nature neuroscience

BRIEF COMMUNICATION

https://doi.org/10.1038/s41593-019-0410-7

Divergence in the functional organization of human and macaque auditory cortex revealed by fMRI responses to harmonic tones

Sam V. Norman-Haignere[1,2,3,4]*, Nancy Kanwisher[2,5,6], Josh H. McDermott[2,5,6,7] and Bevil R. Conway[8,9,10]*

We report a difference between humans and macaque monkeys in the functional organization of cortical regions implicated in pitch perception. Humans but not macaques showed regions with a strong preference for harmonic sounds compared to noise, measured with both synthetic tones and macaque vocalizations. In contrast, frequency-selective tonotopic maps were similar between the two species. This species difference may be driven by the unique demands of speech and music perception in humans.

How similar are the brains of humans and nonhuman primates? Visual cortex is similar between humans and macaque monkeys[1,2], but less is known about audition. Audition is an important test case because speech and music are both central and unique to humans. Speech and music contain harmonic frequency components, which are perceived to have 'pitch'[3]. Humans have cortical regions with a strong response preference for harmonic tones versus noise[4–6]. These regions are good candidates to support pitch perception because their response depends on the presence of low-numbered resolved harmonics known to be the dominant cue to pitch in humans[5,6] (see Supplementary Note). Here, we tested whether macaque monkeys also have regions with a response preference for harmonic tones.

We measured cortical responses to harmonic tones and noise, spanning five frequency ranges (Fig. 1a) in a sparse block design (Fig. 1b) using functional MRI (fMRI) (experiment IA). We tested 3 macaques and 4 human participants. The noise stimuli were presented at a slightly higher sound intensity (73 dB) than the harmonic tone stimuli (68 dB) to equate perceived loudness in humans[6].

To assess tonotopic organization, we contrasted the two lowest and the two highest frequency ranges, collapsing across tone and noise conditions (Fig. 1c). Consistent with prior work, humans showed two mirror-symmetric tonotopic gradients (high→low→high) organized in a V shape around Heschl's gyrus[6,7]. In contrast, macaques showed a straighter and extended version of the same pattern, progressing high→low→high→low from posterior to anterior[8].

We next contrasted responses to harmonic tones versus noise, collapsing across frequency. All humans showed tone-selective voxels that overlapped the low-frequency field of primary auditory cortex and extended into anterior non-primary regions, as expected[4–6] (Fig. 1d). Each human participant showed significant clusters of tone-selective voxels after correction for multiple comparisons (Supplementary Fig. 1; voxel-wise threshold of $P<0.01$, cluster-corrected to $P<0.05$; P values here and elsewhere are two-sided). In contrast, tone-selective voxels were largely absent from macaques (Supplementary Fig. 2 shows maps with a more liberal voxel-wise threshold), and almost never survived cluster-correction. Conversely, macaques showed significant noise-selective voxel clusters, whereas in humans, such voxels were rare and never survived cluster correction.

We quantified these observations using region-of-interest (ROI) analyses. As the human data were more reliable per block, we collected much more data in macaques and, when necessary, subsampled the human data (Fig. 2a,b). ROIs were defined using the same low versus high and tone versus noise contrasts. The ROI size was varied by selecting the top N% of sound-responsive voxels, rank-ordered by the significance of their response preference for the relevant contrast. We used a standard index to quantify selectivity in independent data: (preferred − nonpreferred) / (preferred + nonpreferred) (Fig. 2c–f; Supplementary Figs. 3 and 4 plot the responses for preferred and nonpreferred stimuli separately).

Results from an example ROI size (top 5% of sound-responsive voxels) summarize the key findings (Fig. 2c,d). Low-frequency and high-frequency selectivity were significant in both species (group-level $P\text{s}<0.001$) and were comparable ($P\text{s}>0.112$ between species) ("Ps" indicates multiple tests; for all ROI analyses, significance was evaluated via bootstrapping across participants and runs; see the section "ROI statistics" in the Methods). However, tone-selective responses were only observed in humans (at the group-level, $P<0.001$ for humans and $P=0.776$ for macaques; $P<0.001$ between species). Noise selectivity was significant in macaques but not humans (at the group-level, $P=0.192$ for humans and $P=0.001$ for macaques; $P=0.154$ between species). This pattern was consistent across participants and ROI sizes (Fig. 2e,f). We also confirmed prior observations that tone-selective and low-frequency-selective responses overlap in humans[6]: ROIs defined by low-frequency selectivity were selective for tones compared to noise (at the group level, $P\text{s}<0.002$ for all ROI sizes) (Supplementary Fig. 5). But in macaques, both low-frequency and high-frequency ROIs showed a slight noise preference.

[1] Zuckerman Institute for Mind, Brain and Behavior, Columbia University, New York, NY, USA. [2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA. [3] HHMI Postdoctoral Fellow of the Life Sciences Research Institute, Chevy Chase, MD, USA. [4] Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, PSL University, CNRS, Paris, France. [5] McGovern Institute for Brain Research, Cambridge, MA, USA. [6] Center for Minds, Brains and Machines, Cambridge, MA, USA. [7] Program in Speech and Hearing Biosciences and Technology, Harvard University, Cambridge, MA, USA. [8] Laboratory of Sensorimotor Research, NEI, NIH, Bethesda, MD, USA. [9] National Institute of Mental Health, NIH, Bethesda, MD, USA. [10] National Institute of Neurological Disease and Stroke, NIH, Bethesda, MD, USA. *e-mail: sn2776@columbia.edu; bevil@nih.gov

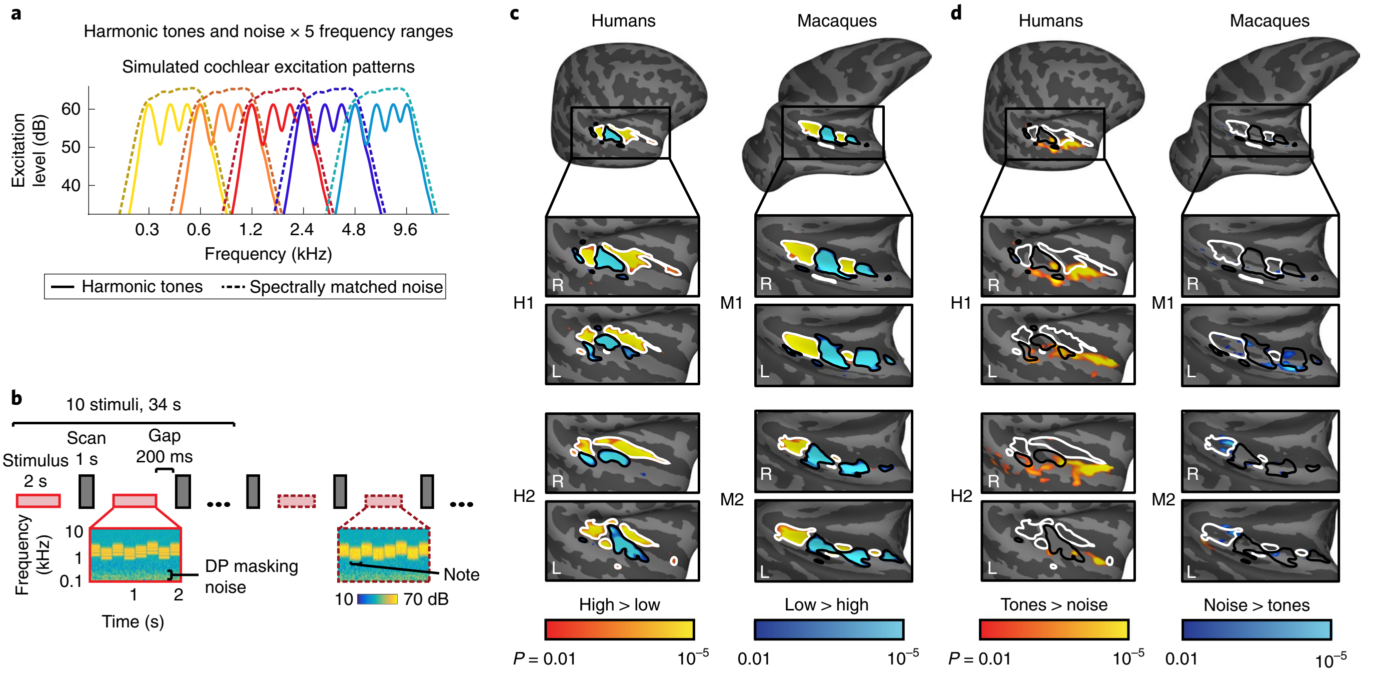

Fig. 1 | Assessing tonotopy and selectivity for harmonic tones versus noise. a, Schematic of the 5 × 2 factorial stimulus design for experiment IA. Stimuli were harmonic tones (harmonics 3–6 of the F_0) and spectrally matched Gaussian noise, each presented in five frequency ranges. Plots show estimated cochlear response magnitudes versus frequency for example notes from each condition. Noise notes had slightly higher intensity (73 versus 68 dB) to approximately equate perceived loudness in humans[6]. **b**, Stimuli were presented in a sparse block design. Scanning and stimulus presentation alternated to avoid scanner noises interfering with stimulus presentation. Each stimulus comprised several notes; each block comprised 10 stimuli of the same condition. The F_0 and frequency range were jittered from note to note to minimize adaptation. Cochleagrams (plotting energy versus time and frequency) are shown for a mid-frequency harmonic tone stimulus (left) and spectrally matched noise stimulus (right). Noise was used to mask distortion products (DPs). **c**, Maps of frequency-selective responses (tonotopy) for 2 human participants (H1, H2) and 2 macaque monkeys (M1, M2), for right (R) and left (L) hemispheres. Voxels showing greater responses to low frequencies (blue, black outlines) versus high frequencies (yellow, white outlines) collapsing across tone and noise conditions (number of blocks per low-frequency or high-frequency condition: M1 = 504, M2 = 408, H1 = 32, H2 = 32). **d**, Maps of tone versus noise responses. Voxels showing greater responses to harmonic tones (yellow) versus noise (blue) collapsing across frequency (number of blocks per tone or noise condition: M1 = 630, M2 = 510, H1 = 40, H2 = 40). Maps are shown for the macaques and humans with the highest response reliability. Maps plot uncorrected voxel-wise significance values (two-sided $P < 0.01$ via a permutation test across stimulus blocks; see Supplementary Fig. 1 for cluster-corrected maps).

