Naturalistic Audiovisual Stimulation Reveals the Topographic Organization of Human Auditory Cortex

Nicholas Hedger [a], Tomas Knapen [b,c]

[a] Centre for Integrative Neuroscience and Neurodynamics, School of Psychology and Clinical Language Sciences, University of Reading, Reading, RG6 6AL, UK

[b] Spinoza Centre for Neuroimaging, Royal Netherlands Academy of Sciences, Meibergdreef 75, 1105 BK Amsterdam, The Netherlands

[c] Cognitive Psychology, Faculty of Behavioural and Movement Sciences, Vrije Universiteit, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

Correspondence can be addressed to either author.

Abstract

Despite the importance of audition in spatial, semantic, and social function, there is no consensus regarding the detailed organisation of human auditory cortex. Using a novel computational model to analyse a high-powered naturalistic audiovisual movie-watching dataset, we simultaneously estimate spectral tuning properties and category selectivity to reveal the modes of organisation and computational motifs that characterise human auditory cortex. We find that regions more remote from the auditory core exhibit more compressive, non-linear response properties and finely-tuned, speech-selective receptive fields in low frequency portions of the tonotopic map. These patterns of organisation mirror aspects of the visual cortical hierarchy, wherein tuning properties progress from a stimulus category-agnostic 'front end' towards more advanced regions increasingly optimised for behaviorally relevant stimulus categories.

1. Introduction

Auditory information processing is fundamental to human spatial, semantic, and social function. The principal dimension in which auditory sensations are organized is sound frequency. The cochlea [1], brainstem [2], thalamic [3,4], and cortical [5] centers of the auditory system all exhibit *tonotopy*: neighbouring neurons with similar spectral sensitivities give rise to a smooth topographic progression of frequency tuning within each region. The primacy of tonotopy as an organizing principle in both the auditory brain [4,5] and artificial neural networks of audition [6] is uncontested, but there is little consensus on the cortical auditory system's tonotopic structure [7,8].

In analogy to the auditory brain, visual brain regions are organized retinotopically: according to the structure of their sensory organ, the retina [9,10]. In contrast to audition's elusive organization, vision science has been able to leverage explicit computational models of the brain's encoding of visual space to chart the visual system's network of more than 25 retinotopic regions [11–13]. The combination of retinotopic mapping with charting selectivity for salient stimulus categories [14] (e.g. faces, objects and environmental scenes) has delivered a relatively detailed understanding of the layered topology of stages that underlie the visual system's processing hierarchy. In this work, our goal was to apply the computational neuroimaging toolkit of vision to the auditory system in order to reveal its topographic organization.

Here, we used naturalistic stimuli to simultaneously chart both low-level tonotopic structure as well as higher level category-selective structure of auditory cortex. To this end, we leveraged a uniquely high-powered 7-Tesla functional imaging dataset acquired in the Human Connectome Project (HCP), in which 174 participants viewed approximately 1 hour of naturalistic audiovisual movies [15,16]. We analyzed these data by fitting a novel nonlinear computational model - the compressive spectral summation (CSS) model that explains the hemodynamic response of cortical locations as resulting from the interaction between an auditory stimulus and the spectral tuning of their population receptive field (pRF) (**Figure 1**, **Methods**).

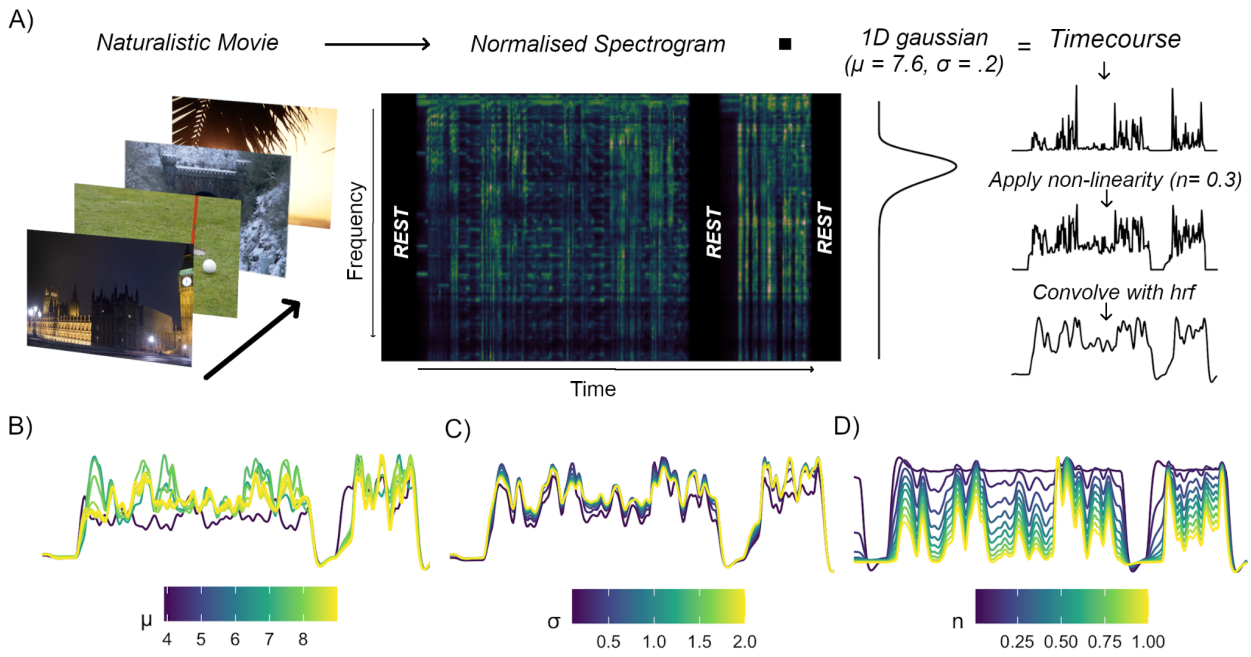

Figure 1. **A)** The movie stimulus is decomposed into a spectrogram that represents the power spectral density of the audio signal. The spectrogram for a 5 minute segment of a movie is shown. A predicted timecourse is generated by taking the dot product of the spectrogram and a Gaussian function defined over log frequency (Hz), parameterised by its peak position (μ) and size (σ). Subsequently a static nonlinearity (n) is applied to the timecourse before convolution with a hemodynamic response function. Panels **B-D** depict a set of model predictions derived from varying the μ, σ and n parameters of the CSS model, as depicted by the colorbars. To visualise the effect of uniquely varying each parameter, the predictions in each plot are generated by fixing the other two parameters. All predictions are peak-normalised.

