Behavioral origin of sound-evoked activity in visual cortex

Célian Bimbard*, Timothy PH Sit[&], Anna Lebedeva[&], Kenneth D Harris, and Matteo Carandini

University College London, London, United Kingdom

*Correspondence: c.bimbard@ucl.ac.uk
[&]These authors contributed equally.

Abstract

Sensory cortices are increasingly thought to encode multisensory information. For instance, primary visual cortex (V1) appears to be influenced by sounds. Here we show that sound-evoked responses in mouse V1 are low-dimensional, similar across neurons and across brains, and can be explained by highly stereotyped uninstructed movements of eyes and body. Thus, neural activity previously interpreted as being sensory or multisensory may have a behavioral origin.

Introduction

Many studies suggest that all of sensory cortex, including primary sensory areas, is multisensory[1]. For instance, mouse primary visual cortex (V1) appears to be influenced by auditory signals (Figure 1, top), which may provide global inhibition[2], modify the neurons' orientation tuning[3,4], boost detection of visual events[5], or even provide tone-specific information, reinforced by prolonged exposure[6] or training[7]. These effects may be due to projections from the auditory system to the visual cortex[3,5,7].

However, there is a possible alternative explanation for these apparent multisensory signals. Sounds can change internal state and evoke uninstructed body movements[8–11]. Internal state and body movements correlate with activity in many brain regions[12–14], including V1[15–17]. It is thus possible that sounds affect visual cortex because they change internal state or behavior (Figure 1, bottom).

To test this possibility, we asked whether sound-evoked signals in mouse V1 could be predicted by uninstructed sound-evoked movements. We recorded the responses of hundreds of V1 neurons to audiovisual stimuli, while filming the mouse. We observed that V1 encoded a low-dimensional representation of sounds, which was tightly correlated to the movements evoked by these sounds. Different sounds evoked different temporal patterns of movement, which were stereotyped across trials and across mice, and these movements could predict the responses of V1 neurons to sounds. Thus, the multisensory activity that has been widely observed across the brain may have a simpler, behavioral origin.

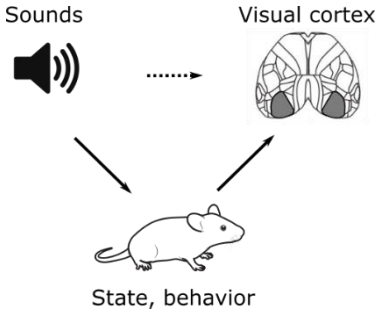


Figure 1. **Explaining auditory effects in visual cortex.** Top. Activity in visual cortex is influenced by sounds. Bottom. Our proposal: sounds influence visual cortex by evoking changes in internal state and behavior.

Results

To explore the influence of sounds on V1 activity, we implanted Neuropixels 1.0 and 2.0 probes[18,19] in 8 mice, and recorded during head fixation while playing artificial and naturalistic audiovisual stimuli (Figure 2a). Here, we illustrate the results with the naturalistic stimuli (results with artificial stimuli were similar; data not shown). We selected eleven 4 s naturalistic movie clips[20], each made of a video (gray-scaled) and a sound (loudness 50-80 dB SPL, Suppl. Figure 1), together with a blank movie (gray screen, no sound). On each trial, we presented a combination of the sound from one clip and the video from another, and repeated each of the 144 combinations 4 times, in random order.

Sounds evoke stereotyped responses in visual cortex

We then identified the visual and auditory components of each neuron's sensory response (Figure 2a-d). A typical V1 neuron responded differently to different combinations of videos and sounds (Figure 2a). To characterize these responses, we used a marginalization procedure similar to factorial ANOVA (see Methods). To measure a neuron's video-related responses (Figure 2b) we computed its mean response to each video (averaged across all concurrent sounds) and subtracted the grand average over all videos and sounds (Figure 2d). Similarly, to characterize the neuron's sound-related responses (Figure 2c) we computed the mean response to each sound (averaged across all concurrent videos) and subtracted the grand average.

Sounds evoked stereotyped, mostly one-dimensional responses in V1 neurons (Figure 2c,e-g). For instance, the time courses of sound responses in two additional example neurons were very similar to that of our first example neuron (Figure 2e). The population activity evoked by sounds was close to one-dimensional. Indeed, applying cross-validated Principal Component Analysis (cvPCA, Ref. [21]) to the sound-related activity of the population (69 neurons in this example animal) revealed that a single dimension explained 54 ± 3% (s.e., n=8 mice) of the sound-related variance (Figure 2f). The time course of population activity projected onto this dimension ("auditory PC1") differed between sounds, but was similar to the responses evoked in individual neurons (Figure 2g). Because this first component captured most of sound-evoked activity in V1, we use its time course to illustrate auditory responses throughout the paper.

The population responses evoked by sounds were also similar across brains (Figure 2h-j). In all mice, the representation of sounds was largely one-dimensional (Figure 2i) and its dependence on sounds was similar across mice (Figure 2j). Indeed, the correlation of auditory PC1 timecourses evoked in different mice was 0.34, close to the test-retest correlation of 0.43 measured within individual mice (Figure 2h). Thus, sounds evoke essentially one-dimensional population activity, which follows a similar time course even across brains.

In contrast, videos elicited responses that were both larger and high-dimensional. Applying the same analysis methods showed that the first visual PC explained more total variance than the first auditory PC (17.3 ± 1.4% vs 1.7% ± 0.3, s.e., n=8 mice; Suppl. Figure 2). Furthermore, higher visual PCs explained substantial amounts of variance, as previously reported[21], while higher auditory PCs did not. This higher-dimensional visual code allowed better decoding of stimulus identity: the accuracy of a template-matching decoder was 94 ± 2% (s.e., n=8 mice) for videos but only 18 ± 2% (s.e., n=8 mice) for sounds, much lower ($p = 0.0313$, paired Wilcoxon sign rank test) yet still significantly above chance ($p = 0.0078$, Wilcoxon sign rank test, Figure 2k).