Could the weak tone selectivity in macaques be attributable to the lower sound intensity of the tones tested in experiment IA (68 dB tones, 73 dB noise)? To address this question, 2 additional monkeys (M4 and M5) were tested using tone and noise stimuli presented at three matched sound levels (70, 75, and 80 dB) (experiment IB). Data obtained in human participants from experiment IA were used for comparison, and did not need to be subsampled because we collected hundreds of repetitions per condition in macaques (Supplementary Fig. 6).

For tones and noise of the same intensity, significant tone-selective voxels were only observed in monkeys for small ROIs (*P*s < 0.002 for the three smallest ROIs at the group level and in individual subjects), and these voxels were substantially less selective than those in humans (Fig. 3a,b; Supplementary Fig. 7a; *P*s < 0.041 for all ROI sizes and all comparisons of every human with every monkey). Noise-selective responses, by contrast, did not differ significantly between species (at the group-level, *P*s > 0.055 for all ROI sizes; Fig. 3b lower panels; Supplementary Fig. 7b). When comparing tones with noises that were 5-dB higher in sound intensity (tones 70 and 75 dB versus noises 75 and 80 dB, respectively), similar to experiment IA, tone-selective responses were even weaker: M5 showed no tone-selective voxels (*P*s > 0.25 for all ROI sizes), and M4 only showed tone-selective responses for the smallest ROI (0.6%, *P* = 0.008). These results suggest that tone-selective voxels in macaques are sensitive to small variations in sound intensity, which we verified by assessing the effect of sound intensity (Fig. 3c; *P*s < 0.049 across all ROI sizes at the group level; *P*s < 0.036 for all but the two smallest ROIs in both individual monkeys). The magnitude of intensity-driven changes was comparable to or larger than the tone versus noise effect, depending on the individual and ROI size (see Methods for quantification).

Frequency-selective responses were significant in both monkeys (*P*s < 0.001 for both low-frequency and high-frequency ROIs in both animals for all but the largest ROI size), and were comparable to the results obtained in humans (Supplementary Figs. 7c,d and 8). For both low-frequency and high-frequency ROIs, the effect of frequency was greater than the effect of intensity (*P*s < 0.002 for the five smallest ROI sizes in both individual monkeys). Responses to the preferred frequency range were always higher than for the non-preferred frequency range for all pairs of intensities (*P*s < 0.002 for the four smallest ROIs for both high-frequency and low-frequency ROIs in both individual monkeys). Thus frequency-selective responses in macaques were tolerant to variations in sound level, whereas tone-selective and noise-selective responses, when evident, were not.

Synthetic tones are familiar to most humans, but perhaps less familiar to macaque monkeys. Were tone-selective responses weak in macaques because the stimuli were not ecologically relevant? To address this question, we measured responses to voiced macaque calls, which contain harmonically organized spectral

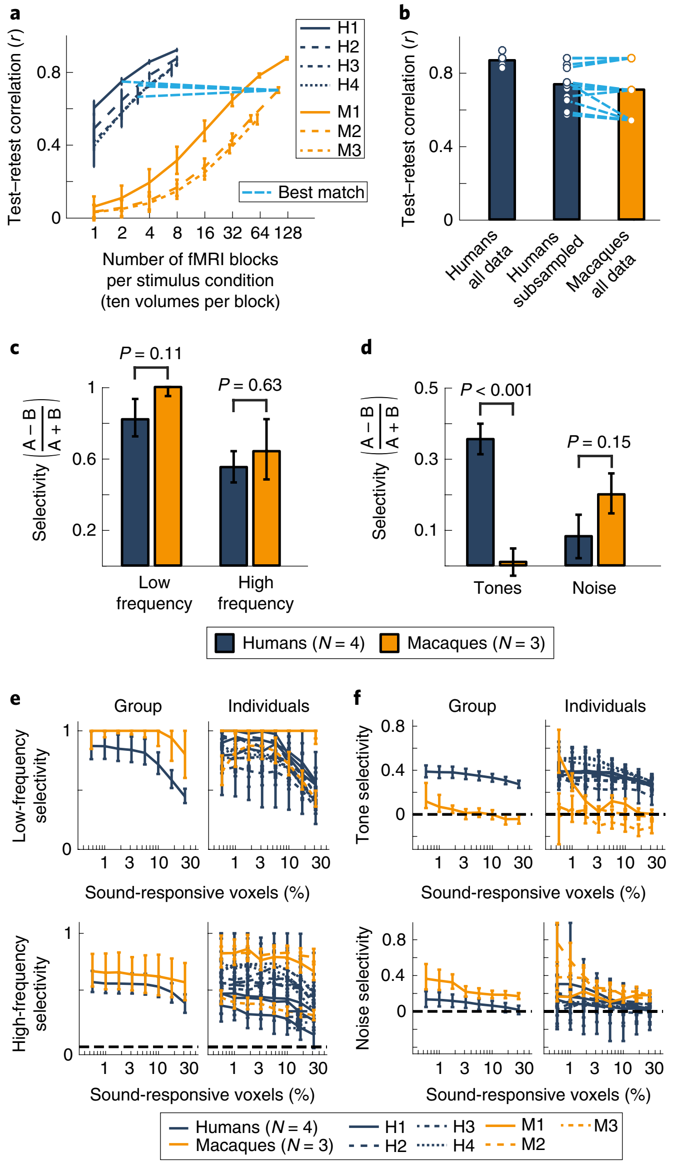

Fig. 2 | ROI analyses controlling for data reliability. a, Test–retest response reliability (Pearson correlation, r) versus data quantity. Dashed light blue lines show the number of blocks in each human (H1–H4) needed to approximately match the response reliability of one monkey (M2). Error bars show 1 s.d. across subsampled sets of runs. **b**, The average response reliability of the macaque and human data (before and after subsampling; circles represent individual participants). **c**,**d**, ROI analyses applied to reliability-matched data. For each human and each monkey, we selected the top 5% of sound-driven voxels with the most significant response preference for low frequencies versus high frequencies (**c**) or tones versus noise (**d**). A standard selectivity metric was applied to the average response of the selected voxels (measured in independent data). **e**,**f**, Same as in **c** and **d**, respectively, but varying the ROI size (percentage of voxels selected) and showing data from individual participants in addition to group-averaged data. Error bars here and elsewhere plot 1 s.e. of the bootstrapped sampling distribution (median and central 68%). Bootstrapping was performed across runs for individual participants and across both participants and runs for group data (each stimulus condition was presented once per run; see the section "ROI statistics" in the Methods).