2. Results

2.1. Global Tonotopic Organisation

The CSS model revealed a population of vertices sensitive to structured auditory information in a large expanse of superior temporal cortex. Preferred frequency (μ) of this population is depicted in **Figure 2A** and **2B**. The population covers Heschl's gyrus (HG) - a landmark traditionally associated with the auditory core [17] and spans portions of the planum temporale (PT), superior temporal gyrus (STG) and terminates ventrally close to the superior temporal sulcus (STS).

The tonotopic structure depicted in Figure **2B** mirrors classical hallmarks of auditory organisation that have been observed in traditional psycho-acoustic paradigms [17,18]. There is a large, 'core'-like, low-frequency region on HG (**i**), surrounded by an antero-medial 'strip' (**ii**) and posterior 'patch' (**iii**)

preferring high frequencies. These high-frequency regions adjoin approximately medially, creating the hallmark 'V' shaped pattern that has been observed consistently in previous investigations [8]. As model performance drops off dorsolaterally, an additional region preferring low frequencies can be observed that roughly aligns with STG (**iv**). The organisation of these features are highly consistent across hemispheres. To quantify the reliability of this tonotopic structure, we assessed the association between the μ estimates obtained from two across-subject folds (the '*early*' and '*late*' subject - see **Methods**). There was robust agreement between these tonotopic maps (Spearman's r_s = .87, p <.001) (**Figure 2C**) and similar levels of agreement with respect to the maps of the pRF size - as defined by full width half maximum (FWHM - r_s = .86, p <.001) and n parameters (r_s = .85, p <.001) (**Supplementary Figure S1**). Although we searched for preferred frequencies up to 8 kHz, we found little selectivity for frequencies above 4 kHz. This range of frequencies represented in our data are in good agreement with previous work involving natural sound stimuli [19–21] and likely reflects the fact that low frequencies tend to dominate natural sounds

2.2. Global Speech Selectivity

We characterised category selectivity via a *speech-selective* model, derived from a linear combination of the CSS-predicted responses to speech and nonspeech stimuli (see **Methods**). **Figure 2D** depicts the improvement in generalisation performance for the speech-selective model over the initial CSS model. Notably, there is very little improvement around HG, but robust improvements in performance can be observed around STG. Speech selectivity (as defined by difference between speech and nonspeech beta-weights) increases ventro-laterally along an axis that roughly aligns and overlaps with STG (**Figure 2E**). **Figure 2F** depicts the μ parameters estimated by the CSS model, with transparency weighted by speech-selective model improvement. Interestingly, this reveals that speech-selective regions primarily occupy the same low-frequency portions of the tonotopic map along STG that were associated with a decline in CSS performance (i.e. region **iv in 2B**).

