# 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]

[1]The Ear Institute, University College London, UK

[2]Department of Biomedical Engineering, University of Rochester, Department of Neuroscience, University of Rochester, Rochester, USA

[3]Institute for Learning and Brain Sciences and Department of Speech and Hearing Sciences, University of Washington, Seattle, USA

7 Figures

59600 characters

Corresponding Author and Lead Contact: Jennifer Bizley j.bizley@ucl.ac.uk

UCL Ear Institute, 332 Gray's Inn Road, London, WC1X 8EE.

## 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, and 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.

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 organisation 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 behavioural 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., 2016).

Visual stimuli can both drive and modulate neural activity in primary and non-primary auditory cortex (Bizley et al., 2007a; Chandrasekaran et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; Perrodin et al., 2015) but the contribution that visual activity in auditory cortex makes to auditory function remains unknown. One possibility is that the integration of cross-sensory information into early sensory cortex provides a bottom-up substrate for the binding of multisensory stimulus features into a single perceptual object (Bizley et al., 2016). We have recently argued that binding is a distinct form of multisensory integration that underpins perceptual object formation. We hypothesise that binding is associated with a modification of the sensory representation and can be identified by demonstrating a benefit in the behavioural or neural discrimination of a stimulus feature orthogonal to the features that link crossmodal stimuli (Fig. 1a). Therefore, in order to demonstrate binding an appropriate crossmodal stimulus should elicit not only enhanced neural encoding of the stimulus features that bind auditory and visual streams (the “binding features”), but that there should be enhancement in the representation of *other* stimulus features (“non-binding features” associated with the source (Fig. 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 (Fig.1). We demonstrate that when visual luminance changes coherently with the amplitude changes of one sound in a mixture, auditory cortex is biased towards representing the temporally coherent sound. Consistent with these effects reflecting cross-modal binding, the encoding of auditory 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.

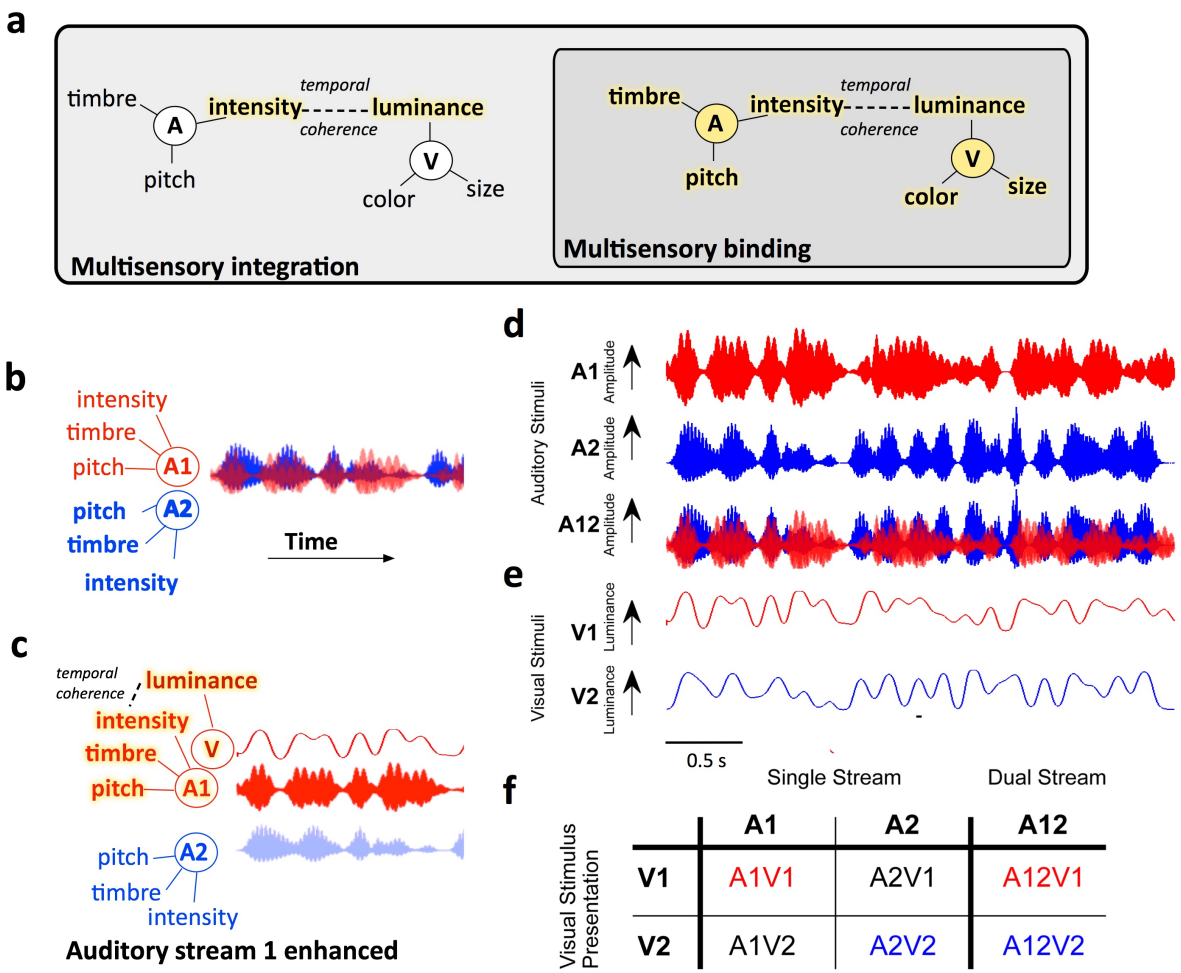



**Figure 1: Hypothesis and experimental design**

**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 colour and size. Other forms of multisensory integration would result in only the features that promote binding being enhanced - here auditory intensity and visual luminance. To identify binding therefore requires a demonstration that non-binding features (e.g. here pitch, timbre, colour or size) are enhanced. **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** Hypothesised 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** illustrates 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.

## Results

We recorded neuronal responses in the auditory cortex of awake passively listening ferrets (n=9 ferrets, 221 single units, 311 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 rendering them suitable for animal models and permitting the separation of general sensory processing mechanisms from language-specific ones. We additionally recorded neural responses in medetomidine-ketamine anesthetised ferrets (n=5 ferrets, 426 single units, 772 multi units) which allowed us to eliminate attentional effects and thus isolate bottom-up processing. 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 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.

The stimuli were two auditory streams comprised of two vowels, each with a distinct pitch and timbre (denoted A1: [u], fundamental frequency (F0) = 175 Hz and A2: [a], F0 = 195 Hz, Fig. 1) and independently amplitude modulated with a low-pass (<7 Hz) envelope (Fig. 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 (Fig. 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, Fig. 1e). We also tested conditions in which a single AV stimulus pair was presented ('single stream' stimuli), where the auditory and visual streams could be temporally coherent (A1V1, A2V2) or independently modulated (A1V2, A2V1).

## 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 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). A visual stimulus is known to enhance the representation of the 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 phenomena 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 to what extent the neural response to the sound mixture was specific to one or other sound stream.

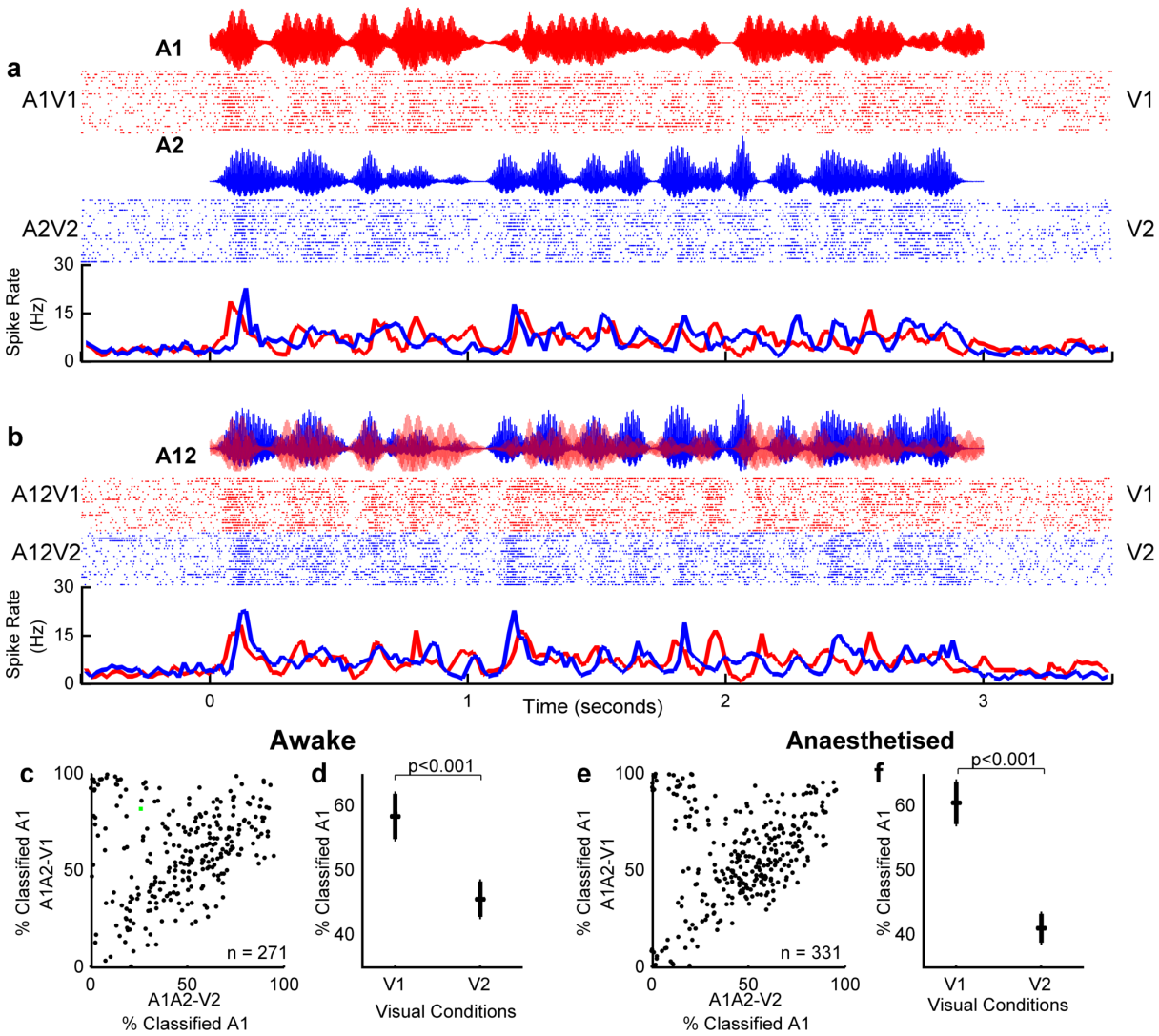



**Figure 2: Visual stimuli can determine which sound stream auditory cortical neurons follow in a mixture.**

Spiking responses from an example unit in response to **a,** single stream AV stimuli used as decoding templates and **b**, dual stream stimuli. In each case rasters and PSTHs are illustrated. When the visual component of the dual stream was V1, the majority of trials were classified as A1V1 (82% (19/23) of trials), and A2V2 when the visual stimulus was V2 (26% (6/23) of responses classified as A1V1 (see also green data point in c). **c-f** population data for awake (**c,d,** 541 units) and anesthetised (**e,f** 661 units) datasets. In each case the left panel (c,e) shows the distribution of decoding values according to the visual condition and the right panel (**d,f**) shows the population mean (± SEM) Pairwise comparisons revealed significant effect of visual condition on decoding in both datasets (p<0.001).

Figure 2 illustrates this approach for a single unit: responses to the single stream AV stimuli (Fig. 2a) formed templates which were used to decode the responses to the dual stream stimuli (Fig. 2b) using a Euclidean distance based spike pattern classifier (additional example: Fig S1). Responses to

142 the dual stream stimuli were more commonly decoded as A1V1 when the visual stimulus was V1,
143 and A2V2 when the visual stimulus was V2. For any unit in which we were able to decode the single
144 stream template responses at a level better than chance ($p<0.05$, permutation test; 541 units, awake
145 dataset; 661 units anesthetised dataset) we proceeded to use these single stream responses to
146 decode dual stream responses. Enhancement of the coherent auditory stimulus representation was
147 visible at the population level (Fig. 2c-f): Auditory cortical responses to dual-stream stimuli most
148 closely resembled responses to the single stream stimulus with the same visual component. This
149 finding was robust in both awake (Fig. 2d, pairwise t-test: $t_{540} = 6.073$, $p<0.001$) and anesthetised
150 animals (Fig. 2f, $t_{660} = 9.5137$, $p<0.001$) suggesting that these effects were not mediated by attention.