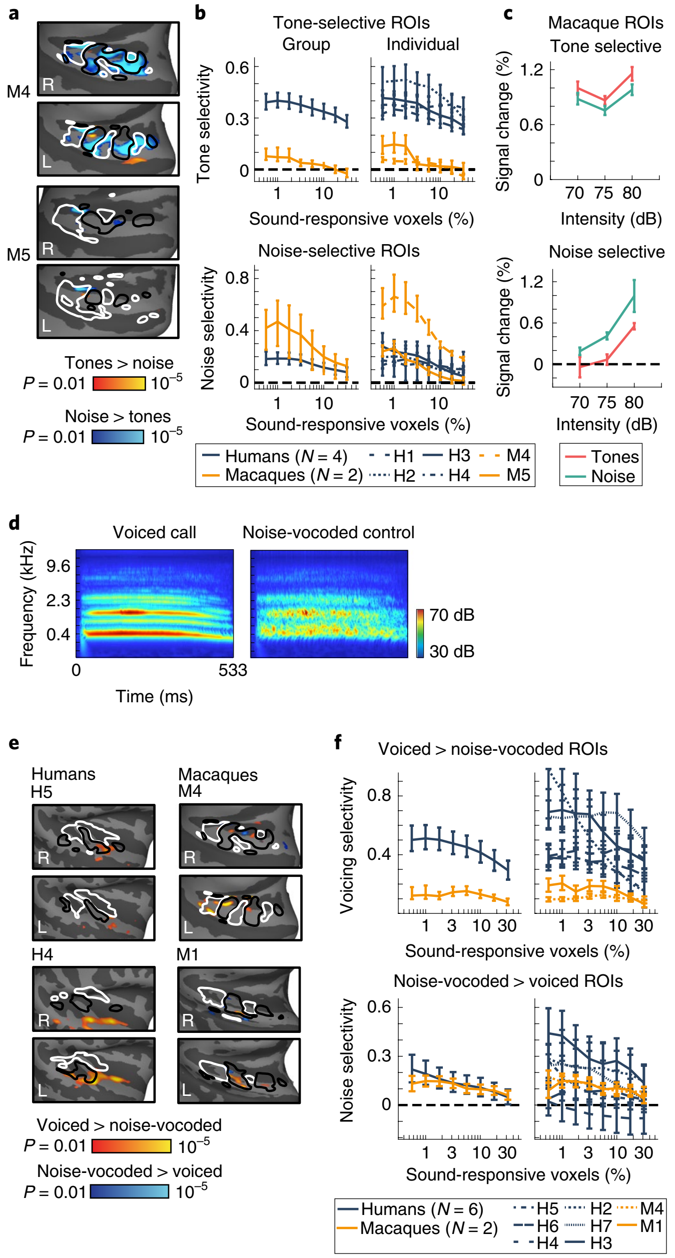

Fig. 3 | Control experiments. a, Experiment IB. Maps of tone versus noise responses averaged across frequency and 3 matched sound intensities (70, 75 and 80 dB) in 2 macaques. Conventions and statistics are the same as in Fig. 1d (number of blocks per tone or noise condition: M4 = 1,395; M5 = 1,380). **b**, ROI analyses for the same tone versus noise contrast. Human data from experiment IA (with non-matched sound intensities) were used for comparison. Conventions and error bars are the same as in Fig. 2f. **c**, ROI responses in macaques broken down by sound intensity for a fixed ROI size (top 1% of sound-driven voxels) (error bars are the same as in **b** and Fig. 2f). **d**, Experiment II. Cochleagrams showing the stimulus conditions: voiced macaque vocalizations, which contain harmonics, and noise-vocoded controls, which lack harmonics but have the same spectrotemporal envelope. **e**, Maps of responses to voiced versus noise-vocoded macaque calls, in 2 humans (H4, H5) and 2 macaque monkeys (M1, M4). Maps plot uncorrected voxel-wise significance values (two-sided $P < 0.01$; see Supplementary Fig. 10 for uncorrected and cluster-corrected maps from all participants). Conventions and statistics are the same as in Fig. 1d (number of blocks per condition being compared: M1 = 288, M4 = 414, H4 = 24, H5 = 22). **f**, ROI analyses for the same voiced versus noise-vocoded contrast. Conventions and error bars the same as in Fig. 2f.

peaks (experiment II). We synthesized noise-vocoded controls by replacing the harmonic frequencies with spectrally shaped noise (Fig. 3d). We note that a preference for voiced versus noise-vocoded calls in macaques could reflect greater familiarity with the voiced stimuli[9,10] rather than a preference for harmonic tones, so this experiment provides a conservative test of whether tone preferences are consistently more selective in humans.

We tested five macaques and six human participants using a range of sound intensities (from 65 to 80 dB). We focus on data from the two macaques with comparable reliability to data from humans, but results were similar using reliability-matched data from all five monkeys (Supplementary Fig. 9). Human participants showed clusters of voxels that responded more strongly to voiced versus noise-vocoded calls of matched sound intensity (Fig. 3e; Supplementary Fig. 10). These clusters had a similar location to the tone-selective voxels identified in experiment IA. Monkeys also showed voxel clusters that responded preferentially to voicing, and these voxels partially overlapped low-frequency tonotopic fields. ROI analyses confirmed the results obtained in both macaques and all human participants (Ps < 0.025 for all but the two largest ROI sizes), but revealed that voice-preferring voxels in macaques were less selective than those in humans (Fig. 3f; Supplementary Fig. 11; Ps < 0.049 for all comparisons between every human participant and both high-reliability macaques for the four smallest ROIs). In contrast, voxels preferentially responsive to noise-vocoded stimuli were similarly selective in humans and macaques (at the group-level, Ps > 0.351 for all ROI sizes between species).

Voice-selective voxels were modulated by sound intensity in macaques (Ps < 0.019 in both monkeys for both tones and noise for all ROI sizes), but not humans (Ps > 0.061 for all participants and ROIs for both tones and noise, except for two ROI sizes from a single participant; Supplementary Fig. 12). These results show that tone selectivity was more pronounced and more intensity-tolerant in humans than macaques, even when assessed with stimuli that are more ecologically relevant to monkeys.

Taken together, these results reveal a species difference in the functional organization of cortical regions implicated in pitch perception. We speculate that the greater sensitivity of the human cortex to harmonic tones is driven in development or evolution by the demands imposed by speech and music perception. While some macaque vocalizations are harmonic or periodic, they are arguably less frequent and varied than human speech or music. Consistent with this hypothesis, humans excel at remembering and discriminating changes in pitch essential to speech and music structure[11], whereas nonhuman primates seem to struggle in this domain[12]. Our results leave open the single-cell basis of the species difference we report. That is, weak voxel selectivity for tones could reflect weak selectivity in individual neurons, or a small fraction of tone-selective neurons within each voxel.

Microelectrode recordings in macaques have not uncovered periodicity-tuned neurons[13,14], which could be related to the weak tone-selective responses we observed. Other nonhuman primates might possess tone-selective regions similar to those present in humans. For example, marmosets show periodicity-tuned neurons that are spatially clustered[15], and are a more vocal species than macaques[16]. Finally, it remains to be seen whether other regions or pathways in the human auditory cortex, such as those selective for speech[17,18] or music[19,20], have counterparts in nonhuman primates[9]. The present results underscore the possibility that the human auditory cortex differs substantially from that of other primates, perhaps because of the centrality of speech and music to human audition.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41593-019-0410-7.

Received: 10 September 2018; Accepted: 19 April 2019;
Published online: 10 June 2019

References

1. Lafer-Sousa, R., Conway, B. R. & Kanwisher, N. G. *J. Neurosci.* **36**, 1682–1697 (2016).
2. Van Essen, D. C. & Glasser, M. F. *Neuron* **99**, 640–663 (2018).
3. de Cheveigné, A. *Oxf. Handb. Audit. Sci. Hear.* **3**, 71 (2010).
4. Patterson, R. D., Uppenkamp, S., Johnsrude, I. S. & Griffiths, T. D. *Neuron* **36**, 767–776 (2002).
5. Penagos, H., Melcher, J. R. & Oxenham, A. J. *J. Neurosci.* **24**, 6810–6815 (2004).
6. Norman-Haignere, S., Kanwisher, N. & McDermott, J. H. *J. Neurosci.* **33**, 19451–19469 (2013).
7. Baumann, S., Petkov, C. I. & Griffiths, T. D. *Front. Syst. Neurosci.* **7**, 11 (2013).
8. Petkov, C. I., Kayser, C., Augath, M. & Logothetis, N. K. *PLoS Biol.* **4**, e215 (2006).
9. Petkov, C. I. et al. *Nat. Neurosci.* **11**, 367–374 (2008).
10. Romanski, L. M. & Averbeck, B. B. *Annu. Rev. Neurosci.* **32**, 315–346 (2009).
11. McPherson, M. J. & McDermott, J. H. *Nat. Hum. Behav.* **2**, 52 (2018).
12. D'Amato, M. R. *Music Percept. Interdiscip. J.* **5**, 453–480 (1988).
13. Schwarz, D. W. & Tomlinson, R. W. *J. Neurophysiol.* **64**, 282–298 (1990).
14. Fishman, Y. I., Micheyl, C. & Steinschneider, M. *J. Neurosci.* **33**, 10312–10323 (2013).
15. Bendor, D. & Wang, X. *Nature* **436**, 1161–1165 (2005).
16. Miller, C. T., Mandel, K. & Wang, X. *Am. J. Primatol.* **72**, 974–980 (2010).
17. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. *Science* **343**, 1006–1010 (2014).
18. Overath, T., McDermott, J. H., Zarate, J. M. & Poeppel, D. *Nat. Neurosci.* **18**, 903–911 (2015).
19. Leaver, A. M. & Rauschecker, J. P. *J. Neurosci.* **30**, 7604–7612 (2010).
20. Norman-Haignere, S. V., Kanwisher, N. G. & McDermott, J. H. *Neuron* **88**, 1281–1296 (2015).

Acknowledgements

The authors thank G. Gagin and K. Bohon for their help in training and scanning animals M1, M2 and M3. They also thank K. Schmidt, D. Yu, T. Haile, S. Eastman and D. Leopold for help in scanning animals M4 and M5. This work was supported by the National Institutes of Health (grant EY13455 to N.K. and grant EY023322 to B.R.C.), the McDonnell Foundation (Scholar Award to J.H.M.), the National Science Foundation (grant 1353571 to B.R.C. and Graduate Research Fellowship to S.N.-H.), the NSF Science and Technology Center for Brains, Minds, and Machines (CCF-1231216) and the Howard Hughes Medical Institute (LSRF Postdoctoral Fellowship to S.N.-H.). The animal work was performed using resources provided by the Neurophysiology Imaging Facility Core (NIMH/NINDS/NEI), as well as the Center for Functional Neuroimaging Technologies at MGH (grant P41EB015896) and a P41 Biotechnology Resource grant supported by the National Institute of Biomedical Imaging and Bioengineering (MGH). The experiments conducted at MGH also involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program and/or High-End Instrumentation Grant Program (grant S10RR021110). The work was also supported by the Intramural Research Program at the NEI, NIMH, and NINDS.

Author contributions

All authors contributed to the experimental design and writing of the paper. S.N.-H. created the stimuli, scanned human participants and performed all analyses. S.N.-H. and B.R.C. scanned macaques in conjunction with the individuals mentioned above.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0410-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.V.N. or B.R.C.