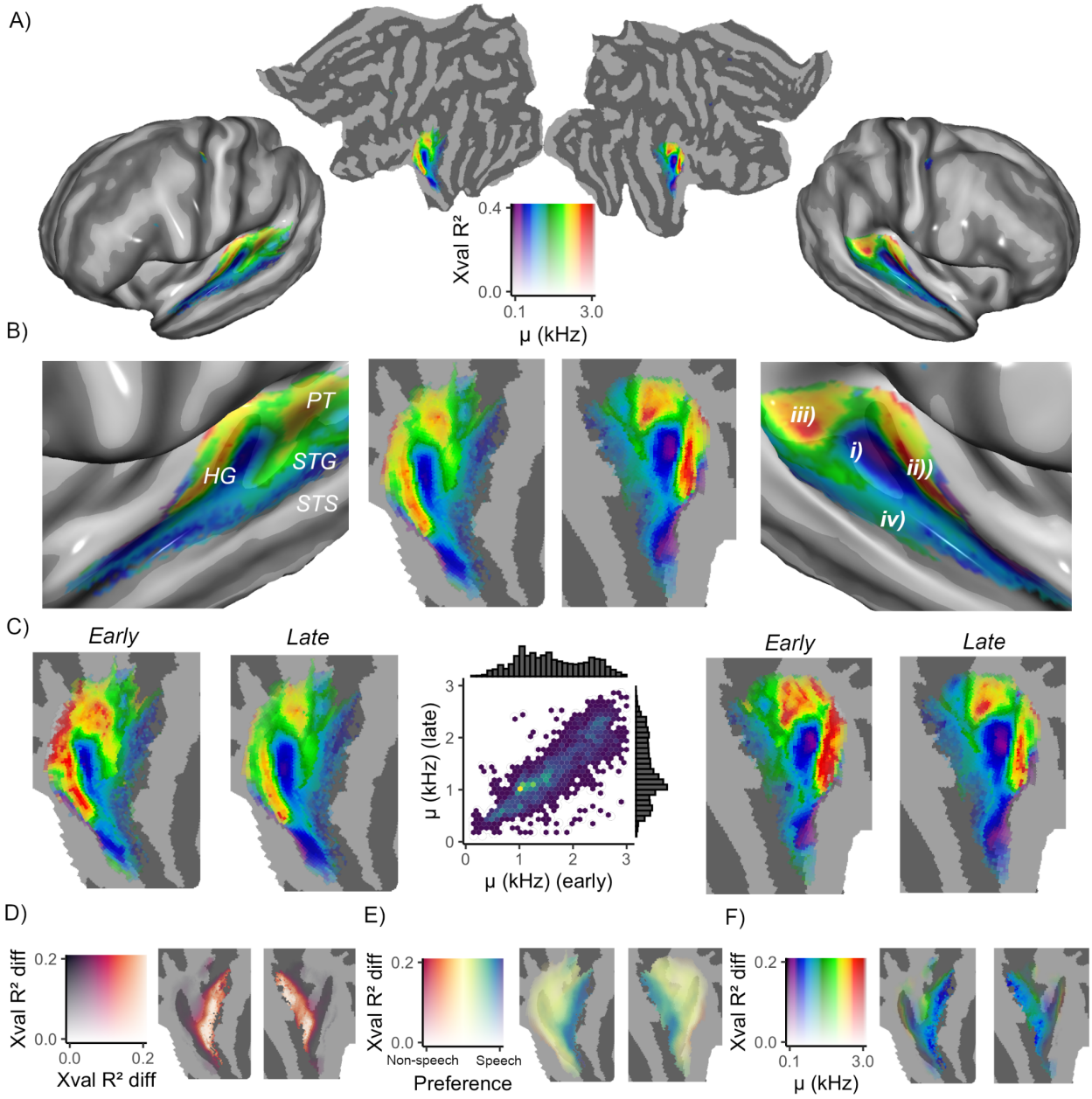

Figure 2. **A)** Shows preferred frequency (μ) of the tonotopic population of vertices that were designated for further analysis (see **Methods**) rendered onto flattened and semi-inflated representations of the cortical surface. The 2D colormap depicts μ across the x axis and generalisation performance (*Xval* R^2) defined by transparency across the y axis. **B)** Depicts zoomed-in views of the data from **A**. **C)** Left, shows the μ data from the left hemisphere of both the 'early' and 'late' subject folds. Right: Shows the data for the right hemisphere of both subjects. Middle: Hexbin plot depicting the association between μ for these two subjects. **D)** Depicts the difference in out-of-sample performance (*Xval* R^2 *diff*) between the speech-selective and CSS models. **E)** Depicts category preferences (speech v nonspeech). **F)** Depicts the μ estimated by the CSS model weighted by model improvement offered by the speech-selective model.

2.3. Decomposition of Functional and Anatomical Parameters.

The preceding modeling has quantified spectral tuning properties, as well as category selectivity in auditory cortex. Since these functional properties are clearly tied to anatomical and myelo-architectural properties (see also [17,22]), we next aggregated our functional quantifications and these anatomical properties into a multimodal dataset to further elucidate auditory cortical topography.

Our estimated functional parameters (the CSS parameters μ, FWHM, n, as well as speech selectivity) are depicted in **Figures 3A-D** alongside structural parameters of the HCP data (i.e. myelin density, cortical thickness, sulcal depth and curvature) in **Figures 3E-H**. We decomposed this wide set of parameters into its low-dimensional structure via weighted principal components analysis [23]. The first 3 components, which together accounted for 83% of the variance explained and showed a high degree of consistency across hemispheres and subject folds are illustrated in **Figures 3I-M**. The feature loadings for each component are summarised in **Figure 3I** above the components themselves in **Figures 3J-M**.

The PCA components reveal separable modes of organisation, ranging from spatially broad to intricate, that differ substantially in their relations to structural and functional parameters. Component 1, for instance, reflects a broad pattern of spatial organisation mostly tied to the structural features, increasing in a roughly postero-medial to antero-lateral direction around the medial tip of HG to STG (**Figure 3K**). Functionally, it also reflects a tendency for μ, FWHM and n to decrease and speech selectivity to increase towards STG. Component 2 (**Figure 3L**) which has a clear resemblance to the preferred frequency data in **Figure 2B,** exhibits a more intricate structure that is most strongly tied to the spectral tuning parameters estimated by the CSS model. In particular, it depicts a region on HG characterised by low μ and FWHM surrounded laterally and medially by regions preferring higher frequencies with more compressive responses. Component 3 (**Figure 3M**) depicts a prominent, heavily myelinated region on HG flanked medially by high FWHM and curvature. These patterns of organisation revealed by the PCA were also highly consistent in the ‘early’ and ‘late’ subjects (**Supplementary Figure S2**).

