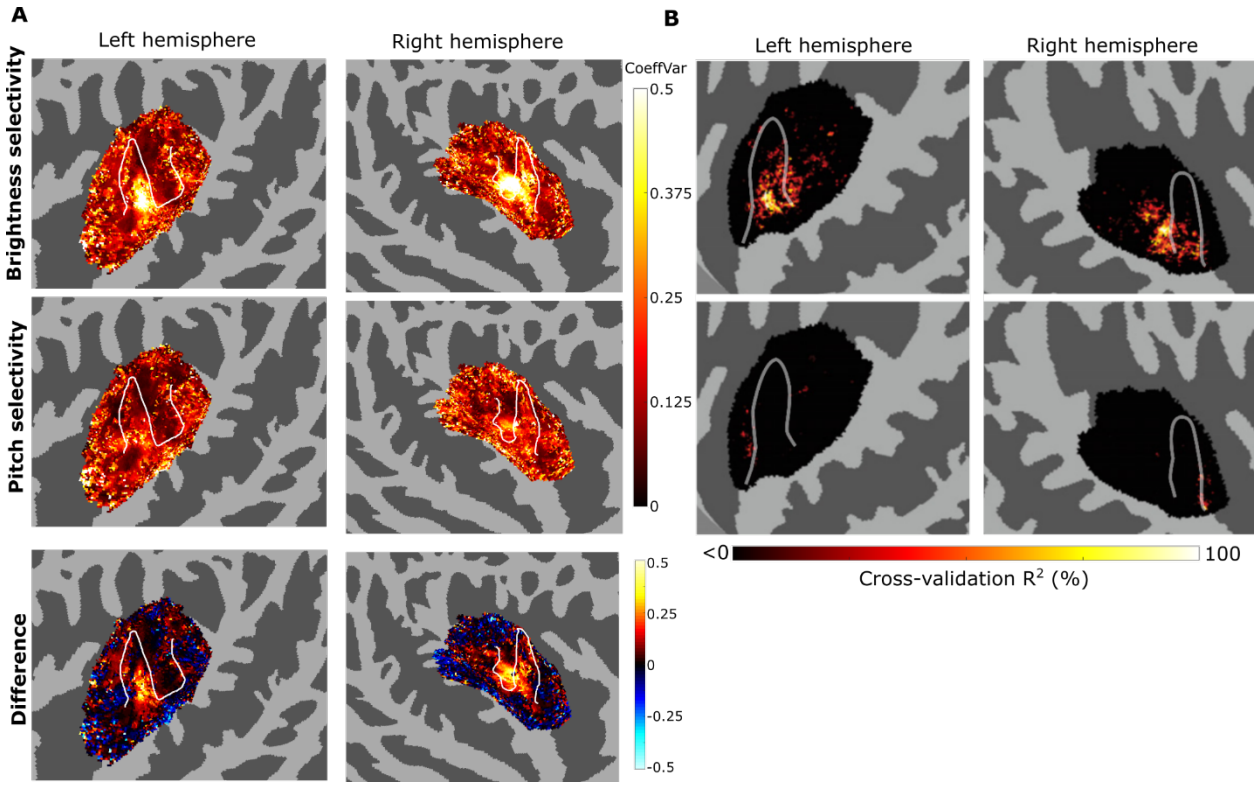

Fig. 8 | Cortical selectivity to brightness is stronger than to pitch in the core regions but somewhat weaker in the surrounding region. (A) Selectivity to pitch and brightness observed in this study. Top row: selectivity to brightness, quantified by the standard deviation of the Fc tuning curve averaged across F0 divided by the mean of the curve. Middle row: selectivity to pitch, quantified by the standard deviation of the F0 tuning curve averaged across Fc divided by the mean of the curve. Bottom row: brightness selectivity subtracted by pitch selectivity. CoeffVar: coefficient of variation. **(B)** Selectivity to brightness (top row) and to pitch (bottom row) observed in ref. (*15*), quantified by the cross-validation R^2 of the fitted Gaussian-shaped Fc tuning curves at an F0 of 200 Hz and that of F0 tuning curves at an Fc of 2400 Hz. Figure adapted from ref. (*15*). Black and white contours delineate the boundaries of HG and HS.

Discussion

In this study, we quantified behavioral confusion between pitch and timbral brightness and investigated the cortical substrates of this confusion in the same participants using fMRI. Behaviorally, participants were not able to discriminate the change in one feature (pitch or timbral brightness) while ignoring the other feature; instead, changes in the non-target feature tended to be mistakenly weighted in the discrimination of the target feature, despite trial-by-trial feedback, so that discrimination was better when pitch and timbre changed congruently and was worse when they changed incongruently. This result indicates confusion between the two features, consistent with previous findings (*7–9*, *11*). Our results extend the previous findings by quantifying the perceptual weight associated with the non-target dimension: a weight of zero would imply no interference or confusion, whereas a weight of 1 would imply equal perceptual weight assigned to the target and non-target dimensions when making discrimination judgments. The mean perceptual weights of 0.51 and 0.30 for the non-target dimensions when discriminating pitch and brightness targets, respectively, imply substantial interference or confusion, but some ability to attend more to the target than the non-target dimension.