Journal peer review information: *Nature Neuroscience* thanks Frederic Theunissen, Kerry Walker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

Methods

Experiment IA: responses to harmonic tones and noises of different frequencies. *Macaque subjects and surgical procedures.* Three male rhesus macaque monkeys (6–10 kg; 5–7 years old) were scanned. Animals were trained to sit in the sphinx position in a custom-made primate chair. Before scanning, animals were implanted with a plastic head-post under sterile surgical conditions[21]. The animals recovered for 2–3 months before they were acclimated to head restraint through positive behavioral reinforcement (for example, juice rewards). All experimental procedures conformed to local and US National Institutes of Health guidelines and were approved by the Institutional Animal Care and Use Committees of Harvard Medical School, Wellesley College, the Massachusetts Institute of Technology (MIT), and the National Eye Institute.

Human participants. Four human participants were scanned (25–33 years old; 3 male, 1 female; all right-handed; one participant (H3) was the author S.N.-H.). Participants had no formal musical training in the 5 years preceding the scan, and were native English speakers, with self-reported normal hearing. Participants had between 2 and 10 years of daily practice with a musical instrument; however, even participants with no musical experience show robust tone-selective voxels[22]. The study was approved by the Committee on the Use of Humans as Experimental Subjects at the MIT. All participants gave informed consent.

Stimuli and procedure. There were ten stimulus conditions organized as a 2 × 5 factorial design: harmonic tones and Gaussian noise each presented in one of five frequency ranges (Fig. 1a).

Each stimulus was 2 s in duration and contained 6, 8, 10, or 12 notes (note durations were 333, 250, 200, or 166 ms, respectively). Linear ramps (25 ms) were applied to the beginning and end of each note. Notes varied in frequency and F_0 (for harmonic notes) to minimize adaptation (Fig. 1b). We have previously found that such variation enhances the overall response to both tones and noise, but does not affect tone selectivity[6]. Tone-selective voxels respond approximately twice as strongly to harmonic tones versus noise, regardless of whether there is variation in F_0. It is conceivable that humans might show a greater response boost with frequency variation than macaques due to melody-specific processing. However, the fact that we also observed more selective responses to tone stimuli using macaque vocalizations (experiment II) demonstrates that our findings cannot be explained by selectivity for melodic processing.

Stimuli were organized into blocks of ten stimuli from the same condition (Fig. 1b). A single scan was collected during a 1.4-s pause after each stimulus (1-s acquisition time)—separating in time the scan acquisition and the stimulus minimizes the impact of scanner sounds.

For each harmonic note, we sampled a F_0 from a uniform distribution with a ten-semitone range. We constrained the note-to-note change in F_0 to be at least three semitones to ensure the changes would be easily detectable (we discarded F_0 values for which the note-to-note change was below three semitones). For the five frequency ranges tested, the mean of the uniform distribution was 100, 200, 400, 800, and 1,600 Hz. All of these F_0 values are within the range of human pitch perception[23,24]. We expected the pitch range of macaques and humans to be similar because they have a similar audible frequency range[25,26] (only slightly higher in macaques) and are able to resolve low-numbered harmonics like those tested here[14]. Although there is growing evidence to indicate that cochlear frequency selectivity differs across species[27], which might affect the extent to which harmonics are resolved[28], these differences appear to be most pronounced between humans and nonprimates[29], and to be modest between macaques and humans[30].

For harmonic conditions, the F_0 and frequency range co-varied such that the power at each harmonic number remained the same. Since the harmonic number primarily determines resolvability, this procedure ensured that each note would be similarly well resolved[6]. Specifically, we bandpass-filtered (in the frequency domain) a complex tone with a full set of harmonics, with the filter passband spanning the third to the sixth harmonic of each note's F_0 (for example, a note with a 100 Hz F_0 would have a passband of 300–600 Hz). Harmonics outside the passband were attenuated by 75 dB per octave on a logarithmic frequency scale (attenuation was applied individually to each harmonic; the harmonics were then summed). We manipulated the harmonic content of each note via filtering (as opposed to including a fixed number of equal-amplitude components) to avoid sharp spectral boundaries, which might otherwise provide a weak pitch cue[31]. Harmonics were added in negative Schroeder phase to minimize distortion products (DPs)[32].

Noise notes were matched in frequency range to the harmonic notes. For each noise note, wideband Gaussian noise was bandpass-filtered (via multiplication in the frequency domain), with the passband set to three to six times a 'reference' frequency, which was sampled using the same procedure used to select F_0 values.

Noise was also used to mask DPs. DPs would otherwise introduce a confound because our stimuli lacked power at low-numbered harmonics (specifically, the fundamental and second harmonic). For harmonic stimuli, DPs produced by cochlear nonlinearities could reintroduce power at these frequencies[33,34], which could lead to greater responses in regions preferentially responsive to low-frequency power for reasons unrelated to pitch. In addition, the Sensimetric earphones used by us and many other neuroimaging laboratories also produce nontrivial DPs[34]. The masking noise was designed to be ~10 dB above the masked threshold of all cochlear and earphone DPs, which should render the DPs inaudible[34]. Specifically, we used a modified version of threshold-equalizing noise (TEN)[35] that was spectrally shaped to have greater power at frequencies with higher-amplitude DPs. The noise had power between 50 Hz (more than half an octave below the lowest F_0) and 15,000 Hz. Frequencies outside this range were attenuated by 75 dB per octave. Within the noise passband, the target just-detectable amplitudes of the shaped threshold-equalizing noise were determined using previously described procedures[34] and were as follows: 50–60 Hz, 59 dB; 80 Hz, 54 dB; 100 Hz, 51 dB; 120 Hz, 49 dB; 150 Hz, 48 dB; 160 Hz, 43 dB; 200 Hz, 41 dB; 240 Hz, 39 dB; 300 Hz, 34 dB; 400–15,000 Hz, 32 dB. The spectrum of the noise was shaped by interpolating these target values (on a log-frequency scale) and multiplying the spectrum of the TEN noise in the frequency domain (using the fast Fourier transform and inverse fast Fourier transform). Masking noise was present throughout the duration of each stimulus, as well as during the 200 ms gaps between stimuli and scan acquisitions (Fig. 1b).

In macaques, noise stimuli in experiment IA were dichotic, with different random samples of Gaussian noise presented to each ear. The use of dichotic noise was an oversight and was remedied in experiment IB. We tested both diotic and dichotic noise in humans and found that tone-selective responses were very similar regardless of the type of noise used (Supplementary Fig. 13). To make our analyses as similar as possible, we only used responses to the dichotic noise in humans for experiment IA.

Each run included one stimulus block per condition and four silence blocks (all blocks were 34 s). The order of stimulus conditions was pseudorandom and counter-balanced across runs: for each subject, we selected a set of condition orders from a large set of randomly generated orders (100,000), such that on average, each condition was approximately equally likely to occur at each point in the run, and each condition was preceded equally often by every other condition in the experiment. For M1 and M3, the first 50 runs had unique condition orders, after which we began repeating orders. For M2, the first 60 runs were unique. Each run lasted 8 min (141 scan acquisitions) in monkeys and 10.8 min (191 scan acquisitions) in humans (human runs were longer because we tested both diotic and dichotic noise). Humans completed as many runs as could be fit in a single 2-h scanning session (between 7 and 8 runs). Macaques completed 126 (M1), 102 (M2), and 60 (M3) runs across 6 (M1), 5 (M2), and 3 (M3) sessions over a period of 15 months. More data were needed in macaques to achieve comparable response reliability, which was in part due to the smaller voxel sizes and greater motion artifacts (macaques were head-posted but could move their body). We did not perform any a priori power analysis, but instead collected as much macaque data as we could given the constraints on amount of scan time available.

Sounds were presented through the same type of MRI-compatible insert earphones in humans and monkeys (Sensimetric S14). Screw-on earplugs (Comply Canal Tips) were used to attenuate scanner noise; slim plugs were used in macaques to accommodate their smaller ear canal. Earphones were calibrated using a Svantek 979 sound meter attached to a GRAS microphone with an ear and cheek simulator (Type 43-AG). During calibration, the earphone tips with earplugs were inserted directly into the model ear canal.

Animals were reinforced with juice rewards to sit calmly, head-restrained, in the scanner. We monitored arousal by measuring fixation as follows: animals received juice rewards for maintaining fixation within ~1 degree of visual angle of a small circle on an otherwise gray screen. Eye movements were tracked using an infrared eye tracker (ISCAN). Human participants were also asked to passively fixate a central circle, but did not receive any reward or feedback.

For the first three scanning sessions of M1 (six sessions in total) and M2 (five sessions in total), visual stimuli were presented concurrently with the audio stimuli with the goal of simultaneously identifying visually selective regions for a separate experiment. Images of faces, bodies, and vegetables were presented in two scans and color gratings were presented in the third. Visual stimuli were never presented in M3 or in any of the other experiments in this study. Since our results were robust across individuals and experiments, the presence of visual stimuli in those sessions cannot explain our findings.

Human MRI scanning. Human data were collected on a 3 Tesla Siemens Trio scanner with a 32-channel head coil (at the Athinoula A. Martinos Imaging Center of the McGovern Institute for Brain Research at the MIT). Functional volumes were designed to provide good spatial resolution and coverage of the auditory cortex. Each functional volume included 15 slices oriented parallel to the superior temporal plane and covering the portion of the temporal lobe superior to and including the superior temporal sulcus (3.4 s TR, 30 ms TE, 90 degree flip angle; 5 discarded initial acquisitions). Each slice was 4-mm thick with an in-plane resolution of 2.1 × 2.1 mm (96 × 96 matrix, 0.4-mm slice gap). iPAT was used to minimize acquisition time (1 s per acquisition). T1-weighted anatomical images were also collected (1-mm isotropic voxels).