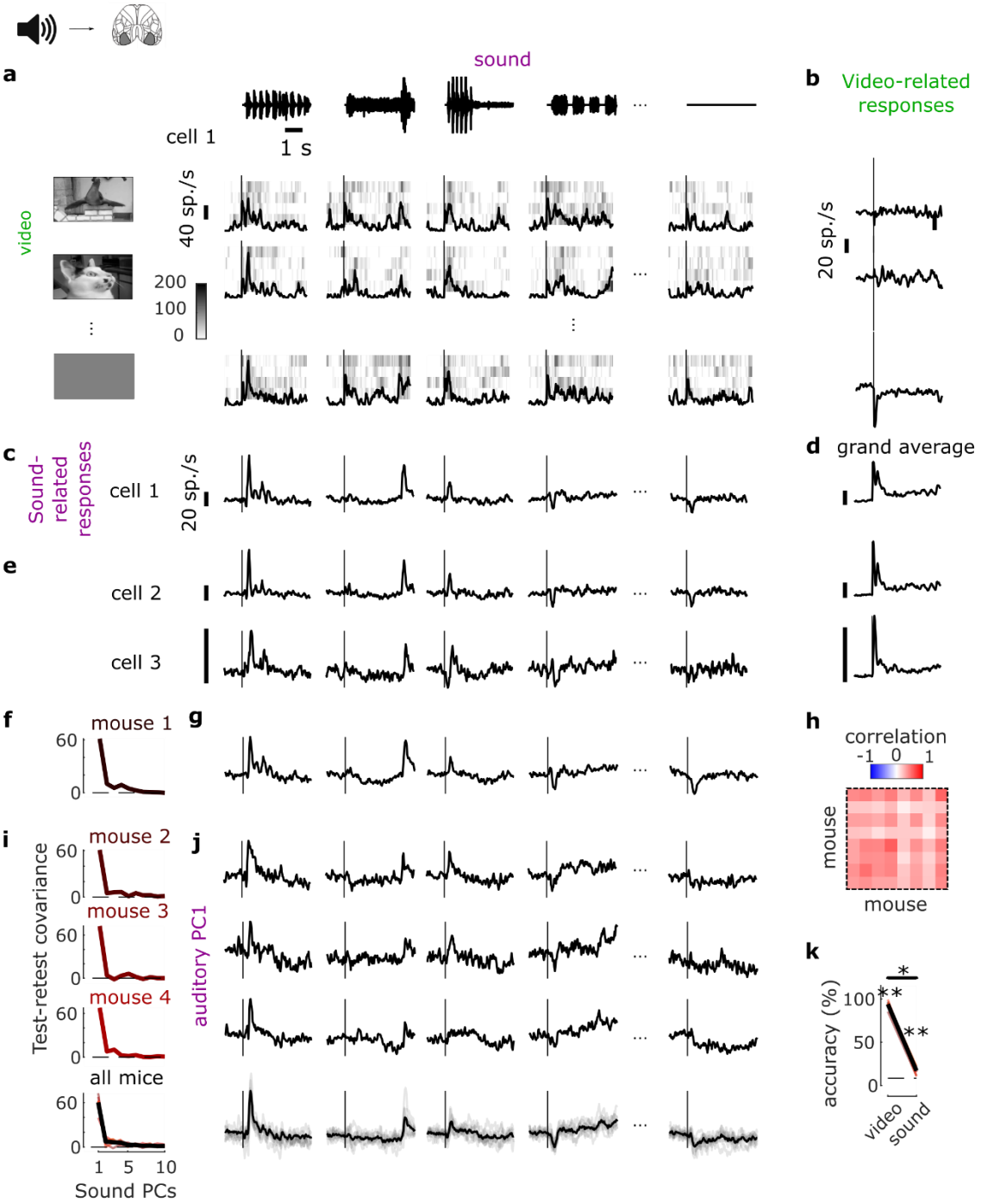


Figure 2. **Sounds evoke stereotyped responses in visual cortex. a**. Responses of an example neuron (cell 1) to combinations of sounds (*columns*) and videos (*rows*). Rasters (*grayscale images*) show single trial responses for each sound-video combination. Curves show the average response over 4 repeats. **b.** Video–related time courses (averaged over all repeats and sound conditions, relative to grand average) for the example cell in **a**. **c.** Sound-related time courses for the same cell, relative to grand average. **d.** Grand average over all conditions for the cell (same scale bars as in **b**,**c**). **e.** Same as **c**-**d** for two other cells. **f.** Cross-validated principal components analysis (cvPCA) of the sound-related population responses for the example mouse shows that sound-related population responses are essentially one-dimensional, i.e. captured by a single time course. **g.** Time courses of the sound-evoked responses along the first auditory dimension ('auditory PC1', arbitrary units) for the example mouse. **h.** Test-retest correlation of the auditory PC1 time courses, within and between all 8 mice. **i,j.** Same as **f** and **g** for three other mice, and for all mice (average in black, each individual mouse in gray). **k.** Decoding accuracy for both video and sound decoding (left) for all 8 mice (colors from brown to orange) and their average (thick black line) (*: p-value < 0.05, **: p-value < 0.01).