Most importantly, we identified potential cortical substrates of this perceptual confusion in two independent aspects of the fMRI responses: one in the tuning to both features within individual voxels or local neural populations, and the other in the topography of preferred pitch and brightness across the auditory cortex. Within individual voxels, we observed a downward shift of their preferred pitch as the stimulus brightness increased, creating a higher predicted pitch percept at brighter timbres, and *vice versa*. Across the auditory cortex, we observed overlapping tonotopic gradients between pitch and brightness, resulting in a shared trajectory between going from a low- to a high-F0-preferred region and going from a low- to a high-Fc-preferred region. Finally, we found cortical activation that directly corresponded to the behavioral confusion: the activation patterns across the auditory cortex were more distinct when pitch and timbre changed congruently than when they changed incongruently. We believe our study is the first to find direct evidence for confusion in the direction of change between these two features in the brain, decades after such confusion was first reported in human behavior (*9*). Additionally, this study is the first to quantify the confusion between pitch and timbral brightness in terms of the perceptual weight of the non-target dimension (either pitch or brightness) on the discrimination of the target dimension (either brightness or pitch). This quantification provides a firmer basis for a phenomenon that has so far been treated mostly qualitatively, and therefore provides novel methodological tools for further investigation of its neural correlates.

Our examination of tuning to pitch (F0) and timbral brightness (Fc) for all the voxels in the auditory cortex covered regions that correspond roughly to anatomical planum polare (PP), HG, HS, planum temporale (PT), and lateral STG (Fig. 2D). Orderly mapping of Fc and F0 were both found across these regions, showing similarity between them in the core region (HG and HS) and differences in the surrounding regions (PT, PP, and lateral STG). In the core region, where joint tuning to pitch and timbral brightness was mostly likely to be observed (Fig. 3C), a low-pitch region and a low-brightness region overlap in anterolateral HG, while two high-pitch and two high-brightness regions overlap in anteromedial HG and posterior HG (Fig. 5). This V-shaped high-low-high gradient matches the location and spatial layout of pure-tone tonotopy (*31*, *35*) and tuning to spectral peaks of complex tones (*15*). In the surrounding regions, where exclusive tuning to pitch or timbre was more likely to be observed (top and middle rows of Fig. 7A), brightness mapping (top row of Fig. 7A) still resembled tonotopy, showing a high-low-high gradient across HG centered in anterolateral HG, as expected based on previous results (*15*). Pitch mapping, on the other hand, had two low-pitch-preferred regions, located in posteromedial HG and lateral STG, forming two high-low-high gradients that surrounded HG (middle row of Fig. 7A).

For the region at the posteromedial HG, some pure-tone tonotopy studies have identified it as a low-frequency-preferred region (*36*, *37*). The region in lateral STG, on the other hand, is near the pitch-sensitive regions located by numerous studies in lateral HG and middle STG where sensitivity to both pitch salience (*19*, *22*, *23*, *38*) and pitch height (*15*, *18*) has been observed. Our results delineate a clearer picture of pitch (F0) and brightness (Fc) mapping in the human auditory cortex, with joint tuning in the core region and distinct gradients in the surrounding regions.

It is worth noting that the brightness mappings had sharper transitions between low- and high-frequency-preferred regions than the pitch mappings (Fig. 5) on the anterior border of HG. A potential explanation may come from the relationship between the frequency range of the stimuli used in this study and the effective frequency ranges for pitch and brightness perception for human listeners. The range of F0s for which listeners can hear a clear melodic pitch extends from roughly 30 to 4,000 Hz, corresponding to about 7 octaves (*39*, *40*), whereas the range of audible frequencies, and hence the range of potential Fc values, for young listeners with normal hearing has a much wider range of 20 to 20,000 Hz, or about 10 octaves. Yet, the range of the stimuli for Fc and F0 in this study was the same, meaning that the proportion of the effective frequency range that our stimuli covered was smaller for Fc than for F0. Assuming a uniform distribution of filters along the frequency axis, we would expect the frequency range of our stimuli to encompass a smaller range of Fc filters than F0 filters, meaning that Fc tuning curves were more likely to appear as a slope (with the preferred Fc at the low or high edge of the sampled frequency range) than a peak (with the preferred Fc within the sampled frequency range). This difference in range may explain why the brightness map in this study appears to be more binarized than the pitch map.

There are debates on whether hemispheric differences exist in the cortical representation of pitch and timbre. The current study found evidence for a small right hemisphere advantage in the cortical *tuning* to pitch and brightness but no evidence for asymmetry in the *size* of cortical regions exhibiting tuning to these two features. For representations of pitch, while bilateral activation is observed in almost all studies with comparable sizes of activated regions between hemispheres (*15*, *22*), the right hemisphere is found to have a small advantage in the magnitude of activation or selectivity, usually in the cases of pitch changes or melodies (*38*, *41–43*). On the other hand, cortical representations of pitch salience, which is usually obtained with a fixed pitch, have generally demonstrated hemispheric symmetry (*19*, *23*, *38*). For representations of timbre, variations in spectral envelope also elicit bilateral activation (*13*, *44*), although a non-significant trend of right lateralization in the sensitivity to spectral variation has been reported (*44*). Thus, the findings of the current study, which found bilateral activation but a small right hemisphere advantage in the joint tuning to pitch and timbral brightness as well as the exclusive tuning to timbral brightness, are in general agreement with previous findings.

Our study also showed that the perceptual and neural confusion between pitch and brightness is robust to language background and musicianship, generalizable across tone and non-tone language speakers as well as across participants with a wide range of musical experience. Although differences in pitch discrimination have been reported between tone and non-tone language speakers (*24*, *25*) and between musicians and non-musicians (*26*, *27*), no study has shown an influence of either language or musical background on the interference between pitch and brightness. Indeed, the only study to compare musicians and non-musicians in a similar task found no evidence that musicians are more or less susceptible to perceptual interference than non-musicians (*8*). Our behavioral results are consistent with this previous finding. Although our sample of 10 fMRI participants is not sufficient to test for an effect of musical or language background on the neural confusion, there is no observable evidence of such

effects in our diverse group of participants (musical training ranged from 0-16 years and included 3 tone-language speakers).