Macaque MRI scanning. Monkey data were also collected on a 3 Tesla Siemens Trio scanner (at the Massachusetts General Hospital Martinos Imaging Center). Images were acquired using a custom-made four-channel receive coil. Functional volumes were similar to those used in humans but had smaller voxel sizes (1-mm isotropic) and more slices (33 slices). Macaque voxels (1 mm^3) were smaller than human voxels (17.4 mm^3), which helped to compensate for their smaller brains[36]. The slices were positioned to cover most of the cortex and to minimize image artifacts caused by field inhomogeneities. The slices always covered the superior temporal plane and gyrus. An AC88 gradient coil insert (Siemens) was used to speed acquisition time and to minimize image distortion. The contrast-enhancing agent MION (monocrystalline iron oxide nanoparticle) was injected into the femoral vein immediately before scanning (8–10 mg per kg of the drug Feraheme, diluted in saline, AMAG Pharmaceuticals). MION enhances fMRI responses, yielding greater percentage signal change values and finer spatial resolution, but the relative response pattern across stimuli and voxels is similar for MION and BOLD (blood oxygenation level-dependent) signals[21,37,38]. MION was not used in humans. Following scanning, the animals received an iron chelator in their home-cage water bottle (deferiprone 50 mg per kg, Ferriprox, ApoPharma) to reduce iron accumulation[39]. T1-weighted anatomical images were also collected for each macaque (0.35-mm isotropic voxels in M1 and M2; 0.5-mm in M3).

Data preprocessing. Human and monkey data were analyzed using the same analysis pipeline and software packages, unless otherwise noted. Functional volumes from each scan were motion-corrected by applying the software MCFLIRT (from FSL) to data concatenated across runs (for one scan in M3, we re-shimmed midway through and treated the data before and after re-shimming as coming from separate scans). Human functional data were then aligned to the anatomical images using a fully automated procedure (FLIRT followed by BBRegister)[40,41]. For macaques, we fine-tuned the initial alignment computed by FLIRT by hand rather than using BBRegister since MION obscures the gray/white-matter boundary upon which BBRegister depends. Manual fine-tuning was performed separately for every scan. Functional volumes were then resampled to the cortical surface (computed by FreeSurfer) and smoothed to improve the signal-to-noise ratio (3-mm full-width at half-maximum kernel in humans; 1-mm full-width at half-maximum in macaques). Results were similar without smoothing. Human data were aligned on the cortical surface to the FsAverage template brain distributed by FreeSurfer. We interpolated the dense surface mesh to a two-dimensional grid (1.5 × 1.5 mm in humans and 0.5 × 0.5 mm in monkeys) to speed up surface-based analyses (grid interpolation was performed using flattened surface maps). Macaque surface reconstructions, using Freesurfer, required manual fine-tuning (described by Freesurfer tutorials) that was not necessary for human anatomical maps, since the automated procedures of the software have been extensively fine-tuned for human data (we also changed the headers of the anatomicals to indicate 1-mm isotropic voxels, thus making the effective brain sizes more similar to human brain sizes).

We excluded runs with obvious image artifacts evident from inspection (one run in M1, five runs in M2, no excluded runs in M3; in M2 four of the five excluded runs were collected with a different phase-encode direction that led to greater artifacts). The run totals mentioned above reflect the amount of data after exclusion.

All analyses were performed in a large constraint region spanning the superior temporal gyrus and plane (defined by hand).

Maps of response contrasts. We contrasted responses to the two lowest and the two highest frequency ranges, averaging across harmonic tones. We also contrasted responses to harmonic tones and noise, averaging across frequency. In humans, tone-selective voxels respond preferentially to tones across a wide range of frequencies, even for high frequencies that only weakly drive responses overall[6]. Thus, averaging across frequency ranges should maximize statistical power.

We computed significance maps for low versus high frequencies and tones versus noise using a standard general linear model (GLM). The design matrix included one regressor for each of the ten conditions in the experiment. Following standard practice, regressors were computed via convolution with a hemodynamic response function (HRF). Because MION inverts and elongates the hemodynamic response relative to BOLD, a different impulse response was used in monkeys and in humans (a finite impulse response model was used to estimate and confirm that our MION HRF was accurate). We used the following HRFs for BOLD and MION:

$$\mathrm{HRF}_{\mathrm{BOLD}}(t) = \left(\frac{t-2.25}{1.25}\right)^{2} e^{-\frac{t-2.25}{1.25}},\ \text{for}\ \ t\ > 2.25\ \text{, otherwise 0} \qquad (1)$$

$$\mathrm{HRF}_{\mathrm{MION}}(t) = -\left(\frac{t}{8}\right)^{0.01} e^{-\frac{t}{8}},\ \text{for}\ \ t > 0 \qquad (2)$$

We included the first ten principal components from white-matter voxels as nuisance regressors in the GLM, similar to standard de-noising techniques[22,42] (using the white-matter segmentation computed by Freesurfer). We found that this procedure improved the test–retest reliability of the estimated responses in macaques. White-matter regressors had little effect on the reliability of human responses, but we included them anyway to make the analysis pipelines as similar as possible. Regression analyses were implemented in MATLAB (using pinv.m) so that we could use a custom permutation test, described below.

For each contrast and voxel, we computed a z-statistic by subtracting the relevant regression beta weights and dividing this difference score by its standard error (estimated using ordinary least squares and fixed effects across runs). We then converted this z-statistic to a measure of significance via a permutation test[22,43]. Specifically, we re-computed the same z-statistic based on 10,000 permuted orderings of blocks (to minimize computation time we used 100 orders per run, rather than 10,000, and for each sample randomly chose one order per run). For each voxel and contrast, we fit the 10,000 z-statistics based on the permuted orders with a Gaussian, and calculated the likelihood of obtaining the observed z-statistic based on the non-permuted condition orders (using a two-sided test). Gaussian fits made it possible to estimate small P values (for example, $P = 10^{-10}$) that would be impossible to approximate by counting the fraction of permuted samples that exceeded the observed statistic.

To correct for multiple comparisons across voxels, we used a variant of cluster-correction suited for the permutation test[22,43]. For each set of permuted condition orders, we transformed the corresponding map of z-statistics into a P value map using the same Gaussian fits described above. We then thresholded this P value map (two-sided $P < 0.01$) and recorded the size of the largest contiguous cluster that exceeded this threshold. Using this approach, we built up a null distribution for cluster sizes across the 10,000 permutations. To evaluate significance, we counted the fraction of times the cluster sizes for this null distribution exceeded that for each observed cluster based on un-permuted orders.

Reliability matching. We believe that our study is the first to compare the selectivity of brain responses between humans and macaques while matching the data reliability. We used the reliability of responses to sounds relative to silence as a measure of data quality (Fig. 2a). First, we estimated the beta weight for each condition in each voxel using two, non-overlapping sets of runs. Second, for each condition we correlated the vector of beta weights across voxels in the superior temporal plane and gyrus for the two datasets. Finally, we averaged the test–retest correlation values across all ten conditions in the experiment. This procedure was performed using different numbers of runs to estimate the response reliability as a function of the amount of data. To estimate the test–retest reliability of the entire dataset, which cannot be measured, we applied the Spearman–Brown correction to the split-half reliability of the complete dataset.

For each human, we selected the number of runs that best matched the reliability of each monkey (using the curves shown in Fig. 2a), subject to the constraint of needing at least two runs per participant (required for the ROI analyses). If the monkey data had higher reliability, we used all of the human runs. The specific runs used for the analysis were randomly selected as part of a bootstrap analysis (see the section “ROI statistics” below).

There is often some variability in the reliability of fMRI voxels across the brain[44]. To assess our sensitivity across brain regions, we calculated the split-half measurement error in the response of each voxel to each condition (Supplementary Fig. 14). Measurement error was calculated as the difference in response between two splits of data (in units of percentage signal change relative to silence). We averaged the absolute value of the error across splits (1,000 random splits) and stimulus conditions (separately for each voxel). For monkeys, we used all of the available data (using half of the runs to compute each split). For humans, we selected the number of runs to match the reliability of the monkey data (as described above). In general, there was no anatomical region in the superior temporal plane and gyrus that had consistently low sensitivity, which suggests that if tone-selective responses were present, we should have been able to detect them.

ROI analysis. We quantified selectivity using ROIs of varying size[22]. Specifically, we selected the top N% of sound-responsive voxels in the auditory cortex (varying N) with the most significant response preference for a given contrast (for example, harmonic tones > noise). Sound-responsive voxels were defined as having a significantly greater average response to all stimulus conditions compared to silence (using a two-tailed, voxel-wise $P < 0.001$ inclusion threshold). We then computed the average response of the selected voxels to each condition using independent data, in units of percentage signal change (computed by dividing the beta weight for each condition/regressor by the mean response of the voxel across time). This analysis was performed iteratively, using one run to measure responses and the remaining run(s) to select voxels (cycling through all runs tested). We used standard ordinary least squares instead of a permutation test to compute the significance values that were then used to select voxels (both for the sound > silence inclusion threshold and to rank-order voxels by the significance of their response preference for a given contrast). We chose not to use a permutation test because the subsampled human datasets did not have many runs, and thus there were not many condition orders to permute. In addition, because we selected the most significantly responsive voxels for a given contrast (after an initial sound > silence screen), the analysis is less sensitive to the absolute significance value of each voxel. Ordinary least squares regression analyses were also implemented in MATLAB. No whitening correction was used since we found that including white-matter regressors substantially whitened the model residuals. Fixed effects was used to pool across runs.