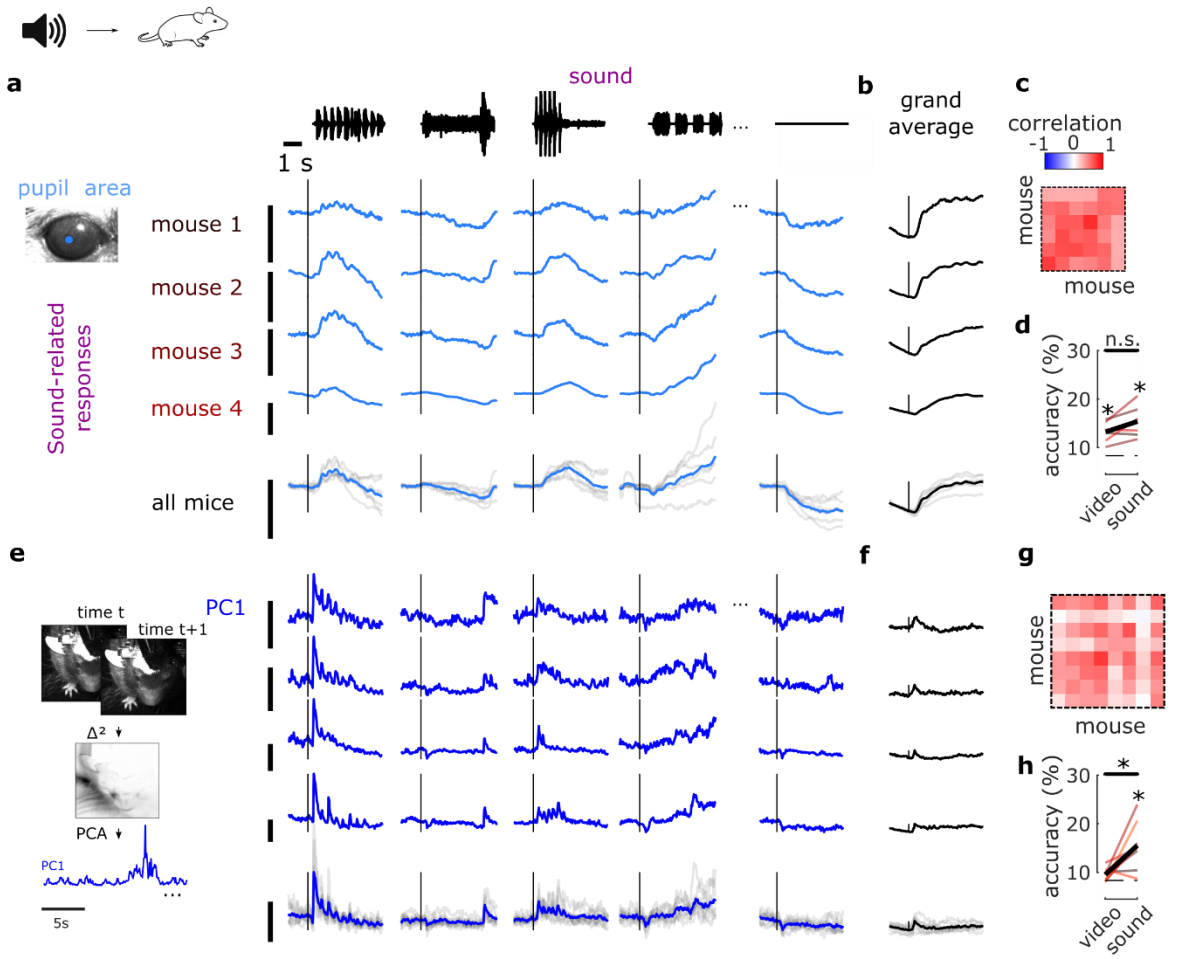


Figure 3. **Sounds evoke stereotyped, uninstructed behavioral responses. a.** Sounds evoked changes in pupil area, whose time course was different across sounds (columns) but similar across mice (rows). Last row shows average across 6 out of 8 mice that were monitored with an eye camera (*inset*) (scale bar: 1 s.d.). **b.** The grand average across sounds was an increase in pupil area (the time courses in **a** are deviations from this grand average). **c.** Cross-validated auto- and cross-correlation of the sound-related pupil responses across 6/8 mice. **d.** Decoding of video and sound identity using pupil area. **e-h.** Same as **a-d** but for body motion, displaying the time course of the first principal component, and for 8/8 mice. (*: p-value < 0.05, **: p-value < 0.01).

Sounds evoke stereotyped, uninstructed behavioral responses

Sounds triggered temporally stereotyped changes in arousal[5] (Figure 3a-b). During the same experiments, in 6 of the 8 mice, we pointed a camera at the eye to measure pupil area, a measure of arousal[12,17,22,23]. Sounds evoked characteristic changes in pupil area, which differed across sounds (Figure 3a) but were consistent across trials and mice, regardless of the video that accompanied the sound (Figure 3a-d).

Moreover, sounds evoked stereotyped, uninstructed eye and body movements[13] (Figure 3e-h). In all mice we measured overall body movements using a wider-angle camera that imaged the head, front paws, and back. We illustrate these movements by plotting the first principal component of facial motion energy[14] (Figure 3e). Sounds evoked large movements, ranging from immediate startle-like responses (<50 ms after sound onset) to more complex, gradual movements (Figure 3e, see Suppl. Figure 3 for all sounds). Sound-evoked body movements were remarkably similar across trials and mice (Figure 3g,h). Similarly, different sounds evoked different eye movements, which were consistent across trials and mice (Suppl. Figure 4). Because sound-evoked movements were different across sounds and similar across trials, we could use them to decode sound identity with 16 ± 2% accuracy (s.e., n=8 mice, Figure 3h).

This accuracy was not statistically different from the 18 ± 2% accuracy of sound decoding from neural activity in visual cortex (p=0.31, paired Wilcoxon sign rank test).

Sound-evoked behaviors predict sound-evoked responses in visual cortex

The body movements evoked by sounds (Figure 3) had a remarkably similar time course to the neural responses evoked by sounds in area V1 (Figure 2), for all sounds (Suppl. Figure 3).

We thus asked to what extent body movements could predict sound-evoked neural activity in V1 (Figure 4a-d). We predicted the sound-evoked population activity in V1 using three models: (1) a purely *auditory* model where the time course of activity depends only on sound identity (equivalent to a test-retest, Figure 4b); (2) a purely *behavioral* model where activity is predicted by pupil area, eye position/motion, and facial movements (Figure 4c); (3) a *full* model where activity is due to an additive combination of both factors (Figure 4d). The models were fitted on the single-trial data but used to predict trial-averages.

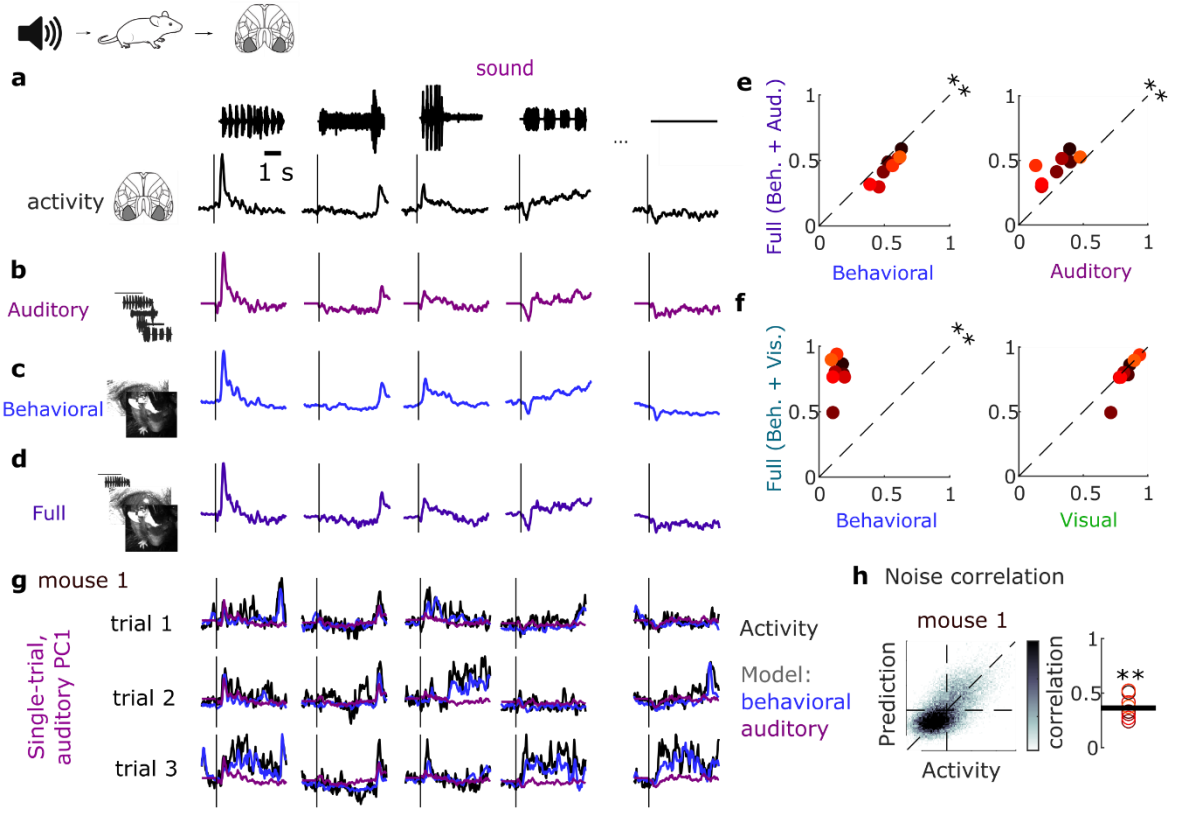


Figure 4. **Sound-evoked behaviors predict sound-evoked responses in visual cortex. a.** Average of the sound-evoked responses along neural auditory PC1 over odd trials from all mice. **b-d.** Cross-validated predictions of neural auditory PC1 responses using only sound stimulus identity (**b.**), eye and body predictors (**c.**), or both at the same time (**d.**), computed using even trials and used to predict odd trials. **e.** Cross-validated correlation of the actual sound responses and their predictions for all mice, comparing different models (full: all predictors, auditory: sounds only, Behavioral: eye and body movements only). **f.** Same as **e**, but for video responses. **h.** Correlation between the single-trial noise in neural activity along auditory PC1 and the single-trial noise in the prediction. Correlation values for all mice are shown on the right. (*: p-value < 0.05, **: p-value < 0.01).