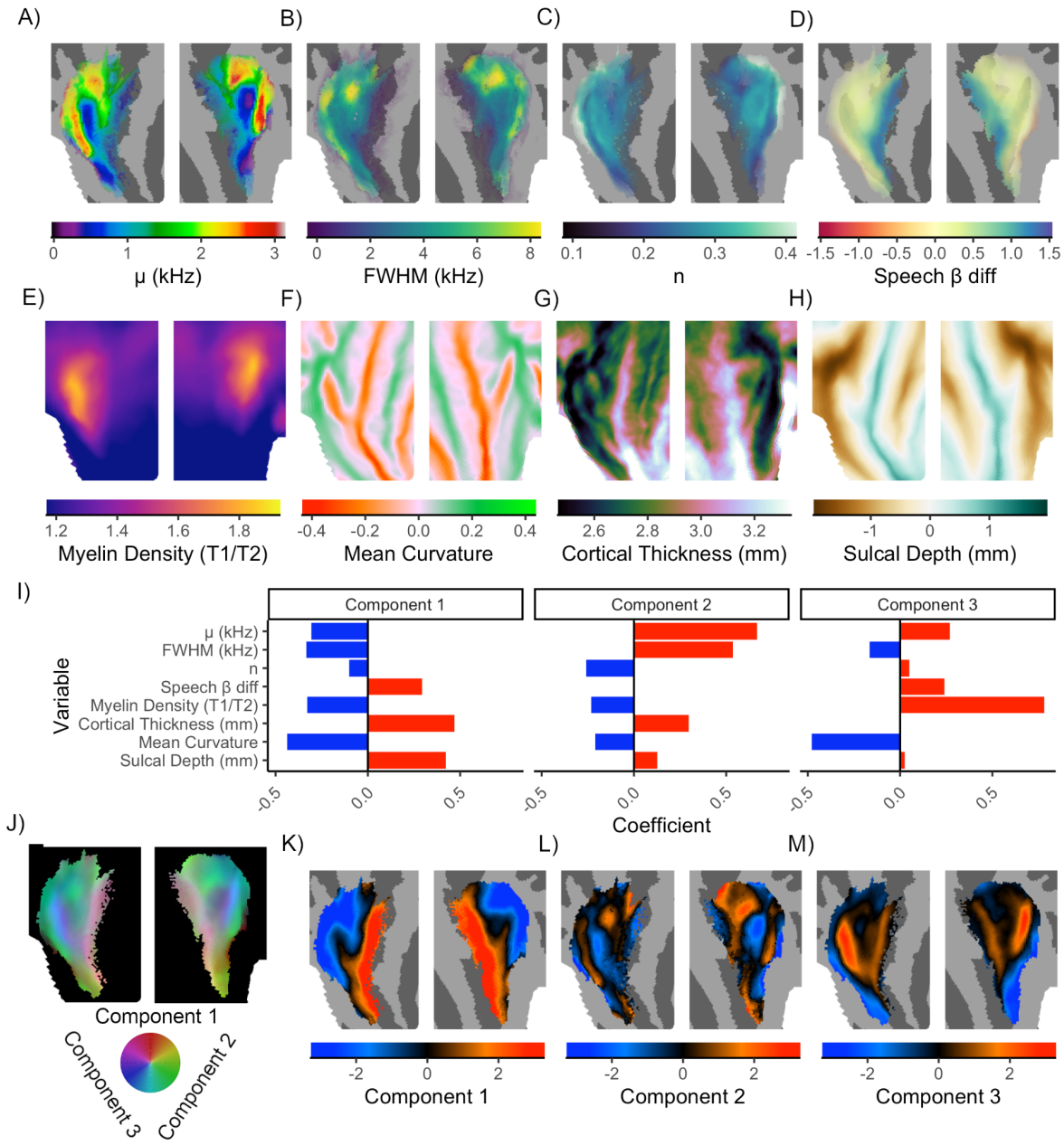

Figure 3. Panels **A-D** depict functional parameters of the data derived from the CSS and speech selective models. Panels **E-H** depict averaged statistics for myelo-architectural and structural parameters. **I)** Depicts the feature loadings onto each of the 3 principal components. **J)** Shows the 3 components overlaid in RGB color space, as depicted by the color-wheel. Panels **K-M** depict the individual components.

2.4. ROI-based Quantification of Functional Properties

To provide spatial precision to our quantifications, we next defined a set of regions of interest (ROIs) based on the presence of tonotopic reversals [18,24], as well as variations in tuning width [19,25] (here

defined as FWHM), speech selectivity [26,27] and myelin density [22] (see **Methods** for details and alternative ROI definitions). The resulting scheme, comprising of 3 'core' ROIs (*A1, R, RT*) and 4 'belt' ROIs (*MBelt, LBelt, P1 and P2*) is depicted in **Figure 4A-B**. To provide a simplified visual representation of the tonotopic gradients present in the data, we fit a linear model to the μ data in each ROI, with the cartesian coordinates of the flattened cortical surface as regressors. These predictions, depicted in **Figures 4A-4B** captured the tonotopic structure of the data very well (R^2 = .71, B =.99, p < .001) .

With ROIs established, we next examined their functional properties. The per-ROI distributions of functional parameters are depicted in **Figures 4C-F**. Between-ROI pairwise differences in these parameters were assessed via weighted-bootstrap procedures (**see Methods**). Since this results in 21 pairwise tests per parameter (with the majority revealing detectable differences), we restrict this section to a brief summary of the salient patterns in the data. The *p* values associated with each pairwise difference are depicted in **Figures 4G-J**.

In terms of variance explained by the CSS model (**Figure 4C**) the poorest model performances are observed in ROIs occupying the peripheral extremes of the tonotopic population (*MBelt, P1 and P2*), likely reflecting the principle that responses more remote from primary sensory regions will be less yoked to the low-level sensory properties of the stimulus. The ROIs vary substantially in speech-preference (**Figure 4D**) with the highest speech preferences observed laterally to the core (e.g. *LBelt, P1, P2*). The FWHM parameter was associated with the least robust differences between ROIs: there was a relatively small range of FWHM values in core regions, but substantial ranges of low and high values in non-core ROIs (**Figure 4E**). This pattern entailed that all core ROIs had detectably lower FWHM relative to *MBelt*, but were not detectably different from, or were larger than ROIs lateral to the core. The most robust differences between ROIs were detected for the exponent parameter n (**Figure 4F***)*. Responses became more compressive with increased distance laterally from the core across *LBelt, P1 and P2*. More medially, *A1, R* and *MBelt* had higher n values than all other regions, whereas *RT* had more compressive responses that were not detectably different from those observed in *P1.*

Previous studies of auditory cortex have indicated that preferred frequency is positively associated with tuning bandwidth [19,28], indicative of a Weber's-law type effect that mirrors the positive relation between pRF size and eccentricity observed in visual cortical field maps [9]. Consistent with these

prior observations, we detected positive relations between μ and FWHM within each ROI, most clearly observed in regions positioned laterally to the core (**Figure 4K**).

Finally, since our initial inspection of the speech-selective model outcomes implied spatial anisotropies in the relationship between preferred frequency and speech selectivity (**Figure 2F**), we formally characterised this relationship within each ROI (**Figure 4L**). In core regions, speech selectivity is low and relatively independent of μ, but this relationship becomes strongly negative in regions lateral to the core. This pattern of results validates our observation that speech-selective regions tend to occupy low frequency portions of the tonotopic map, specifically those residing outside the core.