We used a standard metric to quantify selectivity: (preferred – nonpreferred)/(preferred + nonpreferred). This metric is bounded between −1 and 1 for positive-valued responses and is scale-invariant, which is useful because the overall response magnitude of a voxel is influenced by non-neural factors (for example, MION or vascularization). With negative responses, the metric is no longer easily interpretable. We therefore truncated negative values to zero before applying the selectivity metric. Negative responses were rare, occurring, for example, in highly selective tonotopic ROIs for nonpreferred frequencies (see Supplemental Fig. 3, which separately plots responses to low-frequency and high-frequency stimuli). If responses to both conditions being compared are negative, the selectivity metric is undefined. Such instances were rare, and we simply excluded ROIs when this was the case. Specifically, since we applied bootstrapping to our ROI analyses (described below), we excluded bootstrapped samples where responses were negative for both conditions (bootstrapping analysis described below). We also excluded ROIs where more than 1% of bootstrapped samples were negative, which only occurred in a single human participant (H2) for noise-selective ROIs (specifically, we excluded the two smallest ROIs when H2 was matched to M2 and the five smallest ROIs when H2 was matched to M3).

ROI statistics. Bootstrapping was used for all statistics[45]. For individual participants, we bootstrapped across runs, and for group comparisons, we bootstrapped across both participants and runs. For each statistic of interest, we sampled runs or participants with replacement 10,000 times, and recomputed the desired statistic (this procedure is described in more detail below). To compare conditions, the statistic of interest was the difference in beta weights for those conditions (in units of percentage signal change). To compare species, the statistic of interest was the difference in selectivity values. We then used the distribution of each statistic to compute error bars and to evaluate significance. Significance was evaluated by counting the fraction of the times the sampled statistics fell below or above zero (whichever fraction was smaller), and multiplying by two to arrive at a two-sided P value.

Error bars in all graphs show the median and the central 68% of the bootstrapped sampling distribution, which is equivalent to one standard error for normally distributed distributions (we did not use the standard error because it is inappropriate for asymmetric distributions and sensitive to outliers). When plotting responses to individual conditions (Supplementary Figs. 3 and 4), we used "within-subject" error bars[46], computed by subtracting off the mean of each bootstrapped sample across all conditions before measuring the central 68% of the sampling distribution. We multiplied the central 68% interval by the correction factor shown below to account for a downward bias in the standard error induced by mean-subtraction[46]:

$$\sqrt{\frac{N}{N-1}} \tag{3}$$

where N indicates the number of conditions. We did not use within-subject error bars for selectivity values, since they already reflect a difference between conditions.

To bootstrap across runs for one individual, we sampled N 'test' runs with replacement 10,000 times from those available. 'Test' denotes runs used to evaluate the response of a set of voxels after they have been selected based on their response to a non-overlapping set of $N-1$ 'localizer' runs. We averaged ROI responses across the N sampled test runs to compute a single bootstrapped sample. For macaques, N was always equal to the total number of runs. For subsampled human datasets, N was equal to the number of runs needed to match the reliability of one of the monkey datasets (as described in the section "Reliability matching" described above). In cases where N was smaller than the number of runs available, we selected the $N-1$ localizer runs randomly (if a test run was sampled multiple times, we used the same randomly selected localizer runs).

For group analyses, we sampled K participants with replacement from all K participants available, and then for each participant, bootstrapped across runs, as described in the previous paragraph. For each sampled human participant, we also randomly sampled a specific monkey whose reliability we sought to match. The sampled monkey determined the value of N used in the bootstrapping analysis across runs.

Experiment IB: controlling for sound intensity. *Animal scanning and surgical procedures.* Two macaques were scanned (M4 and M5; female, ~7 kg, 8–9 years old) on a 4.7 Tesla Bruker Biospec vertical bore scanner equipped with a Bruker S380 gradient coil at the Neurophysiology Imaging Facility Core (NIMH/NINDS/NEI). Images were acquired using a custom-made four-channel receive coil. Functional volumes were comprised of 27 slices covering the superior temporal plane and gyrus (1.2-mm isotropic voxels). MION was injected into the saphenous vein immediately before scanning (at ~11.8 mg per kg, using ultrasmall superparamagnetic iron oxide nanoparticles produced by the Imaging Probe Development Center for the NIH intramural program). T1-weighted anatomical images were also collected (0.5-mm isotropic voxels).

Stimuli and procedure. Stimuli were the same as those for experiment IA but with two modifications. First, tone and noise stimuli were played at three sound intensities (70, 75, and 80 dB), and second, the noise was diotic rather than dichotic. Because of the large number of conditions (30 conditions: 5 frequencies × 3 intensities × 2 stimulus types (tones/noise)), we separated sounds with different intensities into different runs (i.e., run 1: 70 dB, run 2: 75 dB, and so on). For analysis purposes, we concatenated data across each set of three consecutive runs. Three boxcar nuisance regressors were included in the GLM to account for run effects (each boxcar regressor consisted of ones and zeros with ones indicating the samples from one of the three concatenated runs; these run regressors were partialled out from white-matter voxel responses before computing principal components). A large amount of data was collected: 279 runs in M4 (18 scanning sessions), and 276 runs in M5 across (17 scanning sessions). Each run lasted 8 min. Scanning took place over a ~2.5-month period.

Data preprocessing. Data were analyzed using the same pipeline as for experiment IA, with one minor difference: manual alignment was not done separately for each scan. Instead, functional data from all scans (after motion correction within a scan) of a given monkey were aligned to the middle functional scan using FLIRT. The middle functional scan was then aligned to the anatomical scan using FLIRT followed by hand-tuning. We chose this approach because of the large number of scans, and because the scan-to-scan functional alignment was of high quality.

Data from one scan session (in M5) were discarded because MION was not properly administered, which was obvious from inspection of the image. Another scan (in M4) was discarded because not enough images were acquired per run. As noted above, we analyzed runs in sets of three, with one run per intensity. We excluded five runs (four in M4, one in M5) because we did not complete a full cycle of three runs. Six runs were excluded (three in M4, three in M5) because they were repeated unintentionally (that is, exactly the same stimuli and stimulus orders as one of the other runs). The run and scan session totals mentioned above are post-exclusion.

Reliability matching. We used human responses to diotic noise from experiment IA for comparison with the monkey data from this experiment. To compare reliability, we averaged responses across the three intensities to make the dataset comparable to the human dataset where only a single intensity was tested. Monkey data were again less reliable per run (Supplementary Fig. 6), but cumulatively, monkey data were slightly more reliable (we collected ~35 times more data in monkeys). Human data were therefore not subsampled.

ROI statistics. We used the same ROI analyses described in experiment IA, averaging across the three intensities tested when identifying frequency and tone- and noise-selective voxels. We again used bootstrapping to test for significant differences between conditions and species. To assess the effect of sound intensity, we used a bootstrapping procedure analogous to a one-way analysis of variance. Specifically, we computed the variance across intensities in the response of each ROI (averaging across the other stimulus factors; that is, frequency and tone or noise), and compared this value with an estimate of the variance under the null, which assumes that there are no differences in the mean response across sound intensities. We used bootstrapping to estimate the null by measuring the variance of each bootstrapped sample across intensities after subtracting off the mean of the bootstrapped samples for each condition.

For each ROI, we compared the magnitude of intensity-driven changes with the selectivity of the ROI for the stimulus contrast used to define it (that is, tones versus noise or low versus high frequencies). For tone-selective ROIs, we measured the response to tone and noise stimuli averaged across frequency for each of the four sound intensities tested. To assess the magnitude of intensity-driven changes, we calculated the response difference between all pairs of sound intensities separately for tones and noises, and averaged the magnitude of these difference scores. To assess the magnitude of the tone versus noise difference, we computed the difference between responses to tones and noises separately for each sound intensity, and averaged the magnitude of these difference scores across intensity. We then subtracted the resulting difference scores for intensity and the tone versus noise comparison, and used bootstrapping to test for a significant difference from zero (indicating a greater effect of intensity or tones versus noise). The same procedure was used to compare the effect of intensity in frequency-selective ROIs, but we used responses to low and high frequency stimuli of different intensities (averaged across tones and noise).

Experiment II: responses to voiced and noise-vocoded macaque vocalizations. *Human participants.* Six participants were scanned (ages were 19, 22, 26, 27, 28, and 37; 5 male, 1 female; all right-handed); three of these participants (H2, H3, and H4) also participated in experiment I.

Animals tested. All five macaques tested in experiments IA and IB were tested in this experiment.

Stimuli. We selected 27 voiced macaque calls (from a collection of 315 previously recorded calls[47]) that were (1) periodic (autocorrelation peak height >0.9, as measured using the software Praat[48]), (2) longer than 200 ms in duration (since very short sounds produce a weaker pitch percept[49]), and (3) had F_0 values below 2 kHz (since very high F_0 values produce a weaker pitch percept[50]). The selected calls ranged in duration from 230 ms to 785 ms (median of 455 ms). Vocalizations were high-pass filtered with a 200 Hz cut-off to remove low-frequency noise present in some recordings (second-order Butterworth filter; the 200 Hz cut-off was above the lowest F_0, which was 229 Hz). Stimuli were downsampled from 50 kHz to 40 kHz to remove frequencies above the range of human hearing (macaques have slightly higher audible frequency ranges[25,26]). Linear ramps (30 ms) were applied to the beginning and end of each vocalization. Vocalizations were RMS normalized.