The body movements evoked by sounds were sufficient to explain sound-evoked responses in visual cortex (Figure 4a-f). Indeed, the behavioral model outperformed the full model in predicting trial-averaged responses to sounds (p = 0.0078, paired Wilcoxon sign rank test, Figure 4e, left). This indicates the extra predictors were unnecessary and led only to overfitting. In contrast, the full model performed better than the auditory model (p = 0.0078, paired Wilcoxon sign rank test, Figure 4e, right). Further analysis indicated that the

main behavioral correlate of these responses were movements of the body (as opposed to movements of the eyes), and especially of the whiskers (Suppl. Figure 5). By contrast, the behavioral model explained a much smaller fraction of the trial-averaged video-evoked neural responses than video stimulus identity, consistent with the fact that visual cortex is, indeed, largely visual (Figure 4f).

Variations in body movement across trials predicted trial-by-trial variations in sound-evoked visual cortical activity (Figure 4g,h). The movements each sound elicited were stereotyped, but not identical across trials. Trial-by-trial variations in visual cortical activity could not be explained by the auditory model (by definition) but were well captured by the behavioral model (Figure 4g). The mean correlation between the trial-by-trial variations of the visual cortex's auditory PC1 and the first principal components of body motion was 0.37, significantly above zero ($p = 0.0078$, Wilcoxon sign rank test, Figure 4h). In other words, the V1 responses evoked by sounds in individual trials followed a similar time course as the body movements observed in those trials.

Discussion

These results confirm the many previous reports of sound-evoked responses in visual cortex[2–7], but provide an alternative interpretation for these responses. We found that sounds evoke highly stereotyped changes in arousal and uninstructed body movements. These behavioral effects, in turn, can explain the responses evoked by sounds in visual cortex.

Our results do not imply that movements themselves cause visual cortical activity; instead, changes in internal state may both drive movements and modulate visual cortical activity. Sound-evoked activity in V1 was low-dimensional, consistent with state modulation, and in contrast to the high-dimensional representation of visual stimuli. A similar mechanism may explain sound-evoked activity in visual cortex under anesthesia[2,3], where movements are not possible but state changes can still occur[24,25].

These observations suggest that other aspects of neural activity previously interpreted as being multisensory might also arise from changes in states or behavior. Stereotyped body movements can be elicited not only by sounds[8–11] but also by images[26–30] and odors[27,31]. Given the extensive correlates of body movement observed throughout the brain[12–14,32,33] these observations reinforce the importance of monitoring behavioral state and body movement when interpreting sensory-evoked activity.

Acknowledgements

We thank Philip Coen and Anwar Nunez-Elizalde for useful conversations and comments on the manuscript, Charu Bai Reddy for help with surgeries, and Yoh Isogai and Daniel Regester for providing the explantable methods. This work was supported by the Wellcome Trust (grant 205093 to MC and KDH), by EMBO (ALTF 740-2019 fellowship to CB), and by the Sainsbury Wellcome Centre PhD program (TS and AL). MC holds the GlaxoSmithKline/Fight for Sight Chair in Visual Neuroscience.

Author contributions

	Bimbard	Sit	Lebedeva	Harris	Carandini
Conceptualization	•				•
Methodology	•	•	•	•	•
Software	•	•			
Formal Analysis	•				
Investigation	•	•	•		
Resources			•		
Writing – Original Draft	•				•
Writing – Review & Editing	•	•	•	•	•
Visualization	•				
Supervision					•
Funding Acquisition	•			•	•

References

1. Ghazanfar, A. A. & Schroeder, C. E. Is neocortex essentially multisensory? *Trends Cogn. Sci.* **10,** 278–285 (2006).

2. Iurilli, G. *et al.* Sound-Driven Synaptic Inhibition in Primary Visual Cortex. *Neuron* **73,** 814–828 (2012).

3. Ibrahim, L. A. *et al.* Cross-Modality Sharpening of Visual Cortical Processing through Layer-1-

Mediated Inhibition and Disinhibition. *Neuron* **89,** 1031–1045 (2016).

4. Meijer, G. T., Montijn, J. S., Pennartz, C. M. A. & Lansink, C. S. Audiovisual Modulation in Mouse Primary Visual Cortex Depends on Cross-Modal Stimulus Configuration and Congruency. *J. Neurosci.* **37,** 8783–8796 (2017).

5. Deneux, T. *et al.* Context-dependent signaling of coincident auditory and visual events in primary visual cortex. *eLife* **8,** e44006 (2019).

6. Knöpfel, T. *et al.* Audio-visual experience strengthens multisensory assemblies in adult mouse visual cortex. *Nat. Commun.* **10,** 5684 (2019).

7. Garner, A. R. & Keller, G. B. A cortical circuit for audio-visual predictions. *bioRxiv* (2020). doi:10.1101/2020.11.15.383471

8. Meyer, A. F., Poort, J., O'Keefe, J., Sahani, M. & Linden, J. F. A Head-Mounted Camera System Integrates Detailed Behavioral Monitoring with Multichannel Electrophysiology in Freely Moving Mice. *Neuron* **100,** 46–60.e7 (2018).

9. Landemard, A. *et al.* Distinct higher-order representations of natural sounds in human and ferret auditory cortex. *bioRxiv* (2020). doi:10.1101/2020.09.30.321695

10. Li, Z. *et al.* Corticostriatal control of defense behavior in mice induced by auditory looming cues. *Nat. Commun.* **12,** 1–13 (2021).

11. Yeomans, P. W. & Frankland, J. S. The acoustic startle reflex: neurons and connections. *Brain Res. Rev.* **21,** 301–314 (1996).

12. Shimaoka, D., Harris, K. D. & Carandini, M. Effects of Arousal on Mouse Sensory Cortex Depend on Modality. *Cell Rep.* **22,** 3160–3167 (2018).

13. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22,** 1677–1686 (2019).

14. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364,** 255 (2019).

15. Niell, C. M. & Stryker, M. P. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron* **65,** 472–479 (2010).

16. Salkoff, D. B., Zagha, E., McCarthy, E. & McCormick, D. A. Movement and Performance Explain Widespread Cortical Activity in a Visual Detection Task. *Cereb. Cortex* **30,** 421–437 (2020).

17. Vinck, M., Batista-Brito, R., Knoblich, U. & Cardin, J. A. Arousal and Locomotion Make Distinct Contributions to Cortical Activity Patterns and Visual Encoding. *Neuron* **86,** 740–754 (2015).

18. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551,** 232–236 (2017).

19. Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science (80-.).* **372,** (2021).