151

152 Modulation of dual stream responses by visual stimulus identity was not simply a consequence of
153 the shared visual component of single stream and dual stream stimuli and was observed in neurons
154 in which visual or auditory identity could be decoded (Fig S1). If this effect was only apparent in
155 visual neurons in auditory cortex then eliminating the visual element of the single stream stimuli
156 should impair decoding performance for the dual stream stimuli. Additional control experiments
157 (n=89 units, awake animals) demonstrated that this was not the case: the enhancement of the
158 temporally coherent sound in the sound mixture was evident whether dual stream stimuli (A12V1
159 and A12V2) were decoded using responses to auditory-only stimuli (A1 or A2) or auditory-visual
160 stimuli (Fig. 3a). We also determined that removing the visual stimulus from the dual-stream
161 eliminated any decoding difference in responses observed. A two-way repeated measures ANOVA
162 on decoded responses with factors of visual stream (V1, V2, no visual), and template type (AV or A)
163 demonstrated a significant effect of visual stream identity on dual stream decoding (Fig. 3d, $F(2,$
164 $528) = 19.320$, $p <0.001$), but there was no effect of template type ( $F(1, 528) = 0.073$, $p = 0.787$) or
165 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.

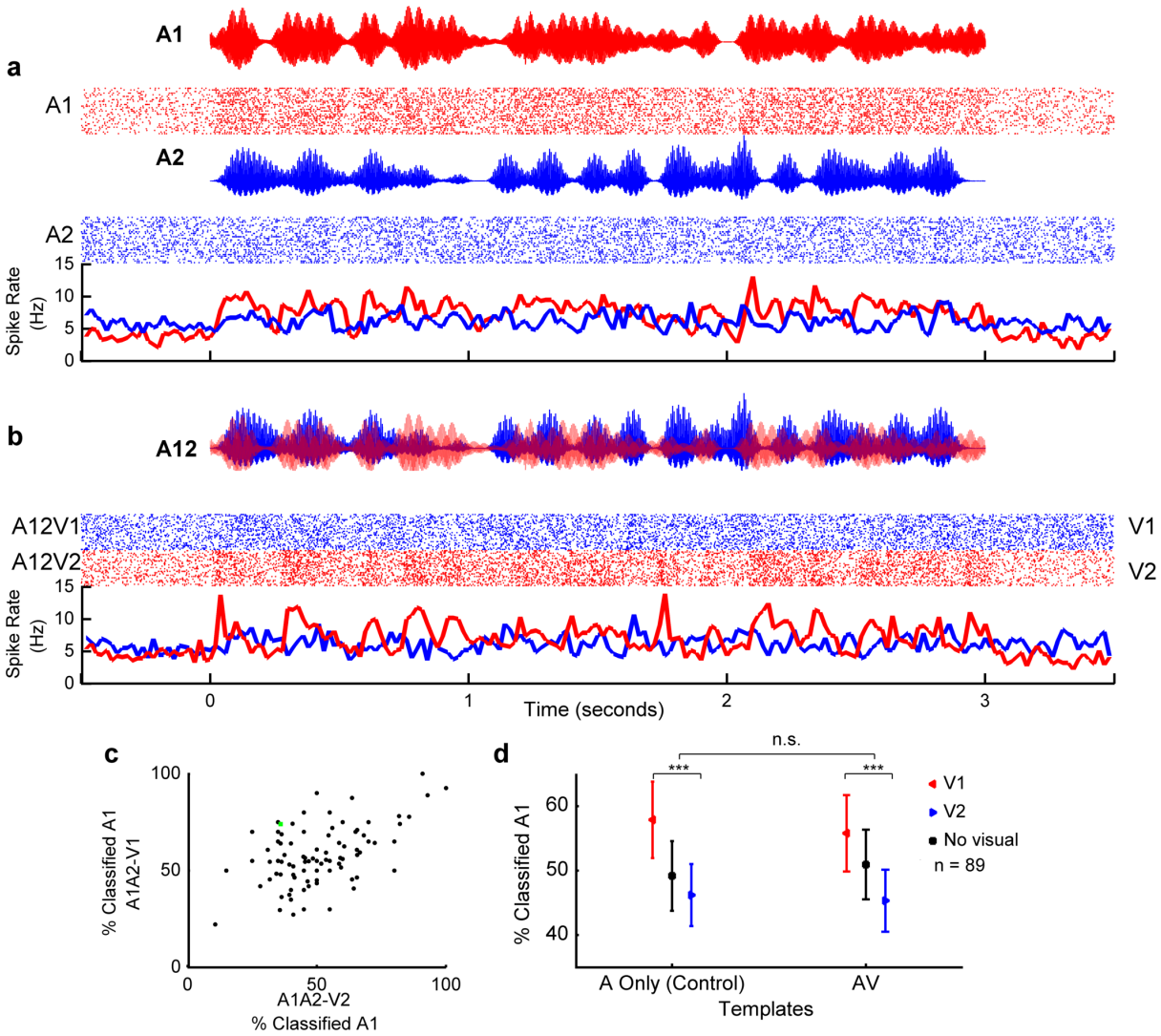



**Figure 3: Visual stimuli shape the neural representation of an auditory scene.**

In an additional control experiment (n=89 units recorded in awake animals), the responses to coherent AV 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). Spiking responses from an example unit in response to **a,** single stream auditory stimuli which 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. **c**, population data showing the proportion of responses classified as A1 when the visual stimulus was V1 or V2 when decoded with auditory-only templates. **d,** Mean (± SEM) values for these units when decoded with A-only templates (i.e. data from **c**), AV templates (as in Fig.2) or in the absence of a visual stimulus. The green data point in **c** depicts the example in **a, b**. Pairwise comparisons revealed significant effect across visual conditions in both datasets (* indicates p<0.001).

182 The ability of auditory-visual temporal coherence to enhance one sound in a mixture was observed
183 across all cortical layers (defined by current source density analysis, see methods, Fig.S2, a,b) and
184 across three tonotopic fields (Fig.S2 c,d) of anaesthetized subjects. While present in all layers, the
185 influence of the visual stimulus was strongest in the supra-granular layers (Fig.S2). Separating units
186 into auditory, visual, and auditory-visual units according to responsiveness to classic
187 neurophysiological stimuli (noise bursts and light flashes) revealed that the impact of a visual
188 stimulus on the representation of a sound mixture was present across functional sub-populations
189 but strongest in visual and auditory-visual units (Fig. S3, anesthetised dataset). Finally, we observed
190 these effects in both single and multi-units (Fig.S4).

191 **Auditory-visual temporal coherence enhances non-binding sound features**

192 A hall-mark of an object-based rather than feature-based representation is that all stimulus features
193 are bound into a unitary perceptual construct, including those features which do not directly
194 mediate binding (Desimone and Duncan, 1995). We predicted that binding across modalities would
195 be promoted via synchronous changes in auditory intensity and visual luminance (Fig. 1b) and
196 observed that the temporal dynamics of the visual stimulus enhanced the representation of
197 temporally coherent auditory streams (Fig. 2e-f and 3d-f). To determine whether temporal
198 synchrony of visual and auditory stimulus components also enhanced the representation of
199 orthogonal stimulus features (Fig. 1c) and thus fulfil a key prediction of binding, we introduced brief
200 timbre perturbations into our acoustic stimuli (two in each of the A1 and A2 streams). Each deviant
201 lasted for 200 ms during which the spectral timbre smoothly transitioned to the identity of another
202 vowel and back to the original vowel. Such deviants could be detected by human listeners and were
203 better detected when the auditory stream in which they were embedded was temporally coherent
204 with an accompanying visual stimulus (Maddox et al., 2015). We hypothesised that, despite
205 containing no information about the timing or occurrence of deviants, a temporally coherent visual
206 stimulus would enhance the representation of timbre deviants in the responses of auditory cortical
207 neurons.

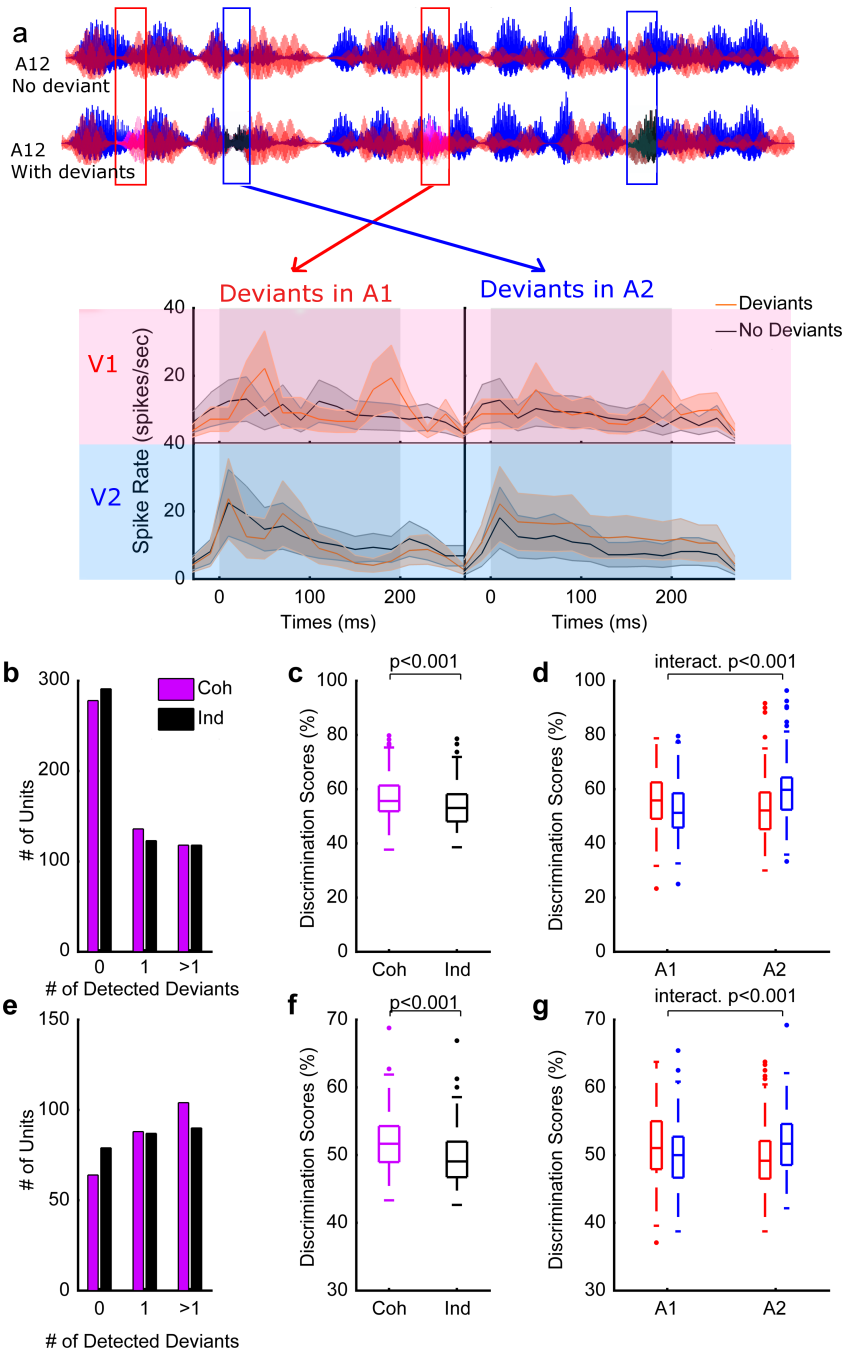



**Figure 4: Temporally coherent changes in visual luminance and auditory intensity enhance the representation of auditory timbre**

**a** Example unit (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. Shaded rectangles indicate the 200 ms window over which the timbre deviant occurred and over which analysis was conducted. **b-d** timbre deviant discrimination in the awake dataset. Two deviants were included in each auditory stream giving a possible maximum of 4per unit **b,** Histogram showing the number of deviants (out of 4) that could be discriminated from spiking responses **c,** Box plots showing the average 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, and the horizontal lines indicates the median. **d,** Discrimination scores for timbre deviant detection in dual stream stimuli. Discrimination scores are plotted according to the auditory

221 stream in which the deviant occurred and the visual stream that accompanied the sound mixture. **e-**
222 **g** show the same as **b-d** but for the anesthetised dataset.