We used the vocoder TANDEM-STRAIGHT to create noise versions of each vocalization by replacing the periodic excitation with a noise excitation[51–53]. To control for minor artifacts of the synthesis algorithm, we used the same algorithm to synthesize voiced versions of each vocalization (using harmonic and periodic excitation). We made two small changes to the published TANDEM-STRAIGHT algorithm[51–53]. First, we used F_0 values computed in Praat[48], which we found were more accurate for macaque vocalizations (TANDEM-STRAIGHT's F_0 tracker is tailored to human speech). Second, for noise-vocoded stimuli, we prevented the algorithm from generating power below the F_0, which would otherwise cause the noise-vocoded stimuli to have greater power at low frequencies. This change was implemented by attenuating frequencies below the F_0 on a frame-by-frame basis based on their distance to the F_0 on a logarithmic scale (75 dB per octave). This attenuation was applied to the spectrotemporal envelope computed by TANDEM-STRAIGHT, and was only applied to frames that were voiced in the original signal (as determined by TANDEM-STRAIGHT; the same attenuation was also applied to the spectrotemporal envelope of the harmonically vocoded stimuli, although the effect of this was minimal since the harmonic excitation had little power below the F_0).

We did not use DP masking noise because vocalizations already have power at low-numbered harmonics. Since DPs have much lower amplitude than stimulus frequency components[33,34], the effect of DPs should be minimal for stimuli with power at low-numbered harmonics.

We created 2-s stimuli by concatenating individual harmonic and noise-vocoded vocalizations. The stimulus set was organized into sets of eight stimuli. Each set included all 27 vocalizations presented once in random order. We created each eight-stimulus set by first stringing together all 27 calls into a longer 16-s stimulus, and then subdividing this longer stimulus into 2-s segments. The average interstimulus interval (ISI) between vocalizations was 142 ms; ISIs were jittered by 40% (the mean ISI was chosen to make the total duration of each eight-stimulus set exactly 16 s). Before dividing the 16-s stimulus into 2-s stimuli, we checked that the cuts did not subdivide individual vocalizations. If they did, we discarded the 16-s stimulus and generated a new one, using a different random ordering of calls and a different jittering of ISIs. We repeated this process to create a large number of 2-s stimuli (1,800 per condition). We used the same ordering and ISIs for voiced and noise-vocoded calls.

We used the same block design described in experiment I (each block included ten 2-s stimuli). New stimuli were presented until all 1,800 stimuli were used, after which we started over. In humans and two monkeys (M4 and M5), the harmonic and noise-vocoded stimuli were each presented at four different sound intensities (65, 70, 75, and 80 dB), yielding 8 conditions in total. For three monkeys (M1, M2, and M3), we only tested three sound intensities per condition and used slightly higher intensities for the harmonic conditions (70, 75, and 80 dB) than the noise conditions (65, 70, and 75 dB) to maximize our chance of detecting tone-selective responses. When combining data across the two designs, we analyzed the matched intensities that were common to both: 70 and 75 dB. For the eight-condition scans (humans, M4, and M5), each run included one block per condition and two blocks of silence. For the six-condition scans (M1, M2, and M3), each run included two blocks per condition and three blocks of silence. Macaques completed 72 runs (M1, two sessions), 30 runs (M2, one session), 67 runs (M3, two sessions), 207 runs (M4, nine sessions over ~1 month), and 35 runs (M5, two sessions). Humans completed 11–12 runs across a single scanning session.

For M1, M2, and M3, we used a different set of earphones to present sounds (STAX SR-003; MR-safe version). STAX earphones and Sensimetrics earphones (used in all other animals and experiments) have different strengths and weaknesses. STAX earphones have less distortion than Sensimetrics earphones[34], and, unlike Sensimetrics, they rest outside the ear canal, which avoids the need to insert an earphone or earplug into the small ear canal of macaques. Sensimetrics earphones provide better sound attenuation due to the use of a screw on earplug (sound attenuating putty was placed around the STAX earphones; specifically, Insta-Putty produced by Insta Molds), and, as a consequence, rest more securely in the ears of the macaques. We observed similar results across animals tested with STAX and Sensimetrics earphones, demonstrating our results are robust to the type of earphone used.

Data acquisition, preprocessing, and analysis. The data collection, preprocessing, and analysis steps were the same as those described in experiment I. As in experiment IB, we only did manual alignment of functionals to anatomicals once per animal, rather than once per scan as in experiment IA.

One scan from M2 was excluded because of large amounts of motion, which resulted in weak or nonsignificant sound-driven responses. One run in M3 was discarded due to image artifacts that produced a prominent grating pattern in the images. The run totals mentioned above are post-exclusion.

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

Data availability

Data are available at the following repository: https://neicommons.nei.nih.gov/#/toneselectivity. We are releasing raw scan data (formatted as NIFTIs), anatomicals and corresponding Freesurfer reconstructions, preprocessed surface data, and timing information indicating the onset of each stimulus block. We also provide the underlying data for all statistical contrast maps and ROI analyses (that is, all data figures) for Figs. 1c,d, 2c–f and 3a–c,e,f and Supplementary Figs. 1–5, 7, 8, 9c,d and 10–13.

Code availability

Our custom MATLAB code mainly consists of wrappers around other FSL and Freesurfer software commands. MATLAB routines are available at https://github.com/snormanhaignere/fmri-analysis. The commit corresponding to the state of the code at the time of publication is tagged as HumanMacaque-NatureNeuro.

References

21. Lafer-Sousa, R. & Conway, B. R. *Nat. Neurosci.* **16**, 1870–1878 (2013).
22. Norman-Haignere, S. V. et al. *J. Neurosci.* **36**, 2986–2994 (2016).
23. Semal, C. & Demany, L. *Music Percept. Interdiscip. J.* **8**, 165–175 (1990).
24. Pressnitzer, D., Patterson, R. D. & Krumbholz, K. *J. Acoust. Soc. Am.* **109**, 2074–2084 (2001).
25. Pfingst, B. E., Laycock, J., Flammino, F., Lonsbury-Martin, B. & Martin, G. *Hear. Res.* **1**, 43–47 (1978).
26. Heffner, R. S. *Anat. Rec. A. Discov. Mol. Cell. Evol. Biol.* **281A**, 1111–1122 (2004).
27. Shera, C. A., Guinan, J. J. & Oxenham, A. J. *Proc. Natl Acad. Sci. USA* **99**, 3318–3323 (2002).
28. Walker, K. M., Gonzalez, R., Kang, J. Z., McDermott, J. H. & King, A. J. *eLife* **8**, e41626 (2019).
29. Sumner, C. J. et al. *Proc. Natl Acad. Sci. USA* **115**, 11322–11326 (2018).
30. Joris, P. X. et al. *Proc. Natl Acad. Sci. USA* **108**, 17516–17520 (2011).
31. Small, A. M. Jr & Daniloff, R. G. *J. Acoust. Soc. Am.* **41**, 506–512 (1967).
32. Schroeder, M. *Inf. Theory IEEE Trans.* **16**, 85–89 (1970).
33. Pressnitzer, D. & Patterson, R. D. In *Proc. 12th International Symposium on Hearing* (eds Breebaart, D. J. et al.) 97–104 (Shaker, 2001).
34. Norman-Haignere, S. & McDermott, J. H. *Neuroimage* **129**, 401–413 (2016).
35. Moore, B. C. J., Huss, M., Vickers, D. A., Glasberg, B. R. & Alcántara, J. I. *Br. J. Audiol.* **34**, 205–224 (2000).
36. Herculano-Houzel, S. *Front. Hum. Neurosci.* **3**, 31 (2009).
37. Leite, F. P. et al. *Neuroimage* **16**, 283–294 (2002).
38. Zhao, F., Wang, P., Hendrich, K., Ugurbil, K. & Kim, S.-G. *Neuroimage* **30**, 1149–1160 (2006).
39. Gagin, G., Bohon, K., Connelly, J. & Conway, B. fMRI signal dropout in rhesus macaque monkey due to chronic contrast agent administration. https://www.abstractsonline.com/Plan/ViewAbstract.aspx?sKey=c1451d63-ca65-4a44-afcc-ce1132062d6e&cKey=efbbc764-4eda-4422-9f70-f6d03b2e2eed&mKey=54c85d94-6d69-4b09-afaa-502c0e680ca7 (Society for Neuroscience, 2014).
40. Jenkinson, M. & Smith, S. *Med. Image Anal.* **5**, 143–156 (2001).
41. Greve, D. N. & Fischl, B. *Neuroimage* **48**, 63 (2009).
42. Kay, K., Rokem, A., Winawer, J., Dougherty, R. & Wandell, B. *Front. Neurosci.* **7**, 247 (2013).
43. Nichols, T. E. & Holmes, A. P. *Hum. Brain Mapp.* **15**, 1–25 (2002).
44. Triantafyllou, C., Polimeni, J. R. & Wald, L. L. *Neuroimage* **55**, 597–606 (2011).
45. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans* (SIAM, 1982).
46. Loftus, G. R. & Masson, M. E. *Psychon. Bull. Rev.* **1**, 476–490 (1994).
47. Hauser, M. D. *Anim. Behav.* **55**, 1647–1658 (1998).
48. Boersma, P. Praat, a system for doing phonetics by computer. *Glot International 5* http://dare.uva.nl/search?arno.record.id=109185 (2002).
49. Gockel, H. E., Moore, B. C. J., Carlyon, R. P. & Plack, C. J. *J. Acoust. Soc. Am.* **121**, 373–382 (2007).
50. Oxenham, A. J., Micheyl, C., Keebler, M. V., Loper, A. & Santurette, S. *Proc. Natl Acad. Sci. USA* **108**, 7629–7634 (2011).
51. Kawahara, H. & Morise, M. *Sadhana* **36**, 713–727 (2011).
52. McDermott, J. H., Ellis, D. P. & Kawahara, H. In *Proc. SAPA-SCALE Conference* (Citeseer, 2012).
53. Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H. & McDermott, J. H.*Nat. Commun.* **9**, 2122 (2018).