20. Gemmeke, J. F. *et al.* Audio Set: An ontology and human-labeled dataset for audio events. *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 776–780 (2017). doi:10.1109/ICASSP.2017.7952261

21. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* (2019). doi:10.1038/s41586-019-1346-5

22. McGinley, M. J., David, S. V. & McCormick, D. A. Cortical Membrane Potential Signature of Optimal States for Sensory Signal Detection. *Neuron* **87,** 179–192 (2015).

23. McGinley, M. J. *et al.* Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron* **87,** 1143–1161 (2015).

24. Klausberger T. *et al.* Brain-state- and cell-type-specific firing of hippocampal interneurons in vivo. *Nature* **421,** 844–848 (2003).

25. Kramis, R., Vanderwolf, C. H. & Bland, B. H. Two types of hippocampal rhythmical slow activity in both the rabbit and the rat: Relations to behavior and effects of atropine, diethyl ether, urethane, and pentobarbital. *Exp. Neurol.* **49,** 58–85 (1975).

26. Yilmaz, M. & Meister, M. Rapid innate defensive responses of mice to looming visual stimuli. *Curr. Biol.* **23,** 2011–2015 (2013).

27. Fink, A. J., Axel, R. & Schoonover, C. E. A virtual burrow assay for head-fixed mice measures habituation, discrimination, exploration and avoidance without training. *eLife* **8,** 1–21 (2019).

28. Procacci, N. M. *et al.* Context-dependent modulation of natural approach behaviour in mice. *Proc. R. Soc. B Biol. Sci.* **287,** (2020).

29. De Franceschi, G., Vivattanasarn, T., Saleem, A. B. & Solomon, S. G. Vision Guides Selection of Freeze or Flight Defense Strategies in Mice. *Curr. Biol.* **26,** 2150–2154 (2016).

30. Socha, K., Whiteway, M., Butts, D. & Bonin, V. Behavioral response to visual motion impacts population coding in the mouse visual thalamus. *bioRxiv* 382671 (2018). doi:10.1101/382671

31. Cohen, L., Rothschild, G. & Mizrahi, A. Multisensory integration of natural odors and

sounds in the auditory cortex. *Neuron* **72,** 357–369 (2011).

32. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576,** 266–273 (2019).

33. Zatka-haas, P., Steinmetz, N. A., Carandini, M. & Harris, K. D. Sensory coding and causal impact of mouse cortex in a visual decision. *bioRxiv* (2021). doi:https://doi.org/10.1101/501627doi

34. Okun, M., Lak, A., Carandini, M. & Harris, K. D. Long term recordings with immobile silicon probes in the mouse cortex. *PLoS One* **11,** 1–17 (2016).

35. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Kenneth D., H. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. *bioRxiv* 061481 (2016). doi:10.1101/061481

36. Wang, Q. *et al.* The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181,** 936–953.e20 (2020).

Methods

Experimental procedures at UCL were conducted according to the UK Animals Scientific Procedures Act (1986) and under personal and project licenses released by the Home Office following appropriate ethics review.

Surgery and recordings

Experiments were performed on 8 mice (6 male and 2 female), between 16 and 38 weeks of age. Visual cortex activity was recorded using Neuropixels 1.0 and 2.0 probes implanted in left primary visual cortex (2.5 mm lateral, 3.5 mm posterior from Bregma). In 5 of the 8 mice the probes were implanted permanently or with a recoverable implant as described in Refs. [19,34] and in the remaining 3 they were implanted with a recoverable implant of a different design (Yoh Isogai and Daniel Regester, personal communication). Results were not affected by the implantation strategy. Sessions were automatically spike-sorted using Kilosort2 (www.github.com/MouseLand/Kilosort2, Ref. [35]) and manually curated to select isolated single cells. Probe location was checked post-hoc by aligning it to the Allen Brain Atlas[36] visually or through custom software (www.github.com/petersaj/AP_histology).

Stimuli

In each session, mice were presented with a sequence of audio, visual or audiovisual movies. The stimuli consisted of all combinations of auditory and visual streams extracted from a set of 11 naturalistic movies depicting the movement of animals such as cats, donkeys and seals, from the AudioSet database[20]. An additional visual stream consisted of a static full-field gray image and an additional auditory stream contained no sound. Movies lasted for 4 s, and were separated by an inter-trial interval of 2 s. The same randomized sequence of movies was repeated 4 times during each experiment, with the second and third repeat separated by a 5 min interval.

The movies were gray-scaled, spatially re-scaled to match the dimensions of a single screen of the display, and duplicated across the three screens. The visual stream was sampled at 30 frames per second. Visual stimuli were presented through three displays (Adafruit, LP097QX1) each with a resolution of 1024 by 768 pixels. The screens covered approximately 270 x 70 degrees of visual angle, with 0 degree being directly in front of the mouse. The screens had a refresh rate of 60 frames per second and were fitted with Fresnel lenses (Wuxi Bohai Optics, BHPA220-2-5) to ensure approximately equal luminance across viewing angles.

Sounds were presented through a pair of Logitech Z313 speakers placed below the screens. The auditory stream was sampled at 44.1 kHz with 2 channels and was scaled to a sound level of -20 decibels relative to full scale.

In situ sound intensity and spectral content was estimated using a calibrated microphone (GRAS 40BF 1/4" Ext. Polarized Free-field Microphone) positioned where the mice sit, and reference loudness was estimated using an acoustic calibrator (SV 30A, Suppl. Figure 1). Mice were systematically habituated to the rig but not to the specific stimuli before the experiment. Presentation of the sounds over

days did not alter the observed behavioral and neural responses.

Videography

Eye and body movements were monitored by illuminating the subject with infrared light (830 nm, Mightex SLS-0208-A). The right eye was monitored with a camera (The Imaging Source, DMK 23U618) fitted with zoom lens (Thorlabs MVL7000) and long-pass filter (Thorlabs FEL0750), recording at 100 Hz. Body movements were monitored with another camera (same model but with a different lens, Thorlabs MVL16M23) situated above the central screen, recording at 40 Hz. Video and stimulus time were aligned using the strobe pulses generated by the cameras, recorded alongside with the output of a screen-monitoring photodiode and the speakers input, all sampled at 2500Hz. Singular Value Decompositions of the face movie and fitting of the pupil area and position were computed using the *facemap* algorithm (www.github.com/MouseLand/facemap).

Data processing

For each experiment, the neural responses constitute a 5-dimensional array $\mathbf{D}$ of size N_t time bins x N_v videos x N_a sounds x N_r repeats x N_c cells. The elements of this matrix are the responses D_{tvarc} measured at time t, in video v, sound a, repeat r, and cell c. $\mathbf{D}$ contains the binned firing rates (30 ms bin size) around the stimulus onset (from 1 s before onset to 3.8 s after onset), smoothed with a causal half gaussian filter (standard deviation of 43 ms), and z-scored for each neuron.