223 To isolate neural responses to the timbre change from those elicited by the on-going amplitude
224 modulation, we extracted the 200ms epochs of the neuronal response during which the timbre
225 deviant occurred and compared these to epochs from responses to otherwise identical stimuli
226 without deviants. We observed that the spiking activity of many units differed between deviant and
227 no-deviant trials (e.g. Fig. 4a) and we were able to discriminate deviant from no-deviant trials with a
228 spike pattern classifier. We first considered the influence of temporal coherence between auditory
229 and visual stimuli on the representation of timbre deviants in the single stream condition (A1V1,
230 A1V2 etc.). We found that a greater proportion of units detected at least one deviant when the
231 auditory stream in which deviants occurred was temporally coherent with the visual stimulus
232 relative to the temporally independent condition. This was true both for awake (Fig. 4b; Pearson chi-
233 square statistic, $\chi^2 = 322.617$, $p < 0.001$) and anesthetised animals (Fig. 4e; $\chi^2 = 288.731$, $p < 0.001$).
234 For units that discriminated at least one deviant, discrimination scores were significantly higher
235 when accompanied by a temporally coherent visual stimulus (Fig.4c, awake dataset, pairwise t-test
236 $t_{300} = 3.599$ $p<0.001$; Fig. 4f, anesthetised data $t_{262} = 4.444$ $p<0.001$).

237

238 We also observed an enhancement in the representation of timbre changes in the context of a
239 sound scene (Fig 4d,g): timbre changes were more reliably encoded when the sound stream in which
240 they were embedded was accompanied by a temporally coherent visual stimuli. We performed a
241 two-way repeated measures ANOVA on deviant discrimination performance with visual condition
242 (V1/V2) and the auditory stream in which the deviants occurred (A1/A2) as factors. We predicted
243 that enhancement of the representation of timbre deviants in the temporally coherent auditory
244 stream would be revealed as a significant interaction term in the dual stream data. Significant
245 interactions were seen in both the awake (Fig. 4d, $F(1, 600) = 29.138$, $p<0.001$) and anesthetised
246 datasets (Fig. 4g, $F(1, 524) = 16.652$, $p<0.001$). We also observed significant main effects of auditory

247 and visual conditions in awake (main effect of auditory stream, $F(1, 600) = 4.565$, $p = 0.033$; main
248 effect of visual condition, $F(1, 600) = 2.650$, $p = 0.010$) but not anesthetised animals (main effect of
249 auditory stream, $F(1,524) = 0.004$, $p = 0.948$; main effect of visual condition, $F(1, 524) = 1.355$, $p =$
250 $0.245$). Thus these data demonstrate the hypothesised enhancement in the neural representation of
251 both binding (i.e. auditory amplitude) and non-binding features (here auditory timbre) that are
252 orthogonal to those that promote binding between auditory and visual streams and so the effects
253 we observe in auditory cortex fulfil our definition of multisensory binding. Next we turn to the
254 question of how these effects are mediated and whether they emerge within or outside of auditory
255 cortex.

256

## 257 Auditory cortical spike patterns differentiate dynamic auditory-visual stimuli more 258 effectively when stimuli are temporally coherent

259 We used the responses to single stream stimuli to classify neurons according to whether they were
260 modulated primarily by auditory or visual temporal dynamics. To determine whether the auditory
261 amplitude envelope reliably modulated spiking we used a spike-pattern classifier to decode the
262 auditory stream identity, collapsed across visual stimuli (i.e. we decoded auditory stream identity
263 from the combined responses to A1V1 and A1V2 stimuli and the combination of A2V1 and A2V2
264 responses). An identical approach was taken to determine if neuronal responses reliably
265 distinguished visual modulation (i.e. we decoded visual identity from the combined responses to
266 A1V1 and A2V1 stimuli and the combined responses elicited by A1V2 and A2V2). Neuronal
267 responses which were informative about auditory or visual stimulus identity at a level better than
268 chance (estimated with a bootstrap resampling) were classified as auditory-discriminating (Fig. 5a-b)
269 and / or visual-discriminating (Fig. 5c-d) respectively.

270

271

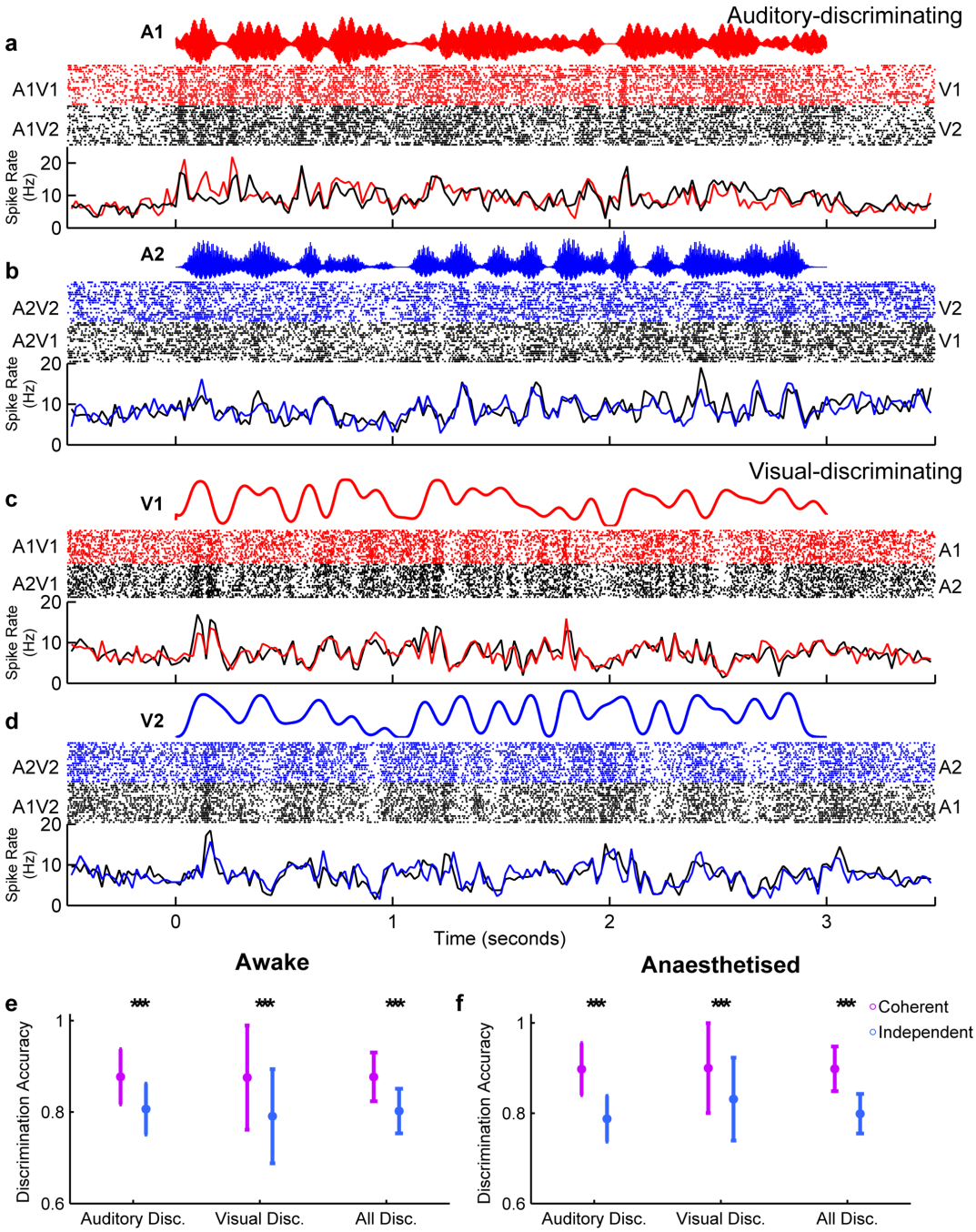



**Figure 5: Auditory-visual temporal coherence enhances neural coding in auditory cortex**.

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, b**, for an auditory discriminating unit) or visual stream (**c, d**, for a visual discriminating unit). In each case the stimulus amplitude (a,b) / luminance (c,d) waveform is shown in the top panel with the resulting raster plots and PSTHs below. **e, f**: Decoder performance (mean ± SEM) for discriminating stimulus identity (coherent: A1V1 vs. A2V2, purple; independent: A1V2 vs. A2V1, blue) in auditory and visual classified units recorded in awake (e) and anaesthetised (f) ferrets. Pairwise comparisons for decoding of coherent versus independent stimuli (*** indicates $p<0.001$ ).

284 In awake animals, 39.5% (210/532) of driven units were auditory-discriminating, 11.1% (59/532)
285 were visual-discriminating, and only 0.4% (2/532) discriminated both auditory and visual stimuli.
286 Overall a smaller proportion of units represented the identity of auditory or visual streams in the
287 anesthetised dataset: 20.2% (242/1198) were auditory-discriminating, 6.8% (82/1198) were visual
288 discriminating, and 0.6% (7/1198) discriminated both. During recordings made under anaesthesia,
289 we also recorded responses to noise bursts and light flashes (both 100 ms duration) presented
290 separately and together to compare how the proportion of auditory / visual discriminating units
291 measured to naturalistic dynamic stimuli compared to more traditional artificial stimuli. Specifically,
292 responsiveness was defined using a two-way ANOVA (factors: auditory stimulus [on/off] and visual
293 stimulus [on/off]) on spike counts measured during stimulus presentation. We defined units as being
294 sound-driven (main effect of auditory stimulus, no effect of visual stimulus or interaction), light-
295 driven (main effect of visual stimulus, no effect of auditory stimulus or interaction) or both (main
296 effect of both auditory and visual stimuli or significant interaction; $p < 0.05$) as in (Bizley et al.,
297 2007b). Using such stimuli revealed that the classification of units by visual / auditory discrimination
298 of single stream stimuli selected a subset of light or sound driven units and that the proportions of
299 auditory, visual and AV units recorded in our sample were in line with previous studies from ferret
300 auditory cortex (65.1% (328/504) of units were driven by noise bursts, 16.1% (81/504) by light
301 flashes and 14.1% (71/504) by both. For units which were classified as auditory or visual
302 discriminating, and for which we recorded responses to noise bursts and light flashes, 53% (160/307)
303 were classified as auditory, 17% (53/307) as visual and 31% (94/307) as auditory-visual when
304 classified with simple stimuli.

305 We hypothesised that the effects we observed in the dual-stream condition might be a consequence
306 of temporal coherence between auditory and visual stimuli enhancing the discriminability of neural
307 responses. We confirmed this prediction by using the same spike pattern decoder to compare our
308 ability to discriminate temporally coherent (A1V1 vs. A2V2) and temporally independent (A1V2 vs.
309 A2V1) stimuli (Fig. 5e, f): Temporally coherent AV stimuli produced more discriminable spike

310 patterns than those elicited by temporally independent ones in both awake (Fig. 5e, pairwise t-test,
311 auditory-discriminating $t_{418}$ = 11.872, p<0.001; visual-discriminating $t_{116}$ = 6.338, p<0.001; All $t_{540}$ =
312 13.610, p<0.001) and anesthetised recordings (Fig.5f, auditory-discriminating $t_{482}$ = 17.754, p<0.001;
313 visual-discriminating $t_{162}$ = 8.186, p<0.001; All $t_{664}$ = 19.461, p<0.001).

314 **Dynamic visual stimuli elicit reliable changes in LFP phase**

315 Temporal coherence between auditory and visual stimulus streams results in more discriminable
316 spike trains in the single stream condition, and an enhancement of the representation of the
317 temporally coherent sound when that sound forms part of an auditory scene. What might underlie
318 the enhanced discriminability observed for temporally coherent cross-modal stimuli? The phase of
319 on-going oscillations determines the excitability of the surrounding cortical tissue (Azouz and Gray,
320 1999; Okun et al., 2010; Szymanski et al., 2011). LFP phase is reliably modulated by naturalistic
321 stimulation (Chandrasekaran et al., 2010; Kayser et al., 2009; Luo and Poeppel, 2007b; Ng et al.,
322 2012; Schyns et al., 2011) and has been implicated in multisensory processing (Golumbic et al., 2013;
323 Lakatos et al., 2007). We hypothesised that sub-threshold visual inputs could modulate spiking
324 activity by modifying the phase of the local field potential such that when auditory and visual stimuli
325 were temporally coherent visual stimuli 'reset' the LFP phase to maximise the likelihood that feed
326 forward sound-evoked activity elicited robust spiking responses. This would provide a mechanism by
327 which neuronal spiking was enhanced when auditory and visual streams are temporally coherent.