Regarding the brain-behavior relationship underlying the confusion between pitch and timbral brightness, we found an asymmetry between pitch and brightness interference in both the behavioral and neural data. Specifically, brightness interferes with pitch more than pitch interferes with brightness, in terms of both the perceptual weighting of the non-target dimension (see Fig. 1B) and the shift in preferred frequencies within individual voxels (brightness changes have a greater effect on voxels' preferred pitch than pitch changes have on voxels' preferred brightness; see Fig. 4D). A limitation of the present study is that due to the limited number of fMRI participants, we cannot investigate whether perceptual weights across participants correlate with levels of neural confusion across participants. Future research involving a larger sample size could be directed at this question. However, we note that there are challenges to establishing such a correlation: to measure the neural confusion, large frequency variations are required to elicit measurable differences in the brain responses to the stimuli; in contrast, perceptual confusion is optimally measurable when frequency variations are near their perceptual threshold. Indeed, we presented the same stimulus set with large frequency steps to all participants in the MRI experiment, but applied near-threshold frequency variations, customized to the participants' F0DL and FcDL in the behavioral experiment. Developing paradigms that enable more comparable measurements of psychophysical performance and neural responses would be valuable.

The existence of confusion between pitch and timbre in both human behavior and the human brain raises the question of its ecological significance. The hypothesis of efficient coding may serve as a potential explanation. It posits that sensory processing is designed to reduce information redundancy and to improve the efficiency of stimulus representation (*45*). To efficiently encode acoustic features that naturally co-vary, perceptual sensitivity to these features should increase in the direction of the covariance and decrease in the direction orthogonal to the covariance. In other words, efficient coding principles predict confusion between the two acoustic features when they change in the same direction, if they typically co-vary under natural conditions. Consistent with this prediction, long-term exposure to correlated acoustic features can lead to better discrimination along the correlated or congruent direction than along the orthogonal or incongruent direction (*12*). In natural sounds, such as those produced by the human voice and by musical instruments, F0 and Fc are often positively correlated (*11*). Therefore, long-term exposure to positively correlated F0 and Fc in natural sounds from birth may lead to optimized representations of the two acoustic features in the brain, which in turn predicts better discrimination when pitch and brightness change congruently and poorer discrimination when they change incongruently. This theory is supported by a recent study investigating the perception of pitch and timbral brightness in 3- and 7-month-old infants. The study tested F0 discrimination when Fc randomly varied, and *vice versa*. Surprisingly, infants' performance was comparable with that of adult musicians and better than that of adult non-musicians (*10*). In other words, infants appeared to be less susceptible to the influence of the variation in non-target dimension than adult non-musicians, showing less confusion between pitch and timbral brightness. This outcome may be due to infants' relatively short exposure to the correlation between F0 and Fc in natural sounds compared with adults. Confusion between pitch and timbre in adult perception and the cortex can thus be seen as an efficient coding strategy, reducing redundancy in the naturally correlated F0 and Fc. Our results provide a potential neural implementation of this efficient coding: cortical activation patterns are more distinguishable for congruent changes in F0 and in Fc than incongruent changes, as indicated by the patterns in the similarity matrix (Fig. 6).

Materials and Methods

MRI Experiment

Participants

Ten participants [five females, four males, one non-binary, mean (SD) age of 30.9 (9.3) years, ranging from 23 to 53 years, all right-handed] were recruited from the University of Minnesota community. Three of the participants were authors (Y. O., E. J. A., and A. J. O.). All participants had normal hearing (pure-tone thresholds no higher than 20 dB hearing level (HL), at octave frequencies between 250 Hz and 8 kHz) except for one who had a slightly elevated pure-tone threshold of 25 dB HL at 4 kHz in their right ear. One participant reported having intermittent tinnitus. Five of the ten participants were identified as non-musicians, having two or fewer years of formal musical training. The other five participants had an average of 6.6 years of musical training on non-percussion instruments. Three of the ten participants were tone-language speakers (Mandarin Chinese or Cantonese). Years of musical training and language background of individual participants are reported in Table S2. The experimental protocol was approved by the University of Minnesota Institutional Review Board (Approval No.: 1207M16625). All participants provided written informed consent before the experiment.

Stimuli

Stimuli were generated digitally in MATLAB (The MathWorks) and presented with its Psychophysics Toolbox extensions (*46*). The stimulus set consisted of 25 harmonic complex tones, each of which was bandpass filtered with 12 dB/oct slopes, then filtered with a 16th-order lowpass filter and a cutoff frequency of 10 kHz (see Fig. 1A). Tones varied both in F0 (between 237 and 747 Hz), affecting pitch, and in the filter's center frequency (Fc; between 840 and 2653 Hz), affecting brightness, in five equal steps (on a logarithmic scale) for both dimensions (i.e., F0 = 237, 315, 420, 560, or 747 Hz; Fc = 840, 1120, 1500, 1990, or 2653 Hz). Each F0 was paired with each Fc in a 5×5 design, for a total of 25 unique stimuli (see Fig. 1B). The ranges of F0 and Fc were selected to ensure that the F0 was always below the Fc (allowing the Fc to be defined by the amplitudes of the harmonic components) and that there were always audible low-numbered, spectrally resolved, harmonics present in all the stimuli to allow for accurate pitch discrimination (confirmed by measuring listeners' sensitivity to F0 and Fc changes at the four corners of the stimulus matrix, as well as at the center; see Fig. S8). The number of steps in F0 and Fc was based on data quality considerations (each condition presented 24 times per session) and the total duration of the experiment (two hours for each of the two sessions). Given these constraints, we determined that five levels per dimension, totaling 25 conditions, provided a good balance between the sampling density of the feature space and the total duration of the experiment.

