PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The initial phase of auditory and visual scene analysis

Jean-Michel Hupé and Daniel Pressnitzer

Phil. Trans. R. Soc. B 2012 **367**, 942-953
doi: 10.1098/rstb.2011.0368

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 942–953
doi:10.1098/rstb.2011.0368

Research

The initial phase of auditory and visual scene analysis

Jean-Michel Hupé[1,*] and Daniel Pressnitzer[2,3]

[1]*Centre de Recherche Cerveau et Cognition, Université de Toulouse and Centre National de la Recherche Scientifique, 31300 Toulouse, France*
[2]*Laboratoire de Psychologie de la Perception, Université Paris Descartes and Centre National de la Recherche Scientifique, 75006 Paris, France*
[3]*Département d'Etudes Cognitives, Ecole Nationale Supérieure, Paris, France*

Auditory streaming and visual plaids have been used extensively to study perceptual organization in each modality. Both stimuli can produce bistable alternations between grouped (one object) and split (two objects) interpretations. They also share two peculiar features: (i) at the onset of stimulus presentation, organization starts with a systematic bias towards the grouped interpretation; (ii) this first percept has 'inertia'; it lasts longer than the subsequent ones. As a result, the probability of forming different objects builds up over time, a landmark of both behavioural and neurophysiological data on auditory streaming. Here we show that first percept bias and inertia are independent. In plaid perception, inertia is due to a depth ordering ambiguity in the transparent (split) interpretation that makes plaid perception tristable rather than bistable: experimental manipulations removing the depth ambiguity suppressed inertia. However, the first percept bias persisted. We attempted a similar manipulation for auditory streaming by introducing level differences between streams, to bias which stream would appear in the perceptual foreground. Here both inertia and first percept bias persisted. We thus argue that the critical common feature of the onset of perceptual organization is the grouping bias, which may be related to the transition from temporally/spatially local to temporally/spatially global computation.

Keywords: multistability; auditory streaming; visual plaid; figure ground organization; competition model

1. INTRODUCTION

Auditory scene analysis leads to the formation of perceptual objects, or 'streams', from the flow of acoustic information reaching the ears [1]. It is what allows us to follow a conversation in a crowded restaurant, in the midst of other conversations, with music in the background and the sound of tinkling glasses. An essential feature of streaming is that it takes time: initially, subjects tend to group all of the acoustic information into one global stream [2,3]. When we first walk into the restaurant, the first impression may be of a 'loud and undifferentiated noise' [4]. Only after some time do streams begin to differentiate, allowing switching of attention to the different sound sources. This is termed the 'build up' of streaming. The aim of this paper is to re-examine the initial build up of streaming in a bistable paradigm and to compare it with a similar paradigm in vision.

Streaming has extensively been studied with sequences of tones akin to simple musical melodies [5–7]. For instance, the subject may hear L and H tones, where L and H represent low and high tone frequencies, repeated in an LHL–LHL–... sequence [8]. For such a stimulus, the first report is usually of one stream ('grouped percept') experienced as a single melody LHL–LHL– .... After a few seconds or tens of seconds, however, perception changes to that of two streams ('split percept'), L–L–L– and –H–H–, which are heard as two concurrent melodies that can be attended selectively, but not simultaneously. Because the first switch to two streams is probabilistic, when averaging over subjects and/or stimulus presentations, one observes a gradual increase in the probability of a two-stream percept over time [3,9]. If any sudden change is introduced in the sequence, such as a change in location, loudness, in the silent pause between tones or even in the attentional focus of the subject, streaming is reset and build up starts again [3,7,10,11].

The build up of streaming has been used as an essential landmark of streaming. In studies measuring objective correlates of streaming, the onset *versus* offset of the streaming sequence is usually contrasted, so performance changes can be attributed to build up and not to acoustic manipulations [12]. Build up is also used to investigate the effect of attention on streaming, with subjective [9] and objective [13] methods. In animal electrophysiology, build up provides a useful tool for accessing the temporal dynamics of streaming,

[*] Author for correspondence (jean-michel.hupe@cerco.ups-tlse.fr).

One contribution of 10 to a Theme Issue 'Multistability in perception: binding sensory modalities'.

as it is a measure that can be averaged across trials and that does not require the co-registration of the perceptual state. Correlates of build up have been found both in the auditory cortex [14] and in the cochlear nucleus [15] of the mammalian auditory system.

Interestingly, in all of the published data, after the initial build up, the probability of hearing two streams stabilizes below 100%. As pointed out by Pressnitzer & Hupé [16], this indicates that there are subsequent perceptual alternations back and forth to a one-stream percept after the initial build up. Perceptual reports for long-lasting sequences confirmed that streaming was indeed a bistable phenomenon [16–18]. The build up of streaming was described [16] as a combination of a systematic bias towards the one-stream interpretation at stimulus onset (even when the two-stream interpretation was later experienced most of the time) and a longer duration of this first percept compared with subsequent one-stream percepts (we shall call this duration effect the 'inertia' of the first percept)—see figs 1*a* and S3 of Pressnitzer & Hupé [16]. Such dynamics are different from those observed in classic examples of visual bistability like binocular rivalry, ambiguous figures or apparent motion. In such instances, when both interpretations are equally likely, which percept is first is random (unless the stimulus was presented a short time before; in that case, 'perceptual stabilization' can occur [19,20]). When the stimulus is biased in favour of one interpretation (for example, higher contrast of the stimulus presented to one eye in binocular rivalry), the first percept typically corresponds to the biased interpretation [21]. Also, percept durations are stochastic but, on average, the duration is rather constant over time for each interpretation (average durations being longer for the preferred interpretation), with no inertia of the first percept. Exceptions to this rule occur when observers are not familiar with different interpretations of the stimulus or are not informed that their perception may change; in that case, the first percept may last much longer than subsequent percepts [22–24], as observed especially for children [25]. Changes in the switching rate over long presentation times have also been reported for some [26] but not all [27] ambiguous visual stimuli. In any case, a constant switching rate should only be observed when subjects are luminance adapted (since luminance adaptation over time corresponds to a weakening of stimulus strength that may lead to a decrease in the switching rate [28]) and when they are able to keep a steady fixation, attentional level and decision criterion over long durations.

Such peculiarities of auditory streaming (first percept bias and inertia) may suggest that the build up of streaming is specific to auditory scene analysis and is not related to the general rules of bistable perception. However, there exists another example of such first-percept bias and inertia: in visual bistability, for ambiguous moving plaids. A plaid is an ambiguous stimulus invented by Wallach [22] (see [23] for the English translation) and characterized as a bistable paradigm by Hupé & Rubin [29]. A plaid is a network of crossing lines seen moving through a circular aperture. It can be perceived either as a single pattern (or 'plaid') moving in a given direction (percept of 'coherent' motion) or as two gratings sliding in opposite directions on top of each other (percept of 'transparent' motion; visit http://cerco.ups-tlse.fr/~hupe/plaid_demo/demo_plaids.html for a demonstration). Over time, perception alternates between coherence and transparency. The first percept for ambiguous plaids tends to be the coherent one whatever the stimulus parameters and the steady-state probability of coherent motion, and it lasts longer than subsequent coherent percepts [16,27,29,30]. There is a formal correspondence between visual plaids and auditory streaming in terms of organization of the sensory scene [31]: a decision has to be made whether to group the scene into one stream/one plaid, or to segment the scene between two streams/two gratings. The coherent visual percept corresponds therefore qualitatively to the one-stream percept in auditory streaming. When averaged across several presentations, as done for the build up of auditory streaming, the percept choice for visual plaids clearly exhibits 'build up' dynamics (figure 1).

Hupé & Rubin [29] were the first to document this surprising behaviour of plaids, but to date there is no explanation for the first percept bias and inertia in vision. Here we evaluate the hypothesis that the build up of auditory streaming and plaid segmentation are due to similar mechanisms. By taking advantage of specifics of each stimulus, we try to achieve a thorough explanation that would have been difficult to test by considering each modality alone.