Pupil area and eye position were baseline-corrected to remove the slow fluctuations and focus on the fast, stimulus-evoked and trial-based fluctuations: the mean value of the pupil area or eye position over the second preceding stimulus onset was subtracted from each trial. Signed eye motion (horizontal and vertical) was computed as the difference of the eye position between time bins. The unsigned motion was obtained as the absolute value of the signed motion. The global eye motion was estimated as the absolute value of the movement in any direction (L2 norm). Eye variables values during identified blinks were interpolated based on their values before and after the identified blink. Body motion variables were defined as the first 128 body motion PCs. Both eye-related and body-related variables were then binned similarly to the neural data. We note that the timing precision for the face motion is limited by both the camera acquisition frame rate (40 fps, not aligned to stimulus onset), and the binning used here (30ms bins, aligned on stimulus onset). Thus, real timings can differ by up to 25ms.

All analyses that needed cross-validation (test-retest component covariance, decoding, prediction) were performed using a training set consisting of half of the trials (odd trials) and a test set based on the other half (even trials). Models were computed on the train set and tested on the test set. Then test and train sets were swapped, and quantities of interest were averaged over the two folds.

To estimate the auto- and cross-correlation of the sound-evoked time courses, the variable of interest was split between training and test set, averaged over all trials (e.g., for sound-related responses, over videos and repeats), and the Pearson correlation coefficient was computed between the training set responses for each mouse and the test set responses of all mice (thus giving a cross-validated estimate of the auto- and the cross-correlation).

Marginalization

To isolate the contribution of videos or sounds in the neural responses we used a marginalization procedure similar to the one used in factorial ANOVA. By D_{tvarc} we denote the firing rate of cell c to repeat r of the combination of auditory stimulus a and visual stimulus v, a time t after stimulus onset. The marginalization procedure decomposes D_{tvarc} into components that are equal across stimuli, related to videos, related to sounds, related to audiovisual interactions, and noise:

$$D_{tvarc} = M_{tc} + V_{tvc} + A_{tac} + I_{tvac} + \epsilon_{tvarc}$$

The first term is the mean of the population responses across videos, sounds, and repeats:

$$M_{tc} = D_{t\cdots c} = \frac{1}{N_v N_a N_r} \sum_v \sum_a \sum_r D_{tvarc}.$$

where dots in the second term indicate averages over the missing subscripts, and N_v,

N_a, N_r denote the total number of visual stimuli, auditory stimuli, and repeats.

The second term, the video-related component, is the average of the population responses over sounds and repeats, relative to this mean response:

$$V_{tvc} = D_{tv\cdot\cdot c} - M_{tc}$$

Similarly, the sound-related component is the average over videos and repeats, relative to the mean response:

$$A_{tac} = D_{t\cdot a\cdot c} - M_{tc}$$

The audiovisual interaction component is the variation in population responses that is specific to each pair of sound and video:

$$I_{tvac} = D_{tva\cdot c} - M_{tc} - V_{tvc} - A_{tac}$$

Finally, the noise component is the variation across trials:

$$\epsilon_{tvarc} = D_{tvarc} - D_{tva\cdot c}$$

In matrix notation, we will call $\mathbf{A}$, $\mathbf{V}$, and $\mathbf{I}$ the arrays with elements A_{tac}, V_{tvc}, and I_{tvac} and size $N_t \times N_a \times N_c$, $N_t \times N_v \times N_c$ and $N_t \times N_v \times N_a \times N_c$.

Dimensionality reduction

The arrays of sound-related responses $\mathbf{A}$, of video-related responses $\mathbf{V}$, and of audiovisual interactions $\mathbf{I}$, describe the activity of many neurons. To summarize this activity, we used cross-validated Principal Component Analysis (cvPCA, Ref. [21]). In this approach, principal component projections are found from one half of the data, and an unbiased estimate of the reliable signal variance is found by computing their covariance with the same projections on a second half of the data.

We illustrate this procedure on the auditory responses. In what follows, all arrays, array elements, and averages (e.g. $\mathbf{A}$, $A_{tac}, A_{t\cdot c}$) refer to training-set data (odd-numbered repeats), unless explicitly indicated with the subscript *test* (e.g. $\mathbf{A}_{test}, A_{tac;test}, A_{t\cdot c;test}$).

We first isolate the auditory responses $\mathbf{A}$ as described above from training set data (odd-numbered trials). We reshape this array to have two dimensions $N_t N_a \times N_c$; and perform PCA:

$$\mathbf{T} = \mathbf{AW}$$

where $\mathbf{T}$ ($N_t N_a \times N_p$) is a set of time courses of the top N_p principal components of $\mathbf{A}$, and $\mathbf{W}$ is the PCA weight matrix ($N_c \times N_p$).

For cvPCA analysis, we took $N_p = N_c$ to estimate the amount of reliable stimulus-triggered variance in each dimension (Fig. 2f,i; Supp. Fig. 2). We computed the projections of the mean response over a test set of even-numbered trials, using the same weight matrix: $\mathbf{T}_{test} = \mathbf{A}_{test}\mathbf{W}$ and evaluated their covariance with the training-set projections:

$$\hat{V}_k = \frac{1}{N_t N_a - 1} \sum_{j=1}^{N_t N_a} \left(T_{jk} - T_{\cdot k}\right)\left(T_{jk;test} - T_{\cdot k;test}\right)$$

This method provides an unbiased estimate of the stimulus-related variance of each component [21]. Analogous methods were used to obtain the signal variance for principal components of the visual response and interaction, by replacing $\mathbf{A}$ with $\mathbf{V}$ or $\mathbf{I}$ (Supp. Fig 2). The cvPCA variances were normalized either by the sum for all auditory dimensions (Figure 2f,i), or the sum for all dimensions from video-related, sound-related and interaction-related decompositions (Suppl. Figure 2).

To determine if a cvPCA dimension had variance significantly above 0, we used a shuffling method. The shuffling was done by changing the labels of both the videos and the sounds for each repeat. We performed this randomization 1,000 times and chose a component to be significant if its test-retest covariance value was above the 99th percentile of the shuffled distribution. We defined the dimensionality as the number of significant components.

For the video-related responses, we found an average of 74 significant components (± 20, s.e., n = 8 mice). As expected, this number grew with the number of recorded neurons[21] (data not shown). For the auditory-related responses, instead, we found only 4 significant components on average (± 0.7, s.e., n = 8 mice). For the interactions between videos and sounds, finally, we found zero significant components (0 ± 0, s.e., n = 8 mice) indicating that the population responses did not reflect significant interactions between videos and sounds.

For visualization of PC time courses (Figure 2g,j), the weight matrices $\mathbf{W}$ were computed from the training set but the projection of the full dataset was used to compute the time courses of the first component.

Decoding

Single-trial decoding for video- or sound-identity was performed using a template-matching decoder applied to neural or behavioral data. In this description, we will focus on decoding sound identity from neural data. The data were again split into training and test sets consisting of odd and even trials.

When decoding auditory-related neural activity (Figure 2k), we took $N_p = 4$, so the matrix $\mathbf{T}$ containing PC projections of the mean training-set auditory responses had size $N_t N_a \times 4$; using more components did not affect the results. To decode the auditory stimulus presented on a given test-set trial, we first removed the video-related component by subtracting the mean response to the video presented on that trial (averaged over all training-set trials) We then projected this using the training-set weight matrix $\mathbf{W}$ to obtain a $N_t \times 4$ timecourse for the top auditory PCs, and found the best-matching auditory stimulus by comparing to the mean training-set timecourses for each auditory stimulus using Euclidean distance. A similar analysis was used to decode visual stimuli, using $N_p = 30$ components.

