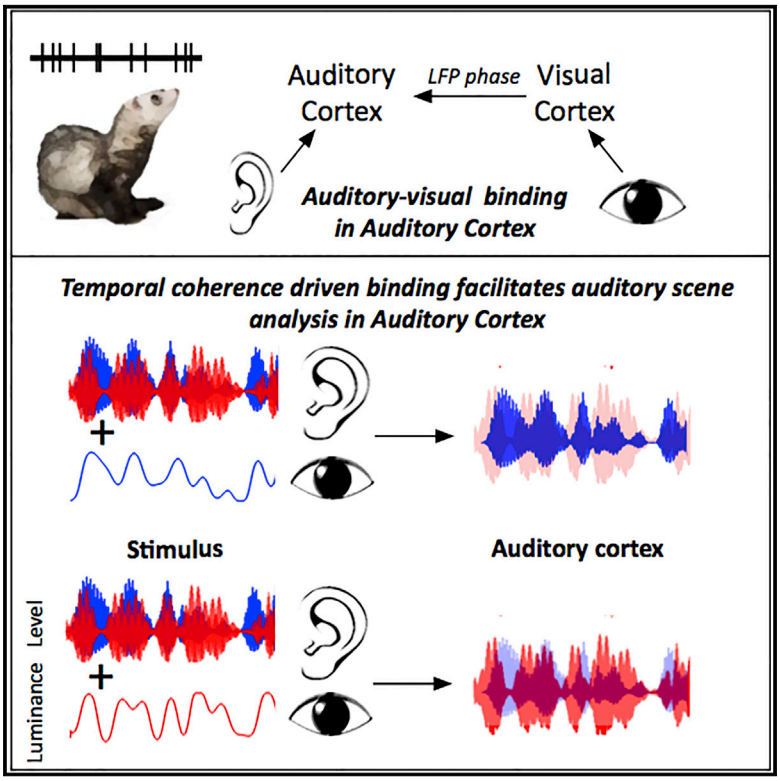
Article

Neuron

Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding

Graphical Abstract



Authors

Huriye Atilgan, Stephen M. Town, Katherine C. Wood, Gareth P. Jones, Ross K. Maddox, Adrian K.C. Lee, Jennifer K. Bizley

Correspondence

j.bizley@ucl.ac.uk

In Brief

Atilgan et al. demonstrate that temporal coherence between auditory and visual stimuli shapes the representation of a sound scene in auditory cortex. Auditory cortex is identified as a site of multisensory binding, with inputs from visual cortex underpinning these effects.

Highlights

- Visual stimuli can shape how auditory cortical neurons respond to sound mixtures
- Temporal coherence between senses enhances sound features of a bound multisensory object
- Visual stimuli elicit changes in the phase of the local field potential in auditory cortex
- Vision-induced phase effects are lost when visual cortex is reversibly silenced

Atilgan et al., 2018, Neuron 97, 640–655
February 7, 2018
https://doi.org/10.1016/j.neuron.2017.12.034

CellPress

CellPress

Neuron
Article

Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding

Huriye Atilgan,[1] Stephen M. Town,[1] Katherine C. Wood,[1] Gareth P. Jones,[1] Ross K. Maddox,[2,3] Adrian K.C. Lee,[3] and Jennifer K. Bizley[1,4,*]

[1]The Ear Institute, University College London, London, UK
[2]Department of Biomedical Engineering and Department of Neuroscience, Del Monte Institute for Neuroscience, University of Rochester, Rochester, NY, USA
[3]Institute for Learning and Brain Sciences and Department of Speech and Hearing Sciences, University of Washington, Seattle, WA, USA
[4]Lead Contact
*Correspondence: j.bizley@ucl.ac.uk
https://doi.org/10.1016/j.neuron.2017.12.034

SUMMARY

How and where in the brain audio-visual signals are bound to create multimodal objects remains unknown. One hypothesis is that temporal coherence between dynamic multisensory signals provides a mechanism for binding stimulus features across sensory modalities. Here, we report that when the luminance of a visual stimulus is temporally coherent with the amplitude fluctuations of one sound in a mixture, the representation of that sound is enhanced in auditory cortex. Critically, this enhancement extends to include both binding and non-binding features of the sound. We demonstrate that visual information conveyed from visual cortex via the phase of the local field potential is combined with auditory information within auditory cortex. These data provide evidence that early cross-sensory binding provides a bottom-up mechanism for the formation of cross-sensory objects and that one role for multisensory binding in auditory cortex is to support auditory scene analysis.

INTRODUCTION

When listening to a sound of interest, we frequently look at the source. However, how auditory and visual information are integrated to form a coherent perceptual object is unknown. The temporal properties of a visual stimulus can be exploited to detect correspondence between auditory and visual streams (Crosse et al., 2015; Denison et al., 2013; Rahne et al., 2008), can bias the perceptual organization of a sound scene (Brosch et al., 2015), and can enhance or impair listening performance depending on whether the visual stimulus is temporally coherent with a target or distractor sound stream (Maddox et al., 2015). Together, these behavioral results suggest that temporal coherence between auditory and visual stimuli can promote binding of cross-modal features to enable the formation of an auditory-visual (AV) object (Bizley et al., 2016b).

Visual stimuli can both drive and modulate neural activity in primary and non-primary auditory cortex (Bizley et al., 2007; Chandrasekaran et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; 2010; Perrodin et al., 2015), but the contribution that visual activity in auditory cortex makes to auditory function remains unclear. One possibility is that the integration of cross-sensory information in early sensory cortex provides a bottom-up substrate for the binding of multisensory stimulus features into a single perceptual object (Bizley et al., 2016b). We have recently argued that binding is a distinct form of multisensory integration that underpins perceptual object formation. We hypothesize that binding is associated with a modification of the sensory representation and can be identified by demonstrating a benefit in the behavioral or neural discrimination of a stimulus feature orthogonal to the features that link crossmodal stimuli (Figure 1A). Therefore, in order to demonstrate binding, an appropriate cross-modal stimulus should not only elicit enhanced neural encoding of the stimulus features that bind auditory and visual streams (the "binding features"), but there should be enhancement in the representation of *other* stimulus features ("non-binding features" associated with the source (Figure 1C).

Here, we test the hypothesis that the incorporation of visual information into auditory cortex can determine the neuronal representation of an auditory scene through multisensory binding (Figure 1). We demonstrate that when visual luminance changes coherently with the amplitude of one sound in a mixture, auditory cortex is biased toward representing the temporally coherent sound. Consistent with these effects reflecting cross-modal binding, the encoding of sound timbre, a non-binding stimulus feature, is subsequently enhanced in the temporally coherent auditory stream. Finally, we demonstrate that the site of multisensory convergence is in auditory cortex and that visual information is conveyed via the local field potential directly from visual cortex.

RESULTS

We recorded neuronal responses in the auditory cortex of awake passively listening ferrets (n = 9 ferrets, 221 single units, 311

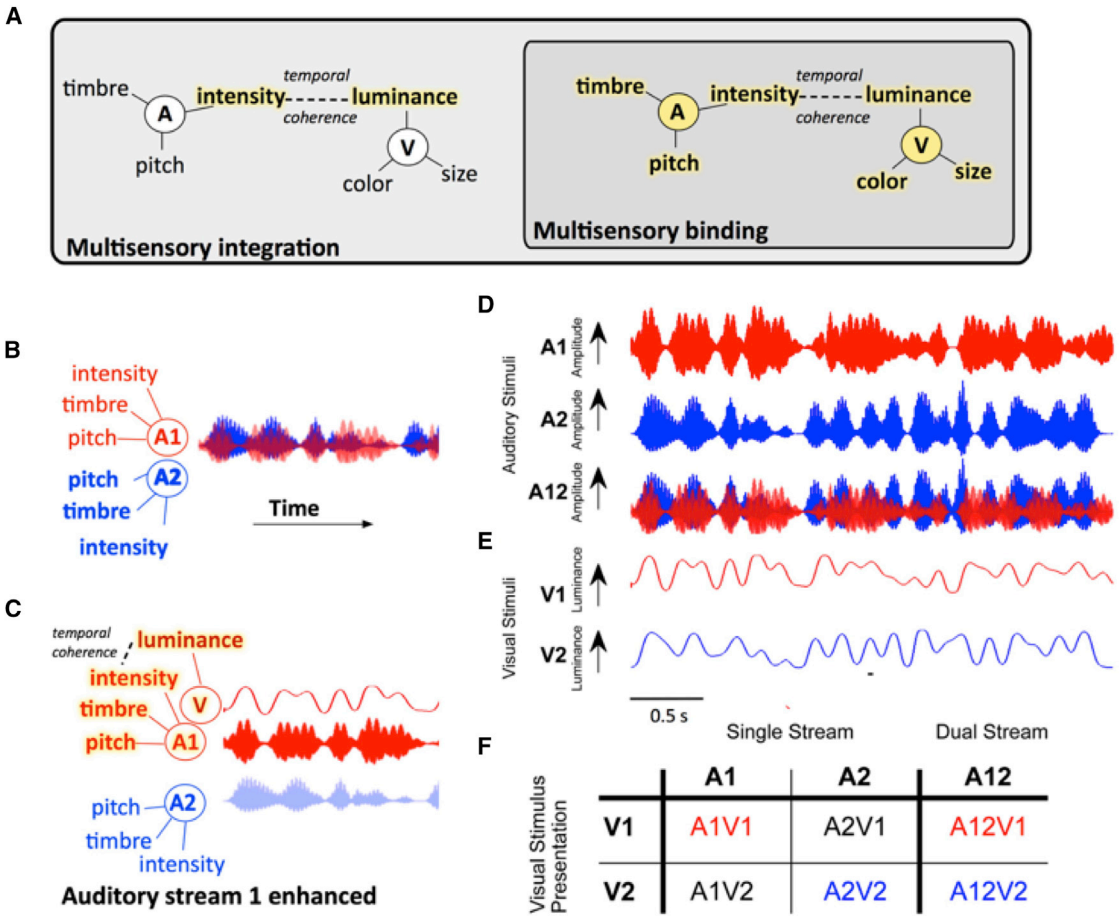


Figure 1. Hypothesis and Experimental Design

(A) Conceptual model illustrating how binding can be identified as a distinct form of multisensory integration. Multisensory binding is defined as a subset of multisensory integration that results in the formation of a crossmodal object. During binding, all features of the audio-visual object are linked and enhanced, including both those features that bind the stimuli across modalities (here temporal coherence between auditory [A] intensity and visual [V] luminance) and orthogonal features such as auditory pitch and timbre, and visual color and size. Other forms of multisensory integration would result in enhancement of only the features that promote binding—here auditory intensity and visual luminance. To identify binding therefore requires a demonstration that non-binding features (e.g., here pitch, timbre, color, or size) are enhanced. Enhanced features are highlighted in yellow.

(B) When two competing sounds (red and blue waveforms) are presented, they can be separated on the basis of their features but may elicit overlapping neuronal representations in auditory cortex.

(C) Hypothesized enhancement in auditory stream segregation when a temporally coherent visual stimulus enables multisensory binding. When the visual stimulus changes coherently with the red sound (A1, top), this sound is enhanced and the two sources are better segregated. Perceptually this would result in more effective auditory scene analysis and an enhancement of the non-binding features.

(D) Stimulus design: auditory stimuli were two artificial vowels (denoted A1 and A2), each with distinct pitch and timbre and independently amplitude modulated with a noisy low pass envelope.

(E) Visual stimulus: a luminance modulated white light was presented with one of two temporal envelopes derived from the amplitude modulations of A1 and A2.

(F) The stimulus combinations that were tested experimentally in single-stream (a single auditory visual pair) and dual-stream (two sounds and one visual stimulus) conditions.

See also Figure S1.

multi-units) in response to naturalistic time-varying auditory and visual stimuli adapted from Maddox et al. (2015). The stimuli are designed to share properties with natural speech; they are modulated at approximately syllable rate and, like competing voices, can be separated on the basis of their fundamental frequency (F0, the physical determinant of pitch). These sounds are devoid of any linguistic content permitting the separation of general sensory processing mechanisms from language-specific ones for human listeners. Maddox et al. (2015) used both pure tones and synthetic vowels as stimuli; here, we used synthetic vowels as these robustly drive auditory cortical responses in the ferret, in neurons with a wide range of characteristic frequencies (Bizley et al., 2009). Ferrets are also well able to distinguish the timbre of artificial vowels (Bizley et al., 2013; Town et al., 2015), and, like human listeners, both ferret behavioral and neural responses show invariant responses to vowel timbre across changes in sound level, location, and pitch (Town et al., 2017). We additionally recorded neural responses

in medetomidine-ketamine anesthetized ferrets (n = 5 ferrets, 426 single units, 772 multi-units), which allowed us to entirely eliminate attentional effects and limit the impact of top-down processing on sensory responses. These experiments also permitted longer recording durations for additional control stimuli and enabled simultaneous characterization of neural activity across cortical laminae. In a subset of these animals, we were able to reversibly silence visual cortex during recording, in order to determine the origin of visual-stimulus elicited neural changes. Recordings were made in awake freely moving animals while they held their head at a drinking spout but were not engaged in a behavioral task and allowed us to measure neural activity free from any confounds associated with pharmacological manipulation and in the absence of task-directed attention, which would likely engage additional neural circuits.