2. BUILD UP OF VISUAL PLAIDS: THE TRISTABILITY HYPOTHESIS

From a phenomenological point of view, plaids are different from other classic bistable paradigms: not only do percepts alternate between coherence and transparency, but they also alternate between two different transparent percepts distinguished by which grating is perceived as foreground and which as background. When put in a pure transparency range (for example, by increasing the angle alpha between the grating directions so the probability of coherence is zero: see fig. 12*a* in [29]), depth ordering bistability is observed [32]. When ambiguity is present between coherence and transparency, there are therefore three possible percepts. These percepts are experienced by all subjects, as shown by the data obtained when they are given the choice of reporting them (figure 2*a* and §2*b*). Plaid perception is therefore bistable when considering only the choice between coherence and transparency (whether one or two moving objects are perceived) but tristable when adding the depth interpretation choice.

We tested whether the build up of plaid segmentation could be due to tristability, since visual stimuli producing purely bistable perception do not seem to show any build up. To do so, we forced plaid perception to bistability by introducing either occlusion cues (figure 2*b*, experiments I and III) or stereoscopic cues (figure 3, experiment II) that removed the ambiguity of depth ordering. A preliminary report of these data has been published in abstract form [33,34].

(a) *Material and methods*

Eighteen observers (including the first author and a research assistant) participated in experiment I (effect

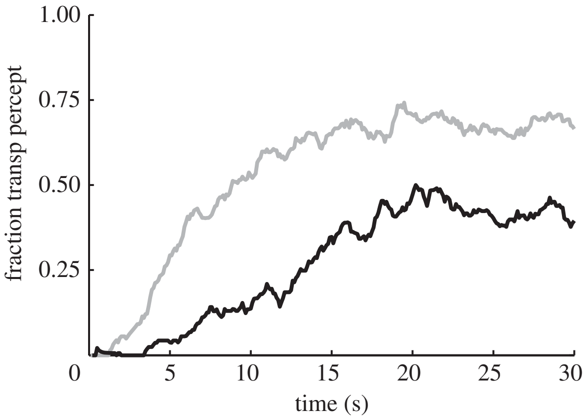

Figure 1. Build up of plaid segmentation for two stimulus parameters (angle alpha between the grating directions) computed across 14 subjects (12 values per subject per condition). 'transp' = transparent interpretation (split percept), perceived as two gratings sliding over each other (J. M. Hupé 2007, unpublished data). Grey line, alpha = 125; black line, alpha = 105.

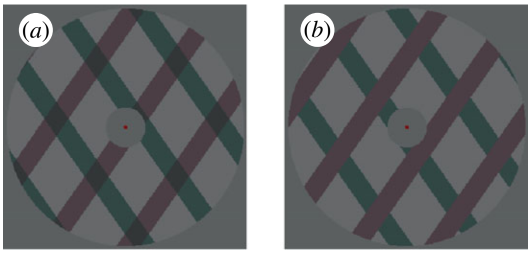

Figure 2. Examples of visual plaids. (*a*) Plaid with transparent intersections. When set in motion upwards, it can be perceived as a single coherent plaid moving upwards or as two gratings sliding over each other, the green one to the right and the red one to the left. Gratings were tinted with red and green, so subjects could easily answer the question: 'When transparent, which grating is in front?' irrespective of the plaid direction of motion. All subjects tested so far ($n = 30$) reported alternations of which grating was perceived in front, in addition to coherent/transparent alternations. (*b*) Plaid with occlusions cues. When transparent motion was experienced, subjects always reported perceiving the occluding grating in front (here, the red one). See http://cerco.ups-tlse.fr/~hupe/plaid_ demo/demo_plaids.html for an interactive demonstration.

of occlusion cues). Five new observers and the first author participated in experiment II (effect of stereoscopic cues). Eight other observers and the first author participated in experiment III (first percept bias). All subjects (except the first author) were naïve about the purpose of the experiments. They had normal or corrected-to-normal eyesight and gave informed consent for their participation. The six subjects in experiment II were verified for correct stereoscopic vision (at least 1 min of arc minimum stereo acuity).

We presented stimuli on a 19 inch Sony Multiscan G400 Trinitron screen (26.25 cm vertical viewable screen size) controlled with a PC running Windows 2000. Resolution was 1024 × 768 pixels and the refresh rate was 75 Hz. The viewing distance was 57 cm. Stimuli were generated on line by custom software written in C++ and using the OpenGL library for anti-aliasing and the SDL library for precise control of timing. Stimuli were ambiguous moving plaids, as used by Hupé & Rubin [29].

The stimuli of experiments I and III comprised two rectangular-wave gratings presented through a circular aperture, 12° in diameter. The luminance of the grey background outside the aperture was 34.7 cd m^{-2} and was similar to the average luminance of the plaid stimulus. The gratings comprised thin dark stripes (duty cycle = 0.3, spatial frequency = 0.3 cycle deg^{-1}, 20.1 cd m^{-2}) on a lighter grey background (37.1 cd m^{-2}), and appeared as figures moving over the background. One of the gratings was reddish (CIE 1931 $x = 0.310$, $y = 0.307$) and the other one greenish (CIE 1931 $x = 0.279$, $y = 0.321$). The intersecting regions were either grey and darker than the gratings (multiplicative transparency, 14.4 cd m^{-2}, CIE 1931 $x = 0.294$, $y = 0.312$), so that it was ambiguous as to which grating was in front, or had the same colour as one of the gratings, so that one of the gratings occluded the other and was perceived as in front. Gratings moved at 1.5° s^{-1} (measured in the direction normal to their orientation) in directions 115° apart (angle alpha) in experiment I or 110, 130 and 150° apart in experiment III. A red fixation point over a 1° radius circular grey mask was added in the middle of the circular aperture to minimize opto-kinetic nystagmus (OKN), and subjects were instructed to fixate this point throughout the stimulus presentation. The pattern could move in eight possible directions (when perceived as coherent), four cardinal (right, left, up, down), the other four oblique (45° from a cardinal axis). There were two repetitions of each stimulus (eight directions of motion by two luminance conditions for the intersections) for a total of 32 (experiment I) or 96 (experiment III) stimuli. For stimuli with occlusion, intersections were red half of the time and green the other half. Presentation time was 60 s in experiment I and 10 s in experiment III.

The stimuli for experiment II were adapted for stereoscopic presentation. We used the 'version 2' stereoscope described by Randolph Blake (http://www.psy.vanderbilt.edu/faculty/blake/Stereoscope/stereoscope.html). Mirrors located close to the eyes allowed each eye to view only half of the computer screen. Each plaid was displayed twice on the screen, one side for each eye. The background outside the aperture was black, and binocular fusion of the circular apertures was helped by horizontal and vertical nonius lines (two colinear segments, only one seen by each eye, which subjects should perceived as aligned) presented before each stimulus, as well as static points displayed all the time around the aperture (figure 3). The circular aperture was 8° in diameter and the central mask was 0.8° in radius. Spatial frequency was 0.5 cycle degree^{-1}. We used only the four cardinal directions of plaid motion to avoid having one of the gratings close to the horizontal (in which case depth cannot be manipulated with stereoscopic cues). The angle alpha took the following values: 80, 85, 90, 95 and 100°. Either the green or the red grating was displaced horizontally by 0.4° (disparity), so it was perceived behind the fixation plane (figure 3). Each subject viewed a total of 40 stimuli. Presentation time was 60 s.