Procedure

Stimuli were presented through MR-compatible Sensimetrics S14 insert earphones with standard foam tips. Custom filters were applied to flatten the frequency response of the earphones. As in ref. (*15*), the stimuli were presented with a "Morse code"-like rhythm to enhance their perceptual salience over the sound of the fMRI pulse sequence. Within each trial, 16 short (50-ms) tones and 16 long (200-ms) tones were presented, with each tone having 20-ms raised-cosine onset and offset ramps. Every 700 ms consisted of two short and two long tones, presented in random order, separated by 50-ms gaps. This process was repeated 8 times, with random shuffling of the long and short tones for each repetition, adding up to a total trial length of 5.6 seconds. All tones presented within a single trial had an identical

539 F0 and Fc. Following one trial, after a silent gap with a random duration of uniform distribution between
540 200 and 600 ms, the next trial was presented.

541 Within one run of ~6 minutes, each of the 25 conditions was presented twice, and the 50 trials were
542 presented in a pseudorandomized order, constrained to prevent two identical stimuli from being
543 presented consecutively. Silent periods of 10 seconds were inserted at the beginning of the run and after
544 every 10 trials. A 15-s silence was further appended at the end of the run. These silent periods were
545 added to establish a baseline hemodynamic response (*43*, *44*).

546 Each participant underwent two experimental sessions of 12 runs each, for a total of 24 runs. Therefore,
547 each condition was presented 24 times per session and 48 times per participant. Multiple repetitions are
548 important for averaging out unwanted measurement variability, and for cross-validation methods of data
549 denoising (*48*, *49*). Participants were asked to press a button whenever a new trial began to help
550 encourage and monitor the alertness of the participants throughout the experiment. After each run,
551 participants were informed of their behavioral performance (percent correct) in that run. Behavioral
552 performance for all participants is reported in Table S3.

553 Before the two experimental sessions, participants performed a loudness-adjustment task in a separate
554 scan session, aiming to minimize any loudness differences between the stimuli when presented in the
555 presence of scanner noise. In this task, participants listened to sequences composed of each of the 25
556 experimental stimuli with the order randomly shuffled, while the scanner was running with the fMRI
557 pulse sequence used in the experimental sessions. Within each sequence, the same stimulus (100 ms)
558 with a silent gap (50 ms) was repeated 20 times with the sound level gradually increasing in 1 dB steps
559 between 43 and 83 dB SPL. Participants were asked to adjust the loudness of the tones to be clearly and
560 comfortably audible over the scanner noise by pressing "1" to decrease the sound level of the sequence
561 and "2" to increase it. Once they were satisfied with the sound level, they could press "3" and the level
562 would remain the same for the rest of the sequence. The task was performed three times for each
563 stimulus. For each participant, the median level of a given stimulus was calculated and later used in that
564 participant's experimental sessions. In addition to the preliminary loudness adjustments, participants
565 were also instructed to listen for any remaining loudness differences between tones at the beginning of
566 each experimental session. Between the runs, if the participant indicated a need for further adjustments
567 to the sound level, the experimenter would change the sound level slightly (1-3 dB) in the next run based
568 on the participant's feedback.

569

570 ***MRI data acquisition***

571 Both anatomical and functional data were acquired using a Siemens 3T Prisma scanner with a 32-
572 channel head coil at the Center for Magnetic Resonance Research (CMRR) at the University of
573 Minnesota. Functional data were acquired in two experimental sessions on separate days using a
574 continuous gradient-echo EPI sequence (TR = 1800 ms; TE = 30 ms; flip angle = 70°; 2-mm isotropic
575 voxels; number of slices = 60; multiband factor = 3; phase partial Fourier = 6/8). Slices were oriented to
576 be approximately 45° to the brainstem to cover most of the brain and to avoid coverage of the eyeballs
577 whose movement could create Nyquist artifacts within auditory regions. A total of 208 volumes were
578 obtained in each run. Four fieldmaps were collected throughout each experimental session for distortion
579 correction (same slab orientation as the gradient-echo EPI; TR = 400 ms; first TE = 2.16 ms; second TE
580 = 4.62 ms; 4-mm slices).

581 Anatomical data were acquired in a different session before the experimental sessions. Three T1-
582 weighted and two T2-weighted volumes were collected. MPRAGE T1-weighted parameters included:

TR = 2400 ms; inversion time (TI) = 1000 ms; TE = 2.22 ms; flip angle = 8°; 0.8-mm isotropic voxels. T2-weighted parameters included: TR = 3200 ms; TE = 563 ms; 0.8-mm isotropic voxels.

MRI data preprocessing

Anatomical and functional data were preprocessed through a series of custom scripts(*50*). For anatomical data, gradient unwarping was performed for T1- and T2-weighted volumes to correct for gradient nonlinearity using the gradient coefficient file obtained from the scanner. Then, three T1-weighted volumes were aligned (rigid-body transformation with six degrees of freedom and cubic interpolation) and averaged. The two T2-weighted volumes were aligned in the same way. The averaged T1 was then used for brain segmentation and cortical reconstruction via Freesurfer(*28*). The averaged T2 was co-registered to the averaged T1 using rigid-body transformation and cubic interpolation.

Functional data were temporally upsampled from 1.8s to 1s. Fieldmaps were spatially upsampled to a slice thickness of 2mm to match that of the EPI and later used for distortion correction. Distortion correction was performed along with motion correction in a single cubic interpolation. The corrected functional volumes were then co-registered to the averaged anatomical T2 using an affine transformation. Finally, the functional data were re-preprocessed such that the data were resampled onto the mid-grey cortical surface using a single cubic interpolation. The analyses in this paper are performed in surface space, instead of volume space, for better visualization of the cortical mappings of F0 and Fc.

The surface-based functional data were analyzed with GLMsingle(*29*), a toolbox for improving the signal-to-noise ratio of estimates of BOLD response amplitudes, especially for experimental designs with a short trial duration and fast-alternating conditions. GLMsingle provides single-trial BOLD response amplitude estimates ("betas"). Upon visual inspection of beta values from each run, the final run from session 1 of P07 was discarded due to quality issues.