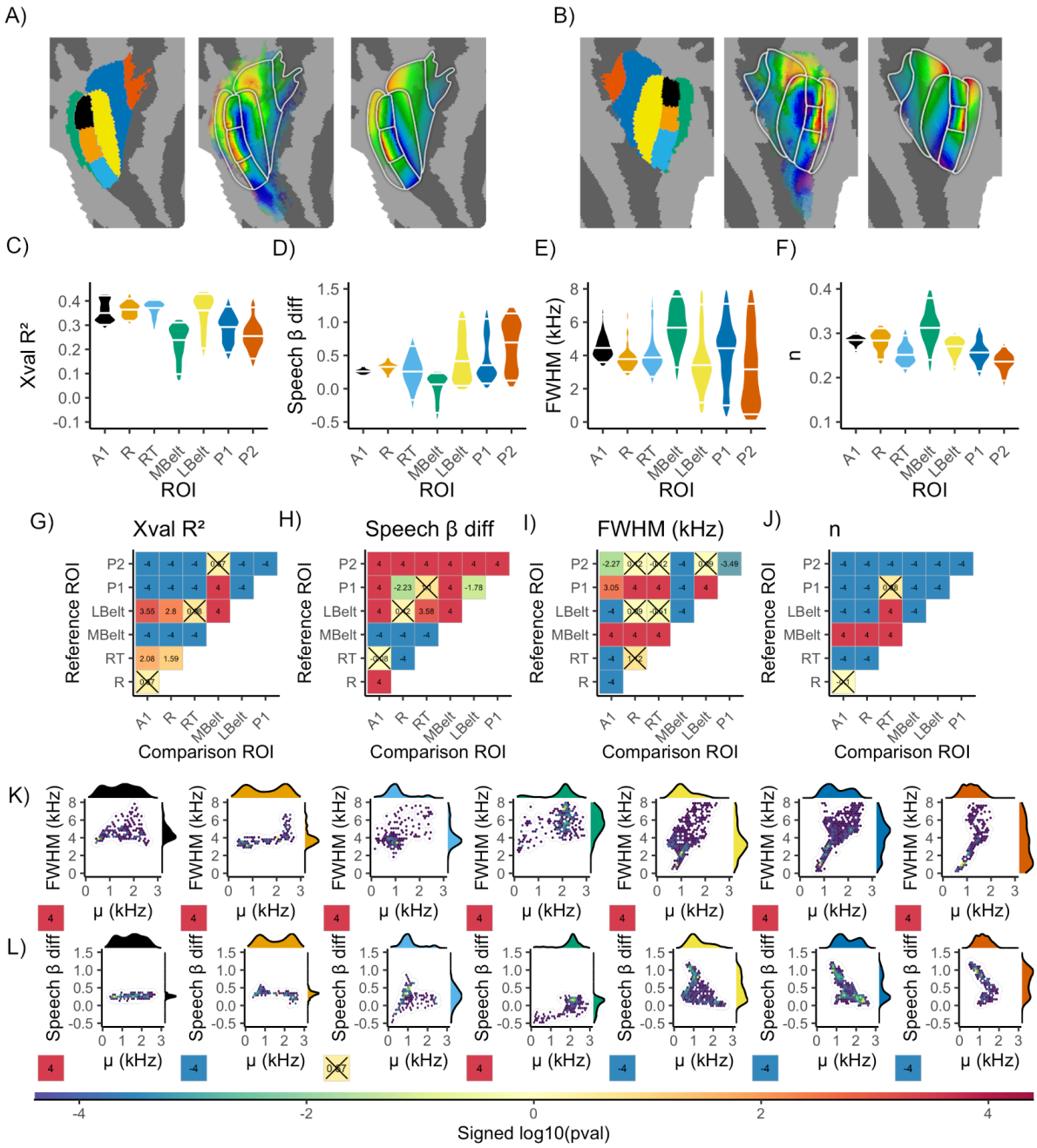

Figure 4. Panel **A** (left to right) shows the ROI definitions, with ROI color corresponding to the colours of the violin /marginal density plots in the remaining figures (left), the μ data (middle) and the linear fit to the μ data per ROI (right) for the left hemisphere. Panel **B** shows the same for the right hemisphere. Panels **C-F** show the per ROI distributions for out-of-sample variance explained by the CSS model (**C**), speech selectivity (**D**), FWHM (**E**) and n (**F**). Violins are normalised to have equal maximum width. The 5th, 50th and 95th percentiles of the distributions are demarcated by white horizontal lines. Panels **G-J** depict the between-ROI pairwise differences for each of the corresponding parameters in panels **C-F**. The values in each cell are signed log10 transformed *p* values. Larger values for a parameter in a reference ROI (y axis) compared to a comparison ROI (x axis) are assigned positive values (p values are -log10 transformed). Smaller values in the reference ROI are assigned negative values (*p* values are log10 transformed). *p* values that do not reach the false discovery rate corrected alpha level of .05 are marked

by a cross in the corresponding cell. Panel **K** shows hexbin plots for the relationship between μ and FWHM in each ROI. The cells at the bottom left of each plot depict the same signed log10 transformed *p* values (now with positive values signifying positive relationships). Panel **L** shows hexbin plots for the relationship between μ and speech selectivity. The corresponding colorbar for all *p* values can be found at the bottom of the figure.

3. Discussion

Drawing inspiration from computational models of vision, our analysis elucidates the computational motifs and modes of organisation that underlie naturalistic auditory processing in human cortex.

Our model mirrored the compressive spatial summation model applied to visual cortex [29,29] by estimating a compressive nonlinearity (n). Interestingly, our range of estimated n values were very close to 0.3 (*M* = 0.29, *SD* = 0.10) - a power-law nonlinearity that is consistent with cochlear amplitude compression [1]. We also revealed a gradient of increased compressivity in regions remote from the core, mimicking a similar profile observed in high-level visual cortex, wherein neurons increasingly tolerate low-level stimulus variations and exhibit more category selecticity [29,30].

We observed finely-tuned, speech-selective receptive fields in low frequency portions of non-primary tonotopic maps. This narrow, low-frequency tuning implies the cortical magnification of spectral features diagnostic of speech and natural soundscapes. By extension, it may reflect information-theoretic efficient coding principles, as illustrated by the observation that learned kernels optimised for encoding speech predict the positive frequency-bandwidth association observed in cochlear filters [31].

The close relationship between category and spectral selectivity would seem to point more to the proposed non-hierarchical structure of auditory cortex [32] , but similar phenomena also characterise high-level sites of visual cortical hierarchy. The fusiform face area (FFA) has a combined preference for faces and central visual field locations, whereas the parahippocampal place area (PPA) prefers environmental scenes and peripheral visual field locations [33] - links between topography and category-selectivity that reflect their computational goals. It is therefore revealing that speech selectivity was relatively independent of preferred frequency in core regions - whereas this relation was strongly negative in regions lateral to the core. This anisotropy in tuning properties agrees with the auditory

processing hierarchy implied by recent neural network models, that progresses from a 'front end' that performs a relatively stimulus category-agnostic analysis of arbitrary sounds towards more advanced regions that are increasingly optimised for salient sound categories, such as speech [6].

It is essential to reflect on potential caveats. Firstly, to enhance signal to noise ratio, our analyses were conducted on across-subject averaged data. Using traditional anatomical co-alignment methods, this would risk regression to the mean via neighbouring tonotopic regions merging together [7]. As such, the vivid tonotopic arrangement we observe speaks both to the robustness of tonotopy as an organising principle and also to the robustness of the HCP cross-subject alignment. Critically, the HCP data were aligned by leveraging 'areal features' (myelin and resting state connectivity maps) explicitly optimised to align primary sensory cortices [34,35], from which our analyses will have benefited greatly relative to traditional alignment based on cortical folding patterns. We must recognise, however, that alignment of regions positioned remotely from primary areas via this method is likely less optimal. This, combined with our conservative criteria for selecting vertices for analysis (see **Methods** and **Supplementary Material S5**), possibly impeded detection of subtle tonotopic arrangements in some non-primary regions. For instance, inconsistent with one recent phase-encoding study [36] and despite our large, naturalistic dataset - the dorsal fields of the STS effectively demarcated the ventral boundary of tonotopic responses (**Supplementary Figure S4)**. It has been proposed that the speech-selective tonotopic portions of STG may represent an initial stage of speech processing wherein the spectral properties of speech sounds begin transformation into their more abstract representations [19]. The neighbouring STS, which exhibits complex multi-modal responses with little visuo-spatial selectivity [37] or spectral selectivity, may represent a site wherein socially relevant features are coded from abstracted forms of audiovisual signals [38].