Subjects were comfortably seated in a darkened room in front of the computer screen with their head

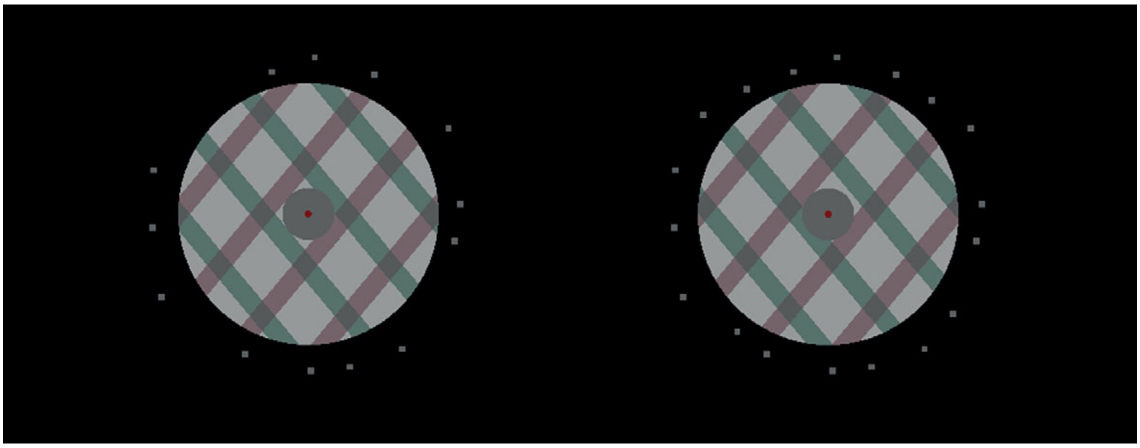

Figure 3. Example of the stimuli displayed on the computer screen for stereoscopic experiment II. By looking at the stimulus through the stereoscope, subjects fused both images within a single plaid (the right eye looked at the right plaid and the left eye at the left plaid). Here the green bars are displaced to the right by 0.4°, making them appear in depth behind the rest of the stimulus.

on a chin rest. In experiment I, binocular eye positions were monitored continuously at 240 Hz using an ISCAN ETL-200 system (Burlington, MA, USA). The cameras were attached to the chin rest just above the eyes and imaged the eyes through semi-transparent mirrors. Subjects looked at the screen through these mirrors. They were tested in two sessions with all 32 stimuli. In the first session, they were first presented with examples of the stimuli and asked about their perception until they spontaneously reported switching. Once they were familiar with the coherent and transparent percepts, we asked half of them to try to indicate, when they perceived transparent motion, whether the red or green grating was in front. For the other half of the subjects, we introduced this question only in the second session. Some subjects found it very easy to report which grating was in front while others found it difficult (their results were, however, not different). In that case, we helped them by indicating that the front grating is typically the one that one pays more attention to [22,23,35] and that captures the sense of motion. Such a phenomenon was validated by informal observations made in the laboratory by many subjects over the past 10 years: when asked about the direction of motion of a plaid moving upwards (coherent direction), naïve subjects often first reported upward motion, then either rightward or leftward motion, then the opposite, until they realized that there were in fact two directions of motion for the transparent percept. All subjects managed to identify the three types of percept. Then subjects were familiarized with the task; depending on the session, they had to continuously report their percept with either a two- or three-button mouse. In the two-choice task, they had to choose between coherence and transparency, whereas in the three-choice task they also had to report, for the transparent percept, whether the red or the green grating was in front. If they were unsure of their percept, they were asked not to press any button. We insisted that they should not hesitate to use this option, especially for subjects who were not that confident about their judgement at the beginning of the experiment. The experimenter sat next to the subject during the whole experiment to control eye movement acquisition and to start the presentation of each stimulus. The fixation point was present throughout the experiment. The experimenter ensured that subjects fixated before starting each trial. After sixteen 1-min trials, subjects were given a rest. They took part in the second session on a different day.

In experiment II, subjects performed the three-choice task; no eyetracking was performed because the stereoscope took the place of the cameras. In experiment III, subjects performed the two-choice task; eye position was not monitored and subjects started each trial themselves by pressing the space bar (11 trials out of 864 were skipped because subjects pressed the space bar twice in a row).

(b) *Results*

We compared the dynamics of perception of bistable and tristable plaids during 1-min trials. In experiment I, 15 (14 naïve) subjects were tested in the two sessions (two- and three-choice tasks) and were included in the analysis. Six of them performed the two-choice task first. There was no effect of task order. For the three-choice task and plaids with ambiguous intersections (figure 2*a*), all of them reported the three possible percepts and were unsure of their percept at most 5% of the time, confirming the reality of the tristable rather than bistable nature of ambiguous plaid perception. On average, the green and red gratings were reported about equally as perceived in front. In the occlusion case (figure 2*b*), however, subjects systematically reported perceiving the occluding grating being in front, and therefore experienced at most two percepts within each trial (bistable perception). In order to compare the results obtained with the two tasks, we analysed the data for the three-choice task as if the task was two-choice—that is, we considered only whether the plaid was perceived as coherent or transparent. In order to compute percept durations, we ignored the short periods where no percept was reported (corresponding, for example, in the

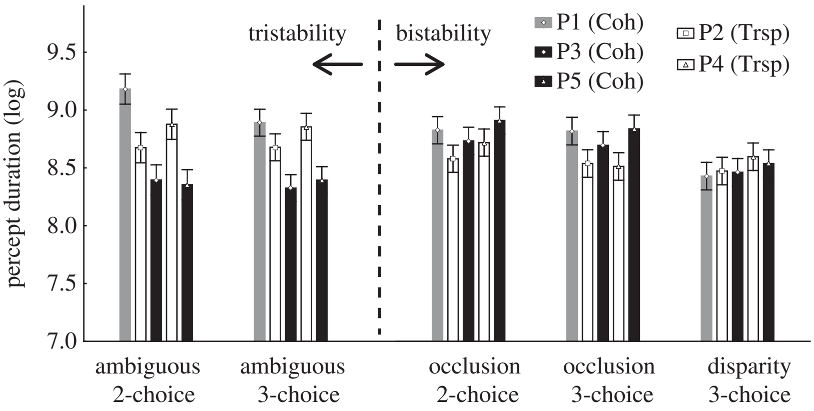

Figure 4. Dynamics of plaid perception for tristable and bistable plaids. First four groups of bars: data from 14 subjects (experiment I). Last group of bars: experiment II (six subjects). In each group, bars from left to right correspond to successive percepts within one trial. Each percept duration was computed in milliseconds (and transformed to its natural logarithm) from the button press indicating a given percept—coherent (Coh) or transparent (Trsp)—to a button press indicating the next percept, thus disregarding which grating was reported in front in the three-choice task, as well as periods with no button press. Errors bars denote 95% confidence intervals computed with ANOVA models including the variable 'subject' as a random factor. We included only trials that started with the coherent percept and with at least five switches between coherence and transparency. The first two groups of bars are for tristable plaids (transparent intersections leading to depth ordering ambiguity). Average percept durations were stable over time, except for the first coherent percept (P1, Coh) which lasted longer than P3 (Coh) and P5 (Coh) for both the two-choice and the three-choice tasks. The next three groups of bars display the results obtained for plaids made bistable either by occlusion or disparity. All coherent percepts had the same average duration, whatever the manipulation (occlusion or disparity) and the task (two-choice or three-choice).

three-choice task to the transition between two depth orderings of the transparent percept). Within each trial we therefore had responses corresponding to a sequence of coherent/transparent percepts. We transformed each percept duration to its logarithm [29] and kept only trials that started with a coherent percept and with at least three complete coherent percepts reported within 60 s (in order to compute the average duration of the first five percepts within the same trials). Fourteen subjects had enough trials for the different conditions. The analysis was performed on 577 trials (out of 32 stimuli × 2 sessions × 14 subjects = 896 trials). Figure 4 shows the average percept durations as a function of percept sequence within each trial, for stimuli with transparent intersections and therefore ambiguous depth relationship (first two groups of bars) and for stimuli with one grating occluding the other (next two groups of bars).

There was no significant effect of the task ('2' *versus* '3' buttons) on percept duration. This shows that there was no response or attentional bias due to the nature of the task. We collapsed the data from the two tasks and computed repeated-measure ANOVAs on the (log) durations of the first three coherent percepts. For tristable perception (first two groups of bars), durations differed significantly: $F(2, 578) = 104$, $p < 10^{-15}$, effect size (partial η^2) ${}_{\mathrm{p}}\eta^2 = 0.27$ (290 trials included). The first coherent percept was longer than the subsequent coherent percepts (what we called the inertia of the first percept), as observed previously [16,27,29,30]. That was not the case for bistable perception due to occlusion: $F(2, 572) = 4.8$, $p = 0.009$, ${}_{\mathrm{p}}\eta^2 = 0.016$ (287 trials included; the p-value is significant but the very small effect size indicates that the difference is in fact negligible; moreover, it is only due to a minor difference between the second and the third coherent percepts). To estimate the strength of the occlusion/bistability manipulation, we computed for each trial the duration difference between the first, coherent, percept and the average of the next two coherent percepts. The effect of occlusion on inertia was very strong ($F(1, 15.3) = 60$, $p < 10^{-6}$, ${}_{\mathrm{p}}\eta^2 = 0.8$) and not significantly different across subjects (interaction effect between occlusion and 'subject' random factor, $F(13, 13) = 1.59$, $p = 0.21$, ${}_{\mathrm{p}}\eta^2 = 0.6$).