ROI definition

One ROI was defined for each hemisphere on the participant's native brain surface space, referred to as the auditory cortex in the results. Aiming to encapsulate the region responsive to the stimuli in the auditory cortex, a polygon was manually drawn per hemisphere per scanning session with both anatomical (to locate HG and HS) and functional (to locate sound-responsive regions) maps as underlays. The anatomical localization of HG and HS was based on a binary curvature map of the participant showing gyri and sulci, and the functional localization of sound-responsive regions was based on a map of the variance explained (R^2) by all experimental stimuli against the baseline (no stimuli) (see Fig. 2D) and a map of mean beta values across all conditions. Importantly, the functional maps were computed based on all experimental trials in the session, reflecting the overall brain response towards all sound stimuli played in the experiment (against the baseline). Thus the ROIs were not biased towards any measure specific to pitch or brightness.

Participant-level ROIs were defined as the union of ROIs defined in the two scanning sessions, respectively. Activation across scanning sessions within individual participants were highly consistent, as demonstrated in Fig. S9. We also defined a group-level ROI for each hemisphere as the intersection of at least 50% of the individual polygons defined above (after mapping each subject to the fsaverage atlas space).

In the ROI-level analyses, aggregating across individual voxels (e.g., Fig. 3B and Fig. 6), we further restricted the ROI to sound-responsive voxels, defined as having a mean beta value significantly larger than 0 ($p < 0.01$).

Correction for the effect of sound intensity

Functional data were analyzed first for each individual session. For each voxel, the beta values for all 24 repetitions (for session 1 of P07 whose final run was discarded, 22 repetitions) of a given condition were averaged, forming a tuning surface of 25 data points in a three-dimensional space, with the height of the surface denoting the beta value of each condition, and the x- and y-axes denoting F0 and Fc of the stimuli, respectively. It was observed that the tuning surface averaged across ROIs showed a bias towards tones with low-frequency spectral content (i.e., tones that have low Fc values tend to have stronger BOLD responses). After examining the sound intensity at which each tone was presented (obtained from the loudness-adjustment task, see Procedure), we found that participants tended to assign higher sound levels to tones with lower Fc values, which may have led to over-compensation for the effects of frequency content on loudness. Therefore, we decided to correct for such a bias by scaling the beta values of a given tone by the sound intensity at which the tone was presented to obtain an equal-sound-intensity tuning surface. To obtain such a tuning surface, an intensity-beta linear fit was calculated for each participant (Fig. S10). First, voxels that were both within the ROI polygon and sound-responsive (see *ROI definition*) were selected as "good voxels". Second, the mean betas across all repetitions for each of the 25 conditions were averaged across these "good voxels" for each session. Third, the 50 intensity-beta pairs from the two functional sessions of one participant were pooled to obtain a linear fit between sound intensity and mean beta from "good voxels" for each participant. Finally, the intensity-corrected beta estimate of every single trial in each of the two sessions was calculated by dividing the original beta of this trial by the predicted beta according to the intensity-beta linear fit of this participant and the sound intensity for the condition of this trial.

Fitting and evaluating encoding models

The following five Gaussian-based encoding models were designed to characterize the tuning of each voxel to F0 and Fc after the two dimensions were logarithmically transformed: 1) The *interaction* model assumes that each voxel is tuned to both F0 and Fc, and the preferred frequency of one dimension is dependent on the value of the other dimension; 2) the *independent* model assumes that each voxel is tuned to both F0 and Fc and that the two dimensions are processed independently (i.e., the preferred F0 of the voxel does not depend on the stimulus Fc); 3) the *Fc-tuned* model assumes that the voxel is tuned only to Fc and is invariant to F0; 4) the *F0-tuned* model assumes that the voxel is tuned only to F0 and is invariant to Fc; and 5) the *untuned* model assumes that the response of the voxel is invariant to both F0 and Fc.

Table **1** provides the equations for each of the models. In models 1-4, f_{Fc} and f_{F0} represent the stimulus F0 and Fc values, CF_{F0} and CF_{Fc} represent the coordinates of the peak of the tuning surface, and σ represents the width (standard deviation) of the Gaussian curve in a given dimension. g represents the height of the tuning surface. Among the five models, the comparison between the *independent* and the *interaction* models is of most interest for this study, as this comparison tests the hypothesis that tuning in one dimension is affected by the other dimension. Mathematically, these two models are highly similar, in that the tuning surface of the *interaction* model is just that of the *independent* model rotated by θ degrees on the Fc-F0 plane (see schematic illustration in Fig. 3A).

Before model fitting, the intensity-corrected beta estimates from two sessions in one participant were combined, forming a participant-level dataset. This was done by averaging the intensity-corrected beta estimates from session 1 and session 2 on the trial level so that the participant-level dataset had the same structure (i.e., number of repetitions for each condition, number of runs, and number of trials) as one single session. To balance the amount of data fed into the participant-level dataset between the two sessions, the last run of session 2 for P07 was discarded, along with session 1 for P07 which had data quality issues. As a result, each voxel had 12 (11 for P07) tuning surfaces, each from a quasi-run in the participant-level dataset consisting of 25 beta estimates for 25 conditions. The tuning surfaces were then averaged across quasi-runs. This averaged tuning surface was then used to fit the five encoding models.

For each voxel, the performance of each of the five encoding models was evaluated using an N-fold leave-one-run-out cross-validation procedure (N is the number of runs in the current participant-level dataset). First, averaged data across N-1 runs were used to find the training parameters. The parameter *g* was imposed to be non-negative. Other parameters were not constrained. To avoid local minima, the parameters were initialized with various sets of initial seeds based on the voxel response. MATLAB's lsqcurvefit function was used to optimize for least-square solutions. The fitted parameters were then tested on the held-out run to obtain an R^2 value. This procedure was repeated N times, each holding out a different run as the testing set, yielding N R^2 values. The average of these N values was then calculated as the cross-validation R^2 for this model.