To decode the sound identity from movement data, we used the z-scored eye variables (pupil area in Figure 3, and eye motion in Suppl. Figure 4), or the 128 first principal components of the motion energy of the face movie, and performed the template-matching the same way as the with the neural data.

The significance of the decoding accuracy (compared to chance) was computed by performing a Wilcoxon sign rank test to compare to chance level ($\frac{1}{12}$), treating each mouse as independent. The comparison between video identity and sound identity decoding accuracy was computed by performing a paired Wilcoxon sign rank test across mice.

Encoding

To predict neural activity from stimuli/behavioral variables ("encoding model"; Figure 4), we again started by extracting audio- or video-related components and performing Principal Component Analysis, as described above, however this time the weight matrices were computed from the full dataset rather than only the training set. Again, we illustrate by describing how auditory-related activity was predicted, for which we kept $N_p = 4$ components; video-related activity was predicted similarly but with $N_p = 30$.

We predicted neural activity using linear regression. The target $\mathbf{Y}$ contained the auditory-related activity on each trial, projected onto the top 4 auditory components: specifically, we compute $D_{tvarc} - M_{tc} - V_{tvc}$, reshape to a matrix of size $N_t N_v N_a N_r \times N_c$, and multiply by the matrix of PC weights $\mathbf{W}$. We predicted $\mathbf{Y}$ by regression: $\mathbf{Y} \approx \mathbf{XB}$, where $\mathbf{X}$ is a feature matrix and $\mathbf{B}$ are weights fit by unpenalized, cross-validated reduced-rank least squares.

The feature matrix depended on the type of prediction being made. To predict from sensory stimulus identity, $\mathbf{X}$ had one column for each combination of auditory stimulus and peristimulus timepoint, making $N_a N_t$ =1,524 columns, $N_t N_v N_a N_r$ rows, and contained 1 during stimulus presentations in a column reflecting the stimulus identity and peristimulus time. With this feature matrix, the weights $\mathbf{B}$ represent the mean activity time course for each dimension and stimulus, and estimation is equivalent to time averaging.

To predict from behavior, we used features for pupil area, pupil position (horizontal and vertical), eye motion (horizontal and vertical -- signed and unsigned), global eye motion (L2 norm of x and y motion, unsigned), blinks (thus 9 eye-related predictors) and the first 128 face motion, with lags from -100 ms to 200 ms (thus 12 lags per predictor, 1644 predictors total). To predict from both stimulus identity and behavior, we concatenated the feature matrices, obtaining a matrix with $3{,}168$ columns. The beginning and end of the time course for each trial were padded with NaNs

(12 – the number of lags – at the beginning and end of each trial, to avoid cross-trial predictions by temporal filters. Thus, the feature matrix has $(N_t + 24)N_v N_a N_r$ rows. A model with the eye variables only, and a model with the face motion variables only was also constructed (Suppl. Figure 5).

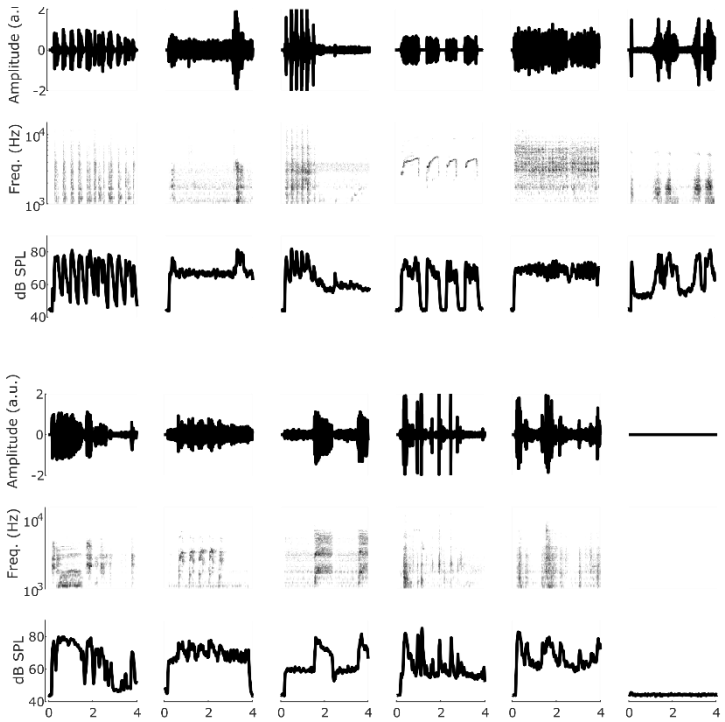
We used Reduced Rank Regression to predict the single trial version of $\mathbf{Y}$ from $\mathbf{X}$ on the training set. The best rank was selected using a 6-fold cross-validation within the training set.

To measure the accuracy of predicting trial-averaged auditory responses (Figure 4a-f), we averaged the $N_t N_v N_a N_r \times N_p$ activity matrix $\mathbf{Y}_{test}$ over all test-set trials of a given auditory stimulus, to obtain a matrix of size $N_t N_a \times N_c$, and did the same for the prediction matrix $\mathbf{X}_{test}\mathbf{B}$, and evaluated prediction quality by the elementwise Pearson correlation of these two matrices.

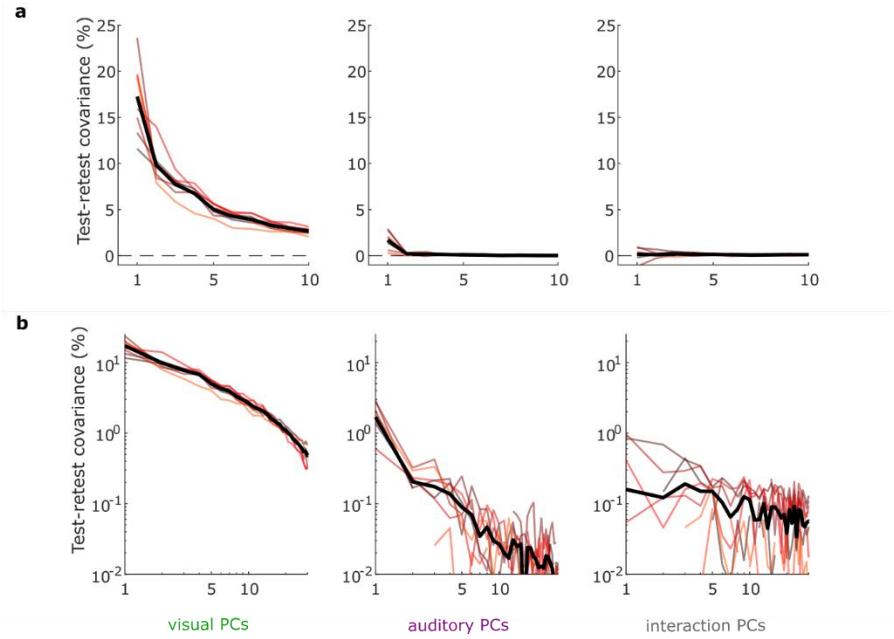
To evaluate predictions of trial-to-trial fluctuations (Figure 4g-h), we computed a "noise" matrix of size $N_t N_v N_a N_r \times N_p$ by subtracting the mean response to each sound: $Y_{tvarp;\,test} - Y_{t\cdot a\cdot p;\,test}$, performed the same subtraction on the prediction matrix $\mathbf{X}_{test}\mathbf{B}$, and evaluated prediction quality by the elementwise Pearson correlation of these two matrices.