We must also consider fundamental disjunctures that heavily constrain the analogies we draw between vision and audition. Firstly, retinotopic mapping models a mobile sensory array that samples locations of the visual field, whereas tonotopic mapping models the spectral signature of auditory signals largely incommensurate with their spatial location. This difference likely explains why pRF sizes are reliably larger in more advanced regions of visual cortex [10,29], whereas a range of narrowly and broadly tuned pRFs can be found outside the auditory core [19,39]. Effective processing of objects, faces and scenes are facilitated by location invariance and pooling signals from large extents of the initial sensory space [29].

Conversely, frequency invariance would impede high level tasks such as speech processing, since speech sounds occupy restricted and predictable portions of the sensory space (primarily low frequencies).

Another divergence is that primary auditory cortex would also occupy a later position in a processing hierarchy than its visual counterpart, since there is a more extensive sequence of auditory than visual nuclei in subcortex [1,2] with comparatively later layers of neural networks best explaining responses in primary auditory cortex [6]. Information in 'primary' auditory regions thus represents a relatively processed format of the initial sensory data. Accordingly, additional explanatory power could be derived from modeling other principled dimensions of auditory data, such as periodicity [40] , selectivity for stimulus categories such as music [27] and spatial cognition or explicit recognition task contrasts [41,42].

To summarise, we used novel model-based procedures combined with naturalistic stimuli to elucidate the detailed topography of human auditory cortex. This clarified the major sensory reference frame that underpins higher level auditory cognition, and revealed computational motifs that mirror classical hallmarks of the visual system. Such similarity may reflect common organisational principles shared across multiple sensory systems, aiding the discovery of novel topographic organisational structures in the representation of other senses.

Methods

Participants and Stimuli

Data were taken from the 174 participants of the HCP movie-watching dataset [15,16]. Participants were scanned while watching short (ranging from 1 to 4.3 minutes in length) independent and Hollywood film clips that were concatenated into movies of 11.9 - 13.7 minutes total length. Before each clip, and after the final clip was displayed, there were 20 second 'rest' periods wherein there was no auditory stimulation and only the word '*REST*' presented on the screen. There were 4 separate functional runs, wherein observers viewed 4 separate movies. All 4 movies contained an identical 83 second 'validation' sequence at the end of the movie. Audio was scaled to ensure that no video clips were too loud or quiet across sessions and was delivered by Sensimetric earbuds that provide high-quality acoustic stimulus

delivery while attenuating scanner noise. Full details of the procedure and experimental setup are reported in the HCP S12000 release reference manual.

Data Format and Preparation.

Ultra-high field fMRI (7T) data from the 174 subjects were used, sampled at 1.6 mm isotropic resolution and a rate of 1 Hz [15]. For all analyses, the Fix independent component analysis-denoised time-course data, sampled to the 59,000 vertex-per-hemisphere areal feature-based cross-subject alignment method (MSMAll- [35]) surface format was used. The MSMAII method is optimised for aligning primary sensory cortices based on variations in myelin density and resting state connectivity maps [34,35] . These data are freely available from the HCP project website.

For the purposes of cross-validation, we ensured that functional runs sampled only unique stimulus content. We did this by removing the final 103 seconds of each movie and corresponding functional data that corresponded to the identical 'validation' sequence and the final rest period. For each run, BOLD time series data were then converted to percent signal change and temporally filtered with a Savitzky-Golay filter (3rd order, 210 seconds in length). Finally, we created across-participant time course averages. Specifically, we created two averages to respect the two slightly different versions of the movies viewed by participants at different phases of data collection. The *'early'* subject was averaged across the participants that viewed the versions of the videos before August 21st, 2014 (N = 42). The *'late'* subject was averaged across the participants that viewed the movie after this date (N = 132). In addition to respecting the minor differences in the videos that were presented, dividing the data in this way also allowed us to characterise the split reliability of parameter estimates and also assess the impact of statistical power, which differs between these two across-participant folds.

Design Matrix

To create a design matrix for our pRF modeling, the original .mp4 file for the movies were converted to a .wav audio file and then submitted to a Fourier transform via scipy's *'signal.spectrogram'* function, to generate spectrograms that represent the power spectral density of the audio signal. The

resulting spectrograms were then mean-downsampled into the TR resolution and normalised via division by the corresponding per-movie, per-frequency standard deviations.

Compressive Spectral Summation pRF Model

A pRF model is a parsimonious encoding model that characterises the relationship between a stimulus and the response from a tuned population of neurons [10]. We modeled the pRF as a 1D Gaussian function defined over log auditory frequency (Hz). The model generates a predicted response by computing a weighted sum of the pRF and the spectrogram and then applying a static power-law nonlinearity. This can be expressed formally as:

$$RESP = [S(f)G(f)]^n$$

$$G(f) = e^{\frac{(f-\mu)^2}{2\sigma^2}}$$

where RESP is the predicted response, *f* is the auditory frequency, S is the spectrogram, G is a Gaussian, n is an exponent parameter and μ and σ are parameters that control the mean and standard deviation of the Gaussian. The resulting timecourses (RESP) generated by the model were convolved with a double-gamma hemodynamic response function and temporally filtered with the same Savitzky-Golay filter as was applied to the functional data. We incorporated a power-law nonlinearity in our modeling to capture a wide range of response properties with a single parameter that is straightforward to interpret. Lower values of n indicate greater compressive nonlinearity and imply that smaller amounts of overlap between the prf and the stimulus produce larger responses. In vision, these more compressive response properties are characteristic of later regions of the visual hierarchy [30]. In application to spatial vision, this model has been referred to as the compressive spatial summation (CSS) model [29]. Here, in its application to audition and the spectral domain, we refer to it as a compressive *spectral* summation model. In such a model, it is important to account for the fact that the effective size of a pRF is influenced not only by σ, but also by n, since nonlinearity implies that pRFs with a small σ may still respond strongly to frequencies remote from μ if n is highly compressive. We therefore performed the following adjustment to σ:

$$\sigma_{adj} = \frac{\sigma}{\sqrt{n}}$$

To provide a final measure of pRF size, we calculated the full width half maximum of these adjusted pRF functions in kHz.