The stimuli were two auditory streams each comprised of a vowel with distinct pitch and timbre (denoted A1: [u], F0 = 175 Hz and A2: [a], F0 = 195 Hz, Figure 1) and independently amplitude modulated with a low-pass (<7 Hz) envelope (Figure 1D). A full-field visual stimulus accompanied the auditory stimuli, the luminance of which was temporally modulated with the modulation envelope from one of the two auditory streams (Figure 1E). We tested stimulus conditions in which both auditory streams were presented (''dual-stream'') and the visual stimulus was temporally coherent with one or other of the auditory streams (A12V1 or A12V2, Figure 1E). We also tested conditions in which a single auditory-visual stimulus pair was presented (''single-stream'' stimuli), where the auditory and visual streams could be temporally coherent (A1V1, A2V2) or independent (A1V2, A2V1), as well as no-visual control conditions (A1, A2).

Auditory-Visual Temporal Coherence Shapes the Representation of a Sound Scene in Auditory Cortex

We first asked whether the temporal dynamics of a visual stimulus could selectively enhance the representation of one sound in a mixture. We therefore recorded responses to auditory scenes composed of two sounds (A1 and A2), presented simultaneously, with a visual stimulus that was temporally coherent with one or other auditory stream (A12V1 or A12V2). It is known that a visual stimulus can enhance the representation of the amplitude envelope of an attended speech stream in auditory cortex (Zion Golumbic et al., 2013; Park et al., 2016). To test whether we could observe a similar phenomenon in single neurons in the absence of selective attention, we used neural responses to temporally coherent single-stream stimuli (i.e., A1V1 and A2V2) to determine the extent to which the neural response to the sound mixture was specific to one or other sound stream.

Figure 2 illustrates this approach for a single unit: responses to the temporally coherent single-stream auditory-visual stimuli (Figure 2A) formed templates that were used to decode the responses to the dual-stream stimuli (Figure 2B) using a Euclidean distance-based spike pattern classifier. Such an approach is ideally suited for classifying neural responses to time-varying stimuli. Auditory cortical responses to the dual-stream stimuli (A12V1 or A12V2) were more commonly decoded as A1V1 when the visual stimulus was V1, and A2V2 when the visual stimulus was V2. Performing this analysis for each neuron in our recorded population yielded similar observations: the coherent auditory stimulus representation was enhanced (Figures 2C, 2D, 2F, and 2G) such that auditory cortical responses to dual-stream stimuli most closely resembled responses to the single-stream stimulus with the shared visual component.

To quantify whether the responses of individual units were significantly influenced by the visual stimulus identity, we first calculated a visual preference index (VPI) as the difference between the percentage of A12V1 trials classified as A1 and the percentage of A12V2 trials classified as A1. Units that were fully influenced by the identity of the visual stimulus would have a VPI score of 100, while those in which the visual stimulus did not influence the response at all would have a score of 0 (Figures 2E and 2H). We assessed the significance of observed VPI scores using a permutation test ($p < 0.05$) to reveal that 33.6% of driven units recorded in awake animals (91/271 units) and 52.9% of units in the anesthetized dataset (175/331 units) had responses to dual-stream stimuli significantly influenced by the visual stimulus.

Modulation of dual-stream responses by visual stimulus identity was not simply a consequence of the shared visual component of single-stream and dual-stream stimuli and was observed in neurons in which visual or auditory identity could be decoded (example response from a unit in which only auditory stimulus identity could be decoded: Figure S2). If this effect was only apparent in visual neurons in auditory cortex, then eliminating the visual element of the single-stream stimuli should prohibit successful decoding for the dual-stream stimuli. Additional control experiments (n = 89 driven units, from 4 awake animals) demonstrated that this was not the case: the enhancement of the temporally coherent sound in the sound mixture was evident whether dual-stream stimuli (A12V1 and A12V2) were decoded using responses to auditory-only stimuli (A1 or A2) or auditory-visual stimuli (A1V1, A2V2 etc.). Within this control data (example unit: Figures 3A and 3B, population data Figures 3C–3H), 32 units (36%) had a significant VPI scores when dual-stream responses were decoded from an auditory-only single-stream templates and 31 units (35%) when decoded with auditory-visual templates. Furthermore, the distribution of VPI values was statistically indistinguishable for decoding dual-stream responses with A-only or auditory-visual templates (Kolmogorov-Smirnov test: all units, $p = 0.9016$; units with visual preference scores significantly >0, $p > 0.9998$), and the distribution of values in Figure 3D was statistically indistinguishable from that in Figure 2E ($p = 0.0864$). We also determined that removing the visual stimulus from the dual-stream condition eliminated any decoding difference in responses observed (Figure 3H). A two-way repeated-measures ANOVA on decoded responses with factors of visual stream (V1, V2, no visual) and template type (auditory-visual or A) demonstrated a significant effect of visual stream identity on dual-stream decoding ($F(2, 528) = 19.320$, $p < 0.001$), but there was no effect of template type ($F(1,528) = 0.073$, $p = 0.787$) or interaction between factors ($F(2,528) = 0.599$, $p = 0.550$). Post hoc comparisons revealed that without visual stimulation, there was no tendency to respond preferentially to either stream, but that visual stream identity significantly influenced the classification of dual-stream responses.

Analysis of recording site locations demonstrated that, in the awake animal, recordings in the posterior ectosylvian gyrus

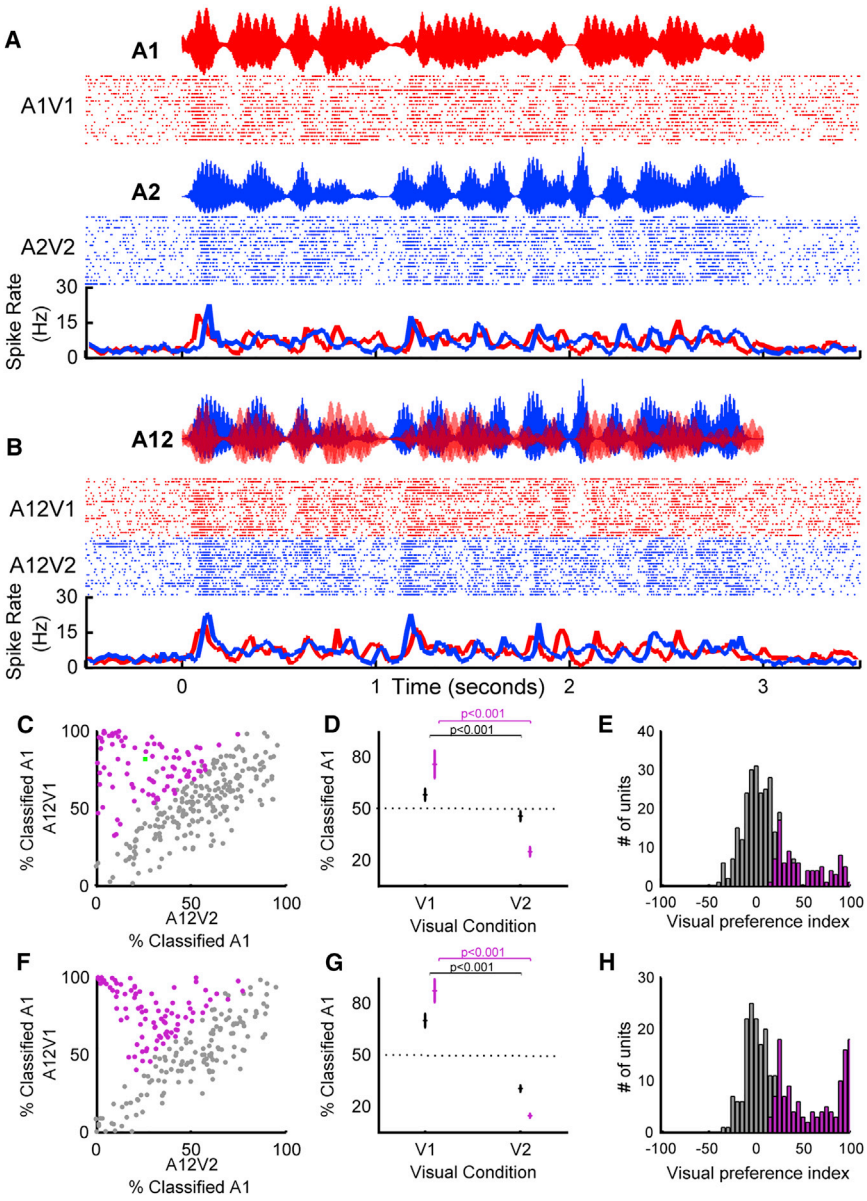


Figure 2. Visual Stimuli Can Determine Which Sound Stream Auditory Cortical Neurons Follow in a Mixture

(A and B) Spiking responses from an example unit in response to (A) single-stream auditory-visual stimuli used as decoding templates and (B) dual-stream stimuli. In each case, rasters and peristimulus time histograms (PSTHs) are illustrated, color coded according to their auditory-visual (A) or visual (B) identity. When the visual component of the dual stream was V1, the majority of trials were classified as A1V1 (82%, 19/23 trials) and whereas when the visual stimulus was V2, only 26% (6/23 trials) were classified as A1V1. (see also green data point in C), yielding a visual preference score of 56%.

(C–H) Population data for awake (C–E 271 units) and anesthetized (F–H 331 units) datasets. In each case, the left panels (C and F) show the distribution of decoding values according to the visual condition, with units in which the VPI was significantly >0 colored purple, whereas those with a VPI value statistically indistinguishable from 0 are colored gray. The middle panels (D and G) show the population mean (±SEM) projecting onto the vertical axis of (C) and (F) for V1 condition and horizontal axis of (C) and (F) for the V2 condition (with purple lines showing data for units with significant VPI values). (E) and (H) show the visual preference index (VPI) color coded according to whether these values were significantly >0. Pairwise comparisons revealed a significant effect of visual condition on decoding in all datasets: awake: All: t_{540} = 6.1, p = 2.3e-09 (n = 271), Sig VPI: t_{180} = 18.8 p = 2.0e-44 (n = 91); anesthetized: All: t_{660} = 9.5, p = 3.3e-20 (n = 331), Sig. VPI: t_{348} = 38.9, p = 1.2e-128 (n = 175).

See also Figures S2–S4.

(PEG, which contains two tonotopic secondary fields) were most strongly influenced by the visual stimulus (Figure S3B). In anesthetized animals, the magnitude of the visual preference scores was similar to that observed in the primary fields of awake animals but was not significantly different across cortical areas (Figure S3E). In both awake and anesthetized animals, units that were classified as "visual discriminating" (see Figure 5; STAR Methods) or "auditory discriminating" were influenced by the visual stimulus, with the magnitude of the effects being greatest in the visual-discriminating units. In anesthetized animals, we confirmed using noise bursts and light flashes that a substantial proportion of both visual-discriminating and auditory-discriminating units were auditory-visual (of 136 visual discriminating units with a significant VPI, 19 were categorized as auditory, 39 as visual, and 78 as auditory-visual; of 39 auditory-discriminating units with significant VPI values, 21 were auditory, 2 were visual, and 16 were auditory-visual Figure S3I). The ability of auditory-visual temporal coherence to enhance one sound in a mixture was observed across all cortical layers (anesthetized dataset; layers defined by current source density analysis, see STAR Methods, Figure S3F) but was strongest in the supra-granular layers (Figure S3G). Finally, we observed these effects in both single and multi-units (Figures S4A and S4B).