nature research

Corresponding author(s): Samuel Norman-Haignere

Last updated by author(s): Mar 23, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Software distributed by Siemens was used for MRI data collection. Stimuli were presented using MATLAB and PsychToolbox (several different versions were used over the several years data was collected; the versions are not relevant for understanding or replicating our results).
Data analysis	Our custom MATLAB code mainly consists of wrappers around other FSL (5.0) and Freesurfer (5.3.0) software commands. MATLAB routines are available here: https://github.com/snormanhaignere/fmri-analysis. The commit corresponding to the state of the code at the time of publication is tagged as HumanMacaque-NatureNeuro.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is available on the following repository:

https://neicommons.nei.nih.gov/#/toneselectivity

We are releasing raw scan data (formatted as NIFTIs), anatomicals and corresponding Freesurfer reconstructions, preprocessed surface data, and timing information indicating the onset of each stimulus block. We also provide the underlying data for all statistical contrast maps and ROI analyses (i.e. all data figures):

Fig 1c-d, 2c-f, 3a-c, 3e-f, S1-S5, S7-S8, S9c-d, S10-S13.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Five macaque were tested in this study. This is a large sample size for macaque research, and was chosen to ensure that our results were robust across animals. Seven humans were tested and we went to considerable lengths to ensure the human and macaque data were matched in reliability. Humans completed as many runs as could be fit in a single 2-hour scanning session (between 7 and 8 runs). Macaques completed 126 (M1), 102 (M2), and 60 (M3) runs across 6 (M1), 5 (M2), and 3 (M3) sessions over a period of 15 months. More data were needed in macaques to achieve comparable response reliability, in part due to the smaller voxel sizes and greater motion artifacts (macaques were head-posted but could move their body). We did not perform any a priori power analysis, but instead collected as much macaque data as we could given the constraints (e.g. amount of scan time available).
Data exclusions	In each experiment, under "Data preprocessing", we indicate the exclusions: Experiment IA: We excluded runs with obvious image artifacts evident from inspection (one run in M1, five runs in M2; no excluded runs in M3; in M2 four of the five excluded runs were collected with a different phase-encode direction that led to greater artifacts). The run totals mentioned above reflect the amount of data after exclusion. Experiment IB: Data from one scan session (in M5) was discarded because MION was not properly administered which was obvious from inspection of the image. Another scan (in M4) was discarded because not enough images were acquired per run. As noted above, we analyzed runs in sets of three, with one run per intensity. We excluded five runs (four in M4, one in M5) because we did not complete a full cycle of three runs (e.g. only tested 70 dB but not 75 and 80 dB). Six runs were excluded (three in M4, three in M5) because they were repeated unintentionally (i.e. exactly the same stimuli and stimulus orders as one of the other runs). The run/scan session totals mentioned above are post-exclusion. Experiment II: One scan from M2 was excluded because of large amounts of motion which resulted in weak/insignificant sound-driven responses. One run in M3 was discarded due to image artifacts that produced a prominent grating pattern in the images. The run totals mentioned above are post-exclusion.
Replication	Experiment IB is a replication a of Experiment IA that controls for a possible confound. Experiment II replicates and extends the key findings from Experiment I, but uses ecologically relevant macaque vocalizations. The human findings with tone and noise stimuli are replications of our prior work (Norman-Haignere et al., 2013, 2016).
Randomization	The order of stimulus conditions was pseudorandom and counter-balanced across runs: for each subject, we selected a set of condition orders from a large set of randomly generated orders (100,000), such that on average each condition was approximately equally likely to occur at each point in the run and each condition was preceded equally often by every other condition in the experiment.
Blinding	This is not relevant to our study. Stimuli were not presented by a person, but by a computer program. It is not clear what it means to analyze the data blind to the conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☐	☒	MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	5 rhesus macaques; 3 male, 2 female; M1 and M2 were 5-7 at the time of scanning. M3 was 6 years old. M4 was 8 and M5 was 9.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field samples.
Ethics oversight	All experimental procedures conformed to local and US National Institutes of Health guidelines and were approved by the Institutional Animal Care and Use Committees of Harvard Medical School, Wellesley College, Massachusetts Institute of Technology, and the National Eye Institute.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Experiment IA: 4 subjects; ages 25-33; 3 male, 1 female; all right-handed; one subject (H3) was author SNH Experiment II: 6 subjects; ages 19, 22, 26, 27, 28, 37; 5 male, 1 female; all right-handed; three subjects (H2,H3,H4) scanned in both Experiments I & II All subjects had no formal musical training in the 5 years preceding the scan, and were native English speakers, with self-reported normal hearing.
Recruitment	The focus of this paper was on testing macaque responses to pitch, and we tested a relatively large number of macaques compared with most studies (5 macaques). Human subjects were primarily recruited from within the lab/department. We always emphasized that participation is completely voluntary and they were compensated for their time. The human findings using tone and noise stimuli are replications of prior work, where we have tested a wider range of subjects (Norman-Haignere et al., 2013, 2016).
Ethics oversight	The study was approved by the Committee On the Use of Humans as Experimental Subjects at MIT. All subjects gave informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Block design with sparse scanning.
Design specifications	This information is detailed in the Methods section.
Behavioral performance measures	No behavioral performance is reported.

Acquisition

Imaging type(s)	fMRI
Field strength	3T and 4.7T as specified in the manuscript.
Sequence & imaging parameters	This information is specified in detail in the manuscript.
Area of acquisition	The functional volumes were designed to provide good spatial resolution in and coverage of the superior temporal plane and gyrus.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Human and monkey data were analyzed using the same analysis pipeline and software packages, unless otherwise noted. Functional volumes from each scan were motion corrected by applying FSL's MCFLIRT software to data concatenated across runs (for one scan in M3, we re-shimmed midway through and treated the data before and after re-shimming as coming from separate scans). Human functional data was then aligned to the anatomical images using a fully automated procedure (FLIRT followed by BBRegister). For macaques, we fine-tuned the initial alignment computed by FLIRT by hand, rather than using BBRegister since MION obscures the gray/white-matter boundary upon which BBRegister depends. Manual fine-tuning was performed separately for every scan. Functional volumes were then

	resampled to the cortical surface (computed by FreeSurfer) and smoothed to improve SNR (3 mm FWHM kernel in humans; 1 mm FWHM in macaques). Results were similar without smoothing. Human data were aligned on the cortical surface to the FsAverage template brain distributed by FreeSurfer. We interpolated the dense surface mesh to a 2-dimensional grid (1.5 x 1.5 mm in humans, and 0.5 x 0.5 mm in monkeys) to speed-up surface-based analyses (grid interpolation was performed using flattened surface maps). Macaque surface reconstructions, using Freesurfer, required manual fine-tuning (described by Freesurfer tutorials) that was not necessary for human anatomicals, since the software's automated procedures have been extensively fine-tuned for human data (we also changed the headers of the anatomicals to indicate 1 mm isotropic voxels, thus making the effective brain sizes more similar to human brain sizes).
Normalization	Human data were normalized to the FsAverage template brain distributed by Freesurfer using surface-based alignment. Monkey data was not aligned to any standard template.
Normalization template	FsAverage for humans.
Noise and artifact removal	Principal components from white-matter, as noted in the Methods.
Volume censoring	None

Statistical modeling & inference

Model type and settings	Our key statistical analyses are based on ROI analyses. Significance was evaluated with bootstrapping. For whole-brain maps, we report both uncorrected and corrected maps (using a cluster-corrected permutation test). Uncorrected maps are shown because one of our key results is that monkeys show weak or absent tone-selective responses.
Effect(s) tested	Our key stats rely are contrasts between two conditions or between species, evaluated with ROI analyses and bootstrapping. When analyzing the effect of sound intensity, we used a bootstrapping analyses analogous to a 1-way ANOVA. The details are described in the Methods.

Specify type of analysis: □ Whole brain □ ROI-based ☒ Both

Anatomical location(s)	We used functional ROIs, identified and tested using independent data. There was a large anatomical constraint region that spanned the superior temporal plane and gyrus.

Statistic type for inference (See Eklund et al. 2016)	We report both uncorrected and corrected maps. Cluster-correction was implemented using a permutation test, and is described in the Methods.
Correction	Cluster-correction via a permutation test.

Models & analysis

n/a	Involved in the study
☒	□ Functional and/or effective connectivity
☒	□ Graph analysis
☒	□ Multivariate modeling or predictive analysis