Model Fitting and Vertex Selection

CSS model

Model fits for each vertex in the brain were obtained by finding parameter combinations that maximized the correlation between the predicted and observed fMRI time-courses. This parameter search consisted of a coarse-to-fine strategy. We first performed an initial coarse grid search of μ and σ values, based on the initial parameter space sampled in a previous pRF study of auditory cortex [39] . μ was bounded between .088 kHz and 8 kHz and σ had an upper bound of 4. The best fitting parameters from this set were then used as starting parameters for a nonlinear search algorithm (scipy's *minimize* function, using the L-BFGS-B method) which uses nonlinear minimization to find the pRF model parameters μ and σ that maximize the correlation between the pRF-predicted time-series and the observed fMRI time-course. These parameters for μ and σ were then submitted as starting parameters to a final iterative search, wherein the full set of model parameters - μ, σ and additionally the exponent parameter (n) were estimated.

To assess the generalisation performance of the pRF models, we used a leave-one-run-out cross-validation strategy. Models were fit to four separate folds of data. In each fold, the training data were concatenated across 3 movies and the fitted parameters were used to generate predictions for the left out movie. The out-of-sample performance of these predictions was then assessed by the coefficient of determination (R^2). All pRF model fitting procedures were carried out using procedures from a dedicated python package: *prfpy.* To summarise the fold-wise parameter estimates, we calculated across-fold averages of the parameters in a manner that was weighted by the corresponding out-of-sample R^2. This procedure was carried out for both the 'early' and 'late' subject before the parameters obtained from each were again weighted-averaged to provide final parameters.

To focus our analysis and select vertices for further analyses, we used a conservative, null-model based criterion to exclude non-tonotopic vertices (**Supplementary Material S5 and Supplementary Figure S9**). These procedures resulted in a 'tonotopic population' consisting of 3760 from the total 118584 cortex vertices being included in further analyses. We did not perform any spatial-smoothing of data or model parameters at any point in analysis or presentation of our data.

Speech-selective model

In addition to purely tonotopic responses captured by the CSS model, it is important to characterise the combination of both tonotopic and speech-selective responses. To this end, we additionally fit a simple *speech-selective* model that involved a linear combination of the separate CSS model-predicted responses to speech and nonspeech stimuli. HCP movie files were privately uploaded to YouTube. We employed YouTube's auto-captioning algorithm to provide an initial estimate of the onset and offset times of speech sounds. These initial estimates were then manually fine-tuned by NH. Corrections mostly involved adding brief speech sounds that occured in the presence of ambient noise, or adjusting offset times, which were occasionally estimated to occur beyond the actual duration of the speech. With onset and offset times refined, we created two new design matrices. The *speech* design matrix was created by coding parts of the spectrogram overlapping with nonspeech sounds as 0. The *nonspeech* design matrix was the inverse - with all instances of speech coded as 0. We then used the parameters estimated by the CSS model to generate predictions based on these new *speech* and *non-speech* design matrices. These predictions were then entered as regressors for the design matrix of the speech-selective model, wherein we estimated the linear combination of the speech and non-speech regressors that minimised the error between the predicted and observed timecourses for every cortical location separately. Generalisation performance was evaluated with the same cross-validation strategy as reported for the CSS model.

ROI Definitions

Approaches to delineating distinct regions of auditory cortex are heterogenous, ranging from a simple, binary distinction between 'core' and 'non-core' regions [27,39], to extrapolating detailed non-human

primate atlases onto human data [24,43], or defining multiple regions as remote as the STS based solely on tonotopic reversals [36] . This diversity may reflect the preference for 'lumping' versus 'splitting' of distinct cortical areas [9]. To respect this range of perspectives, we used 3 strategies to define regions of interest (ROIs) within which to provide spatial precision to our quantifications. A detailed account of each of these strategies is provided in **Supplementary Material S3 and Supplementary Figures S3-S4**. Here in the main text, we report quantifications for ROIs defined in our own '*splitting'* scheme, wherein we delineate between 7 core and belt regions based on the presence of tonotopic reversals [18,24], variations in tuning width [19,25] (here defined as FWHM), degree of speech selectivity [26,27] and myelination [22]. For completeness, in **Supplementary Material S4 and Supplementary Figures S5-S8**, we provide equivalent quantifications for ROIs defined in two alternative schemes: i) a simple *'lumping'* scheme that binarizes the tonotopic population of vertices into 'core' and 'non-core' regions and ii) a *'neutral'* scheme, that ignores our own model parameters and uses the existing HCP multi-modal parcellation of auditory cortices [34].

Bootstrapping Procedure

p-values were computed using 10^4 fold bootstrap procedures, with the estimation of parameters (mean differences, regression slopes) for each bootstrap sample weighted by out-of-sample R^2. To test whether bootstrapped distributions differed from a certain threshold, *p*-values were defined as the ratio of bootstrap samples below versus above that threshold multiplied by 2 (all reported *p*-values are two-tailed). All *p* values were then corrected for multiple comparisons across ROIs using a false discovery rate correction. For any comparisons wherein all bootstrapped samples fell on one side of the threshold, the *p* value was coded as $1/10^4$. For presentation, we $-\log^{10}$ transformed *p* values resulting from a positive difference between a given reference and comparison ROI (resulting in positive values) and log10 transformed *p* values resulting from negative differences (resulting in negative values). The same transformation was applied to *p* values resulting from positive v negative associations between variables. All reported *p* values thus reside on a scale between -4 and 4.

Data Access Statement

We have provided our estimated model parameters and ROI definitions in 'gifti' surface format (both in original HCP 59k space and upsampled to 'fsaverage' 164k space) in the following figshare repository: https://figshare.com/projects/Data_from_Naturalistic_Stimulation_Reveals_the_Topographic_Organization_of_Human_Auditory_Cortex_/117288 . We also provide the HCP 'pycortex' subject that was used to produce visualisations.