328 Stimulus evoked changes in the local field potential (LFP) were evident from the recorded voltage
329 traces and analysis of inter-trial phase coherence demonstrated that there were reliable changes in
330 phase across repetitions of identical AV stimuli (Fig. 6 a, b). To isolate the influence of visual activity
331 on the LFP at each recording site, and address the hypothesis that visual stimuli elicited reliable
332 changes in the LFP, we calculated phase and power dissimilarity functions for stimuli with identical
333 auditory signals but differing visual stimuli (Luo and Poeppel, 2007b). Briefly, this analysis assumes
334 that if the phase (or power) within a particular frequency band differs systematically between

335 responses to two different stimuli, then inter-trial phase coherence (ITPC) across repetitions of the
336 same stimulus will be greater than across randomly selected stimuli. For each frequency band in the
337 LFP, we therefore compared "within-stimulus" ITPC for responses to each stimulus (A1 stream Fig.
338 6c; A2 stream Fig. 3d) with "across-stimulus" ITPC calculated from stimuli with *identical auditory*
339 *components* but randomly selected visual stimuli (e.g. randomly drawn from A1V1 and A1V2; Fig.
340 6c). The difference between within-stimulus and across-stimulus ITPC was then calculated across
341 frequency and described as the phase dissimilarity index (PDI) (Fig. 6e, f) with positive PDI values
342 indicating reliable changes in phase coherence elicited by the visual component of the stimulus.
343 Importantly, because both test distributions and the null distribution contain identical sounds any
344 significant PDI value can be attributed directly to the visual component of the stimulus.

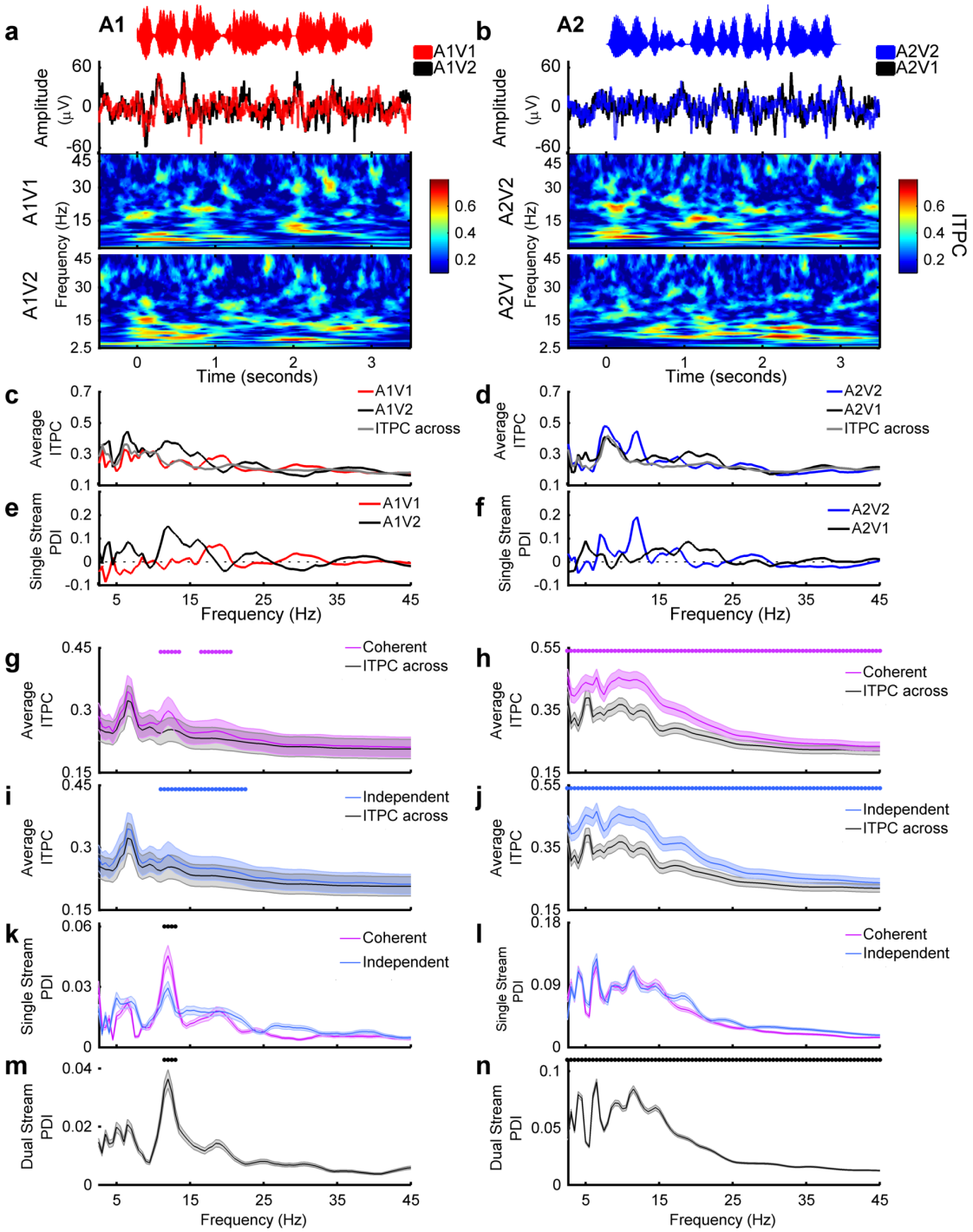



**Figure 6: Visual stimuli elicit reliable changes in the phase of the local field potential**

**a, b** Example LFP responses to single stream stimuli across visual conditions. Data obtained from the recording site at which multiunit spiking activity discriminated auditory stream identity in Fig. 5 a and b. The amplitude waveforms of the stimuli are shown in the top panel, with the evoked LFP underneath (mean across 21 trials). The resulting inter-trial phase coherence (ITPC) values are shown in the bottom two panels. **c, d** ITPC was calculated for coherent and independent AV stimuli separately and compared to a null distribution (ITPC across). Single stream phase dissimilarity values (PDI) were calculated by comparing ITPC within values to the ITPC across null distribution (**e, f**). **g,I** Population mean inter-trial phase coherence (ITPC) values across frequency for coherent (**g**) and independent (**i**) conditions. 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:** Mean (±SEM) single stream phase dissimilarity index (PDI) values for coherent and independent stimuli in animals. Black dots indicate frequencies at which the temporally coherent single stream PDI is significantly greater than in the independent conditions ($p<0.001$). **h,j,l** as **g,l,k** for anesthetised dataset. **m, n,** mean ±SEM dual stream PDI values for awake (m) and anesthetised (n) datasets.

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 what frequencies the across-trial phase reliability was significantly positive, we compared the within-stimulus values with the across-stimulus (null distribution) 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 and 20 Hz where visual stimuli enhanced the phase reliability (Fig. 3g, h). In anesthetised animals, average PDI values were larger than in awake animals and all frequencies tested had single stream PDI values that were significantly non-zero (Fig. 3j,k). We therefore conclude that visual stimulation elicited reliable changes in the LFP phase in auditory cortex. In contrast to LFP phase, a parallel analysis of across trial power reliability showed no significant effect of visual stimuli on LFP power in any frequency band (Fig.S5).

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. While 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 temporally coherent AV stimuli compared to temporally independent AV stimuli. To address this we performed a pairwise comparison of single stream PDI values yielded 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 11-14 Hz band where coherent stimuli elicited significantly greater phase coherence (Fig. 6i). In anaesthetised animals, the single stream PDI did not differ between coherent and independent stimuli at any frequency (Fig. 6l). Together these data

384 suggests that visual inputs modulate the phase of the low frequency field potential in auditory
385 cortex largely independently of any temporal coherence between auditory and visual stimuli. This
386 finding supports the conjecture that multisensory binding occurs within auditory cortex.

387 To understand whether similar mechanisms could underlie the visual-stimulus induced
388 enhancement of a temporally coherent sound in a mixture we performed similar analysis on the
389 data collected in response to the dual stream stimuli. We generated within-stimulus ITPC values for
390 each dual-stream stimulus (i.e. A12V1 and A12V2, Fig. 6a) and across-stimulus ITPC by randomly
391 selecting responses across visual conditions (Fig. 6b). We then expressed the difference as the dual
392 stream phase dissimilarity index (dual stream PDI, Fig. 6c). Since the auditory components were
393 identical in each dual stream stimulus, the influence of the visual component on LFP phase could be
394 isolated as non-zero dual stream PDI values (paired t-test, bonferoni corrected, $\alpha$ = 0.0012). In
395 awake animals, the dual stream PDI was significantly non- zero at 11-14Hz and 16-19 Hz (Fig. 6d, e)
396 whereas in anesthetised animals, we found positive dual stream PDI values across all frequencies
397 tested (Fig. 6f, g). In anesthetised animals where we were able to use the responses of units to noise
398 and light flashes to categorise units as auditory, visual or auditory-visual, and were able to confirm
399 significant PDI values in the LFP recorded on the same electrode as units in each of these
400 subpopulations (Fig. S3c). In awake animals, we tested auditory visual stimuli presented at three
401 different modulation rates (7, 12 and 17Hz) and confirmed that significant PDI values were obtained
402 at very similar frequencies across these modulation rates consistent with these being evoked phase
403 resets rather than stimulus-induced oscillations (Fig. S6).

404

405 **Visual cortex mediates visual-stimulus induced LFP changes in auditory cortex**

406 Visual inputs to auditory cortex potentially originate from many sources: In the ferret visual multiple
407 visual cortical fields are known to innervate auditory cortex (Bizley et al., 2007b), but frontal and
408 thalamic areas are additional candidates for sources of top-down and bottom-up multisensory

409 innervation. Therefore, in order to determine the origin of the visual effects that we observe in
410 auditory cortex, we performed an additional experiment in which we cooled the gyral surface of the
411 posterior suprasylvian sulcus where visual cortical fields SSY (Cantone et al., 2006) and Area 21
412 (Innocenti et al., 2002) are located (Fig. 7a). Neural tracer studies have demonstrated that these
413 areas directly project to auditory cortex in the ferret (Bizley et al., 2007b). We used a cooling loop
414 cooled to 9-10°C to reversible silence neural activity within <500 um of the loop (see Fig. 3 (Wood et
415 al., 2017). Using simple noise bursts and light flashes we verified that cooling visual cortex did not
416 alter the response to noise bursts in auditory cortex (repeated measures ANOVA $F_{(2,164)} = 0.42$
417 $p=0.88$) but did reversibly attenuate the spiking response to light flashes in visual cortical sites
418 >500um from the cooling loop (repeated measures ANOVA $F_{(2,92)} = 6.83$ $p=0.001$, post-hoc
419 comparisons shows pre-cool and cool-post were significantly different, pre-post were not
420 significantly different indicating the effects were reversible). We measured responses to the single
421 stream stimuli in auditory and visual cortex before and during cooling, while the cooling loop was
422 held at 9-10°C. From the LFP, we calculated the across-trial-phase coherence and phase dissimilarity
423 indexes (as in Fig. 6). Cooling significantly decreased the magnitude of the single stream PDI values in
424 auditory cortex (Fig 7e,h). A 3-way repeated measures ANOVA with factors of visual condition
425 (coherent/independent), frequency, and cortical temperature (warm/cooled) on the SS PDI values
426 obtained in auditory cortex showed a main effect of frequency ($F_{(88,22605)} = 47.91$, $p<0.1*10^9$)
427 and temperature ($F_{(1,22605)} = 1072$, $p<0.1*10^9$,) but not visual condition ($p=0.49$). In contrast, those
428 in visual cortex (away from the loop) were unaffected by cooling (3-way ANOVA demonstrated that
429 the magnitude of the SS PDI value was influenced by frequency ($F_{(88,17265)} = 24.73$, $p<0.1*10^9$), but
430 not temperature ($p=0.75$) or visual condition ($p=0.29$)). From these data we conclude that the
431 influence of visual stimuli on the auditory cortical field potential phase is mediated, at least in part,
432 by inputs from visual cortical areas SSY and 21. While cooling does not allow us to confirm that these
433 are direct mono-synaptic inputs the observation that the phase effects in other areas of visual cortex

are uninfluenced suggests that we are influencing these areas selectively rather than supressing visual processing generally.