To visualize the facial areas important to explain neural activity (Suppl. Figure 5b), we reconstructed the weights of the auditory PC1 prediction in pixel space. Let $\mathbf{b}_0^{\text{body}}$ (1×128) be the weights predicting neural auditory PC1 at lag 0 from each of the 128 body motion PCs. Let $\boldsymbol{\omega}$ $(128 \times$ total number of pixels in the video) be the weights of each of these 128 face motion PCs in pixel space (as an output of the facemap algorithm). We obtained an image $\mathbf{I}$ of the pixel-to-neural weights by computing $\mathbf{I} = \mathbf{b}_0^{\text{body}}\boldsymbol{\omega}$.

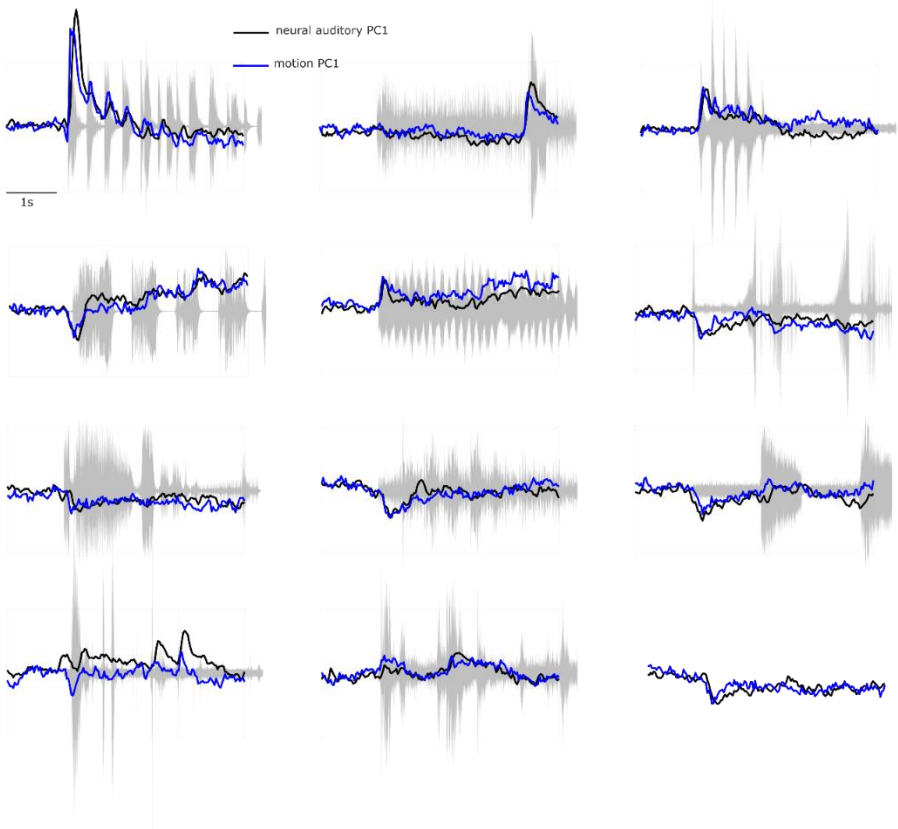
Supplementary Figures



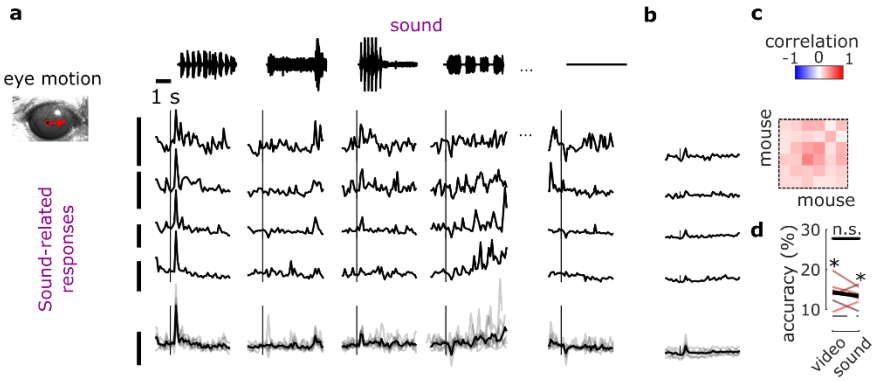
Suppl. Figure 1. **Naturalistic sounds used in this study: spectral content and loudness.** For each sound is displayed: top: amplitude; middle: frequency spectrum; bottom: loudness.



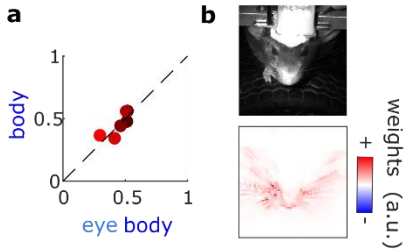
Suppl. Figure 2. **Neural responses are largely visually driven. a.** Test-retest covariance for videos PCs (left), auditory PCs (middle) and interactions PCs (right), for all 8 mice (colors from brown to orange) and their average (black). Test-retest covariance was normalized by the total amount of test-retest covariance across components and video/sound/interaction conditions to show comparable proportions. **b.** Same as **a** but with a logarithmic scale for both x- and y-axes. Negative values are not displayed.



Suppl. Figure 3. **Neural and behavioral responses differ across sounds but resemble each other.** Responses along neural auditory PC1 (black), and behavioral PC1 (blue) sampled at 30ms time bins for all sounds. Responses are averaged over trials, videos, and mice, and z-scored. On the background is visible the envelope of the corresponding sound.



Suppl. Figure 4. **Sounds trigger eye-movements. a.** Sound-related responses eye movement for each sound condition, for 4 example mice, and all mice (6 out of 8 were monitored with an eye camera). **b.** Grand average. **c.** Cross-validated auto- and cross-correlation of the pupil area sound-related responses across all 6 mice. **d.** Decoding of video and sound identity using eye movement.



Suppl. Figure 5. **Sound-evoked V1 responses were mainly driven by whisker movements**. **a**. Correlation of the actual data and their predictions for all mice, comparing different models (eye/body: a model containing both eye and body movements predictors; body: a model containing body movements predictors only). **b**. Top: image of a front camera recording from an example mouse. Bottom: Reconstruction weights of the auditory PC1's best prediction for the same mouse. Most of the weights are related to the whiskers.