Experiment II confirmed the absence of inertia for bistable plaids by measuring the dynamics of perception of plaids made bistable by stereo depth cues. The cues were efficient for the six tested subjects: when they perceived transparency, they systematically reported the red grating being in front when the green grating was stereoscopically presented behind the fixation plane, and *vice versa*. They were unsure of their percept for less than 3% of the total time. Within the 223 trials that started with the coherent percept, 165 had at least three complete (not interrupted by the end of the 60 s trial) coherent percepts reported within 60 s (at least 13 trials for every subject). There was no inertia: coherent percept durations were stable over time (figure 4, last groups of bars; $F(2, 318) = 0.66$, $p = 0.42$, ${}_{\mathrm{p}}\eta^2 = 0.005$).

Inertia of the first coherent percept was therefore due to tristability. But the first percept bias was still present for bistable plaids. The first percept was the transparent one for only 29 out of 448 trials in experiment I (occlusion manipulation) and 17 out of 240 trials in experiment II (disparity manipulation), suggesting that the first percept bias was not related to tristability. However, in the occlusion case, the average steady-state probability of coherence for bistable plaids was close to 50% or even slightly above (figure 4: the coherent percepts—dark bars—lasted slightly longer than the transparent ones—white

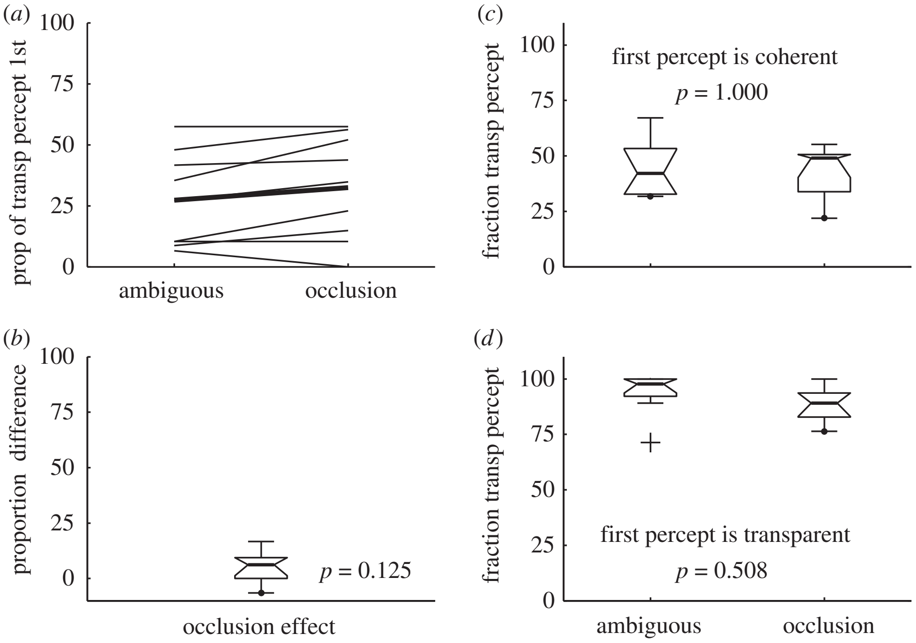

Figure 5. Experiment III: the first percept bias does not depend on occlusions cues. (*a*) Proportion of trials for which subjects reported the transparent percept first, for tristable ('ambiguous') and bistable ('occlusion') plaids. Thick line: average data for nine subjects; thin lines: data for each subject based on 48 trials for each condition. (*b*) The proportion difference was not statistically significant (sign test, d.f. = 8). Median, lower and upper quartiles are displayed. Whiskers mark the most extreme values within 1.5 times the interquartile range. Crosses denote outlier values. We computed within each trial the fraction of time the percept was reported as transparent, as a function of whether the first percept was coherent ((*c*) $n = 596$) or transparent ((*d*) $n = 257$). The luminance of the intersections had no effect (paired sign test, d.f. = 8).

bars). For tristable plaids, the first percept was the coherent one even when the transparent percept was dominant over long presentation times [29]. In experiment III, we tested whether that was also the case for bistable plaids. We modified the plaid parameters to make perception more within the transparency range and focussed on the first percept choice with short-duration (10 s) trials. We computed the proportion of trials for which the transparent percept was reported first. This happened relatively often because of our choice of parameters, thus avoiding possible floor effects. The proportion of first transparent percepts was not different when intersections were transparent (tristable plaid) or when they were opaque (bistable plaid), as shown in the left panels of figure 5. When the first percept was transparent, coherence was not or rarely experienced during the next 10 s (meaning that the stimuli were outside of the ambiguity range), while when the first percept was coherent, transparency was still experienced on average 50% of the time (right panels). Importantly, this was the case for both tristable and bistable plaids.

These results clearly demonstrate that the build up of plaid segmentation depends on two independent mechanisms, first percept bias and inertia, and only one of them (inertia) depends on whether perception is bistable or tristable.

3. TRISTABILITY FOR AUDITORY STREAMING

Auditory streaming shares some common features with plaid perception, such as whether the scene is grouped into one stream (corresponding to the coherent percept for plaids) or is split into two streams (corresponding to the transparent percept). We wondered whether the build up of auditory streaming could also be due to a tristable perception (we call this the 'tristability hypothesis'). For plaids, the visual system has to decide which grating is in front, since there cannot be two objects at the same place, unless one is in front of the other. Is there an analogous figure/ground or depth ordering organization in audition, once the sequence of sounds is interpreted as intertwined sounds coming from different sources? In vision, subjects sometimes found it difficult to be aware of which grating was in front, even though such perceptual decision should be mandatory. However, they all noted that their attention was spontaneously driven towards either the red or the green grating, especially following a switch from coherence; they may even have monitored which grating they were paying attention to rather than which one they 'perceived' as in front. We hypothesized that similar attentional switches may exist for auditory streaming: in the two-stream percept, one may focus on the L–L–L– stream or on the –H–H– one. We collected data in a streaming paradigm, but asked subjects to indicate not only the one-stream *versus* two-stream decision, but also which stream was perceptually dominant. We used a novel three-percept task for streaming and attempted to bias which stream subjects would pay attention to, by making one set of tones, L or H, higher in level (and therefore louder) than the other.

(a) *Material and methods*

Twelve subjects (11 were naïve) participated in the auditory experiment. All reported normal hearing and no history of neurological disorder. Two subjects were musically trained (more than 10 years). The results for one subject had to be excluded *a posteriori* because of technical issues (defective headphone leading to monaural presentation on a few trials).

Because of the asymmetry in the L–H–L– stimulus (the L tone occurs twice as often as the H tone), we tested two sequence types: low–high–low (LHL) and high–low–high (HLH). The exact frequencies, L and H, were roved over about two octaves from trial to trial. This was intended to reduce across-trial adaptation effects [36]. For each trial, the frequency of tone L was drawn from a uniform distribution on a log-scale starting at 440 Hz and with a range of two octaves (up to 1760 Hz). The frequency of tone H was then computed with an interval relative to L of three or five semitones (18.9 or 33.5%, respectively). Three level conditions were used: equal sound pressure level between A and B, +18 dB difference between L and H with H being higher in level, and −18 dB difference between L and H with L being higher in level. Level conditions ($n = 3$) and sequence types ($n = 2$) were crossed, so there were six different types of stimuli. The frequency interval was five semitones in the equal-level conditions and three semitones in the unequal-level conditions. This was set after pilot data to match approximately the overall probability of two-stream percepts in all conditions. The L and H tones had a 120 ms duration, including 10 ms onset and offset ramps (raised cosine). They were presented without any intervening silence during the triplet. The silent gap between triplets was 120 ms. The overall sound level was calibrated for a 440 Hz tone in the equal-level condition and set to 65 dBA. All tones were first normalized to the same root-mean-square value. The level differences were then applied as appropriate, by scaling the individual tones while still ensuring that the total level in a sequence was held constant across conditions. For instance, LHL +18 dB corresponded to an L tone at 56 dBA and an H tone at 74 dBA. The total sequence duration was 30 s. Twenty repeats for each of the six stimulus types were used for each subject, spread over four experimental blocks (1.5 to 2 h). To facilitate perceptual reports, the sequence type (LHL or HLH) was fixed during a block, but it was counterbalanced across subjects.