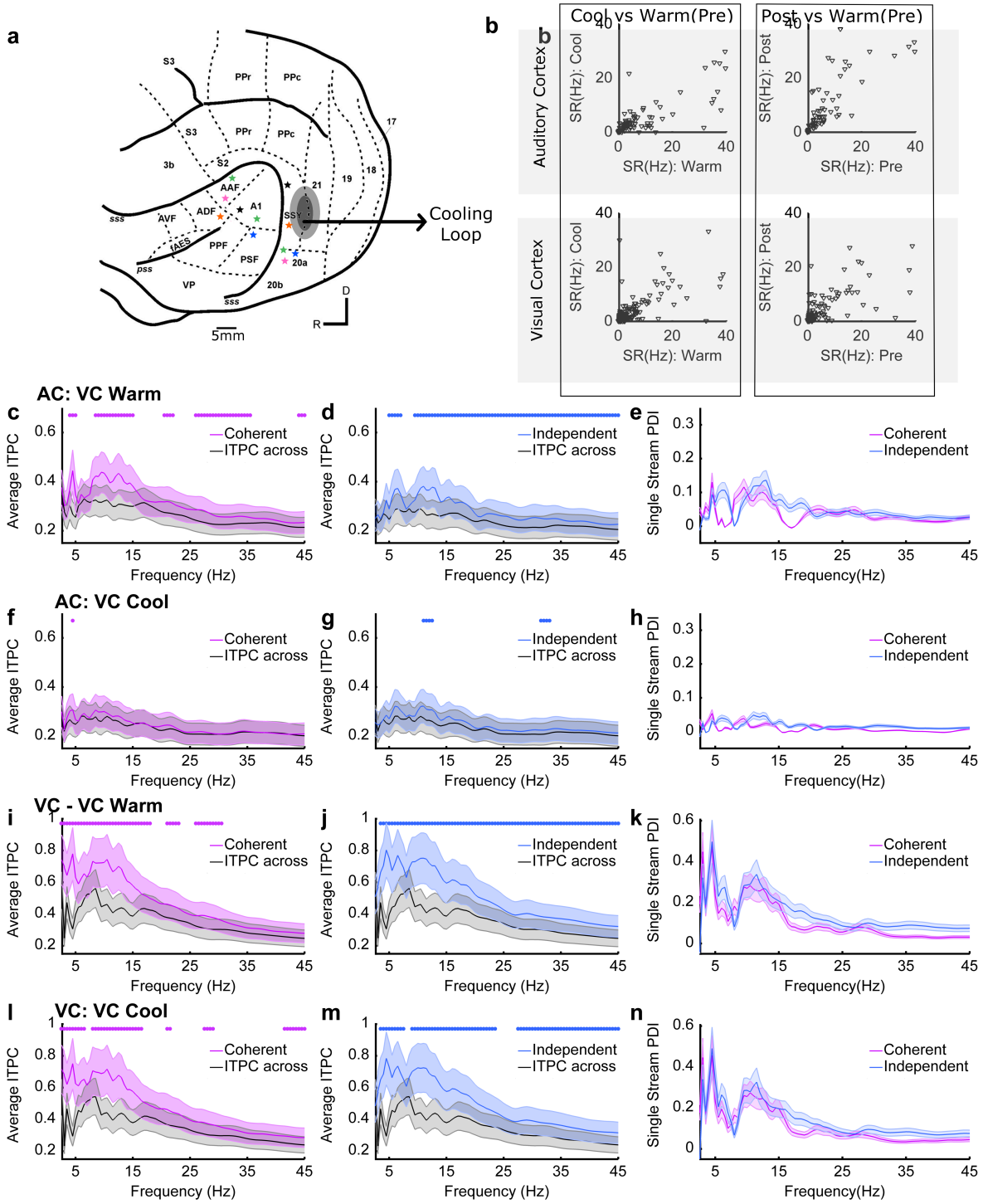



**Figure 7: Visual stimulus induced LFP phase changes in auditory cortex are mediated in visual cortex**

**a** Schematic showing the location of auditory cortical recording sites and the location of a cooling loop (black, grey line marks the 500 um radius over which cooling is effective (Wood et al., 2017) which was used to inactivate visual cortex. **b** Spike rate responses in Auditory Cortex (top row) and visual cortex (bottom row, sites >500 um from the loop) in response to noise bursts or light flashes before, during and after cooling. **c,d**, inter-trial phase coherence values for the coherent (**c**) and independent stimuli (**d**) AV stimuli recorded in auditory cortex (AC) prior to cooling visual cortex (VC) compared to the shuffled null distribution (ITPC across). Asterisks indicate the frequencies at which the ITPC values are significantly different from the shuffled ITPC-across distribution **e** single stream Phase Dissimilarity Index values calculated from the ITPC values in **c** and **d**. **f-h** - as c-e but while visual cortex was cooled to 9 degrees. **i-n** as c-h but for sites in visual cortex >500um from the cooling loop. c-h includes data from 83 sites from 6 electrode penetrations, i-n includes data from 47 sites from 5 penetrations.

## Discussion

Here we provide insight into how and where auditory-visual binding occurs, and provide evidence that this effect is mediated by cortical-cortical interactions between visual and auditory cortex. Visual stimuli elicit reliable changes in the phase of the local field potential in auditory cortex, 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 phase resets elicited by the visual stimulus via the LFP interact with the feed-forward sound evoked activity and result in an enhanced spiking output from 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 in auditory cortex which 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 pertubation 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 anesthetised animals ruling out any top-down effect 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 rhythm information and information about the amplitude of the speech waveform which may help listeners by cueing them to pay attention to the auditory envelope (Peelle and Sommers, 2015) as well as provide 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. Therefore, in order to probe more general principles we chose to employ continuous naturalistic non-speech stimuli, that utilized modulation rates that fell within the range of syllable rates in human speech but lacked any linguistic content(Chandrasekaran et al., 2009). 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; Crosse et al., 2016; Luo et al., 2010)'(Park et al., 2016), but functional imaging methods make if difficult to demonstrate enhanced neural encoding of features beyond the amplitude envelope. The implication of our findings is that the representation of the spectro-temporal features that allow speech recognition would be enhanced in auditory cortex when a listener views a talker's face, including features such as pitch that help pull apart voices in an auditory stream.

Visual speech information is hypothesised 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 local field potential. Within this framework, cross-modal binding results from an interaction between the bottom-up auditory spiking information and the visual stimulus induced phase changes of the LFP. Indeed low-frequency entrainment to modulations in an on-going stimulus are observed in the human brain including to lip movements(Park et al.,

494 2016) and have been shown to optimize listening performance for non-speech stimuli (Henry and
495 Obleser, 2012). In contrast, later integration is likely to underlie information about speech gestures
496 that might be used to constrain lexical identity (Peelle and Sommers, 2015).

497 Consistent with previous studies, our analysis of local field potential activity revealed that visual
498 information reliably modulated the phase of oscillatory activity in auditory cortex (Chandrasekaran
499 et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; Perrodin et al., 2015). This occurred
500 independently of the modulation frequency of the stimulus suggesting that rather than entraining
501 oscillations at the stimulus modulation frequency particular features within the stream (presumably
502 points at which the luminance changed rapidly from low-high amplitude) triggered relatively broad-
503 band phase resets. Neuronal excitability varies with LFP phase (Jacobs et al., 2007; Klimesch et al.,
504 2007; Lakatos et al., 2013; Lőrincz et al., 2009) and such phase resets may be the physiological
505 mechanism through which cross-sensory information is integrated. Our analysis allowed us to isolate
506 changes in LFP phase that were directly attributable to the visual stimulus and identified reliable
507 changes in the LFP phase irrespective of whether the visual stimulus was temporally coherent with
508 the auditory stimulus. Such a finding is consistent with the idea that the LFP phase synchronization
509 arises from fluctuating inputs to cortical networks(Lakatos et al., 2007; Mazzoni et al., 2008;
510 Szymanski et al., 2011), and the loss of these effects with cooling suggests that visual cortex is a key
511 component of this network. Our finding that visual stimulation elicited reliable phase modulation in
512 both awake and anesthetised animals suggests that bottom-up cross-modal integration interacts
513 with selective attention, which also modulates phase information in auditory cortex(Golumbic et al.,
514 2013),(Park et al., 2016). While our data suggest that cross-modal binding can occur in the absence
515 of attention, it is likely that the effects we observe in auditory cortex are the substrates on which
516 selective attention acts to further boost the representation of cross-modal objects.

517 In the awake animal the impact of visual stimulation on LFP phase reliability was smaller than in the
518 anesthetised animal and was restricted to a narrower range of frequencies, consistent with a

519 dependence of oscillatory activity on behavioural state (Tukker et al., 2007; Voloh and Womelsdorf,
520 2016; Wang, 2010). Since the neural correlates of multisensory binding are evident in the
521 anesthetised animal, the specific increase in alpha phase reliability that occurred in awake animals in
522 response to temporally coherent auditory-visual stimulus pairs (Fig. 3c, 6e) may indicate an
523 attention-related signal triggered by temporal coherence between auditory and visual signals, or an
524 additional top-down signal conveying cross-modal information. Phase resetting or synchronisation of
525 alpha phase has been associated both with enhanced functional connectivity (Voloh and
526 Womelsdorf, 2016) and as a top-down predictive signal for upcoming visual information (Samaha et
527 al., 2015). Disambiguating these possibilities would require simultaneous recordings in auditory and
528 visual cortex and/or recording during the performance of a task designed to explicitly manipulate
529 attention.

530 Temporal coherence between sound elements has been proposed as a fundamental organising
531 principle for auditory cortex (Elhilali et al., 2009; O'Sullivan et al., 2015b) and here we extend this
532 principle to the formation of cross-modal constructs. Our data provide evidence that one role for the
533 early integration of visual information into auditory cortex is to resolve competition between
534 multiple sound sources within an auditory scene and that these neural computations occur pre-
535 attentively. While some proponents of a temporal coherence based model for auditory streaming
536 have stressed the importance of attention in auditory stream formation (Lu et al., 2017), neural
537 signatures of temporal-coherence based streaming are present in passively listening subjects
538 (O'Sullivan et al., 2015a; Teki et al., 2016). Previous studies have demonstrated a role for visual
539 information in conveying lip movement information to auditory cortex (Chandrasekaran et al., 2013;
540 Crosse et al., 2015; Ghazanfar et al., 2005; Golumbic et al., 2013), but such stimuli make it difficult to
541 separate sensory information from linguistic cues. Our data generated using non-speech stimuli
542 provide evidence that at least part of the boost provided by visualising a speaker's mouth arises
543 from a more general (language-independent) phenomenon whereby visual temporal cues facilitate
544 auditory scene analysis through the formation of cross-sensory objects. Our data are supportive of

visual cortical providing at least one source of information. Other visual cortical fields and sub-cortical structures innervate tonotopic auditory cortex (Bizley et al., 2007a; 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)..

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.

# Figure Legends

**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 colour and size. Other forms of multisensory integration would result in only the features that promote binding being enhanced - here auditory intensity and visual luminance. To identify binding therefore requires a demonstration that non-binding features (e.g. here pitch, timbre, colour or size) are enhanced. **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** Hypothesised 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** illustrates 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.

**Figure 2: Visual stimuli can determine which sound stream auditory cortical neurons follow in a mixture.**

Spiking responses from an example unit in response to **a,** single stream AV stimuli used as decoding templates and **b**, dual stream stimuli. In each case rasters and PSTHs are illustrated. When the visual component of the dual stream was V1, the majority of trials were classified as A1V1 (82% (19/23) of trials), and A2V2 when the visual stimulus was V2 (26% (6/23) of responses classified as A1V1 (see also green data point in c). **c-f** population data for awake (**c,d,** 541 units) and anesthetised (**e,f** 661 units) datasets. In each case the left panel (c,e) shows the distribution of decoding values according to the visual condition and the right panel (**d,f**) shows the population mean (± SEM) Pairwise comparisons revealed significant effect of visual condition on decoding in both datasets ($p<0.001$).

**Figure 3: Visual stimuli shape the neural representation of an auditory scene.**

In an additional control experiment (n=89 units recorded in awake animals), the responses to coherent AV 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). Spiking responses from an example unit in response to **a,** single stream auditory stimuli which 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. **c**, population data showing the proportion of responses classified as A1 when the visual stimulus was V1 or V2 when decoded with auditory-only templates. **d,** Mean (± SEM) values for these units when decoded with A-only templates (i.e. data from **c**), AV templates (as in Fig.2) or in the absence of a visual stimulus. The green data point in **c** depicts the example in **a, b**. Pairwise comparisons revealed significant effect across visual conditions in both datasets (* indicates p<0.001).