Behavioral Experiment

Participants

All ten participants from the MRI experiment took part in the behavioral experiment, along with 10 newly recruited participants [8 females, 2 males, mean (SD) age of 24.8 (4.2) ranging from 19 to 30 years] from the University of Minnesota community. All of the newly recruited participants had normal hearing according to the criteria in the MRI experiment. Three participants were identified as non-musicians, having two or less years of formal musical training. The other seven participants had an average of 9.4 years of musical training on non-percussion instruments. Six of the ten new participants were tone-language speakers (Mandarin Chinese and/or Cantonese). The experimental protocol was approved by the University of Minnesota Institutional Review Board (Approval No.: 0605S85872). All participants provided written informed consent before the experiment.

Stimuli

The stimuli were harmonic complex tones, digitally generated in the same way as in the MRI Experiment, except for the duration of the tones. Each tone in the behavioral experiment had a duration of 200 ms, including 20-ms raised-cosine onset and offset ramps. Each trial contained two tones played sequentially, separated by a silent gap of 300 ms. All the stimuli were converted to analog at a sampling rate of 48 kHz using a Lynx E22 sound card with 24-bit resolution and were presented diotically via open-ear headphones (Sennheiser HD 650) in a sound-attenuating booth at a level of 60 dB SPL.

Before the main experiment, participants' F0 difference limens (F0DLs) and Fc difference limens (FcDLs) were measured for the middle tone in the 5×5 stimulus set used in the MRI Experiment (F0 = 420 Hz and Fc = 1500 Hz). In the F0DL measurement, Fc was held constant at 1500 Hz while the F0 of the two tones was different and geometrically centered around 420 Hz. In the FcDL measurement, F0 was held constant at 420 Hz while the Fc of the two tones was different and geometrically centered

around 1500 Hz. The individual F0DLs and FcDLs from this preliminary experiment were used to set the DLs for each listener in the main experiment.

In the main experiment, as in the DL measurements, all variations in F0 and Fc were geometrically centered around F0 = 420 Hz and Fc = 1500 Hz. For the up-down judgment of pitch, the ΔF0 between the two tones was always set to one F0DL for each listener individually, whereas ΔFc was set to -2, -1, 0, +1, or +2 FcDLs, where positive values imply congruence between the F0 and Fc direction, negative values imply incongruence, and 0 implies no change in Fc. For the up-down judgment of timbral brightness, ΔFc between the two tones was always set to one FcDL for each listener individually, whereas the ΔF0 was set to -2, -1, 0, +1, or +2 F0DLs.

Procedure

Both the preliminary DL measurement and the main experiment adopted a two-interval two-alternative forced-choice paradigm delivered through the AFC software package (*51*). Two virtual buttons labeled "1" and "2" were displayed on the computer screen, indicating the first and the second tones, respectively. Each button flashed when the corresponding tone was played. Participants were asked to select which tone was higher in pitch or brightness (depending on the condition), and responded by clicking on one of the two buttons displayed on the screen with a mouse or by pressing "1" or "2" on the keyboard. After each response, immediate feedback of "Correct" or "Incorrect" was displayed on the screen before proceeding to the next trial. Feedback was provided in both the initial DL measurements and the main experiment.

The F0DL and FcDL measurements used a standard two-down one-up tracking procedure that tracks the 70.7% correct point on the psychometric function (*52*). Each track began with a ΔF (ΔF0 in the F0DL measurement and ΔFc in the FcDL measurement) of 20%, which decreased or increased by a factor of two, depending on whether the response of the previous trial was correct or not. After the first reversal from an "up" to a "down" direction of the tracking procedure, ΔF changed by a factor of 1.41. After the second such reversal, ΔF changed by a factor of 1.19. The tracking stopped after a further four reversals, and the DL of this track was calculated as the geometric mean of the ΔF values at the last four reversal points.

Before the measurements, participants were given examples of tones with low and high pitches, as well as tones with duller or brighter timbres. They were also encouraged to use the immediate feedback given in the experiment to establish their concepts of pitch and timbral brightness. Each participant completed four measurements of F0DL and four measurements of FcDL in a pseudorandomized order in about 10 minutes. Each participant's final F0DL and FcDL were calculated as the geometric mean of the four measurements.

The main experiment included 10 blocks of the up-down judgments of pitch and 10 blocks of the up-down judgments of timbral brightness in a pseudorandomized order. As in the DL measurements, participants were asked to indicate which of the two tones had a higher pitch (for the blocks of pitch judgment) or brighter timbre (for the blocks of timbre judgment). Each block had five ΔFs (-2, -1, 0, 1, and 2 DLs) in the non-target dimension (ΔFc in pitch judgments and ΔF0 in timbre judgments) each with 10 trials presented in a random order, and the ΔF in the target dimension (ΔF0 in pitch judgment and ΔFc in timbre judgment) was held constant at the participant's DL in this dimension. There were also 10 practice trials at the beginning of each block, where the ΔF in the target dimension was set to 2 DLs instead of one, to help participants find the perceptual cues needed to succeed in the task.

Therefore, each block contained a total of 60 trials (two practice trials and 10 test trials at each value of ΔF in the non-target dimension). Each participant took about 60 minutes to complete the main experiment.