References

1. McDermott, J. H. Audition. in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (ed. Wixted, J. T.) 1–57 (John Wiley & Sons, Inc., 2018). doi:10.1002/9781119170174.epcn202.
2. Sitek, K. R. *et al.* Mapping the human subcortical auditory system using histology, postmortem MRI and in vivo MRI at 7T. *eLife* **8**, e48932 (2019).
3. De Martino, F. *et al.* Spatial organization of frequency preference and selectivity in the human inferior colliculus. *Nat. Commun.* **4**, 1386 (2013).
4. Moerel, M., De Martino, F., Uğurbil, K., Yacoub, E. & Formisano, E. Processing of frequency and location in human subcortical auditory structures. *Sci. Rep.* **5**, 17048 (2015).
5. Saenz, M. & Langers, D. R. M. Tonotopic mapping of human auditory cortex. *Hear. Res.* **307**, 42–52 (2014).
6. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98**, 630-644.e16 (2018).
7. Brewer, A. A. & Barton, B. Maps of the Auditory Cortex. *Annu. Rev. Neurosci.* **39**, 385–407 (2016).
8. Moerel, M., De Martino, F. & Formisano, E. An anatomical and functional topography of human auditory cortical areas. *Front. Neurosci.* **8**, (2014).
9. Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual Field Maps in Human Cortex. *Neuron* **56**, 366–383 (2007).

10. Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *NeuroImage* **39**, 647–660 (2008).
11. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic Maps of Visual Topography in Human Cortex. *Cereb. Cortex N. Y. NY* **25**, 3911–3931 (2015).
12. Knapen, T. Topographic connectivity reveals task-dependent retinotopic processing throughout the human brain. *Proc. Natl. Acad. Sci.* **118**, (2021).
13. van Es, D. M., van der Zwaag, W. & Knapen, T. Topographic Maps of Visual Space in the Human Cerebellum. *Curr. Biol.* **29**, 1689-1694.e3 (2019).
14. Brewer, A. A., Liu, J., Wade, A. R. & Wandell, B. A. Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nat. Neurosci.* **8**, 1102–1109 (2005).
15. Uğurbil, K. *et al.* Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *NeuroImage* **80**, 80–104 (2013).
16. Glasser, M. F. *et al.* The Minimal Preprocessing Pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124 (2013).
17. Da Costa, S. *et al.* Human primary auditory cortex follows the shape of Heschl's gyrus. *J. Neurosci. Off. J. Soc. Neurosci.* **31**, 14067–14075 (2011).
18. Langers, D. R. M. & van Dijk, P. Mapping the Tonotopic Organization in Human Auditory Cortex with Minimally Salient Acoustic Stimulation. *Cereb. Cortex N. Y. NY* **22**, 2024–2038 (2012).
19. Moerel, M., Martino, F. D. & Formisano, E. Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *J. Neurosci.* **32**, 14205–14216 (2012).
20. Sood, M. R. & Sereno, M. I. Areas activated during naturalistic reading comprehension overlap topological visual, auditory, and somatotomotor maps. *Hum. Brain Mapp.* **37**, 2784–2810 (2016).
21. Dick, F. *et al.* In Vivo Functional and Myeloarchitectonic Mapping of Human Primary Auditory Areas. *J. Neurosci.* **32**, 16095–16105 (2012).
22. De Martino, F. *et al.* High-Resolution Mapping of Myeloarchitecture In Vivo: Localization of Auditory Areas in the Human Brain. *Cereb. Cortex N. Y. N 1991* **25**, 3394–3405 (2015).
23. Delchambre, L. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Mon. Not. R. Astron. Soc.* **446**, 3545–3555 (2015).

24. Humphries, C., Liebenthal, E. & Binder, J. R. Tonotopic organization of human auditory cortex. *NeuroImage* **50**, 1202–1211 (2010).
25. Moerel, M. *et al.* Processing of Natural Sounds: Characterization of Multipeak Spectral Tuning in Human Auditory Cortex. *J. Neurosci.* **33**, 11888–11898 (2013).
26. Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P. & Pike, B. Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312 (2000).
27. Norman-Haignere, S., Kanwisher, N. G. & McDermott, J. H. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron* **88**, 1281–1296 (2015).
28. Cheung, S. W., Bedenbaugh, P. H., Nagarajan, S. S. & Schreiner, C. E. Functional organization of squirrel monkey primary auditory cortex: responses to pure tones. *J. Neurophysiol.* **85**, 1732–1749 (2001).
29. Kay, K. N., Winawer, J., Mezer, A. & Wandell, B. A. Compressive spatial summation in human visual cortex. *J. Neurophysiol.* **110**, 481–494 (2013).
30. Kay, K. N., Weiner, K. S. & Grill-Spector, K. Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr. Biol. CB* **25**, 595–600 (2015).
31. Smith, E. C. & Lewicki, M. S. Efficient auditory coding. *Nature* **439**, 978–982 (2006).
32. Staeren, N., Renvall, H., De Martino, F., Goebel, R. & Formisano, E. Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. *Curr. Biol.* **19**, 498–502 (2009).
33. Finzi, D. *et al.* Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. *Nat. Commun.* **12**, 2278 (2021).
34. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
35. Robinson, E. C. *et al.* MSM: a new flexible framework for Multimodal Surface Matching☆. *NeuroImage* **100**, 414–426 (2014).
36. Striem-Amit, E., Hertz, U. & Amedi, A. Extensive Cochleotopic Mapping of Human Auditory Cortical Fields Obtained with Phase-Encoding fMRI. *PLOS ONE* **6**, e17832 (2011).
37. Pitcher, D. *et al.* The Human Posterior Superior Temporal Sulcus Samples Visual Space Differently From Other Face-Selective Regions. *Cereb. Cortex* **30**, 778–785 (2020).

38. Pitcher, D. & Ungerleider, L. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends Cogn. Sci.* **25**, (2021).
39. Thomas, J. M. *et al.* Population receptive field estimates of human auditory cortex. *NeuroImage* **105**, 428–439 (2015).
40. Barton, B., Venezia, J. H., Saberi, K., Hickok, G. & Brewer, A. A. Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proc. Natl. Acad. Sci.* **109**, 20738–20743 (2012).
41. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
42. Rauschecker, J. P. & Tian, B. Mechanisms and streams for processing of 'what' and 'where' in auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 11800–11806 (2000).
43. Leaver, A. M. & Rauschecker, J. P. Functional Topography of Human Auditory Cortex. *J. Neurosci.* **36**, 1416–1428 (2016).

Acknowledgements

Data were provided by the Human Connectome Project, Washington University - Minnesota Consortium (Principal Investigators D. Van Essen and K. Uğurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis. NH is supported by an Leverhulme Trust Early Career Fellowship (ECF 2019-305).