Subjects were seated in a double-walled sound insulated booth (Industrial Acoustics). Sounds were presented over an RME Fireface UC sound card and Sennheiser 250 Linear II headphones. Subjects were asked to report continuously their dominant percept by means of a computer keyboard. Four choices were possible: a single stream; a two-stream percept with the fast stream dominant; a two-stream percept with the slow stream dominant; ‘don’t know’. Training was provided to familiarize the subjects with the task. They could manipulate a graphical interface where they could change the frequency difference between the L and H tones over a broad range (1–16 semitones), thus favouring the perception of one or two streams. They could choose to hear both tones, only the fast stream or only the slow stream. When subjects reported being confident with the instructions and task, they were run in a short training block without feedback to familiarize themselves with the response interface, and data collection then began.

(b) *Results*

All 11 subjects were able to indicate which stream they thought was dominant in their perception: they indicated being unsure of their percept less than 1% of the total listening time. When all tones had equal level, all subjects reported the three possible percepts: one stream (mean = 47 ± 5.0% SEM), two streams and attending to the fast stream (35 ± 3.3%) or to the slow stream (18 ± 2.6%). The bias in favour of the fast stream was especially strong for four subjects who reported it between three and five times more often than the slow stream (reported only between 6 and 11% of the total time). The ratio in favour of the fast stream was between 1.2 and 2.4 for the other seven subjects. The sequence type had a significant influence on the proportions of two-stream percepts (LHL reported 7% more often as a fast stream, $p < 0.003$) but without any interaction with the level conditions ($p > 0.09$), so the results were pooled across LHL and HLH sequences.

We analysed the data for percept duration. For this analysis, the two main types of percepts (one-stream *versus* two-stream) were contrasted, in the same way as we did for plaids. We kept only trials that started with a one-stream percept and with at least three complete one-stream percepts (meaning at least five percept switches) reported within the 30 s. Seven subjects produced enough trials for this analysis in all conditions. The analysis was performed on 418 trials (out of 120 stimuli × 7 subjects = 896 trials). There were, respectively, 158, 105 and 155 trials for each level condition shown in figure 6*a*, each subject contributing between 30 and 118 trials. Figure 6*a* shows the average percept durations within each trial, for the different level conditions. The first one-stream percept was systematically longer than subsequent one-stream percepts, displaying the expected inertia.

The three level conditions were designed to favour either tristable or bistable perception. We estimated the presence of tristability by measuring the proportion of time each percept lasted for the seven subjects included in the duration analysis. For this new analysis, we used all of their 120 trials and not only the ones with at least five switches, in order to get more reliable estimates. Subjects reported more often attending to the louder tone (figure 6*b*). However, the level manipulation was not as effective in producing bistability as the occlusion or disparity cue in vision: for all level conditions, the three response categories were used.

Perhaps because of the bias in favour of the fast stream, there was no condition for which there was a clear balance between two-stream percepts (fast or slow dominant stream), as observed for plaids for transparent percepts (red or green grating in front). Moreover, the link between inertia and tristability did not seem to be very tight: inertia was as strong when the fast tone was louder as for the other conditions, even though the percept was on average

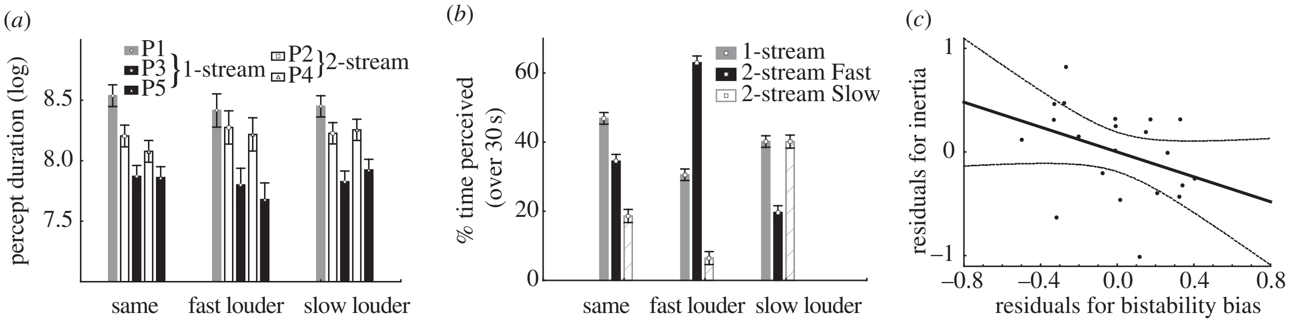

Figure 6. Test of the hypothesis that the build up of auditory streaming is due to tristable perception. (*a*) First percept inertia (longer P1) was present when the fast and the slow tones had the same intensity, but also when one was higher in level (+18 dB) than the other and thus perceived as louder. (*b*) The loudness manipulation had the expected effect on the reported percepts: when two streams were perceived, the louder tone was perceived more often as dominant ('2-stream Fast' means that the fast stream was dominant). However, on average, some tristability was always experienced. (*c*) When analysing the data at the single-subject level, we observed a (non-significant) negative relationship between first percept inertia and the bias towards bistability (21 values, regression line with 95% confidence intervals). See text.

much closer to bistability. There were, however, important differences between subjects that are not apparent in the average data. In order to capture these, we computed the first percept inertia for each condition and subject (using the data averaged in figure 6*a*): P1 – (P3 + P5)/2. Differences varied between 0.2 and 5.4 s. We then computed a bistability index using all trials for each condition and subject (using the data averaged in figure 6*b*):

$$\text{bistability bias} = \frac{|\text{2-stream Fast} - \text{2-stream Slow}|}{\text{2-stream Fast} + \text{2-stream Slow}}$$

This index varied between 0.02 and 1, with 0 indicating balanced tristability and 1 indicating pure bistability. We computed an analysis of covariance (ANCOVA) between bistability index and first percept inertia, including the random variable 'subject'. Figure 6*c* shows the residual correlation: inertia was negatively related to bistability bias, as predicted. The correlation was not significant ($F(1, 13) = 2.02$, $p = 0.18$, ${}_{\mathrm{p}}\eta^2 = 0.13$), but we did not have much power to test it (power = 0.26).

The final analysis we performed on the streaming data was related to the first percept bias (one-stream or two-stream). The first percept was overwhelmingly one-stream (two-stream percepts were reported first in only 180 out of 1319 trials). Importantly, for the unequal-level condition, stimuli for trials starting with the one-stream percept ($n = 728$) were subsequently experienced as two streams for 59% of the whole trial duration, while for trials starting with the two-stream percept ($n = 152$) stimuli were experienced as two streams for 88% of the trial duration.

These results confirmed the presence of both first percept bias and inertia for auditory streaming, similar to those observed in previous experiments [16]. Subjects were able to perform a tristability task for auditory streaming when instructed to, so, just as we argued for plaids, first percept inertia may be due to tristability dynamics. However, the tristable *versus* bistable manipulation was less successful for streaming than for plaids. Perhaps, as a result, we observed no significant relationship between a bistability index and inertia. In addition, we observed a large variability in reports between subjects, possibly suggesting a qualitative difference between the visual and auditory cases. In the visual case, the depth ordering was defined by optical cues and putatively independent of attention, whereas for auditory streaming, the dominant percept may have been selected by voluntary attentional biasing, with potentially different strategies across subjects.

4. DISCUSSION

In this series of experiments, we addressed the question of the initial phase of multistable perception for both visual plaids and auditory tone sequences (streaming paradigm). These two types of stimuli show two intriguing common characteristics: the first perceptual report is usually of a single perceptual object (coherent plaid or one stream), which we termed the first percept bias, and this first report lasts longer than subsequent reports of the same kind, which we termed the first percept inertia. As a result, the perceptual segregation of both stimuli builds up over time. The present results show that different processes may be at work for first percept bias and inertia. For visual plaids, inertia could be eliminated by having a truly bistable paradigm (only two possible percepts) instead of a tristable one (one coherent percept *versus* two transparent percepts). However, the first percept bias persisted even for bistable plaids. For auditory streaming, in spite of being able to manipulate the amount of tristability, we failed to abolish either the first percept bias or its inertia. In the following, we discuss how those results inform our understanding of the perceptual competition between percepts in the initial phase of multistable perception.

(a) *First percept inertia*

For plaids, inertia is clearly related to tristability. Tristable perception is experienced for plaids because there is a conflict between different cues for the transparent percept: the two gratings are perceived as moving in different directions, yet they share intersections. These intersections, when in the luminance transparency range, may be assigned equally well to one or the other grating, but not to both, as long as the stimulus is interpreted within a single plane. Visual objects cannot be in the same plane, share intersections, but move in different directions. Therefore, a resolution of the conflict requires a three-dimensional interpretation of the visual scene with one object above the other. For plaids, this leads to depth ordering ambiguity and perceptual alternations. Why inertia is related to depth ordering competition is not intuitive, but a plausible hypothesis can be formalized in a standard model of multistable perception (appendix: mechanisms of plaid multistability).

For auditory streaming, the cause of first percept inertia remains unclear. We failed to induce clear bistability by manipulating the relative level of the L and H tones. This could be because there is no inherent conflict in having two concurrent streams in hearing, with or without different levels. The dominant percept (fast split or slow split) may thus emerge not because of the resolution of an inherent foreground/background incompatibility, but rather because of attentional switches largely under the volitional control of the subject. In this case, each subject may chose a different criterion and decide to be influenced or not by the level differences. This is consistent with the variability observed in the auditory data. The present experiment can then be interpreted in two ways: either there is no real tristability for stream segregation and the first percept inertia needs some other explanation, or there is tristability (and percept inertia can be explained in the same way as for plaids) but we failed to eliminate it with our stimulus manipulation.

(b) *First percept bias*

Both the visual and auditory experiments show that the first percept bias (the tendency to start with a grouped percept after stimulus onset) is robust to several stimulus manipulations and independent of first percept inertia. Thus, in all cases studied here, a build up of perceptual organization was observed in the averaged data. This first percept bias could be due to different causes in vision and audition, if, for example, for plaids it was due to eye movements (Moreno-Bote 2011, personal communication). Eye movements were minimized by having the fixation point always present before the onset of the moving stimulus, as well as by the presence of a circular grey mask around the fixation point. We measured eye position in experiment I as well as in the experiments used for figure 1 (906 trials in total, first percept reported as coherent in 892 trials). Three subjects had almost no detectable eye movement at all and yet they experienced the first percept bias. The other subjects showed almost no large saccade but made small (less than or around 1° magnitude) OKN. Ocular drifts, when present, followed the plaid ('two-dimensional' motion, see below) as often as the gratings' directions ('one-dimensional' motion, see below). Importantly, two-dimensional OKN were present during the first 1 s of stimulus presentation in only 18.5% of the trials. We conclude that the first percept bias cannot be due to eye movements in the case of plaids and, therefore, a similar explanation to the first percept bias may be sought for both auditory streaming and plaid perception.

A plausible cause for the first percept bias may be easier to describe for the auditory case. The streaming stimulus consists of a sequence of tones. In order to realize that there is a repetition of L and H tones, and thus two potential streams L and H, at least a few triplets must be experienced. This could be why the percept starts with a grouped interpretation: it takes at least the time for the stimulus to unfold to discover that there could be an alternative interpretation to the one-stream percept. Bregman [1] suggested that this was an 'accumulation of evidence' process. A slightly different account has been suggested by Denham & Winkler [17], who described the grouped percept as predicted by a local rule (grouping of successive sounds) and the split percept by a global rule (grouping sounds that are separated by longer silent gaps). In the same vein, Shamma *et al.* [37] suggested that correlated changes across cues are what allow streams to form. Measuring the correlation between L tones on the one hand and H tones on the other requires some time, leading again to the prediction of a first percept bias.

Could a similar explanation apply to visual plaids? There is no sequential presentation in time for plaids, but the hypothesis could be made that visual cues favouring the coherent percept are analysed faster than visual cues favouring the transparent percept. Such a hypothesis can be given an intuitive explanation: at presentation onset, depth ordering (whether ambiguous or not) must be resolved before the transparent percept can be experienced, and this may take some time. More formally, moving plaids contain many local motion cues, that is, cues within the size of receptive fields at the earlier stages of visual processing. Some of these local motion cues (one-dimensional cues) are carried by the lines that make up the plaids (excluding the intersections). The direction of those lines is ambiguous, because of the 'aperture problem': neurons with small receptive fields can only determine the motion component that is perpendicular to the lines, so they cannot disambiguate between the many line motion directions that are compatible with the same local line displacement [22,23,35]. Such local cues are also conflicting (the two gratings move in opposite directions). Other motion cues (two-dimensional cues) are carried by the plaid's intersections, and their direction is locally not ambiguous. Plaid perception requires integration of those local motion cues one way or the other over a broad spatial range to achieve a global interpretation. The direction of the 'coherent' global interpretation is the same as that of the intersections; this interpretation may therefore rely on *locally unambiguous* cues, while the alternative interpretation of two gratings sliding over each other must rely on *globally unambiguous* computations. Global computations must take some time, so the earlier (because local) availability of unambiguous cues, compatible with the coherent interpretation, may explain why it is achieved first.

Similar explanations may therefore be proposed for the percept bias in audition due to temporal properties

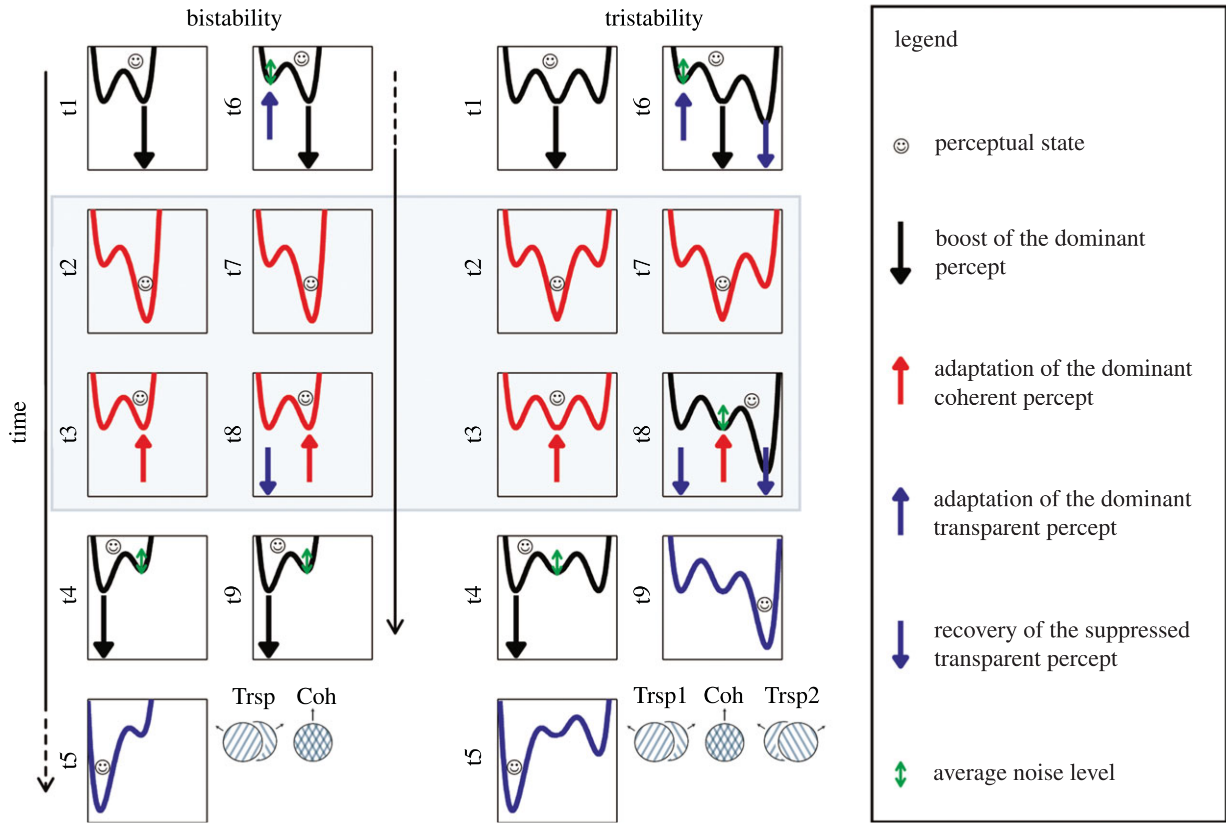
bistability
tristability
time
t1
t2
t3
t4
t5
t6
t7
t8
t9
Trsp
Coh
Trsp1
Coh
Trsp2
legend
perceptual state
boost of the dominant percept
adaptation of the dominant coherent percept
adaptation of the dominant transparent percept
recovery of the suppressed transparent percept
average noise level

Figure 7. (*Caption opposite*.)

and in vision due to spatial properties. In both cases, the resolution of the competition between several possible interpretations reveals the local to global dynamic mechanisms of perceptual organization.

5. CONCLUSIONS

The present study applied the same analysis and theoretical framework to visual plaid perception and auditory streaming. For both stimuli, the initial phase of perception is different from the on-going multistable perception. In vision, the first percept for ambiguous moving plaids is the coherent (grouped) percept and it lasts longer than subsequent coherent percepts. This is what we termed here first percept bias and inertia, respectively. Bias and inertia are not observed for other visual bistable paradigms. However, both first percept bias and inertia are observed for auditory streaming, where the first percept is also biased towards one stream (grouped) and it lasts longer than subsequent one-stream percepts. Together, these effects correspond to the classic and much studied build up of auditory streaming [7]. Here we showed that, for visual plaids, first percept bias and inertia corresponded to two independent phenomena: parametric manipulations could influence first percept inertia without affecting first percept bias. The first percept inertia was clearly due to plaid perception being tristable rather than bistable. We favour a similar explanation for first percept inertia in audition, but the current data did not warrant such a definite conclusion. Perceptual tristability in auditory streaming has yet to be demonstrated. Our results however indicate that the aggregate measure of build up could be complemented usefully by separate analyses of bias and inertia. As for first percept biases in the two modalities, we argue that they are due to different but conceptually related causes: in vision, the bias may reflect the transition from spatially local to spatially global computations, whereas in audition it seems to be caused by the transition from temporally local to temporally global computations.

This research was supported by a grant from Agence Nationale de Recherche ANR-08-BLAN-0167-01. Valérie Juillard and Cynthia Niggel assisted in the visual data collection; Francisco Cervantes Constantino collected the auditory data.

APPENDIX. MECHANISMS OF PLAID MULTISTABILITY

Why the first coherent percept lasts longer than subsequent coherent percepts (‘inertia’) for tristability but not bistability requires an explanation. In this appendix, we outline a conceptual model accounting for such behaviours. The model is based on standard accounts of perceptual competition, with additional constraints that may pertain to the general mechanisms of multistability, notably concerning the dynamics of adaptation and a specific role of attention and/or conscious perception (suggested notably by the involvement of fronto-parietal areas [38]) in the switching mechanisms (figure 7). The model uses the framework of non-linear dynamics, as has often been the case for multistability [39–41]. Details are kept to a minimum in this short appendix, in favour of a qualitative description.

Figure 7. (*Opposite*.) An illustration of a conceptual model to explain the dynamics of bistable and tristable perception, accounting in particular for the first percept inertia observed for tristable plaids. The representation is that of 'energy landscapes' [39–41], which represent the energy state of the system associated with various percepts. The gist of the model is that the system will attempt to find equilibrium in the energy minima (wells of the landscape), but this equilibrium can be perturbed by noise. The depth of a well is thus indicative of the average duration spent in the corresponding percept. Left-hand side: model illustration for bistable perception. We use the cartoon illustration proposed by Kim *et al.* [42] of a double-well landscape. The figure shows the evolution of the energy landscape at nine key times (t1 to t9, sampled at constant intervals) after stimulus onset (note that the second column is the continuation of the first column). Red colour indicates that the coherent percept is experienced, blue represent the transparent percept and black denotes perceptual transition phases. The position of the smiley face also indicates the perceptual state. In this model, the energy landscape is transformed over time and by the perceptual state. Arrows indicate the directions of changes in the landscape. Those changes are assumed here to be governed by three main mechanisms: adaptation of the dominant percept and noise [40,42–44], plus 'boosting' of the percept reaching perceptual awareness. Adaptation occurs only for the currently dominant percept, so the suppressed percept recovers from its previously adapted state ('adaptation driven oscillatory regime' [40]). Noise becomes a significant cause of switching only when the dominant percept has adapted. The 'boosting' hypothesis provides a role for attention and consciousness [38] in perceptual organization, since it states that a potential perceptual organization is strengthened by being experienced. Importantly, it is sufficient to explain the phenomenon of 'perceptual stabilization' whereby the bistable interpretation that was experienced before the interruption of a stimulus is more likely to dominate when the stimulus re-appears [19, 20]. Imagine that, as is the case for many ambiguous stimuli, which percept is experienced first is random (both interpretations are equally likely, i.e. same depth of the energy wells at 't1'); once a percept is experienced, it gets reinforced (deeper well at 't2'); if the stimulus is removed at that time and then presented again, the energy landscape favours the previously experienced percept. The only specificity of the present model for plaids is the first percept bias: whatever the probability of coherence and transparency, the first percept is supposed to be the coherent one. Reasons for this additional hypothesis are detailed in the main text, in the discussion of local *versus* global cues. The bistable model predicts no inertia, consistent with the experimental observations ($t2 + t3 = t7 + t8$). Right-hand side: model illustration for tristable perception. For tristable plaids, we added a third energy well in the landscape representation. Importantly, we placed the 'coherent' well between the two 'transparent' wells, because switches between two transparency states are typically interleaved with a coherent percept [34]. The grey area on the figure highlights the key difference between bistable and tristable plaids. For bistable plaids, the first ($t2 + t3$) and second ($t7 + t8$) coherent percepts (Coh) last on average the same time. For tristable plaids, however, the second coherent percept is shorter (t7) because the competing transparent percept 'Trsp2' has a low energy (the 'Trsp2' right well at t7 is deeper than any of the 'Trsp' wells at t2). The reason for this deeper well is what we called the 'recovery of the suppressed transparent percept' at t6 (downward blue arrow). Here we represented only the key events of the dynamics. For example, we did not show the recovery of adaptation for the coherent percept. Formal modelling will be required to specify the respective strengths and time constants of these mechanisms [45]. We also made some implicit assumptions that should be tested or modelled formally. For example, the recovery from adaptation of the previously experienced percept (blue arrow at t8, left-hand side) must finish before the new dominant percept has finished being boosted and adapted (otherwise, in the bistable case, $t2 + t3$ may not have the same duration as $t7 + t8$: the energy landscapes of t3 and t8 must be identical). Also, our supposed mechanism of 'recovery of the suppressed transparent percept' (which we also called 'recovery from adaptation'), which is a key mechanism to explain inertia in tristability, may in fact correspond to several mechanisms. In the tristable case described here, we applied such 'recovery' to Trsp2 (at t6 and t8) even though Trsp2 had not been experienced yet, so Trsp2 cannot recover from adaptation yet. Here, this mechanism is better understood as some form of coupling between the neural representations of the percepts, with mutual inhibition between competing interpretations [27]. Such coupling would imply that whenever a percept gets strengthened the alternative percept gets weakened, and reciprocally ('push–pull' mechanism). The downward blue arrow at t6 would be linked to the upward blue arrow at t6 (adaptation of the dominant transparent percept), and this therefore corresponds to the competition mechanism for figure/ground assignment. The downward blue arrows at t8 would correspond to the competition mechanism between integration and segmentation. In our model, some form of inhibitory coupling appears as one of the mechanisms necessary to explain inertia for tristable perception.

REFERENCES

1 Bregman, A. S. 1990 *Auditory scene analysis: the perceptual organization of sound*. Cambridge, MA: Bradford Books, MIT Press.

2 Bregman, A. S. 1978 Auditory streaming is cumulative. *J. Exp. Psychol. Hum. Percept. Perform.* **4**, 380–387. (doi:10.1037/0096-1523.4.3.380)

3 Anstis, S. & Saida, S. 1985 Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.* **11**, 257–271. (doi:10.1037/0096-1523.11.3.257)

4 Shamma, S. 2008 On the emergence and awareness of auditory objects. *PLoS Biol.* **6**, e155. (doi:10.1371/journal.pbio.0060155)

5 Miller, G. A. & Heise, G. A. 1950 The trill threshold. *J. Acoust. Soc. Am.* **22**, 637–638. (doi:10.1121/1.1906663)

6 Bregman, A. S. & Campbell, J. 1971 Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol.* **89**, 244–249. (doi:10.1037/h0031163)

7 Moore, B. C. J. & Gockel, H. E. 2012 Properties of auditory stream formation. *Phil. Trans. R. Soc. B* **367**, 919–931. (doi:10.1098/rstb.2011.0355)

8 Van Noorden, L. P. A. S. 1975 Temporal coherence in the perception of tone sequences. Doctoral dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands.

9 Cusack, R., Deeks, J., Aikman, G. & Carlyon, R. P. 2004 Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J. Exp. Psychol. Hum. Percept. Perform.* **30**, 643–656. (doi:10.1037/0096-1523.30.4.643)

10 Roberts, B., Glasberg, B. R. & Moore, B. C. J. 2008 Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. *J. Exp. Psychol. Hum. Percept. Perform.* **34**, 992–1006. (doi:10.1037/0096-1523.34.4.992)

11 Rogers, W. L. & Bregman, A. S. 1998 Cumulation of the tendency to segregate auditory streams: resetting by changes in location and loudness. *Percept. Psychophys.* **60**, 1216–1227. (doi:10.3758/BF03206171)

12 Micheyl, C. & Oxenham, A. J. 2010 Objective and subjective psychophysical measures of auditory stream integration and segregation. *J. Assoc. Res. Otolaryngol.* **11**, 709–724. (doi:10.1007/s10162-010-0227-2)

13 Thompson, S. K., Carlyon, R. P. & Cusack, R. 2011 An objective measurement of the build-up of auditory streaming and of its modulation by attention. *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 1253–1262. (doi:10.1037/a0021925)

14 Micheyl, C., Tian, B., Carlyon, R. P. & Rauschecker, J. P. 2005 Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* **48**, 139–148. (doi:10.1016/j.neuron.2005.08.039)

15 Pressnitzer, D., Sayles, M., Micheyl, C. & Winter, I. M. 2008 Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* **18**, 1124–1128. (doi:10.1016/j.cub.2008.06.053)

16 Pressnitzer, D. & Hupé, J. M. 2006 Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* **16**, 1351–1357. (doi:10.1016/j.cub.2006.05.054)

17 Denham, S. L. & Winkler, I. 2006 The role of predictive models in the formation of auditory streams. *J. Physiol. Paris* **100**, 154–170. (doi:10.1016/j.jphysparis.2006.09.012)

18 Kashino, M., Okada, M., Mizutani, S., Davis, P. & Kondo, H. M. 2007 The dynamics of auditory streaming: psychophysics, neuroimaging and modeling. In *Hearing—from basic research to applications* (eds B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Upperkamp & J. Verhey*)*, pp. 275–283. Heidelberg: Springer.

19 Leopold, D. A., Wilke, M., Maier, A. & Logothetis, N. K. 2002 Stable perception of visually ambiguous patterns. *Nat. Neurosci.* **5**, 605–609. (doi:10.1038/nn0602-851)

20 Pearson, J. & Brascamp, J. 2008 Sensory memory for ambiguous vision. *Trends. Cogn. Sci.* **12**, 334–341. (doi:10.1016/j.tics.2008.05.006)

21 Mamassian, P. & Goutcher, R. 2005 Temporal dynamics in bistable perception. *J. Vis.* **5**, 361–375. (doi:10.1167/5.4.7)

22 Wallach, H. 1935 Uber visuell wahrgenommene Bewegungsrichtung. *Psychol. Forsch.* **20**, 325–380. (doi:10.1007/BF02409790)

23 Wuerger, S., Shapley, R. & Rubin, N. 1996 'On the visually perceived direction of motion' by Hans Wallach: 60 years later. *Perception* **25**, 1317–1367. (doi:10.1068/p251317)

24 Rock, I. & Mitchener, K. 1992 Further evidence of failure of reversal of ambiguous figures by uninformed subjects. *Perception* **21**, 39–45. (doi:10.1068/p210039)

25 Rock, I., Gopnik, A. & Hall, S. 1994 Do young children reverse ambiguous figures? *Perception* **23**, 635–644. (doi:10.1068/p230635)

26 Long, G. M. & Toppino, T. C. 2004 Enduring interest in perceptual ambiguity: alternating views of reversible figures. *Psychol. Bull.* **130**, 748–768. (doi:10.1037/0033-2909.130.5.748)

27 Rubin, N. & Hupé, J. M. 2005 Dynamics of perceptual bi-stability: plaids and binocular rivalry compared. In *Binocular rivalry* (eds A. Alais & R. Blake), pp. 137–154. Cambridge: MIT Press.

28 Levelt, W. J. M. 1968 *On binocular rivalry.* The Hague, Paris: Mouton.

29 Hupé, J. M. & Rubin, N. 2003 The dynamics of bi-stable alternation in ambiguous motion displays: a fresh look at plaids. *Vision Res.* **43**, 531–548. (doi:10.1016/S0042-6989(02)00593-X)

30 Hupé, J. M. & Rubin, N. 2004 The oblique plaid effect. *Vision Res.* **44**, 489–500. (doi:10.1016/j.visres.2003.07.013)

31 Hupé, J. M., Joffo, L. M. & Pressnitzer, D. 2008 Bistability for audiovisual stimuli: perceptual decision is modality specific. *J. Vis.* **8**(7), 1. (doi:10.1167/8.7.1)

32 Moreno-Bote, R., Shpiro, A., Rinzel, J. & Rubin, N. 2008 Bi-stable depth ordering of superimposed moving gratings. *J. Vis.* **8**(7), 20. (doi:10.1167/8.7.20)

33 Hupé, J. M. & Juillard, V. A. 2009 Buildup of visual plaid segmentation and auditory streaming may be explained by the perception of these ambiguous stimuli being tristable rather than bistable. *Soc. Neurosci. Abstr.* 652.16.

34 Hupé, J. M. 2010 Dynamics of ménage à trois in moving plaid ambiguous perception. *J. Vis.* **10**(7), 1217. (doi:10.1167/10.7.1217)

35 Wallach, H. 1976 *On perception*. New York: Quadrangle.

36 Snyder, J. S., Carter, O. L., Hannon, E. E. & Alain, C. 2009 Adaptation reveals multiple levels of representation in auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 1232–1244. (doi:10.1037/a0012741)

37 Shamma, S. A., Elhilali, M. & Micheyl, C. 2011 Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34**, 114–123. (doi:10.1016/j.tins.2010.11.002)

38 Kleinschmidt, A., Sterzer, P. & Rees, G. 2012 Variability of perceptual multistability: from brain state to individual trait. *Phil. Trans. R. Soc. B* **367**, 988–1000. (doi:10.1098/rstb.2011.0367)

39 Suzuki, S. & Grabowecky, M. 2002 Evidence for perceptual "trapping" and adaptation in multistable binocular rivalry. *Neuron* **36**, 143–157. (doi:10.1016/S0896-6273(02)00934-0)

40 Moreno-Bote, R., Rinzel, J. & Rubin, N. 2007 Noise-induced alternations in an attractor network model of perceptual bistability. *J. Neurophysiol.* **98**, 1125–1239. (doi:10.1152/jn.00116.2007)

41 Kelso, J. A. S. 2012 Multistability and metastability: understanding dynamic coordination in the brain. *Phil. Trans. R. Soc. B* **367**, 906–918. (doi:10.1098/rstb.2011.0351)

42 Kim, Y. J., Grabowecky, M. & Suzuki, S. 2006 Stochastic resonance in binocular rivalry. *Vision Res.* **46**, 392–406. (doi:10.1016/j.visres.2005.08.009)

43 Blake, R., Sobel, K. V. & Gilroy, L. A. 2003 Visual motion retards alternations between conflicting perceptual interpretations. *Neuron* **39**, 869–878. (doi:10.1016/S0896-6273(03)00495-1)

44 Shpiro, A., Moreno-Bote, R., Rubin, N. & Rinzel, J. 2009 Balance between noise and adaptation in competition models of perceptual bistability. *J. Comput. Neurosci.* **27**, 37–54. (doi:10.1007/s10827-008-0125-3)

45 Huguet, G., Hupé, J. M. & Rinzel, J. 2011 A model for dynamical switching in tristable perception for visual plaids. *Soc. Neurosci. Abstr.* 722.15.