Data and statistical analysis

As described in *Stimuli*, the geometric means of the DL measurements were used as the F0DL and FcDL in the main experiment. For the main experiment, values of d' for each of the five conditions ($\Delta F_{\text{non-target dimension}}$ = -2, -1, 0, 1, or 2 DLs) in the two tasks (pitch and timbre direction discrimination) were calculated for each participant. A linear function was then fitted to the five d' values from each the two tasks, with the x-axis being $\Delta F_{\text{non-target dimension}}$ and the y-axis being the d' in this condition. The slope of the function therefore reflected the effect of the variation in the non-target dimension on the discriminability of the target dimension. To evaluate the contribution of the non-target dimension to discrimination of the target dimension, the slope of the linear fit was divided by the theoretical d' that corresponds to one DL, i.e., d' of 1.09 for an accuracy of 70.7% (*53*). Two-tailed Wilcoxon signed-rank tests were performed to compare the weights of the non-target dimension with a zero-slope null hypothesis (i.e., no influence of the non-target dimension). Group averaged d' values and their linear fits were also calculated and are shown in Fig. 1A.

Table 1 | Equations of the five encoding models.

Model No.	Model name	Equation
1	Interaction	$R(f_{F0}, f_{Fc}) = ge^{\frac{X^2}{-2\sigma_X^2}+\frac{Y^2}{-2\sigma_Y^2}}$ where $\binom{X}{Y} = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix} * \begin{pmatrix} f_{Fc}-CF_{Fc} \\ f_{F0}-CF_{F0} \end{pmatrix}$ (Eq. 1)
2	Independent	$R(f_{F0}, f_{Fc}) = ge^{\frac{(f_{F0}-CF_{F0})^2}{-2\sigma_{F0}^2}+\frac{(f_{Fc}-CF_{Fc})^2}{-2\sigma_{Fc}^2}}$ (Eq. 2)
3	Fc-tuned	$R(f_{F0}, f_{Fc}) = ge^{\frac{(f_{Fc}-CF_{Fc})^2}{-2\sigma_{Fc}^2}}$ (Eq. 3)
4	F0-tuned	$R(f_{F0}, f_{Fc}) = ge^{\frac{(f_{F0}-CF_{F0})^2}{-2\sigma_{F0}^2}}$ (Eq. 4)
5	Untuned	$R(f_{F0}, f_{Fc}) = g$ (Eq. 5)

References

1. A. J. M. Houtsma, Pitch and timbre: Definition, meaning and use. *Journal of New Music Research* **26**, 104–115 (1997).

2. A. J. Oxenham, Pitch perception. *J Neurosci* **32**, 13335–13338 (2012).

3. A. Caclin, S. McAdams, B. K. Smith, S. Winsberg, Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. *J Acoust Soc Am* **118**, 471–482 (2005).

4. T. M. Elliott, L. S. Hamilton, F. E. Theunissen, Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *The Journal of the Acoustical Society of America* **133**, 389–404 (2013).

5. J. M. Grey, Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America* **61**, 1270–1277 (1977).

6. P. G. Singh, I. J. Hirsh, Influence of spectral locus and F0 changes on the pitch and timbre of complex tones. *The Journal of the Acoustical Society of America* **92**, 2650–2661 (1992).

7. F. A. Russo, W. F. Thompson, An interval size illusion: The influence of timbre on the perceived size of melodic intervals. *Perception & Psychophysics* **67**, 559–568 (2005).

8. E. J. Allen, A. J. Oxenham, Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America* **135**, 1371–1379 (2014).

9. R. D. Melara, L. E. Marks, Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & Psychophysics* **48**, 169–178 (1990).

10. B. K. Lau, A. J. Oxenham, L. A. Werner, Infant Pitch and Timbre Discrimination in the Presence of Variation in the Other Dimension. *JARO* **22**, 693–702 (2021).

11. K. Siedenburg, S. Jacobsen, C. Reuter, Spectral envelope position and shape in sustained musical instrument sounds. *The Journal of the Acoustical Society of America* **149**, 3715–3726 (2021).

12. C. E. Stilp, T. T. Rogers, K. R. Kluender, Rapid efficient coding of correlated complex acoustic properties. *Proceedings of the National Academy of Sciences* **107**, 21914–21919 (2010).

13. 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).

14. E. J. Allen, P. C. Burton, J. Mesik, C. A. Olman, A. J. Oxenham, Cortical correlates of attention to auditory features. *J. Neurosci.* **39**, 3292–3300 (2019).

15. E. J. Allen, J. Mesik, K. N. Kay, A. J. Oxenham, Distinct representations of tonotopy and pitch in human auditory cortex. *J. Neurosci.* **42**, 416–434 (2022).

16. D. Bendor, X. Wang, The neuronal representation of pitch in primate auditory cortex. *Nature* **436**, 1161–1165 (2005).

17. D. Bendor, X. Wang, Cortical representations of pitch in monkeys and humans. *Current Opinion in Neurobiology* **16**, 391–399 (2006).

18. V. De Angelis, F. De Martino, M. Moerel, R. Santoro, L. Hausfeld, E. Formisano, Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. *NeuroImage* **180**, 291–300 (2018).

19. P. E. Gander, S. Kumar, W. Sedley, K. V. Nourski, H. Oya, C. K. Kovach, H. Kawasaki, Y. Kikuchi, R. D. Patterson, M. A. Howard, T. D. Griffiths, Direct electrophysiological mapping of human pitch-related processing in auditory cortex. *NeuroImage* **202**, 116076 (2019).

20. T. D. Griffiths, The Neural Processing of Complex Sounds. *Annals of the New York Academy of Sciences* **930**, 133–142 (2001).

21. K. L. Hyde, I. Peretz, R. J. Zatorre, Evidence for the role of the right auditory cortex in fine pitch resolution. *Neuropsychologia* **46**, 632–639 (2008).

22. 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).

23. 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. *Journal of Neuroscience* **24**, 6810–6815 (2004).

24. I. Peretz, S. Nguyen, S. Cummings, Tone Language Fluency Impairs Pitch Discrimination. *Front. Psychol.* **2** (2011).

25. P. Q. Pfordresher, S. Brown, Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics* **71**, 1385–1398 (2009).

26. S. M. K. Madsen, M. Marschall, T. Dau, A. J. Oxenham, Speech perception is similar for musicians and non-musicians across a wide range of conditions. *Sci Rep* **9**, 10404 (2019).