Auditory-Visual Temporal Coherence Enhances Nonbinding Sound Features

A hallmark of an object-based rather than feature-based representation is that all stimulus features are bound into a unitary perceptual construct, including those features that do not directly mediate binding (Desimone and Duncan, 1995). We predicted that binding across modalities would be promoted via synchronous changes in auditory intensity and visual luminance (Figures 1B and S1) and observed that the temporal dynamics of the visual stimulus enhanced the representation of temporally

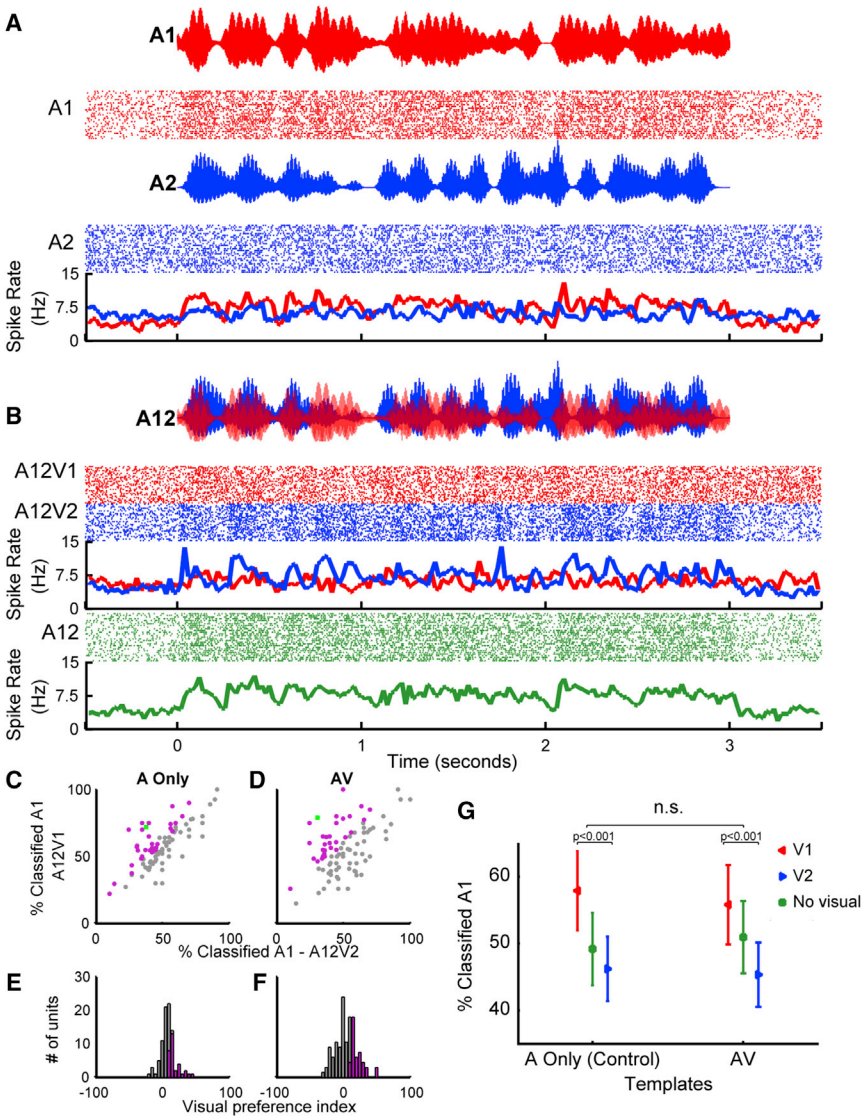


Figure 3. Visual Stimuli Shape the Neural Representation of an Auditory Scene

(A and B) In an additional control experiment (n = 89 units recorded in awake animals), the responses to coherent auditory-visual and auditory-only (A Only) single-stream stimuli were used as templates to decode dual-stream stimuli either accompanied by visual stimuli (V1/V2) or in the absence of visual stimulation (no visual). Shown are spiking responses from an example unit in response to (A) single-stream auditory stimuli that were used as decoding templates to decode the responses to dual-stream stimuli in (B); in each case, the auditory waveform, rasters, and PSTHs are shown. In this example, when decoded with auditory-visual templates: 79% (22/28) of responses were classified as A1 when the visual stimulus was V1, and 32% of responses (9/28) were classified as A1 when the visual stimulus was V2, yielding a VPI score of 47%. When decoded with A-only templates, the values were 75% when V1 (22/28) and 35% when V2 (10/28), yielding a VPI of 40%. For comparison, the auditory-only condition (A12) is shown in green.

(C and D) Population data showing the proportion of responses classified as A1 when the visual stimulus was V1 or V2 when decoded with auditory-only templates (C) or auditory-visual templates (D).

(E and F) Resulting VPI scores from auditory-only decoding (E) or auditory-visual decoding (F).

(G) Mean (±SEM) values for these units when decoded with A-only templates, auditory-visual templates (as in Figure 2), or in the absence of a visual stimulus.

The green data point in (C) and (D) depicts the example in (A) and (B).

coherent auditory streams (Figures 2C–2H and 3D–3F). To determine whether temporal synchrony of visual and auditory stimulus components also enhanced the representation of orthogonal stimulus features and thus fulfill a key prediction of binding (Bizley et al., 2016b), we introduced brief timbre perturbations into our acoustic stimuli (two in each of the A1 and A2 streams). Each deviant lasted for 200 ms during which the spectral timbre smoothly transitioned to the identity of another vowel and back to the original. It is important to note that neither the amplitude of the auditory envelope nor the visual luminance were informative about whether, or when, a change in sound timbre occurred (Figure S1). Such timbre deviants could be detected by human listeners and were better detected when embedded in an auditory stream that was temporally coherent with an accompanying visual stimulus (Maddox et al., 2015). We hypothesized that a temporally coherent visual stimulus would enhance the representation of timbre deviants in the responses of auditory cortical neurons.

To isolate neural responses to the timbre change from those elicited by the on-going amplitude modulation, we extracted 200-ms epochs of the neuronal response during the timbre deviant and compared these to epochs from stimuli without deviants that were otherwise identical (Figure S1). We observed that the spiking activity of many units differed between deviant and no-deviant trials (e.g., Figures 4A and S6), and we were able to discriminate deviant from no-deviant trials with a spike pattern classifier. For each neuron, our classifier reported both the number of deviants that could be detected (i.e., discriminated better than chance as assessed with a permutation test, the maximum is 4, two per auditory stream), and a classification score (where 100% implies perfect discrimination, and 50% chance discrimination, averaged across all deviants for any unit in which at least one deviant was successfully detected). We first considered the influence of temporal coherence between auditory and visual stimuli on the representation of timbre deviants in the single-stream condition (A1V1, A1V2 etc.). We found that a greater proportion of units detected at least one deviant when the auditory stream in which deviants occurred was temporally coherent with the visual stimulus, relative to the temporally independent

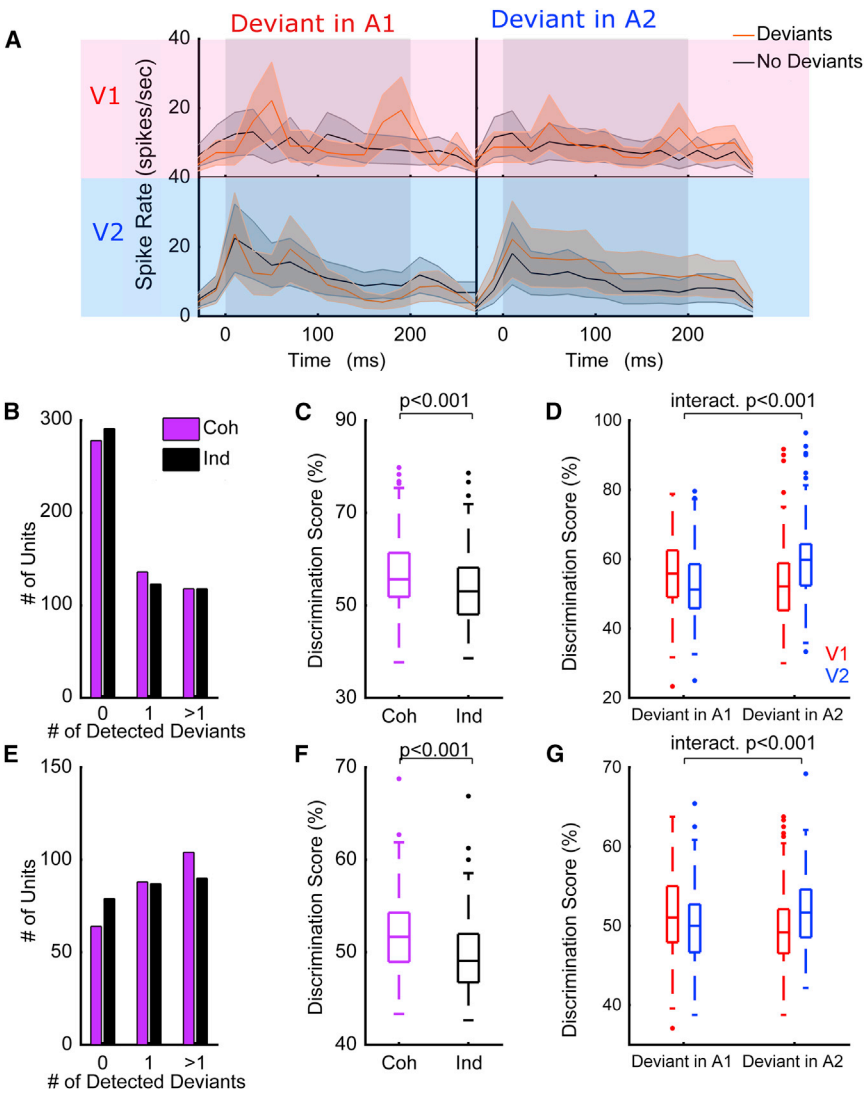


Figure 4. Temporally Coherent Changes in Visual Luminance and Auditory Intensity Enhance the Representation of Auditory Timbre

(A) Example unit response (from the awake dataset) showing the influence of visual temporal coherence on spiking responses to dual-stream stimuli with (red PSTH) or without (black PSTH) timbre deviants.

(B and C) Timbre deviant discrimination in the awake dataset. Two deviants were included in each auditory stream giving a possible maximum of 4 per unit (B), histogram showing the number of deviants (out of 4) that could be discriminated from spiking responses (C), and boxplots showing the timbre deviant discrimination scores in the single-stream condition across different visual conditions (Coh: coherent, ind: independent). The boxes show the upper- and lower-quartile values, the horizontal lines indicate the median, and the whiskers depict the most extreme data points not considered to be outliers (which are marked as individual symbols).

(D) Discrimination scores for timbre deviant detection in dual-stream stimuli in awake animals. Discrimination scores are plotted according to the auditory stream in which the deviant occurred and the visual stream that accompanied the sound mixture. V1 stimuli are plotted in red, and V2 stimuli in blue; therefore, the boxplots at the far left and right of the plot represent the cases in which the deviants occurred in an auditory stream that was temporally coherent with the visual stimulus, while the central two boxplots represent the discrimination of deviants occurring in the auditory stream that was temporally independent of the visual stimulus.

(E–G) The same as (B)–(D) but for the anesthetized dataset.

See also Figure S6.

condition. This was true both for awake (Figure 4B; Pearson chi-square statistic, $\chi^2 = 322.617$, $p < 0.001$) and anesthetized animals (Figure 4E; $\chi^2 = 288.731$, $p < 0.001$). For units that detected at least one deviant, discrimination scores were significantly higher when accompanied by a temporally coherent visual stimulus (Figure 4C, awake dataset, pairwise t test $t_{300} = 3.599$ $p < 0.001$; Figure 4F, anesthetized data $t_{262} = 4.444$ $p < 0.001$).

We also observed an enhancement in the representation of timbre changes in the context of a sound scene (Figures 4D and 4G): timbre changes were more reliably encoded when the sound stream in which they were embedded was accompanied by a temporally coherent visual stimulus. We performed a two-way repeated-measures ANOVA on deviant discrimination performance with visual condition (V1/V2) and the auditory stream in which the deviants occurred (A1/A2) as factors. We anticipated that enhancement of the representation of timbre deviants in the temporally coherent auditory stream would be revealed as a significant interaction term in the dual-stream data. Significant interaction terms were seen in both the awake (Figure 4D, $F(1,600) = 29.138$, $p < 0.001$) and anesthetized datasets (Figure 4G, $F(1,524) = 16.652$, $p < 0.001$). We also observed significant main effects of auditory and visual conditions in awake (main effect of auditory stream, $F(1,600) = 4.565$, $p = 0.033$; main effect of visual condition, $F(1,600) = 2.650$, $p = 0.010$) but not anesthetized animals (main effect of auditory stream, $F(1,524) = 0.004$, $p = 0.948$; main effect of visual condition, $F(1,524) = 1.355$, $p = 0.245$).

Finally, to determine whether a temporally coherent visual stimulus enhanced the representation of non-binding features relative to auditory-alone stimuli, we collected additional control data (3 animals, 39 driven units) in which single-stream stimuli were presented with, or without a temporally coherent visual stimulus. These data (Figures S6A–S6C) confirmed that the presence of a visual stimulus enhanced the encoding of timbre deviants relative to the auditory-only condition. The magnitude of the influence of auditory-visual temporal coherence on timbre deviant encoding was equivalent in single and multi-units (Figures S4C and S4D).

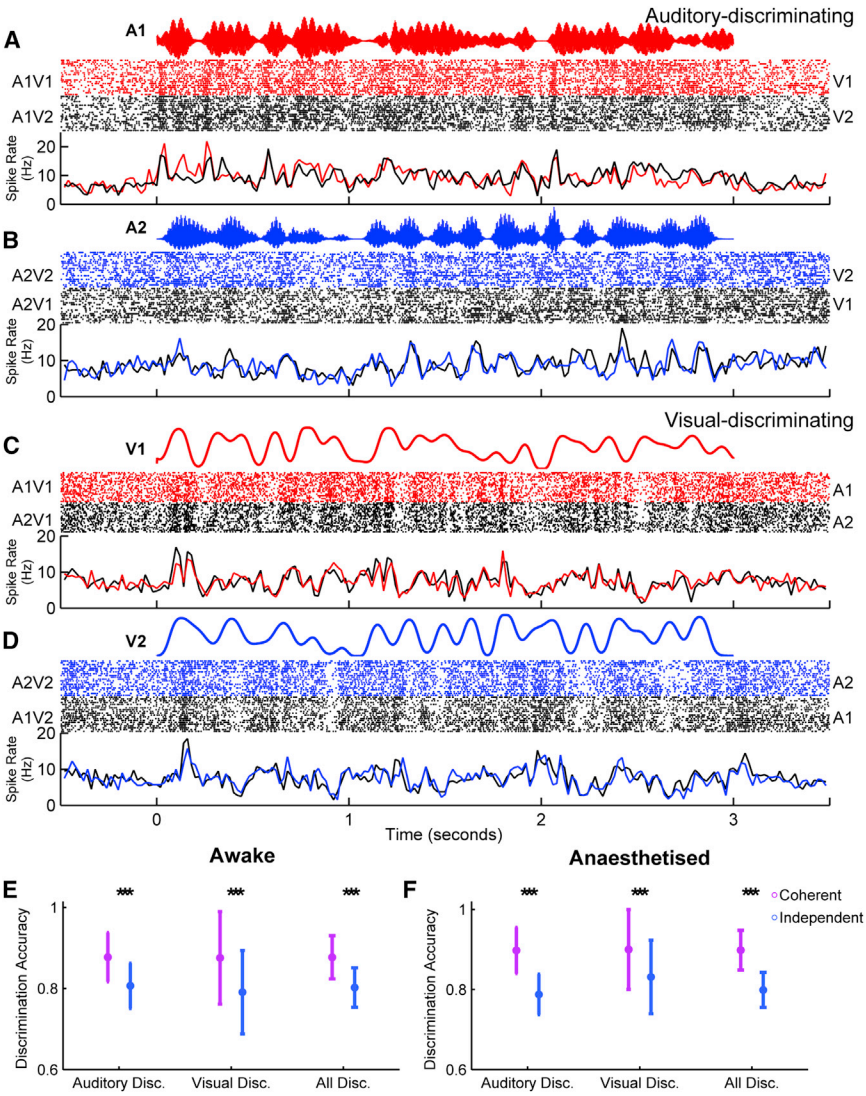


Figure 5. Auditory-Visual Temporal Coherence Enhances Neural Coding in Auditory Cortex

(A–D) A pattern classifier was used to determine whether neuronal responses were informative about auditory or visual stimuli. The responses to single-stream stimuli are shown for two example units, with responses grouped according to the identity of the auditory (A and B, for an auditory discriminating unit) or visual stream (C and D, for a visual discriminating unit). In each case, the stimulus amplitude (A and B)/luminance (C and D) waveform is shown in the top panel with the resulting raster plots and PSTHs below.

(E and F) Decoder performance (mean ± SEM) for discriminating stimulus identity (coherent: A1V1 versus A2V2, purple; independent: A1V2 versus A2V1, blue) in auditory and visual classified units recorded in awake (E) and anesthetized (F) ferrets. Pairwise comparisons for decoding of coherent versus independent stimuli (***$p < 0.001$).

Together these data demonstrate the predicted enhancement in the neural representation of both binding (i.e., auditory amplitude) and non-binding features (here auditory timbre) that are orthogonal to those that promote binding between auditory and visual streams, meaning the effects we observe in auditory cortex fulfill our definition of multisensory binding. Next, we turn to the question of how these effects are mediated and whether they emerge within or outside of auditory cortex.

Auditory Cortical Spike Patterns Differentiate Dynamic Auditory-Visual Stimuli More Effectively When Stimuli Are Temporally Coherent

We used the responses to single-stream stimuli to classify neurons according to whether they were dominantly modulated by auditory or visual temporal dynamics. To determine whether the auditory amplitude envelope reliably modulated spiking, we used a spike-pattern classifier to decode the auditory stream identity, collapsed across visual stimulus (i.e., we decoded auditory stream identity from the combined responses to A1V1 and A1V2 stimuli and the combination of A2V1 and A2V2 responses). An identical approach was taken to determine whether neuronal responses reliably distinguished visual modulation (i.e., we decoded visual identity from the combined responses to A1V1 and A2V1 stimuli and the combined responses elicited by A1V2 and A2V2). Neuronal responses that were informative about auditory or visual stimulus identity at a level better than chance (estimated with a bootstrap resampling) were classified as auditory discriminating (Figures 5A and 5B) and/or visual discriminating (Figures 5C and 5D), respectively.

In awake animals, 39.5% (210/532) of driven units were auditory discriminating, 11.1% (59/532) were visual discriminating, and only 0.4% (2/532) discriminated both auditory and visual stimuli. Overall a smaller proportion of units represented the identity of auditory or visual streams in the anesthetized dataset: 20.2% (242/1198) were auditory discriminating, 6.8% (82/1198) were visual discriminating, and 0.6% (7/1198) discriminated both. Using simple noise bursts and light flashes in anesthetized animals revealed that the classification of units as visual/auditory discriminating based on the single-stream stimuli selected a subset of light and/or sound driven units and that the proportions of auditory, visual, and auditory-visual units recorded in our sample were in line with previous studies from ferret auditory cortex (65.1% [328/504] of units were driven by noise bursts, 16.1% [81/504] by light flashes and 14.1% [71/504] by both). When considering the units that were classified as auditory or visual discriminating based on single-stream stimuli, and for which we recorded responses to noise bursts and light flashes, 53% (160/307) were classified as auditory, 17% (53/307) as visual and 31% (94/307) as auditory-visual when classified with simple stimuli (see also Figure S3I).

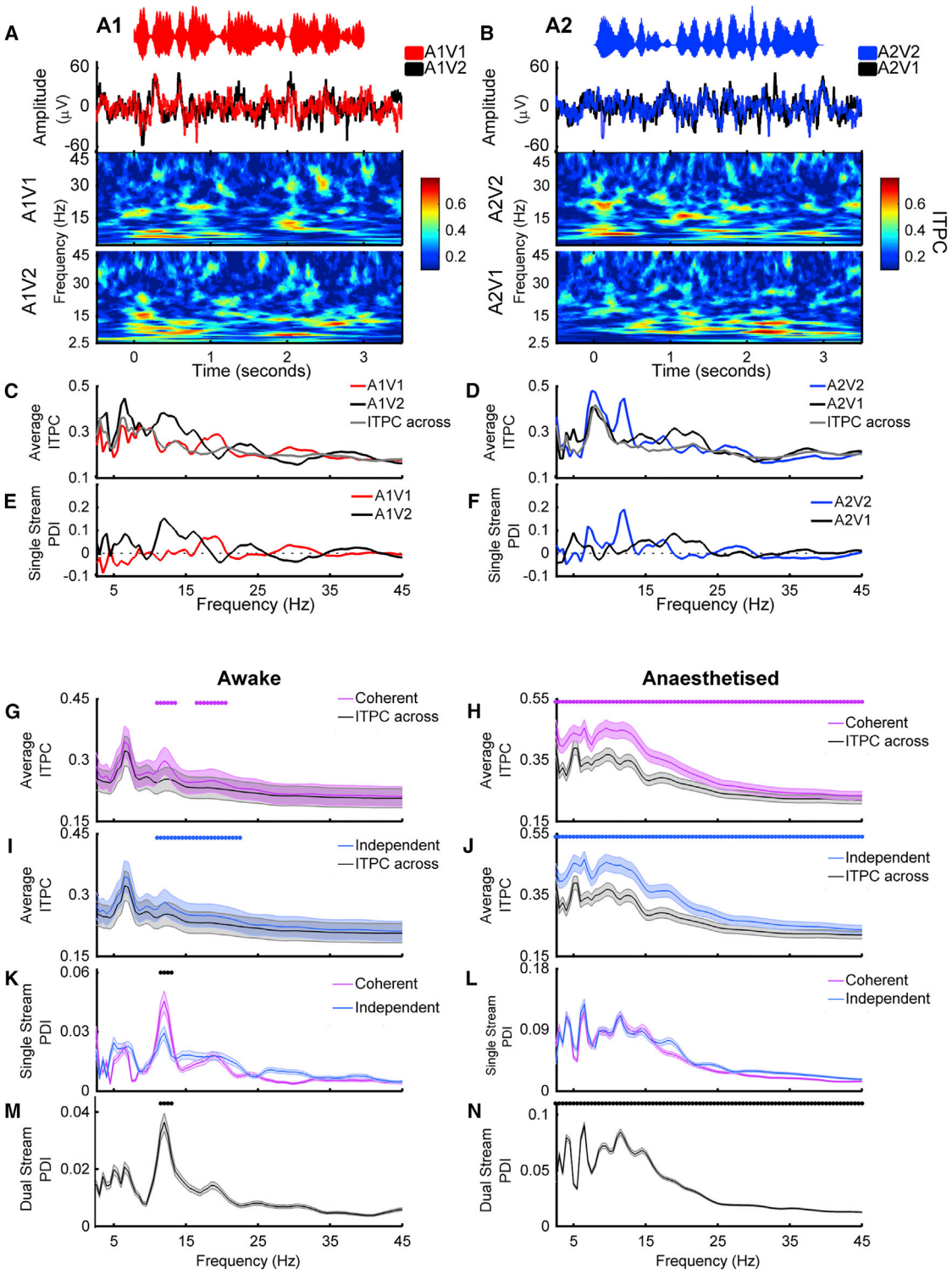


Figure 6. Visual Stimuli Elicit Reliable Changes in the Phase of the LFP

(A and B) Example LFP responses to single-stream auditory stimuli (A, A1 stream; B, A2 stream) across visual conditions. Data obtained from the recording site at which multi-unit spiking activity discriminated auditory stream identity in Figures 5A and 5B. The amplitude waveforms of the stimuli are shown in the top row, with the evoked LFP underneath (mean across 21 trials). The resulting inter-trial phase coherence (ITPC) values are shown in the bottom two rows, top row showing temporally coherent auditory-visual stimuli, bottom row showing temporally independent auditory-visual stimuli.

(legend continued on next page)

We hypothesized that the effects we observed in the dual-stream condition might be a consequence of temporal coherence between auditory and visual stimuli enhancing the discriminability of neural responses. We confirmed this prediction by using the same spike pattern decoder to compare our ability to discriminate temporally coherent (A1V1 versus A2V2) and temporally independent (A1V2 versus A2V1) stimuli (Figures 5E and 5F): temporally coherent auditory-visual stimuli produced more discriminable spike patterns than those elicited by temporally independent ones in both awake (Figure 5E, pairwise t test, auditory discriminating $t_{418} = 11.872$, $p < 0.001$; visual discriminating $t_{116} = 6.338$, $p < 0.001$; All $t_{540} = 13.610$, $p < 0.001$) and anesthetized recordings (Figure 5F, auditory discriminating $t_{482} = 17.754$, $p < 0.001$; visual discriminating $t_{162} = 8.186$, $p < 0.001$; all $t_{664} = 19.461$, $p < 0.001$). We further determined that neither the mean nor maximum evoked spike rates were different between trials in response to temporally coherent and temporally independent auditory visual stimuli (Figure S5). We also observed that the impact of auditory-visual temporal coherence was stronger in single units than multi-units in the awake dataset (Figure S4E). Therefore, the improved discrimination ability observed in response to temporally coherent auditory-visual stimuli is most likely to arise due to an increase in the reliability with which a spiking response occurred.

Dynamic Visual Stimuli Elicit Reliable Changes in LFP Phase

Temporal coherence between auditory and visual stimulus streams results in more discriminable spike trains in the single-stream condition and an enhancement of the representation of the temporally coherent sound when that sound forms part of an auditory scene. What might underlie the increased discriminability observed for temporally coherent cross-modal stimuli? The phase of on-going oscillations determines the excitability of the surrounding cortical tissue (Azouz and Gray, 1999; Okun et al., 2010; Szymanski et al., 2011). Local field potential (LFP) phase is reliably modulated by naturalistic stimulation (Chandrasekaran et al., 2010; Kayser et al., 2009; Luo and Poeppel, 2007; Ng et al., 2012; Schyns et al., 2011) and has been implicated in multisensory processing (Zion Golumbic et al., 2013; Lakatos et al., 2007). We hypothesized that sub-threshold visual inputs could modulate spiking activity by modifying the phase of the LFP such that, when visual-stimulus-induced changes in LFP phase coincided with auditory-stimulus evoked activity, the spiking precision in auditory cortex was enhanced.

Stimulus-evoked changes in the LFP were evident from the recorded voltage traces, and analysis of inter-trial phase coherence demonstrated that there were reliable changes in phase across repetitions of identical auditory-visual stimuli (Figures 6A and 6B). To isolate the influence of visual activity on the LFP at each recording site and address the hypothesis that visual stimuli elicited reliable changes in the LFP, we calculated phase and power dissimilarity functions for stimuli with identical auditory signals but differing visual stimuli (Luo and Poeppel, 2007). Briefly, this analysis assumes that if the phase within a particular frequency band differs systematically between responses to two different stimuli, then inter-trial phase coherence (ITPC) across repetitions of the same stimulus will be greater than across randomly selected stimuli. For each frequency band in the LFP, we therefore compared "within-stimulus" inter-trial phase coherence for responses to each stimulus (A1 stream Figure 6C; A2 stream Figure 6D) with "across-stimulus" inter-trial phase coherence calculated from stimuli with identical auditory components but randomly selected visual stimuli (e.g., randomly drawn from A1V1 and A1V2). The difference between within-stimulus and across-stimulus inter-trial phase coherence was then calculated across frequency and described as the phase dissimilarity index (PDI) (Figures 6E and 6F, single site example, Figures 6K and 6L, population data) with positive PDI values indicating reliable changes in phase coherence elicited by the visual component of the stimulus. Importantly, because both test distributions and the null distribution contain identical sounds, any significant PDI value can be attributed directly to the visual component of the stimulus.

We calculated PDI values for each of the four single-stream stimuli and grouped conditions by coherency (coherent: A1V1/A2V2, or independent: A1V2/A2V1). To determine at which frequencies the across-trial phase reliability was significantly positive, we compared the within-stimulus values with the across-stimulus values for each frequency band (paired t test with Bonferroni correction for 43 frequencies, $\alpha = 0.0012$). In awake subjects, we identified a restricted range of frequencies between 10.5 and 20 Hz where visual stimuli enhanced the phase reliability (Figures 6G and 6I). In anesthetized animals, average PDI values were larger than in awake animals, and all frequencies tested had single-stream PDI values that were significantly non-zero (Figures 6H and 6J). We therefore conclude that visual stimulation elicited reliable changes in the LFP phase in auditory cortex. In contrast to LFP phase, a parallel

(C and D) ITPC averaged across stimulus presentation time was calculated for each stimulus separately (C, A1V1 and A1V2; D, A2V2 and A2V1) and for trials with a randomly selected visual stimulus (ITPC across).

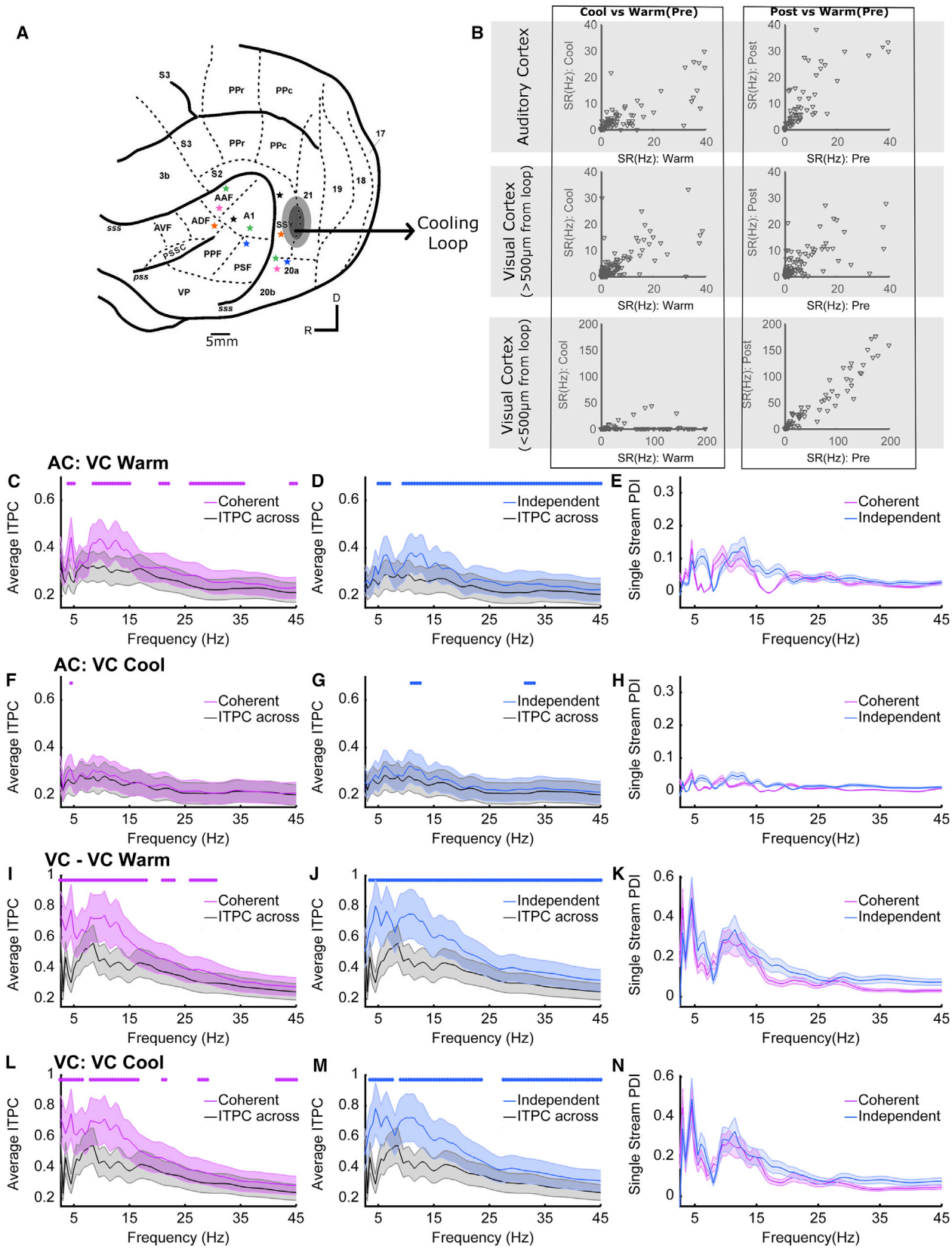
(E and F) Single-stream phase dissimilarity values (PDI) were calculated by comparing ITPC within values to the ITPC across null distributions for each stimulus class (E, A1V1 and A1V2; F, A2V2 and A2V1).

(G and H) Population mean ITPC values across frequency for temporally coherent stimuli (G, awake dataset, significant frequencies 10.5–13, 16–20 Hz; H, anesthetized dataset).

(I and J) Population mean ITPC values across frequency for temporally independent stimuli (I, awake dataset, significant frequencies 10.5–22 Hz; J, anesthetized dataset, no frequencies significantly different). Dots indicate frequencies at which the ITPC-within values were significantly greater than the ITPC-across values (pairwise t test, $\alpha = 0.0012$, Bonferroni corrected for 43 frequencies).

(K and L) Mean (±SEM) single-stream phase dissimilarity index (PDI) values for coherent and independent stimuli in awake animals (K, significant frequencies 10.5–12.5) and anesthetized (L) animals. Black dots indicate frequencies at which the temporally coherent single-stream PDI is significantly greater than in the independent conditions ($p < 0.001$).

(M and N) Mean (±SEM) dual-stream PDI values for awake (M, significant frequencies 10.5–12.5) and anesthetised (N) datasets.



(legend on next page)

analysis of across trial power reliability showed no significant effect of visual stimuli on LFP power in any frequency band (Figures S7A and S7C).

If visual information was only conveyed in the case of temporally coherent stimuli, this might indicate that the locus of binding was outside of auditory cortex and that the information being provided to auditory cortex already reflected an integrated auditory-visual signal. The LFP is thought to reflect the combined synaptic inputs to a region (Viswanathan and Freeman 2007), and so significant single-stream PDI values for both temporally independent and coherent stimuli suggest that the correlates of binding observed in auditory cortex were not simply inherited from its inputs. Since there were significant PDI values for both temporally independent and coherent stimuli, we next asked whether there were any frequencies at which phase coherence was significantly greater in auditory-visual stimuli that were temporally coherent compared to temporally independent. We performed a pairwise comparison of single-stream PDI values obtained from temporally coherent and independent stimuli, for all frequency points. In awake animals, PDI values were similar for temporally coherent and temporally independent stimuli, except in the 10.5- to 12.5-Hz band where coherent stimuli elicited significantly greater phase coherence (Figure 6K). In anesthetized animals, the single-stream PDI did not differ between coherent and independent stimuli at any frequency (Figure 6I). Together these data suggest that visual inputs modulate the phase of the field potential in auditory cortex largely independently of any temporal coherence between auditory and visual stimuli. This finding supports the conjecture that multisensory binding occurs within auditory cortex.

To understand whether the same mechanisms could underlie the visual-stimulus-induced enhancement of a temporally coherent sound in a mixture, we performed similar analyses on the data collected in response to the dual-stream stimuli. We generated within-stimulus inter-trial phase coherence values for each dual-stream stimulus (i.e., A12V1 and A12V2) and across-stimulus inter-trial phase coherence by randomly selecting responses across visual conditions. We then expressed the difference as the dual-stream phase dissimilarity index (dual-stream PDI, Figures 6M and 6N). Since the auditory components were identical in each dual-stream stimulus, the influence of the visual component on LFP phase could be isolated as non-zero dual-stream PDI values (paired t test, Bonferroni corrected, $\alpha = 0.0012$). In awake animals, the dual-stream PDI was significantly non-zero at 10.5–12.5 (Figure 6M), whereas in anesthetized animals, we found positive dual-stream PDI values across all frequencies tested (Figure 6N). In anesthetized animals, where we could use the responses of units to noise and light flashes to categorize units as auditory, visual, or auditory-visual, we confirmed significant dual-stream PDI values in the LFP recorded on the same electrode as units in each of these subpopulations (Figure S3L). In awake animals, we tested auditory-visual stimuli presented at three different modulation rates (7, 12, and 17 Hz) and confirmed that values were obtained at very similar LFP frequencies across these modulation rates—consistent with these being evoked phase alignments rather than stimulus-entrained oscillations (Figure S7I). Additional evidence for this hypothesis comes from the fact that, in the awake data, the frequencies at which the single- and dual-stream PDI values are significant are entirely non-overlapping with the modulation rate of the stimulus, which was band limited to 7 Hz.

Visual Cortex Mediates Visual-Stimulus-Induced LFP Changes in Auditory Cortex

Visual inputs to auditory cortex potentially originate from many sources: in the ferret, multiple visual cortical fields are known to innervate auditory cortex (Bizley et al., 2007), but frontal and thalamic areas are additional candidates for sources of top-down and bottom-up multisensory innervation. To determine the origin of the visual effects that we observe in auditory cortex, we performed an additional experiment in which we cooled the gyral surface of the posterior suprasylvian sulcus where visual cortical fields suprasylvian visual area (SSY) (Cantone et al., 2006) and area 21 (Innocenti et al., 2002) are located (Figure 7A). Neural tracer studies have demonstrated that these areas directly project to auditory cortex in the ferret (Bizley et al., 2007). We used a cooling loop cooled to 9°C–10°C to reversibly silence neural activity within <500 μm of the loop (see Figure 7B, see also Wood et al., 2017). Using simple noise bursts and light flashes at each site that we cooled, we verified that cooling visual cortex did not alter the response to noise bursts in auditory cortex (repeated-measures ANOVA on spike rates in response to a noise burst pre-cooling, during cooling, after cooling, $F(2,164) = 0.42$ $p = 0.88$) but did reversibly attenuate the spiking response to light flashes in visual cortical sites >500 μm from the cooling loop (repeated-measures ANOVA $F_{(2,92)} = 6.83$ $p = 0.001$, post hoc comparisons shows pre-cool and cool-post were significantly different, pre-post were not significantly different indicating the effects were reversible) and under the loop

Figure 7. Visual-Stimulus-Induced LFP Phase Changes in Auditory Cortex Are Mediated by Visual Cortex

(A) Schematic showing the location of auditory cortical recording sites and the location of a cooling loop (black, gray line marks the 500-μm radius over which cooling is effective; Wood et al., 2017), which was used to inactivate visual cortex. Individual recording sites contributing to (C)–(N) are shown with stars (simultaneous recordings are marked in the same color).

(B) Spike rate responses in auditory cortex (top row) and visual cortex (bottom row, sites >500 μm from the loop) in response to noise bursts or light flashes before, during, and after cooling.

(C and D) Inter-trial phase coherence values (mean ± SEM) for the coherent (C) and independent (D) auditory-visual stimuli recorded in auditory cortex (AC) prior to cooling visual cortex (VC) compared to the shuffled null distribution (inter-trial phase coherence across). Asterisks indicate the frequencies at which the inter-trial phase coherence values are significantly different from the shuffled inter-trial phase coherence-across distribution.

(E) Single-stream phase dissimilarity index values calculated from the inter-trial phase coherence values in (C) and (D).

(F–H) As in (C)–(E) but while visual cortex was cooled to 9 degrees.

(I–N) As in (C)–(H) but for sites in visual cortex >500 μm from the cooling loop. (C)–(H) include data from 83 sites from 6 electrode penetrations; (I)–(N) include data from 47 sites from five penetrations.

(F(2,210) = 30.2586; p = 2.8350e-12, pre-cool versus cooled, cooled versus post-cooled significantly different, pre-post not significantly different). We measured responses to the single-stream stimuli in auditory and visual cortex before and during cooling. From the LFP, we calculated the across-trial-phase coherence and phase dissimilarity indexes (as in Figure 6). Cooling visual cortex significantly decreased the magnitude of the single-stream PDI values in auditory cortex (Figures 7E and 7H). A 3-way repeated-measures ANOVA with factors of visual condition (coherent/independent), frequency, and cortical temperature (warm/cooled) on the single-stream PDI values obtained in auditory cortex showed a main effect of frequency ($F_{(88,22605)}$ = 47.91, $p < 0.1*10^{-9}$) and temperature ($F_{(1,22605)}$ = 1072, $p < 0.1*10^{-9}$) but not visual condition (p = 0.49). In contrast, LFP at recording sites in visual cortex away (>500 μm) from the loop were unaffected by cooling (3-way ANOVA demonstrated that the magnitude of the single-stream PDI value was influenced by frequency ($F_{(88,17265)} = 24.73$, $p < 1*10^{-9}$), but not temperature (p = 0.75) or visual condition (p = 0.29), Figures 7I–7N). From these data, we conclude that the influence of visual stimuli on the auditory cortical field potential phase is mediated, at least in part, by inputs from visual cortical areas SSY and 21. While cooling does not allow us to confirm that visual inputs are direct mono-synaptic connections (Bizley et al., 2016a), the observation that the phase effects in other areas of visual cortex are unaffected suggests that cooling selectively influenced communication between auditory and visual cortices rather than suppressing visual processing generally.

DISCUSSION

Here, we provide insight into how and where auditory-visual binding occurs and provide evidence that this effect is mediated by cortico-cortical interactions between visual and auditory cortex. While numerous studies have reported the incidence of auditory-visual interactions in auditory cortex over the past decade (Bizley et al., 2007; Chandrasekaran et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; 2010; Perrodin et al., 2015), evidence for their functional role has remained less apparent. Here, we show that one role for the early integration of auditory and visual information is to support auditory scene analysis. We observe the influence of visual stimuli in auditory cortex as reliable changes in the phase of the LFP, which occur irrespective of auditory-visual temporal coherence, indicating that the inputs to auditory cortex reflect the unisensory stimulus properties. When the visual and auditory stimuli are temporally aligned, activity elicited by the visual stimulus interacts with feedforward sound-evoked activity and results in spiking output that more precisely represents the temporally coherent sound within auditory cortex. These results are consistent with the binding of cross-modal information to form a multisensory object because they result in a modification of the representation of the sound that is not restricted to the features that link auditory and visual signals but extends to other non-binding features. These data provide a physiological underpinning for the pattern of performance observed in human listeners performing an auditory selective attention task, in which the detection of a perturbation in a stimulus stream is enhanced or impaired when a visual stimulus is temporally coherent with the target or masker auditory stream respectively (Maddox et al., 2015). The effects of the visual stimulus on the representation of an auditory scene can be observed in anesthetized animals suggesting that these effects can occur independently of attentional modulation.

Previous investigations of the impact of visual stimuli on auditory scene analysis have frequently used speech stimuli. Being able to see a talker's mouth provides listeners with information about the rhythm and amplitude of the speech waveform that may help listeners by cueing them to pay attention to the auditory envelope (Peelle and Sommers, 2015) as well as by providing cues to the place of articulation that can disambiguate different consonants (Sumby and Pollack, 1954). However, the use of speech stimuli makes it difficult to dissociate general multisensory mechanisms from speech-specific ones when testing in human subjects. Therefore, in order to probe more general principles across both human (Maddox et al., 2015) and non-human animals (here), we chose to employ continuous naturalistic non-speech stimuli that utilized modulation rates that fell within the range of syllable rates in human speech (Chandrasekaran et al., 2009) but lacked any linguistic content. Previous work has demonstrated that a visual stimulus can enhance the neural representation of the speech amplitude envelope both in quiet and in noise (Crosse et al., 2015, 2016; Luo et al., 2010; Park et al., 2016), but functional imaging methods make it difficult to demonstrate enhanced neural encoding of features beyond the amplitude envelope. The implication of our findings is that representation of the spectro-temporal features that allow speech recognition such as voice pitch would be enhanced in auditory cortex when a listener views a talker's face, even though such spectro-temporal features may not be represented by the visual stimulus.

Visual speech information is hypothesized to be relayed to auditory cortex through multiple routes in parallel to influence the processing of auditory speech: our data support the idea that early integration of visual information occurs (Möttönen et al., 2004; Okada et al., 2013; Peelle and Sommers, 2015; Schroeder et al., 2008) and is likely to reflect a general phenomenon whereby visual stimuli can cause phase-entrainment in the LFP. Within this framework, cross-modal binding potentially results from the temporal coincidence of evoked auditory responses and visual-stimulus elicited inputs that we observe as phasic changes of the LFP.

Consistent with previous studies, our analysis of LFP activity revealed that visual information reliably modulated LFP phase in auditory cortex (Chandrasekaran et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; Perrodin et al., 2015). This occurred independently of the modulation frequency of the stimulus suggesting that, rather than entraining oscillations at the stimulus modulation rate, relatively broadband phase resets are triggered by particular features within the stream (presumably points at which the luminance changed rapidly from low-high amplitude). The LFP reflects the synaptic inputs to a region and LFP phase synchronization is thought to arise from fluctuating inputs to cortical networks (Lakatos et al., 2007; Mazzoni et al., 2008; Szymanski et al., 2011). Since neuronal excitability varies with LFP phase (Jacobs et al.,

2007; Klimesch et al., 2007; Lakatos et al., 2013; Lőrincz et al., 2009), synaptic inputs from visual cortex may provide a physiological mechanism through which temporally coincident cross-sensory information is integrated. Our analysis allowed us to isolate changes in LFP phase that were directly attributable to the visual stimulus and to identify that reliable changes in LFP phase occurred irrespective of whether the visual stimulus was temporally coherent with the auditory stimulus. Such results suggest that the observed effects of cross-modal temporal coherence were not simply inherited within the inputs to auditory cortex. This is consistent with observations that unisensory visual stimuli can elicit reliable phase (but not power) effects in auditory cortex (Mercier et al., 2015). Moreover, the effects that we observed in the LFP were lost when we silenced visual cortex, indicating that inputs from visual cortex are a key contributor to the effects of auditory-visual temporal coherence that we observed in auditory cortex. Our finding that visual stimulation elicited reliable phase modulation in both awake and anesthetized animals suggests that bottom-up cross-modal integration interacts with selective attention, which has also been associated with modulation of phase information in auditory cortex (Zion Golumbic et al., 2013; Park et al., 2016). While our data suggest that cross-modal binding can occur in the absence of attention, it is likely that the additional neural pathways engaged during selective attention act to further enhance the representation of attended cross-modal objects.

In both awake and anesthetized animals, we observed three key findings: (1) that visual stimuli elicit robust effects on the LFP phase; (2) that auditory-visual temporal coherence shapes the response to a sound mixture such that temporally coherent auditory-visual stimuli are more reliably represented in the spiking response; and (3) that the spiking response to auditory timbre deviants (a non-binding feature) was enhanced. While these key findings were recapitulated in both states, there were some important differences. First, in the awake animal, the phase alignment in the LFP was generally smaller in magnitude and was only significantly modulated across a smaller range of frequencies (10.5–20 Hz as opposed to 4–45 Hz). Such differences are consistent with a dependence of oscillatory activity on behavioral state (Tukker et al., 2007; Voloh and Womelsdorf, 2016; Wang, 2010). Second, in the awake animal only, we observed a significant increase in the phase reliability (at 10.5–12.5 Hz) for temporally coherent auditory-visual stimuli when compared to temporally independent stimuli. Since the neural correlates of multisensory binding are evident in the anesthetized animal, the specific increase in alpha phase reliability that occurred only in awake animals in response to temporally coherent auditory-visual stimulus pairs (Figures 6K and 6M) may indicate an attention-related signal triggered by temporal coherence between auditory and visual signals, or an additional top-down signal conveying cross-modal information. Phase resetting and synchronization of alpha phase has been associated both with enhanced functional connectivity (Voloh and Womelsdorf, 2016) and as a top-down predictive signal for upcoming visual information (Samaha et al., 2015). Understanding how attention engages additional brain networks and disambiguating these possibilities would require simultaneous recordings in auditory and visual cortex recording while trained animals performed the auditory selective attention task which motivated this study. Finally, in awake and anesthetized animals, we observed that the impact of auditory-visual temporal coherence on the representation of sound mixtures (as assessed by visual preference scores) was of a similar magnitude in the primary areas (A1 and anterior auditory field [AAF], located in the middle ectosylvian gyrus [MEG]). In contrast, in the awake animal, neurons in the PEG, where secondary tonotopic fields posterior pseudosylvian field (PPF) and posterior suprasylvian field (PSF) are located, had significantly higher VPI scores than those in the MEG, while in anesthetized animals VPI scores were statistically indistinguishable across cortical fields. This suggests that in the awake animal additional mechanisms exist to enhance the effects that are present in the primary areas. These results were mirrored in the impact of auditory-visual temporal coherence on non-binding features (as assessed by the impact of auditory-visual temporal coherence on deviant detection ability) where the visual stimulus had a stronger influence in PEG than MEG in the awake animal and did not differ across regions (and was overall of a smaller magnitude) in anesthetized animals. Our cooling studies (in anesthetized animals) do not allow us to determine whether this enhancement reflects the greater variety of inputs from visual cortex that terminate in secondary as opposed to primary auditory cortex (Bizley et al., 2007), top-down inputs from higher areas (e.g., parietal or frontal cortex), or are a consequence of intracortical processing within auditory cortex.

Temporal coherence between sound elements has been proposed as a fundamental organizing principle for auditory cortex (Elhilali et al., 2009; O'Sullivan et al., 2015), and here we extend this principle to the formation of cross-modal constructs. Our data provide evidence that one role for the early integration of visual information into auditory cortex is to resolve competition between multiple sound sources within an auditory scene and that these neural computations occur pre-attentively. While some proponents of a temporal coherence-based model for auditory streaming have stressed the importance of attention in auditory stream formation (Lu et al., 2017), neural signatures of temporal-coherence-based streaming are present in passively listening subjects (O'Sullivan et al., 2015; Teki et al., 2016). Previous studies have demonstrated a role for visual information in conveying lip movement information to auditory cortex (Chandrasekaran et al., 2013; Crosse et al., 2015; Ghazanfar et al., 2005; Zion Golumbic et al., 2013), but such stimuli make it difficult to separate sensory information from linguistic cues. Our data obtained using non-speech stimuli provide evidence that at least part of the boost provided by visualizing a speaker's mouth arises from a more general (language-independent) phenomenon whereby visual temporal cues facilitate auditory scene analysis through the formation of cross-sensory objects. Our data are supportive of visual cortical areas providing at least one source of information. Other visual cortex fields and sub-cortical structures innervate tonotopic auditory cortex (Bizley et al., 2007; Budinger et al., 2006) and may potentially provide additional visual inputs to auditory cortex. Further

dissecting the origin of visual innervation requires experiments that allow pathway specific manipulation of neuronal activity (for example, by silencing the terminal fields of neurons that project from a candidate area into auditory cortex, Bizley et al., 2016a).

Finally, the neural correlates of multisensory binding were apparent in units that best discriminated either the auditory or visual characteristics of single auditory-visual streams, although the magnitude of the effects was stronger in visual-discriminating units. Nevertheless, both classes of neurons showed enhanced encoding of temporally coherent versus temporally independent auditory visual streams, suggesting that both subgroups could be described as "auditory-visual." This was confirmed in anesthetized animals where neurons were additionally characterized with simple stimuli (noise bursts and light flashes) and revealed that 54% of visual-discriminating stimuli and 41% of auditory-discriminating neurons were classified as auditory-visual—that is, they either responded to both modalities or had their response to one modality modulated by the other. Together, these results suggest that multisensory processing is prevalent throughout auditory cortex and that cross-sensory processing has the potential to have a significant impact on the representation of acoustic features in auditory cortex.

In summary, activity in auditory cortex was reliably affected by visual stimulation in a manner that enhanced the representation of temporally coherent auditory information. Enhancement of auditory information was observed for sounds presented alone or in a mixture and for sound features that were related to (amplitude) and orthogonal to (timbre) variation in visual input. Such processes provide mechanistic support for a coherence-based model of cross-modal binding in object formation and indicate that one role for the early integration of visual information in auditory cortex is to support auditory scene analysis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Animal preparation
  - Cortical cooling
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Spiking responses
  - Classification as auditory or visual with simple stimuli
  - Analysis of responses during cooling

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and can be found with this article online at https://doi.org/10.1016/j.neuron.2017.12.034.

ACKNOWLEDGMENTS

This work was funded by grants to the following authors: J.K.B.: Wellcome Trust / Royal Society (WT098418MA), Biotechnology and Biological Sciences Research Council (BB/H016813/1), and an Action on Hearing Loss Studentship (596: UEI: JB); R.K.M.: NIH (R00DC014288) and a Hearing Health Foundation Emerging Research grant; A.K.C.L.: NIH (R01DC013260); and J.K.B. and A.K.C.L.: an International Exchanges Scheme award from the Royal Society.

AUTHOR CONTRIBUTIONS

Conception and Design, H.A., R.K.M., A.K.C.L., and J.K.B.; Data Acquisition, H.A., S.M.T., K.C.W., G.P.J., and J.K.B.; Data Analysis and Interpretation, H.A. and J.K.B., Drafting or Revising the Article, H.A., S.M.T., R.K.M., A.K.C.L., and J.K.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 26, 2017
Revised: October 28, 2017
Accepted: December 22, 2017
Published: January 25, 2018

REFERENCES

Antunes, F.M., and Malmierca, M.S. (2011). Effect of auditory cortex deactivation on stimulus-specific adaption in the medial geniculate body. J Neurosci. *31*, 17306–17316.

Azouz, R., and Gray, C.M. (1999). Cellular mechanisms contributing to response variability of cortical neurons in vivo. J. Neurosci. *19*, 2209–2223.

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. J. Stat. Software *31*. Published online September 23, 2009. https://doi.org/10.18637/jss.v031.i10.

Bizley, J.K., Nodal, F.R., Bajo, V.M., Nelken, I., and King, A.J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb. Cortex *17*, 2172–2189.

Bizley, J.K., Walker, K.M.M., Silverman, B.W., King, A.J., and Schnupp, J.W.H. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J. Neurosci. *29*, 2064–2075.

Bizley, J.K., Walker, K.M.M., King, A.J., and Schnupp, J.W.H. (2013). Spectral timbre perception in ferrets: Discrimination of artificial vowels under different listening conditions. J. Acoust. Soc. Am. *133*, 365–376.

Bizley, J.K., Jones, G.P., and Town, S.M. (2016a). Where are multisensory signals combined for perceptual decision-making? Curr. Opin. Neurobiol. *40*, 31–37.

Bizley, J.K., Maddox, R.K., and Lee, A.K. (2016b). Defining auditory-visual objects: Behavioral tests and physiological mechanisms. Trends Neurosci. *39*, 74–85.

Brosch, M., Selezneva, E., and Scheich, H. (2015). Neuronal activity in primate auditory cortex during the performance of audiovisual tasks. Eur. J. Neurosci. *41*, 603–614.

Budinger, E., Heil, P., Hess, A., and Scheich, H. (2006). Multisensory processing via early cortical stages: Connections of the primary auditory cortical field with other sensory systems. Neuroscience *143*, 1065–1083.

Cantone, G., Xiao, J., and Levitt, J.B. (2006). Retinotopic organization of ferret suprasylvian cortex. Vis. Neurosci. *23*, 61–77.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. PLoS Comput. Biol. *5*, e1000436.

Chandrasekaran, C., Turesson, H.K., Brown, C.H., and Ghazanfar, A.A. (2010). The influence of natural scene dynamics on auditory cortical activity. J. Neurosci. *30*, 13919–13931.

Chandrasekaran, C., Lemus, L., and Ghazanfar, A.A. (2013). Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. Proc. Natl. Acad. Sci. USA *110*, E4668–E4677.

Crosse, M.J., Butler, J.S., and Lalor, E.C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. *35*, 14195–14204.

Crosse, M.J., Di Liberto, G.M., and Lalor, E.C. (2016). Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J. Neurosci. *36*, 9888–9895.

Denison, R.N., Driver, J., and Ruff, C.C. (2013). Temporal structure and complexity affect audio-visual correspondence detection. Front. Psychol. *3*, 619.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. *18*, 193–222.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., and Shamma, S.A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron *61*, 317–329.

Foffani, G., and Moxon, K.A. (2004). PSTH-based classification of sensory stimuli using ensembles of single neurons. J. Neurosci. Methods *135*, 107–120.

Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., and Logothetis, N.K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J. Neurosci. *25*, 5004–5012.

Innocenti, G.M., Manger, P.R., Masiello, I., Colin, I., and Tettoni, L. (2002). Architecture and callosal connections of visual areas 17, 18, 19 and 21 in the ferret (Mustela putorius). Cereb. Cortex *12*, 411–422.

Jacobs, J., Kahana, M.J., Ekstrom, A.D., and Fried, I. (2007). Brain oscillations control timing of single-neuron activity in humans. J. Neurosci. *27*, 3839–3844.

Kaur, S., Rose, H.J., Lazar, R., Liang, K., and Metherate, R. (2005). Spectral integration in primary auditory cortex: Laminar processing of afferent input, in vivo and in vitro. Neuroscience *134*, 1033–1045.

Kayser, C., Petkov, C.I., and Logothetis, N.K. (2008). Visual modulation of neurons in auditory cortex. Cereb. Cortex *18*, 1560–1574.

Kayser, C., Petkov, C.I., and Logothetis, N.K. (2009). Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. Hear. Res. *258*, 80–88.

Kayser, C., Logothetis, N.K., and Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. Curr. Biol. *20*, 19–24.

Kelly, J.B., Judge, P.W., and Phillips, D.P. (1986). Representation of the cochlea in primary auditory cortex of the ferret (Mustela putorius). Hear Res. *24*, 111–115.

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. Brain Res. Brain Res. Rev. *53*, 63–88.

Lakatos, P., Chen, C.-M., O'Connell, M.N., Mills, A., and Schroeder, C.E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron *53*, 279–292.

Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C., and Schroeder, C.E. (2013). The spectrotemporal filter mechanism of auditory selective attention. Neuron *77*, 750–761.

Lőrincz, M.L., Kékesi, K.A., Juhász, G., Crunelli, V., and Hughes, S.W. (2009). Temporal framing of thalamic relay-mode firing by phasic inhibition during the alpha rhythm. Neuron *63*, 683–696.

Lu, K., Xu, Y., Yin, P., Oxenham, A.J., Fritz, J.B., and Shamma, S.A. (2017). Temporal coherence structure rapidly shapes neuronal interactions. Nat. Commun. *8*, 13900.

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron *54*, 1001–1010.

Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. *8*, e1000445.

Maddox, R.K., Atilgan, H., Bizley, J.K., and Lee, A.K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. eLife *4*, e04995.

Mazzoni, A., Panzeri, S., Logothetis, N.K., and Brunel, N. (2008). Encoding of naturalistic stimuli by local field potential spectra in networks of excitatory and inhibitory neurons. PLoS Comput. Biol. *4*, e1000239.

Mercier, M.R., Molholm, S., Fiebelkorn, I.C., Butler, J.S., Schwartz, T.H., and Foxe, J.J. (2015). Neuro-oscillatory phase alignment drives speeded multisensory response times: An electro-corticographic investigation. J. Neurosci. *35*, 8546–8557.

Möttönen, R., Schürmann, M., and Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. Neurosci. Lett. *363*, 112–115.

Ng, B.S.W., Schroeder, T., and Kayser, C. (2012). A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. J. Neurosci. *32*, 12268–12276.

O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. J. Neurosci. *35*, 7256–7263.

Okada, K., Venezia, J.H., Matchin, W., Saberi, K., and Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. PLoS ONE *8*, e68959.

Okun, M., Naim, A., and Lampl, I. (2010). The subthreshold relation between cortical local field potential and neuronal firing unveiled by intracellular recordings in awake rats. J. Neurosci. *30*, 4440–4448.

Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. eLife *5*. Published online May 5, 2016. https://doi.org/10.7554/eLife.14521.

Peelle, J.E., and Sommers, M.S. (2015). Prediction and constraint in audiovisual speech perception. Cortex *68*, 169–181.

Perrodin, C., Kayser, C., Logothetis, N.K., and Petkov, C.I. (2015). Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. Proc. Natl. Acad. Sci. USA *112*, 273–278.

Quiroga, R.Q., Nadasdy, Z., and Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. *16*, 1661–1687.

Rahne, T., Deike, S., Selezneva, E., Brosch, M., König, R., Scheich, H., Böckmann, M., and Brechmann, A. (2008). A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming. Brain Res. *1220*, 118–131.

Samaha, J., Bauer, P., Cimaroli, S., and Postle, B.R. (2015). Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. Proc. Natl. Acad. Sci. USA *112*, 8439–8444.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. *12*, 106–113.

Schyns, P.G., Thut, G., and Gross, J. (2011). Cracking the code of oscillatory activity. PLoS Biol. *9*, e1001064.

Sumby, W.H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. *26*, 212–215.

Szymanski, F.D., Rabinowitz, N.C., Magri, C., Panzeri, S., and Schnupp, J.W. (2011). The laminar and temporal structure of stimulus information in the phase of field potentials of auditory cortex. J. Neurosci. *31*, 15787–15801.

Teki, S., Barascud, N., Picard, S., Payne, C., Griffiths, T.D., and Chait, M. (2016). Neural correlates of auditory figure-ground segregation based on temporal coherence. Cereb. Cortex *26*, 3669–3680.

Town, S.M., Atilgan, H., Wood, K.C., and Bizley, J.K. (2015). The role of spectral cues in timbre discrimination by ferrets and humans. J. Acoust. Soc. Am. *137*, 2870–2883.

Town, S.M., Wood, K.C., and Bizley, J.K. (2017). Neural correlates of perceptual constancy in Auditory Cortex. BioXriv https://www.biorxiv.org/content/early/2017/07/15/102889.

Tukker, J.J., Fuentealba, P., Hartwich, K., Somogyi, P., and Klausberger, T. (2007). Cell type-specific tuning of hippocampal interneuron firing during gamma oscillations in vivo. J. Neurosci. *27*, 8184–8189.

Viswanathan, A., and Freeman, R.D. (2007). Neurometabolic coupling in cerebral cortex reflects synaptic more than spiking activity. Nat. Neurosci. *10*, 1308–1312.

Voloh, B., and Womelsdorf, T. (2016). A role of phase-resetting in coordinating large scale neural networks during attention and goal-directed behavior. Front. Syst. Neurosci. *10*, 18.

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. Physiol. Rev. *90*, 1195–1268.

Wood, K.C., Town, S.M., Atilgan, H., Jones, G.P., and Bizley, J.K. (2017). Acute inactivation of primary auditory cortex causes a sound localisation deficit in ferrets. PLoS ONE *12*, e0170264.

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J. Neurosci. *33*, 1417–1426.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Ferret (wild type)	Highgate Farms	N/A
Software and Algorithms		
(MATLAB) Circular Statistic Toolbox	Berens, 2009	https://philippberens.wordpress.com/code/circstats/
OpenEx and OpenDeveloper	Tucker Davis Technologies	http://www.tdt.com/downloads.html
MATLAB (versions 2012b, 2014b)	MathWorks	RRID:SCR_001622
Preference Index/ Inter trial Phase Coherence	Luo and Poeppel, 2007	N/A
Current Density Analysis	Kaur et al., 2005	N/A
Euclidean distance based pattern classifier	Foffani and Moxon, 2004	N/A
WaveClus spike-sorting algorithm	Quiroga et al., 2004	https://github.com/csn-le/wave_clus
Other		
TDT System 3 hardware	Tucker Davis Technologies	http://www.tdt.com

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents (data and MATLAB code) should be directed to and will be fulfilled by the Lead Contact, Jennifer Bizley (j.bizley@ucl.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The experiments were approved by the Animal Welfare and Ethical Review Board of University College London and The Royal Veterinary College, and performed under license from the UK Home Office (PPL 70/7267) and in accordance with the Animals Scientific Procedures Act 1986.

Neural responses were recorded in a total of 19 awake pigmented adult female ferrets (Mustela putorius furo; 1-5 years old). Fourteen of these animals contributed data to the awake dataset: Data from 9 of these animals was used for the main experiment (532 units), data from 11 other animals (6/9 in the main experiment, plus five other ferrets, totalling 128 units) was collected for additional control analysis (Figures 3E and S6). Females (700-1500 g, wild-type) were co-housed in groups of 2-9. These animals were trained in a variety of psychoacoustic tasks unrelated to the current study prior to and after the implantation of recording electrodes. Animals were tested for this study on days when they were not participating in psychoacoustic testing. Five adult females were used to record responses under anesthesia.

METHOD DETAILS

Animal preparation

Full methods for recording under anesthesia can be found in Bizley et al., (2009). Briefly, ferrets were anesthetized with medetomidine (Domitor; 0.022mg/kg/h; Pfizer, Sandwich, UK) and ketamine (Ketaset; 5mg/kg/h; Fort Dodge Animal Health, Southampton, UK). The animal was intubated and the left radial vein was cannulated in order to provide a continuous infusion (5 mL/h) of a mixture of medetomidine and ketamine in lactated ringers solution augmented with 5% glucose, atropine sulfate (0.06 mg/kg/h; C-Vet Veterinary Products) and dexamethasone (0.5 mg/kg/h, Dexadreson; Intervet, UK). The ferret was placed in a stereotaxic frame in order to implant a bar on the skull, enabling the subsequent removal of the stereotaxic frame. The left temporal muscle was largely removed, and the suprasylvian and pseudosylvian sulci were exposed by a craniotomy, revealing auditory cortex (Kelly et al., 1986). The dura was removed over auditory cortex and the brain protected with 3% agar solution. The eyes were protected with zero-refractive power contact lenses. The animal was then transferred to a small table in a sound-attenuating chamber. Body temperature, end-tidal CO_2, and the electrocardiogram were monitored throughout the experiment. Experiments typically lasted between 36 and 56 h. Neural activity was recorded with multisite silicon electrodes (Neuronexus Technologies) in a 1x 16, 2x 16 or 4x 8 (shank x number of sites) configuration. For experiments in which visual cortex was cooled, we extended the craniotomy caudally to expose visual cortex and placed a cooling loop over the posterior suprasylvian gyrus. Details of the manufacture of the cooling loop and validation of its efficacy in the ferret animal model are provided in full in (Wood et al., 2017).

Full surgical methods for implanting recording electrode arrays to facilitate recording from awake animals are available in Bizley et al. (2013). Briefly, animals were bilaterally implanted with WARP-16 drives (Neuralynx, Montana, USA) loaded with high impedance tungsten electrodes (FHC, Bowdoin, USA) under general anesthesia (medetomidine and ketamine induction, as above, isoflurane maintenance 1%–3%). Craniotomies were made over left and right auditory cortex, a small number of screws were inserted into the skull for anchoring and grounding the arrays, and the WARP-16 drive was anchored with dental acrylic and protected with a capped well. Recording electrodes in awake animals targeted tonotopic auditory cortex (area MEG, containing fields A1 and AAF, and PEG, tonotopic belt areas PPF and PSF are located). Auditory fields were estimated prior to implantation based on known sulcal landmarks and confirmed with regular assessments of frequency tuning and post-mortem histology. Animals were allowed to recover for a week before the electrodes were advanced into auditory cortex. Pre-operative, peri-operative and post-operative analgesia were provided to animals under veterinary advice. Recordings were made over the next 1-2.5 years, with electrodes individually advanced every few weeks until the thickness of auditory cortex was traversed. Recordings were made while animals were passively listening/watching stimuli and holding their head at a waterspout. During the recording a continuous stream of water was delivered from the spout.

Stimulus Presentation

All stimuli were created using TDT System 3 hardware (Tucker-Davis Technologies, Alachua, FL) and controlled via MATLAB (Mathworks, USA). For recordings in awake animals, sounds were presented over two loud speakers (Visaton FRS 8). Water deprived ferrets were placed in a dimly lit testing box (69 × 42 × 52 cm length x width x height) and received water from a central reward spout located between the two speakers. Sound levels were calibrated using a Brüel and Kjær (Norcross, GA) sound level meter and free-field 1/2-inch microphone (4191). Auditory streams were presented at 65 dB SPL (Figure 1A). Visual stimuli were delivered by illuminating the spout with a white LED which provided full field illumination (Precision Gold N76CC Luxmeter, 0 to 36.9 lux). The animals were not required to do anything other than maintain their heads in position at the spout where they were freely rewarded. Recording was terminated when animals were sated.

For anesthetised recordings, acoustic stimuli were presented using Panasonic headphones (Panasonic RP-HV297, Bracknell, UK) at 65 dB SPL. Visual stimuli were presented with a white Light Emitting Diode (LED) which was placed in a diffuser at a distance of roughly 10 cm from the contralateral eye so that it illuminated virtually the whole contralateral visual field.

Stimuli and data acquisition

Auditory stimuli were artificial vowel sounds that were created in MATLAB (MathWorks, USA). In the behavioral experiment that motivated this study (Maddox et al., 2015), stimuli were 14 s in duration. However, we adapted the stimulus duration in awake recordings to 3 s in order to collect sufficient repetitions of all stimuli, and to ensure animals maintained their head position facing forward for the whole trial duration. The animals were observed constantly via a webcam and recording was terminated / paused if the animal's head moved from the center spout. In the anesthetised recording, stimulus streams were 14 s long, as in the human psychophysics but we only analyzed the first 3 s to ensure datasets were directly comparable (see also Figure S7 e-h which replicates analysis for 3 s and 14 s stimuli).

Stimulus A1 was the vowel [u] (formant frequencies F1-4: 460, 1105, 2857, 4205 Hz, F0 = 195Hz), A2 was [a] (F1-4: 936, 1551, 2975, 4263 Hz, F0 = 175Hz). Streams were amplitude modulated with a noisy lowpass (7 Hz cut-off) envelope. Unless specifically noted, the timbre of the auditory stream remained fixed throughout the trial. However, we also recorded responses to auditory streams that included brief timbre deviants. As in our previous behavioral study, deviants were 200ms epochs in which the identity of the vowel was varied by smoothly changing the first and second formant frequencies to and from those identifying another vowel. Stream A1 was morphed to/from [ɛ] (730, 2058, 2857, 4205 Hz) and A2 to/from [i] (437, 2761, 2975, 4263 Hz).

Visual stimuli were generated using an LED whose luminance was modulated with dynamics that matched the amplitude modulation applied to A1 or A2. In single stream conditions a single auditory and single visual stream were presented (e.g., A1V1, A1V2, A2V1, or A2V2) whereas in dual stream conditions both auditory streams were presented simultaneously, accompanied by a single visual stimulus (A12V1, A12V2) (Figure 1E). Auditory streams were always presented from both speakers so that spatial cues could not facilitate segregation, and stimulus order was varied pseudo-randomly. In the anesthetised recordings each stimulus was presented 20 times. In the awake dataset, recording duration was determined by how long the ferret remained at the central location (mean repetitions: 20, minimum: 14, maximum: 34).

During anaesthetised recordings, pure tone stimuli (150 Hz to 19 kHz in 1/3-octave steps, from 10 to 80 dB SPL in 10 dB, 100 ms in duration, 5 ms cosine ramped) were also presented. These allowed us to characterize individual units and determine tonotopic gradients, so as to confirm the cortical field in which any given recording was made. Additionally broadband noise bursts and diffuse light flashes (100 ms duration, 70 dB SPL) were presented and used to classify a unit as auditory, visual or auditory visual. LFPs were subjected to current source density analysis to identify sources and sinks as described by Kaur et al. (2005).

Cortical cooling

During these experiments we made joint recordings in visual cortex (usually > 500 μm from the cooling loop, in order to determine whether visual cortical processing was impaired generally) and auditory cortex simultaneously. We recorded responses to the single stream stimuli before and during cooling, and, at each site additionally recorded responses to noise bursts and light flashes before, during, and after cooling. We used the responses to simple stimuli such as these to show that we could recover the original spiking responses (and data were excluded from any recording sites in which did not return to within 20% of their pre-cooling spike rates

(a common criterion used in cooling studies: e.g., Antunes and Malmierca, 2011). We did not record responses to the longer stimuli used in this study in the post-cooling condition as the additional recording time for these stimuli would have compromised our ability to record across several different sites in each animal.

QUANTIFICATION AND STATISTICAL ANALYSIS

Electrophysiological data were analyzed offline. Spiking activity and local field potential signals were extracted from the broadband voltage waveform by filtering at 0.3-5kHz and 1-150 Hz respectively. Spikes were detected, extracted, and then sorted with a spike-sorting algorithm (WaveClus, Quiroga et al., 2004).

Spiking responses

We used a Euclidean distance based pattern classifier (Foffani and Moxon, 2004) with leave-one-out cross validation to determine whether the neuronal responses to different stimuli could be discriminated. Spiking responses to a given stimulus were binned into a series of spike counts from stimulus onset (0 s) to offset (3 s) in 20 ms bins. The average across-repetition responses to each stimulus (excluding the to-be-classified response) was calculated and used as templates for decoding. Each single trial response was then classified by calculating the Euclidean distance between itself and the templates and assigning it to the stimulus class with the closest template. To determine whether the classifier performed significantly better than expected by chance, a 1000 iteration permutation test was performed where trials were drawn (with replacement) from the observed data and randomly assigned to stimulus classes, before decoding was repeated. A neural response was considered to be significantly informative about stimulus identity if the observed decoding value exceeded the 95th percentile of the distribution of decoding values from the randomly drawn data.

This approach allowed us to classify units according to their functional properties: auditory units discriminated two auditory stimuli based on the amplitude modulation of sound (A1 versus A2) regardless of visual dynamics, (Figures 5A and 5B), visual units discriminated visual presentations based on temporal envelope of visual stimuli (V1 versus V2) regardless of auditory presentation (Figures 5C and 5D) and auditory-visual units could do both. This approach was extended to classify dual stream responses by using the average response to each of the temporally coherent single stream stimuli (A1V1 or A2V2) as templates (Figures 2, 3, and S2–S4). Performance was (arbitrarily) expressed as the proportion of responses classified as being from the A1, and compared for the two dual stream stimuli with different visual conditions (Figure 5). All units in which either auditory or visual stimulus identity could be decoded were included in the main dual-stream analysis (Figure 2). For the control no-visual single stream case (Figure 3) all driven units were included as it was not possible to additionally collect responses to all possible permutations of the single stream stimuli that were necessary for the classification as auditory/visual discriminating. A Visual Preference Index was derived from this measure as the difference between the percentage of A12V1 trials labeled A1 and the percentage of A12V2 trials labeled A1. Therefore units which were fully influenced by the identity of the visual stimulus would have a visual preference score of 100, while those in which the visual stimulus did not influence the response at all would have a score around 0 (Figures 2E and 2H). We then assessed the significance of observed VPI scores using a permutation test ($p < 0.05$) in which the identity of single stream trials used to generate classifier templates was shuffled and the VPI recalculated for 1000 iterations.

Timbre deviant analysis

In order to determine how a visual stimulus influenced the ability to decode timbre deviants embedded within the auditory streams we used the cross-validated pattern classifier described above for analyzing single stream stimuli to discriminate deviant from no-deviant trials. Responses were considered over the 200 ms time window that the deviant occurred (or the equivalent time point in the no-deviant stimulus) binned with a 10 ms resolution. Significance was assessed by a 1000 iteration permutation test in which trials were randomly drawn with replacement from deviant and no-deviant responses. The discrimination score was calculated as the proportion of trials correctly classified.

Classification as auditory or visual with simple stimuli

During recordings made under anesthesia, we also recorded responses to noise bursts and light flashes (both 100 ms duration) presented separately and together to compare how the proportion of auditory / visual discriminating units measured to naturalistic dynamic stimuli compared to more traditional artificial stimuli. Specifically, responsiveness was defined using a two-way ANOVA (factors: auditory stimulus [on/off] and visual stimulus [on/off]) on spike counts measured during stimulus presentation. We defined units as being sound-driven (main effect of auditory stimulus, no effect of visual stimulus or interaction), light-driven (main effect of visual stimulus, no effect of auditory stimulus or interaction) or auditory-visual (main effect of both auditory and visual stimuli or significant interaction; $p < 0.05$) as in Bizley et al., 2007.

Phase/power dissimilarity analysis

Local field potential recordings were considered for all sites at which there was a significant driven spiking response, irrespective of whether that response could discriminate auditory or visual stream identity. For the single stream trials, we computed a single stream Phase Dissimilarity Index (PDI), which characterizes the consistency and uniqueness of the temporal phase/power pattern of neural responses to continuous auditory stimuli (Luo and Poeppel, 2007). This analysis compares the phase (or power) consistency across repetitions of the same stimulus against a baseline of phase-consistency across trials in which different stimuli were presented.

In the first stage of PDI analysis, we obtained a time-frequency representation of each response using wavelet decomposition with complex 7-cycle Morlet wavelets in 0.5 steps between 2.5–45 Hz, resulting in 86 frequency points. Next, we calculated the inter-trial phase-coherence value (ITPC; Equation 1) at each time-frequency point, across all trials in which the same stimulus was presented. For each frequency band, the ITPC time-course was averaged over the duration of the analysis window and across all repetitions to obtain the average *within-stimulus ITPC*.

$$ITPC_{t,f} = \left| \frac{\sum_{k=1}^{N} e^{i\theta_{k,t,f}}}{N} \right| \quad \text{Equation 1}$$

in which N is equal to the number of trials, and θ is the phase of trial k at a given frequency (f) and time (t). The *across-stimuli ITPC* was estimated using the same approach but using shuffled data, such that the ITPC was computed across trials with the same auditory stimulus but randomly drawn visual stimuli. The single stream phase dissimilarity index (Single stream PDI) was computed as the difference between the ITPC value calculated for *within* trials and the ITPC values calculated *across* visual trials (Equation 2). The dissimilarity function for each frequency bin i was defined as;

$$Single\ Stream\ PDI_i = \frac{\sum_{j=1}^{N} ITPC_{ij}\ within_{vis}}{N} - \frac{\sum_{j=1}^{N} ITPC_{ij},\ across_{vis}}{N}. \quad \text{Equation 2}$$

Large positive PDI indicate that responses to individual stimuli have a highly consistent response on single trials. Single stream PDI values were calculated for each stimulus type and then averaged across stimuli to calculate values for temporally coherent and temporally independent auditory visual stimuli. Single stream PDI was positive if within stimulus ITPC was larger than across-stimulus ITPC (pairwise t test, $p < 0.05$ Bonferroni correction for 86 frequencies points) and was considered significant if a minimum of 2 adjacent bins exceeded the corrected threshold. PDI magnitude values were calculated by summing the PDI values across all significant frequencies.

Dual stream phase dissimilarity index (dual stream PDI) values were calculated by extending this approach for dual stream stimuli with the goal of determining how the temporal envelope of the visual stimulus influenced the neural response to a sound mixture. To this end, we calculated the *within-dual ITPC* from the A12V1 trials and A12V2 trials separately and *across-dual ITPC* by randomly selecting trials from both stimuli (Equation 3). The within-dual and across-dual ITPCs were then averaged over time and subtracted to yield the dual stream PDI (Equation 3).

$$Dual\ Stream\ PDI_i = \frac{\sum_{j=1}^{N} ITPC_{ij}\ within_{dual}}{N} - \frac{\sum_{j=1}^{N} ITPC_{ij}\ across_{dual}}{N}. \quad \text{Equation 3}$$

Positive dual stream PDI values indicate that the time course of the neural responses was influenced by visual input, despite the identical acoustic input. We determined whether the dual stream PDI was greater if the *within_dual ITPC* was significantly larger than *across_dual ITPC* (pairwise t test, $p < 0.05$ Bonferroni correction, as above). PDI magnitude values were calculated by summing the PDI values across all significant frequencies.

Analysis of responses during cooling

To measure the effects of cooling at different distances from the cooling loop we used control stimuli (light flashes and noise bursts as described above) before, during, and after lowering the temperature of the cooling loop to 8-10°C. We used a repeated-measures ANOVA to determine whether there was an impact of cooling on light-evoked spike rates in visual cortex and sound-evoked firing rates in auditory cortex. Our physiological recordings confirmed that within the vicinity of the loop the inactivation spanned all cortical layers. As the temperature change dropped off with distance, at distances further from the loop the cooling was more restricted to superficial layers. These data are presented in full in Wood et al. (2017).