**Figure 4: Temporally coherent changes in visual luminance and auditory intensity enhance the representation of auditory timbre**

**a** Example unit (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. Shaded rectangles indicate the 200 ms window over which the timbre deviant occurred and over which analysis was conducted. **b-d** timbre deviant discrimination in the awake dataset. Two deviants were included in each auditory stream giving a possible maximum of 4per unit **b,** Histogram showing the number of deviants (out of 4) that could be discriminated from spiking responses **c,** Box plots showing the average 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, and the horizontal lines indicates the median. **d,** Discrimination scores for timbre deviant detection in dual stream stimuli. Discrimination scores are plotted according to the auditory stream in which the deviant occurred and the visual stream that accompanied the sound mixture. **e-g** show the same as **b-d** but for the anesthetised dataset.

**Figure 5: Auditory-visual temporal coherence enhances neural coding in auditory cortex**.

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, b**, for an auditory discriminating unit) or visual stream (**c, d**, for a visual discriminating unit). In each case the stimulus amplitude (a,b) / luminance (c,d) waveform is shown in the top panel with the resulting raster plots and PSTHs below. **e, f**: Decoder performance (mean ± SEM) for discriminating stimulus identity (coherent: A1V1 vs. A2V2, purple; independent: A1V2 vs. A2V1, blue) in auditory and visual classified units recorded in awake (e) and anaesthetised (f) ferrets. Pairwise comparisons for decoding of coherent versus independent stimuli (*** indicates p<0.001 ).

**Figure 6: Visual stimuli elicit reliable changes in the phase of the local field potential**

**a, b** Example LFP responses to single stream stimuli across visual conditions. Data obtained from the recording site at which multiunit spiking activity discriminated auditory stream identity in Fig. 5 a and b. The amplitude waveforms of the stimuli are shown in the top panel, with the evoked LFP underneath (mean across 21 trials). The resulting inter-trial phase coherence (ITPC) values are shown in the bottom two panels. **c, d** ITPC was calculated for coherent and independent AV stimuli separately and compared to a null distribution (ITPC across). Single stream phase dissimilarity values (PDI) were calculated by comparing ITPC within values to the ITPC across null distribution (**e, f**). **g,I** Population mean inter-trial phase coherence (ITPC) values across frequency for coherent (**g)** and independent (**i**) conditions. 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:** Mean (±SEM) single stream phase dissimilarity index (PDI) values for coherent and independent stimuli in animals. Black dots indicate frequencies at which the temporally coherent single stream PDI is significantly greater than in the independent conditions (p<0.001). **h,j,l**

646 as **g,l,k** for anesthetised dataset. **m, n,** mean ±SEM dual stream PDI values for awake (m) and
647 anesthetised (n) datasets.

648

649 **Figure 7: Visual stimulus induced LFP phase changes in auditory cortex are mediated in visual**
650 **cortex**

651 **a** Schematic showing the location of auditory cortical recording sites and the location of a cooling
652 loop (black, grey line marks the 500 um radius over which cooling is effective (Wood et al., 2017)
653 which was used to inactivate visual cortex. **b** Spike rate responses in Auditory Cortex (top row) and
654 visual cortex (bottom row, sites >500 um from the loop) in response to noise bursts or light flashes
655 before, during and after cooling. **c,d**, inter-trial phase coherence values for the coherent (**c**) and
656 independent stimuli (**d**) AV stimuli recorded in auditory cortex (AC) prior to cooling visual cortex (VC)
657 compared to the shuffled null distribution (ITPC across). Asterisks indicate the frequencies at which
658 the ITPC values are significantly different from the shuffled ITPC-across distribution **e** single stream
659 Phase Dissimilarity Index values calculated from the ITPC values in **c** and **d**. **f-h** - as c-e but while
660 visual cortex was cooled to 9 degrees. **i-n** as c-h but for sites in visual cortex >500um from the
661 cooling loop. c-h includes data from 83 sites from 6 electrode penetrations, i-n includes data from 47
662 sites from 5 penetrations.

663

## 664 Acknowledgments

665 This work was funded by grants to each author: JKB: Wellcome Trust / Royal Society WT098418MA;

666 Biotechnology and Biological Sciences Research Council (BB/H016813/1), and an Action on Hearing

667 Loss Studentship (596: UEI: JB); RKM: NIH K99DC014288 and Hearing Health Foundation Emerging

668 Research Grant; AKCL: NIH R01DC013260; and an International Exchanges Scheme award from the

669 Royal Society to JKB and AKCL.

## 670 Author contributions

671 HA, RKM, AKCL, JKB Conception and design, HA, SMT, KCW, GPJ, JKB Acquisition of data, HA, JKB

672 Analysis and interpretation of data, HA, SMT, RKM, AKCL, JKB Drafting or revising the article.

## 673 References

674 Azouz, R., and Gray, C.M. (1999). Cellular mechanisms contributing to response variability of cortical
675 neurons in vivo. The Journal of neuroscience 19, 2209-2223.
676 Bizley, J.K., Maddox, R.K., and Lee, A.K. (2016). Defining Auditory-Visual Objects: Behavioral Tests
677 and Physiological Mechanisms. Trends in Neurosciences.

Bizley, J.K., Nodal, F.R., Bajo, V.M., Nelken, I., and King, A.J. (2007a). Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17, 2172-2189.
Bizley, J.K., Nodal, F.R., Bajo, V.M., Nelken, I., and King, A.J. (2007b). Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17, 2172-2189.
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., 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. Journal of Neuroscience 29, 2064-2075.
Brosch, M., Selezneva, E., and Scheich, H. (2015). Neuronal activity in primate auditory cortex during the performance of audiovisual tasks. European Journal of Neuroscience 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. Visual neuroscience 23, 61-77.
Chandrasekaran, C., Lemus, L., and Ghazanfar, A.A. (2013). Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. Proceedings of the National Academy of Sciences 110, E4668-E4677.
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. The Journal of Neuroscience 30, 13919-13931.
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. The Journal of Neuroscience 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. The Journal of Neuroscience 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.
Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual review of neuroscience 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. The Journal of Neuroscience 25, 5004-5012.
Golumbic, E.Z., Cogan, G.B., Schroeder, C.E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. The Journal of Neuroscience 33, 1417-1426.
Henry, M.J., and Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proceedings of the National Academy of Sciences 109, 20095-20100.
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. The Journal of neuroscience 27, 3839-3844.
Kaur, S., Rose, H., 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. Cerebral Cortex 18, 1560-1574.
Kayser, C., Petkov, C.I., and Logothetis, N.K. (2009). Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. Hearing Res 258, 80-88.
Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition–timing hypothesis. Brain Research Reviews 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. Nature communications 8, 13900.
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.
Luo, H., and Poeppel, D. (2007a). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001-1010.
Luo, H., and Poeppel, D. (2007b). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001-1010.
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.
Möttönen, R., Schürmann, M., and Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. Neuroscience letters 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. The Journal of Neuroscience 32, 12268-12276.
O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015a). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. J Neurosci 35, 7256-7263.
O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015b). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. The Journal of Neuroscience 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. The Journal of neuroscience 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.
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. Proceedings of the National Academy of Sciences 112, 273-278.
Rahne, T., Deike, S., Selezneva, E., Brosch, M., Konig, R., Scheich, H., Bockmann, M., and Brechmann, A. (2008). A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming. Brain Research 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. Proceedings of the National Academy of Sciences 112, 8439-8444.
Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends in cognitive sciences 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. The journal of the acoustical society of america 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. The Journal of Neuroscience 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.
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. The journal of neuroscience 27, 8184-8189.
Voloh, B., and Womelsdorf, T. (2016). A Role of Phase-Resetting in Coordinating Large Scale Neural Networks During Attention and Goal-Directed Behavior. Frontiers in systems neuroscience 10.
Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. Physiological reviews 90, 1195-1268.
Wood, K.C., Town, S.M., Atilgan, H., Jones, A.K., and Bizley, J.K. (2017). Acute Inactivation of Primary Auditory Cortex Causes a Sound Localisation Deficit in Ferrets. PLoS One DOI: 10.1371/journal.pone.0170264.
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.

822 ## Methods

823 Animal preparation: The experiments were approved by the Animal Welfare and Ethical Review
824 Board of University College London and The Royal Veterinary College, and performed under license
825 from the UK Home Office (PPL 70/7267) and in accordance with the Animals Scientific Procedures
826 Act 1986. Neural responses were recorded in a total of 11 awake pigmented adult female ferrets
827 (Mustela putorius furo; 1-5 years old). Data from 9 of these animals was used for the main
828 experiment (532 units), data from 8 other animals (6/9 plus two other ferrets, 89 units) was
829 collected for additional control analysis (Fig. 3e, Fig S3, S6). Animals were chronically implanted with
830 recording electrodes and passively listening/watching stimuli while holding their head at a water
831 spout and receiving continuous water reward. These animals were trained in various listening tasks
832 for other studies. An additional 5 adult females were used to record responses under anaesthesia.
833 Recording electrodes in awake animals targeted tonotopic auditory cortex (fields A1 and AAF and
834 their low frequency boarders with tonotopic belt areas PPF and PSF).

835

836 Full methods for recording under anaesthesia can be found in Bizley et al., (2009). Briefly, ferrets
837 were anesthetized with medetomidine (Domitor; 0.022mg/kg/h; Pfizer, Sandwich, UK) and ketamine
838 (Ketaset; 5mg/kg/h; Fort Dodge Animal Health, Southampton, UK). The animal was intubated and
839 the left radial vein was cannulated in order to provide a continuous infusion (5 ml/h) of a mixture of
840 medetomidine and ketamine in lactated ringers solution augmented with 5% glucose, atropine
841 sulfate (0.06 mg/kg/h; C-Vet Veterinary Products) and dexamethasone (0.5 mg/kg/h, Dexadreson;
842 Intervet, UK). The ferret was placed in a stereotaxic frame in order to implant a bar on the skull,
843 enabling the subsequent removal of the stereotaxic frame. The left temporal muscle was largely
844 removed, and the suprasylvian and pseudosylvian sulci were exposed by a craniotomy, revealing
845 auditory cortex (Kelly et al., 1986). The dura was removed over auditory cortex and the brain
846 protected with 3% agar solution. The eyes were protected with zero-refractive power contact lenses.

847 The animal was then transferred to a small table in a sound-attenuating chamber. Body
848 temperature, end-tidal CO2, and the electrocardiogram were monitored throughout the experiment.
849 Experiments typically lasted between 36 and 56 h. Neural activity was recorded with multisite silicon
850 electrodes (Neuronexus Technologies) in a 1x 16, 2x 16 or 4x 8 (shank x number of sites)
851 configuration. For experiments in which visual cortex was cooled we extended the craniotomy
852 caudally to expose visual cortex and placed a cooling loop over the posterior suprasylvian gyrus.
853 Details of the manufacture of the cooling loop and validation of its efficacy in the ferret animal
854 model are provided in full in (Wood et al., 2017).

855

856 Full surgical methods for recording implanting electrode arrays to facilitate recording from awake
857 animals are available in Bizley et al. (2013). Briefly, animals were bilaterally implanted with WARP-16
858 drives (Neuralynx, Montana, USA) loaded with high impedance tungsten electrodes (FHC, Bowdoin,
859 USA) under general anaesthesia (medetomidine and ketamine induction, as above, isoflurane
860 maintenance 1-3%). Craniotomies were made over left and right auditory cortex, a small number of
861 screws were inserted into the skull for anchoring and grounding the arrays, and the WARP-16 drive
862 was anchored with dental acrylic and protected with a capped well. Animals were allowed to recover
863 for a week before the electrodes were advanced into auditory cortex. Pre-operative, peri-operative
864 and post-operative analgesia were provided to animals under veterinary advice. Recordings were
865 made over the next 1-2.5 years, with electrodes individually advanced every few weeks until the
866 thickness of auditory cortex was traversed.

867

868 Stimulus presentation: All stimuli were created using TDT System 3 hardware (Tucker-Davis
869 Technologies, Alachua, FL) and controlled via MATLAB (Mathworks, USA). For recordings in awake
870 animals, sounds were presented over two loud speakers (Visaton FRS 8). Water deprived ferrets
871 were placed in a dimly lit testing box (69 x 42 x 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 ½-inch microphone (4191). Auditory streams were presented at 65 dB SPL (Fig. 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 behavioural experiment that motivated this study(Maddox et al., 2015), stimuli were 14 seconds in duration. However, we adapted the stimulus duration in awake recordings to 3 seconds in order to collect sufficient repetitions of all stimuli, and to ensure animals maintained their head position facing forwards 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 centre spout. In the anesthetised recording stimulus streams were 14 seconds long, as in the human psychophysics but we only analysed the first 3 seconds to ensure datasets were directly comparable (see also Fig. S7 which replicates analysis for 3 second and 14 second 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 behavioural study, deviants were 200ms epochs in

897 which the identity of the vowel was varied by smoothly changing the first and second formant
898 frequencies to and from those identifying another vowel. Stream A1 was morphed to/from [ɛ] (730,
899 2058, 2857, 4205 Hz) and A2 to/from [i] (437, 2761, 2975, 4263 Hz).

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

909 During anaesthetised recordings, pure tone stimuli (150 Hz to 19 kHz in 1/3-octave steps, from 10 to
910 80 dB SPL in 10 dB, 100 ms in duration, 5 ms cosine ramped) were also presented. These allowed us
911 to characterize individual units and determine tonotopic gradients, so as to confirm the cortical field
912 in which any given recording was made. Additionally broadband noise bursts and diffuse light flashes
913 (100 ms duration, 70 dB SPL) were presented and used to classify a stimulus as auditory, visual or
914 auditory visual. LFPs were subjected to current source density analysis to identify sources and sinks
915 as described by Kaur et al. (2005)

916

917 Data Analysis: Electrophysiological data were analysed offline. Spiking activity and local field
918 potential signals were extracted from the broadband voltage waveform by filtering at 0.3-5kHz and
919 1-150 Hz respectively. Spikes were detected, extracted and then sorted with a spike-sorting
920 algorithm (WaveClus, Quiroga et al., 2004).

921 We used a Euclidean distance based pattern classifier (Foffani and Moxon, 2004) with leave-one-out
922 cross validation to determine whether the neuronal responses to different stimuli could be
923 discriminated. Spiking responses to a given stimulus were binned into a series of spike counts from
924 stimulus onset (0 s) to offset (3s) in 20 ms bins. The average across-repetition response to each
925 stimulus (excluding the to-be-classified response) were used as templates and the response to a
926 single stimulus presentation was classified by calculating the Euclidean distance between itself and
927 the template sweeps and assigning it to the closest template. To determine whether the classifier
928 performed significantly better than expected by chance, a 1000 iteration permutation test was
929 performed where trials were drawn (with replacement) from the observed data and randomly
930 assigned to a stimulus that was then used for template formation / decoding. A neural response was
931 considered to be significantly informative about stimulus identity if the observed value exceeded the
932 95th percentile of the randomly-drawn distribution.

933 This approach allowed us to classify units according to their functional properties: auditory units
934 discriminated two auditory stimuli based on the amplitude modulation of sound (A1 versus A2)
935 regardless of visual dynamics, (Fig. 2a, b), visual units discriminated visual presentations based on
936 temporal envelope of visual stimuli (V1 versus V2) regardless of auditory presentation (Fig. 2c, d)
937 and AV units could do both. This approach was extended to classify dual stream responses by using
938 the average response to each of the temporally coherent single stream stimuli (A1V1 or A2V2) as
939 templates. Performance was (arbitrarily) expressed as the proportion of responses classified as being
940 from the A1, and compared for the two dual stream stimuli with different visual conditions (Figure
941 5). To be considered in this analysis the response of a unit had to be informative about the single
942 stream stimuli (i.e. classified as either auditory and/or visual discriminating).

943

944 Phase/power dissimilarity analysis: Local field potential recordings were considered for all sites at
945 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, 2007a). This analysis compares the phase (or power) consistency across repetitions of the same stimulus with 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; Equ.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*.

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

In which N is equal to the number of trials, and $\theta$ 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 (Equ.2). y. The dissimilarity function for each frequency bin *i* was defined as;

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

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

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

981

982 $$Dual\ Stream\ PDI_i = \frac{\sum_{j=1}^{N} ITPC_{ij}\ within_{dual}}{N} - \frac{\sum_{j=1}^{N} ITPC_{ij}\ across_{dual}}{N} \qquad \text{Equ.3}$$

983

984 Positive dual stream PDI values indicate that the time course of the neural responses was influenced
985 by visual input, despite the identical acoustic input. We determined whether the dual stream PDI
986 was greater if the *within_dual ITPC* was significantly larger than *across_dual ITPC* (pairwise t-test,
987 $p<0.05$ Bonferroni correction, as above). PDI magnitude values were calculated by summing the PDI
988 values across all significant frequencies.

989

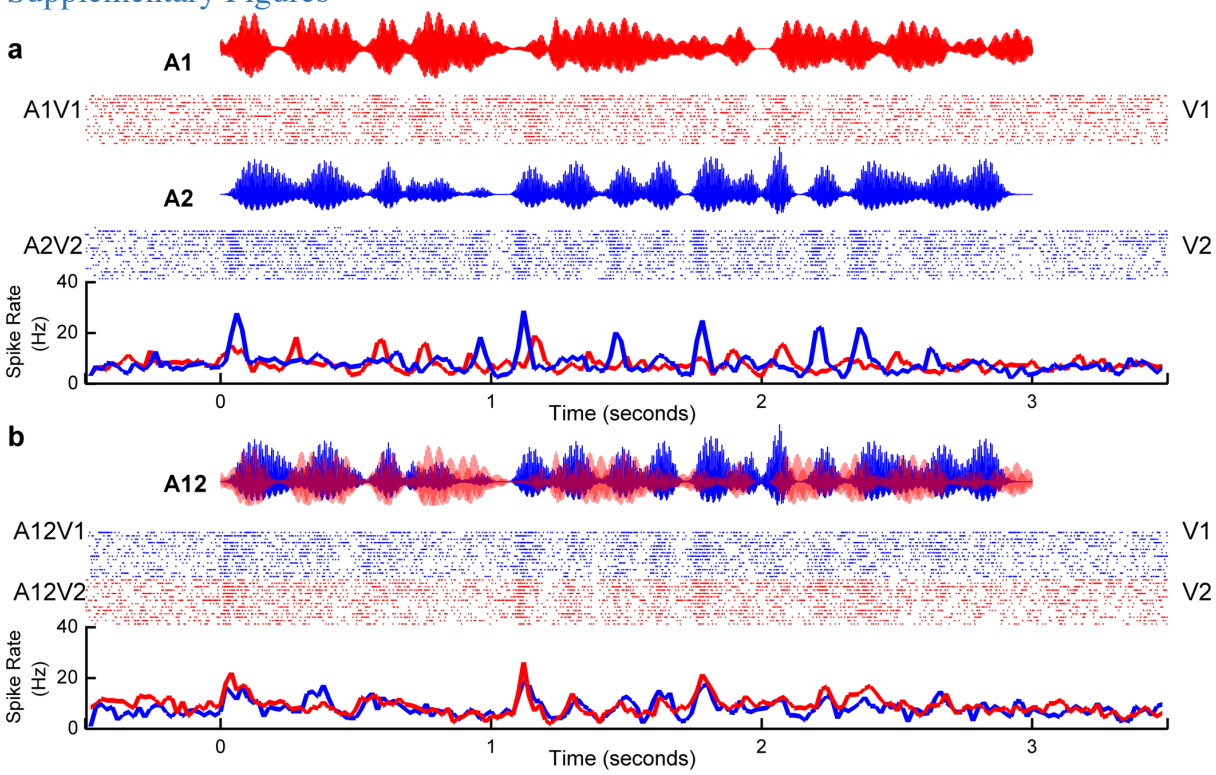
990 Timbre deviant analysis*:* In order to determine how a visual stimulus influenced the ability to
991 decode timbre deviants embedded within the auditory streams we used the cross-validated pattern
992 classifier described above for analysing single stream stimuli to discriminate deviant from no-deviant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

998 **Code and Data Availability**

999 Data and MATLAB code is available from the authors upon request.

1000

1001 ## Supplementary Figures



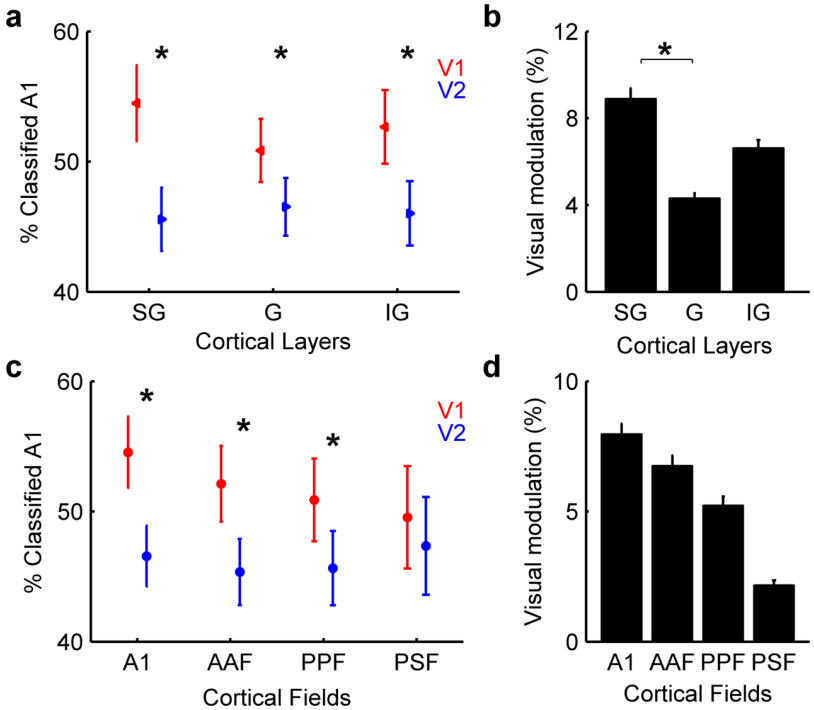

1002

1003 **Supplementary figure 1: related to Figure 2**

1004 **Visual stimuli can determine which sound stream auditory cortical neurons follow in a mixture**
1005 **in auditory-discriminating units**

1006 The spiking responses of an example unit are shown to coherent single stream (**a**) and dual stream
1007 stimuli (**b**). This example unit was an auditory discriminating unit recorded from an awake animal.

When the visual stimulus was V1, 67% of trials were classified as A1V1, when the visual stimulus was V2 only 29% of trials were classified as A1V1.

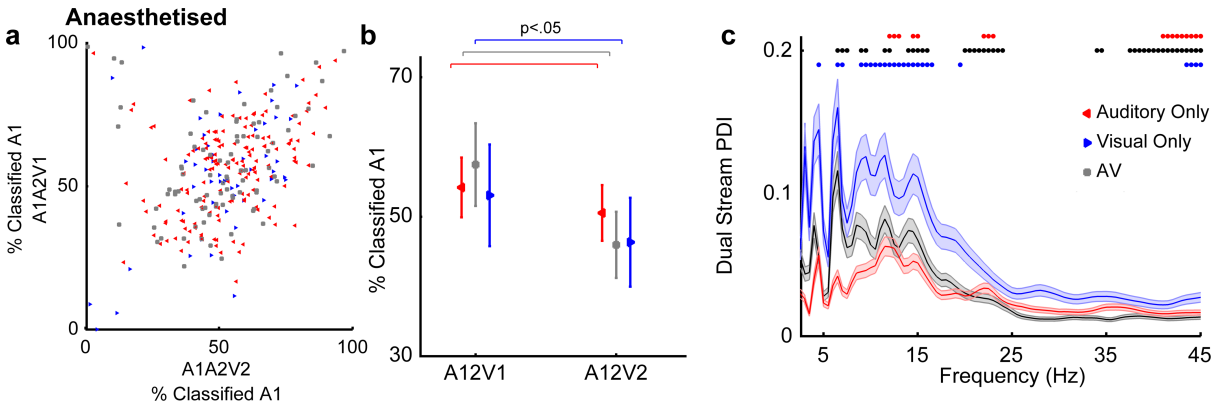



**Supplementary figure 2: Related to Figure 2**

**Preference to the auditory stream with which the visual stimulus was coherent observed across all cortical layers and in three tonotopic fields.**

In the anesthetised animal, we determined tonotopic gradients from responses to pure tone stimuli to accurately confirm the cortical field in which any given recording was made. Recordings were made with linear shank electrodes, facilitating current source density analysis to identify the cortical layers. **a** Proportion of responses classified as A1 when the visual stimulus was V1 or V2 are shown for supragranular layers (SG, $t_{700}$=5.686, p<0.001), granular layer (G, $t_{878}$=3.481 p<0.001) and infragranular layers (IG, $t_{690}$=4.418, p<0.001). Data shown as mean (± SEM) across units. A two-way ANOVA across visual condition and cortical layers showed only a significant effect of visual condition (F (1, 2273) =64.288, p<0.001) but not of layers (F (1, 2273) =0.91, p =0.404) and no interaction was observed (F (1, 2273) =2.679, p=0.068). Significant post-hoc comparisons (p<0.05) are shown with an asterisk. **b,** In order to compare the magnitude of the effect of the visual stimulus across cortical layers we calculated visual modulation as the decoding score when the visual stimulus was V1 – the decoder score when the visual stimulus was V2. Plotted are the mean (± SEM) visual modulation across different layers. A one-way ANOVA across cortical layers showed a significant effect of cortical layer (F (2, 1134) =3.15, p =0.043). Post hoc comparison revealed that visual modulation in SG is greater than G. **c**, as for a but for different cortical fields (A1, $t_{798}$ =5.435, p<0.001; AAF, $t_{636}$ =4.302, p<0.001; PPF, $t_{510}$ =3.609, p<0.001; PSF, $t_{317}$ =0.932, p=0.352). A two way ANOVA across visual condition and cortical field showed only a significant effect of visual condition (Visual condition: F (1, 2267) =40.301, p<0.001; fields: F (1, 2267) =1.937, p=0.121; visual condition x fields: F (1, 2267) =1.804, p=0.144). Significant post-hoc comparisons (p<0.05) are shown with an asterisk. In order to compare the magnitude of the effects across cortical fields we again calculated a visual modulation value. **b,** Mean (± SEM) visual modulation across different

fields. A one-way ANOVA across cortical layers showed a no significant effect of cortical field (F (3, 1130) =2.19, p =0.0877).

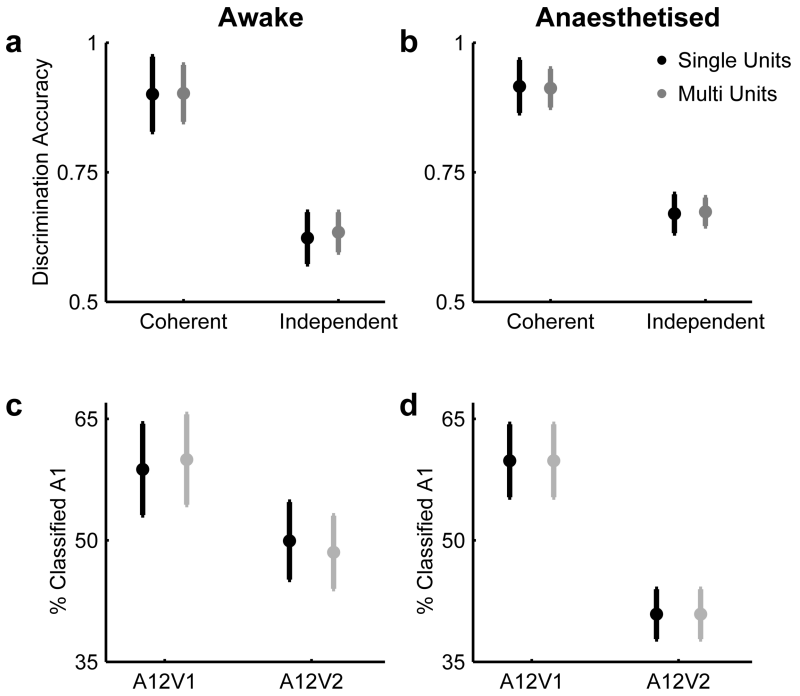



**Supplementary figure 3: Related to Figure 3**

**Auditory, visual and auditory-visual units are influenced by visual stimuli.**

In anesthetised animals we also recorded responses to 100 ms noise bursts and/or LED flashes in order to apply conventional classification of units as either auditory, visual or auditory-visual. Unit classification was based on a two way ANOVA of spike counts calculated over a 200ms window following stimulus onset with auditory and visual stimuli as factors. Units in which both auditory and visual stimuli significantly modulated spiking or in which there was a significant interaction between stimuli were classified as auditory-visual (AV). This analysis yielded recording sites at which units were classified as auditory only (A; n=160 units), visual only (V, n=53) and AV (grey, n=94). This allowed us to explore the visual modulation effects according to independently determined functional definitions of units as auditory, visual or auditory visual.

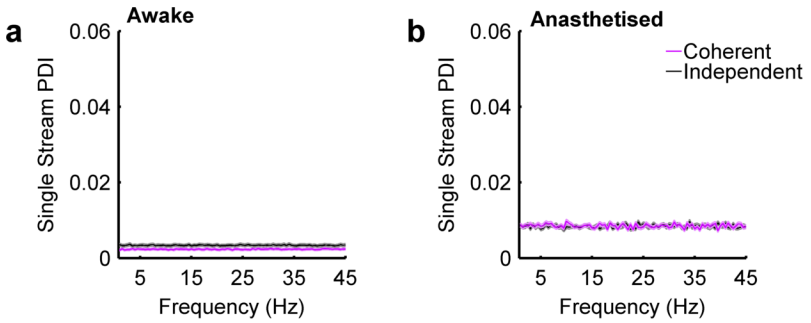
**a**, The proportion of responses classified as A1 when visual stimuli as V1 or V2 is plotted with each unit color coded according classification as A, V or AV. **b**, Mean (± SEM) values across the population. A two-way ANOVA revealed a significant effect of visual condition (F (1, 608) = 18.142, p<0.001), but no effect of unit type (F (2,608) = 0.765, p=0.4657), or interaction between factors (F (1,608) =2.474, p=0.0851) on the proportion of responses classified as A1. Pairwise post-hoc comparisons confirmed a significant effect of visual stimulus in all subgroups (p<0.05). **c**, Mean (± SEM) dual stream phase dissimilarity index (PDI) values for recording sites categorised according to the spiking responses recorded there. Symbols indicate the frequencies at which the dual stream PDI index was significant (pairwise t-test, p<0.001 with correction). While the phase effects are greatest at the sites where visual activity was recorded, significant dual stream PDI values were observed in all three unit types. In all three cases significant phase coherence was seen at 12Hz, 13.5Hz-14.5Hz and 42.5-44.5Hz. Modulation at 10-12 Hz was only observed at sites in which AV and V responses were recorded.




**Supplementary figure 4: Related to Figure 3**

**The effects of temporal coherence on single stream decoding and of visual identity on dual stream decoding are evident in both single and multi-units.**

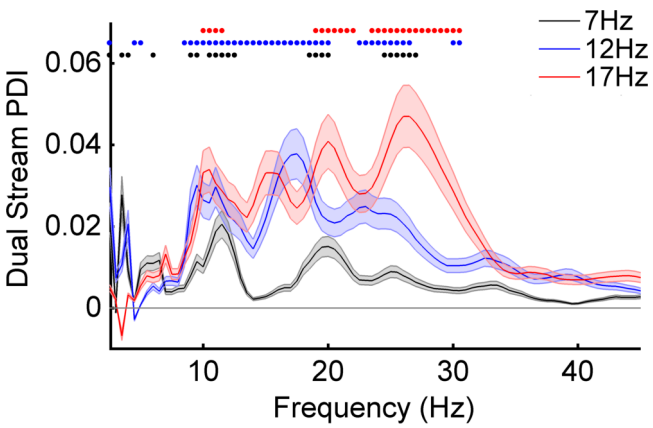
**a,b** Discrimination accuracy for single stream decoding is indistinguishable for single units and multi-unit activity. Two way ANOVA revealed a significant effect of visual condition and no effect of unit type in discrimination accuracy values in awake recordings (a, Visual condition: $F(1,851) = 14.6607$ $p<0.01$, Unit Type: $F(1,851)= 1.02$, $p=0.312$; interaction: $F(1,851)=0.005$, $p=0.449$) and anaesthetised recordings (b, Visual condition: $F(1,1847) = 24.514$, $p<0.01$, Unit Type: $F(1,1847)= 0$, $p=0.986$; interaction: $F(1,1847)=0.66$, $p=0.418$) **c, d**. Impact of visual stimulus on dual stream decoding in single units and multi units. A two way ANOVA revealed a significant effect of visual condition but no effect of unit type in discrimination accuracy values in awake recordings (**c**, Visual condition: $F(1,449) = 19.41$ $p<0.01$, Unit Type: $F(1,851)= 1.07$, $p=0.414$; interaction: $F(1,851)=0.33$, $p=0.568$ ) and anaesthetised recordings (**d**, Visual condition: $F(1,1847) = 15.41$, $p<0.01$, Unit Type: $F(1,1847)= 1.1$, $p=0.506$;interaction: $F(1,1847)=0.66$, $p=0.418$).




**Supplementary figure 5, related to Fig.6**

**Dynamic visual stimuli do not elicit reliable changes in LFP power.**

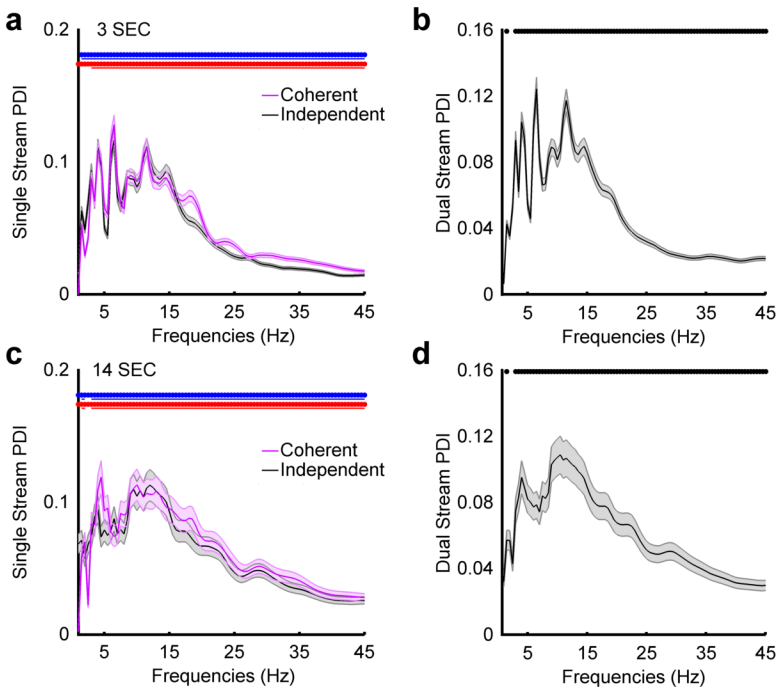
**a, b:** Mean (±SEM) single stream phase dissimilarity index values for coherent (A1V1 and A2V2) and independent (A1V2 and A2V1) stimuli at all recording sites at which there was a driven spiking response for awake (a) and anaesthetised (b) animals.




**Supplementary figure 6: related to figure 6**

**Dual stream PDI values in the alpha range are independent of amplitude modulation rate.**

Stimuli were generated with three different amplitude modulation rates (<7Hz, as in the main experiment, <12Hz and <17Hz, values picked to avoid harmonics of 7 Hz) and responses to these were recorded in 92 units. Symbols indicate where the dual stream phase selectivity index was significant (pairwise t-test, $p<0.05$ with correction). In all three cases significant phase coherence is seen between 10Hz-11.5Hz, 19Hz-20Hz and 24-26 Hz.




**Supplementary figure 7: related to experimental methods**

**Stimulus duration does not influence the frequency range across which significant single stream PDI values and dual stream PDI values are observed.**

We observed that phase coherence values were higher for longer duration stimuli and therefore in order to directly compare anesthetised and awake datasets, all analysis in the anesthetised dataset was restricted to the first three seconds of the stimulus. At longer stimulus durations the ITPC profile and resulting PDI varied more smoothly with frequency (a, c); However at both durations phase values were significantly different from zero at all frequencies. The pattern of significant phase selectivity values was also preserved across stimulus durations. (**b, d**). Frequency points at which the single stream PDI value and dual stream PDI values were similar in 3 second length (**a, b**) and 14 second length (**c, d**) Blue, red and black symbols indicate where the PDI was significant (pairwise t-test, a = 0.0012 with correction).