27. C. Micheyl, K. Delhommeau, X. Perrot, A. J. Oxenham, Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research* **219**, 36–47 (2006).

28. B. Fischl, FreeSurfer. *NeuroImage* **62**, 774–781 (2012).

29. J. S. Prince, I. Charest, J. W. Kurzawski, J. A. Pyles, M. J. Tarr, K. N. Kay, Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife* **11**, e77599 (2022).

30. C. J. Plack, *The Sense of Hearing* (Routledge, Taylor & Francis Group, London ; New York, Third edition., 2018).

838 31. M. Saenz, D. R. M. Langers, Tonotopic mapping of human auditory cortex. *Hearing Research* **307**,
839 42–52 (2014).

840 32. G. Langner, M. Sams, P. Heil, H. Schulze, Frequency and periodicity are represented in orthogonal
841 maps in the human auditory cortex: evidence from magnetoencephalography. *J Comp Physiol A*
842 **181**, 665–676 (1997).

843 33. M. Moerel, F. D. Martino, E. Formisano, Processing of natural sounds in human auditory cortex:
844 tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* **32**, 14205–14216 (2012).

845 34. B. Barton, J. H. Venezia, K. Saberi, G. Hickok, A. A. Brewer, Orthogonal acoustic dimensions
846 define auditory field maps in human cortex. *Proceedings of the National Academy of Sciences* **109**,
847 20738–20743 (2012).

848 35. M. Moerel, F. De Martino, E. Formisano, An anatomical and functional topography of human
849 auditory cortical areas. *Front Neurosci* **8**, 225 (2014).

850 36. S. Da Costa, W. van der Zwaag, J. P. Marques, R. S. J. Frackowiak, S. Clarke, M. Saenz, Human
851 Primary Auditory Cortex Follows the Shape of Heschl's Gyrus. *J. Neurosci.* **31**, 14067–14075
852 (2011).

853 37. T. M. Talavage, M. I. Sereno, J. R. Melcher, P. J. Ledden, B. R. Rosen, A. M. Dale, Tonotopic
854 Organization in Human Auditory Cortex Revealed by Progressions of Frequency Sensitivity.
855 *Journal of Neurophysiology* **91**, 1282–1296 (2004).

856 38. R. D. Patterson, S. Uppenkamp, I. S. Johnsrude, T. D. Griffiths, The Processing of Temporal Pitch
857 and Melody Information in Auditory Cortex. *Neuron* **36**, 767–776 (2002).

858 39. F. Attneave, R. K. Olson, Pitch as a Medium: A New Approach to Psychophysical Scaling. *The*
859 *American Journal of Psychology* **84**, 147–166 (1971).

860 40. D. Pressnitzer, R. D. Patterson, K. Krumbholz, The lower limit of melodic pitch. *J Acoust Soc Am*
861 **109**, 2074–2084 (2001).

862 41. P. Albouy, L. Benjamin, B. Morillon, R. J. Zatorre, Distinct sensitivity to spectrotemporal
863 modulation supports brain asymmetry for speech and melody. *Science* **367**, 1043–1047 (2020).

864 42. A. Brechmann, H. Scheich, Hemispheric shifts of sound representation in auditory cortex with
865 conceptual listening. *Cerebral Cortex* **15**, 578–587 (2005).

866 43. R. Zatorre, A. Evans, E. Meyer, A. Gjedde, Lateralization of phonetic and pitch discrimination in
867 speech processing. *Science* **256**, 846–849 (1992).

868 44. J. D. Warren, A. R. Jennings, T. D. Griffiths, Analysis of the spectral envelope of sounds by the
869 human brain. *NeuroImage* **24**, 1052–1057 (2005).

870 45. H. B. Barlow, Possible principles underlying the transformation of sensory messages. *Sensory*
871 *communication* **1**, 217–233 (1961).

46. M. Kleiner, D. Brainard, D. Pelli, What's new in Psychtoolbox-3? *Perception* **36**, 1–16 (2007).

47. K. N. Kay, S. V. David, R. J. Prenger, K. A. Hansen, J. L. Gallant, Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Hum Brain Mapp* **29**, 142–156 (2008).

48. K. N. Kay, J. Winawer, A. Mezer, B. A. Wandell, Compressive spatial summation in human visual cortex. *J Neurophysiol* **110**, 481–494 (2013).

49. T. Naselaris, E. Allen, K. Kay, Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences* **40**, 45–51 (2021).

50. K. Kay, K. W. Jamison, L. Vizioli, R. Zhang, E. Margalit, K. Ugurbil, A critical assessment of data quality and venous effects in sub-millimeter fMRI. *NeuroImage* **189**, 847–869 (2019).

51. S. D. Ewert, "AFC—A modular framework for running psychoacoustic experiments and computational perception models" in *Proceedings of the International Conference on Acoustics AIA-DAGA* (Deutsche Gesellschaft für Akustik eV Berlin, 2013)vol. 2013, pp. 1326–1329.

52. H. Levitt, Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America* **49**, 467–477 (1971).

53. M. J. Hautus, N. A. Macmillan, C. D. Creelman, *Detection Theory: A User's Guide* (Routledge, 2021).

Acknowledgments

We thank Jaeeun Lee for assistance in data collection and Juraj Mesik for helpful advice.

Funding:
National Institutes of Health grant R01 DC005216 (AJO)

Author contributions:

Conceptualization: EJA, AJO
Methodology: EJA, KNK, AJO
Investigation: YO, EJA
Formal analysis and visualization: YO
Supervision: EJA, KNK, AJO
Writing—original draft: YO, AJO
Writing—review and editing: YO, EJA, KNK, AJO

Competing interests: All other authors declare they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials.