PHILOSOPHICAL TRANSACTIONS B

Phil. Trans. R. Soc. B.
doi:10.1098/not yet assigned

Auditory and visual scene analysis: an overview

Hirohito M. Kondo[1*], Anouk M. van Loon[2,3*], Jun-Ichiro Kawahara[4], and Brian C. J. Moore[5]

[1] *Human Information Science Laboratory, NTT Communication Science Laboratories, NTT Corporation, Atsugi, Kanagawa 243-0198, Japan, ORCID ID: 0000-0002-7444-4996*

[2] *Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, 1081 BT, Amsterdam, The Netherlands, ORCID ID: 0000-0002-9015-7647*

[3] *Institute of Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, 1081 BT, Amsterdam, The Netherlands*

[4] *Department of Psychology, Graduate School of Letters, Hokkaido University, Sapporo 060-0810, Japan, ORCID ID: 0000-0002-4096-3923*

[5] *Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, United Kingdom, ORCID ID: 0000-0001-7071-0671*

Keywords: scene analysis, perceptual organization, stream formation, attention, salience, individual differences

Summary

We perceive the world as stable and composed of discrete objects even though auditory and visual inputs are often ambiguous due to spatial and temporal occluders and changes in the conditions of observation. This raises important questions regarding where and how 'scene analysis' is performed in the brain. Recent advances from both auditory and visual research suggest that the brain does not simply process the incoming scene properties. Rather, top-down processes such as attention, expectations, and prior knowledge facilitate scene perception. Thus, scene analysis is linked not only with the extraction of stimulus features and formation and selection of perceptual objects, but also with selective attention, perceptual binding, and awareness. This special issue covers novel advances in scene-analysis research obtained using a combination of psychophysics, computational modelling, neuroimaging, and neurophysiology, and presents new empirical and theoretical approaches. For integrative understanding of scene analysis beyond and across sensory modalities, we provide a collection of 15 articles that enable comparison and integration of recent findings in auditory and visual scene analysis.

Introduction

Imagine you are walking on a big busy square. Cars are crossing, pedestrians are walking past and towards you, someone rings their bicycle bell to warn you that they want to pass, you hear people chatting, a taxi-driver shouting, and the bell of the nearby school is indicating that school just finished. Meanwhile you notice a

* Joint first authorship. Authors for correspondence (kondo.hirohito@lab.ntt.co.jp or anouk.vanloon@gmail.com).

beautiful coloured tree with its leaves turning orange because autumn is setting in, and you start to think about your next holiday destination. Our brain is very well equipped to rapidly convert such a mixture of sensory inputs – both visual and auditory – into coherent scenes so as to perceive meaningful objects and guide navigation [1, 2], and also to imagine visual and auditory scenes and distinguish them from 'real' scenes.

The task of analysing a mixture of sounds so as to arrive at percepts corresponding to the individual sound sources was termed 'auditory scene analysis' by Albert Bregman [3]. The task is also known as the 'cocktail party problem' [4], which refers to the ability to follow one conversation when many people are talking at the same time. The auditory system has to determine whether a sequence of sounds all came from a single source, and should be perceived as a single "stream", or whether there were multiple sources [5]. In the latter case, each sound in the sequence has to be allocated to its appropriate source and multiple streams should be heard. Similarly in vision, the visual system has to partition a visual scene into one or more objects and a background, determining which elements in the scene 'belong' to which object or to the background. Visual scene analysis research was initially impelled by a compelling idea of David Marr [6]. He postulated that the purpose of the visual system is to provide a representation of what is present in the outside world. Although the sensation of seeing complex scenes is seemingly effortless and occurs extremely rapidly, the sensory input is highly complex and dynamic. It takes only a few hundred milliseconds to activate a large cascade of different brain regions, each performing a different transformation of the sensory input [7]. The underlying neural mechanisms of these complex spatiotemporal processes pose major conceptual and methodological challenges for researchers in cognitive neuroscience [8, 9].

Our main aim for this special issue was to present an overview of work on auditory and visual scene analysis from a multi-disciplinary perspective, emphasising new approaches and developments. While early work on auditory and visual scene analysis focussed on relatively simple artificial scenes, recently research has been extended to more realistic situations, such as simulated cocktail parties [10, 11] and natural visual scenes [12-14]. Furthermore, scene analysis is facilitated by the use of statistical regularities. Humans can rapidly and automatically learn complex sensory statistics and use them to improve perceptual inference, even when there is no conscious awareness of the statistical regularities [15]. There are distinct and consistent individual differences in scene analysis, especially among special populations, such as those with autism spectrum disorder (ASD) [16] and these individual differences can help to reveal the underlying mechanisms of scene analysis [17-19]. Many published papers on scene analysis have focussed exclusively on the auditory and visual domains, making it difficult to obtain a bird's-eye view of scene-analysis research or to appreciate links between auditory and visual scene analysis [20]. This issue provides an overview of all of these developments from the viewpoint of different disciplines and considering both vision and hearing. The papers in the issue cover a wide range of experimental techniques: psychophysics (in audition, vision, and for multimodal interactions); functional neuroimaging; the measurement of evoked potentials; computational modelling; and single-cell recording.

Revising the Hierarchical Framework for Scene Analysis

Sensory information processing has often been considered in a hierarchical framework, that is, as a series of

discrete stages that successively produce increasingly abstract representations. This is sometimes called 'bottom-up' processing and the stages can range from low-level to high-level. However, it is now acknowledged that the flow of processing is not unidirectional [21, 22], and that there are strong 'top-down' influences from factors such as attention, the goals of the individual in the specific situation, expectations, and prior knowledge [9, 23-25]. The relative influence of bottom-up and top-down processes and the way that they interact remain unclear. For visual perception, the underlying neural architecture consists of multiple hierarchical stages of processing from the retina through sub-cortical structures to the cortex, where multiple distinct visual areas have been defined. Even scene-selective visual areas have been identified, in particular with functional magnetic resonance imaging (fMRI) studies in humans. These areas respond more when viewing natural scenes than when viewing objects or faces [2]. How these regions fit into the larger hierarchical framework of visual processing, is, however, a topic of considerable debate.

Animals as well as humans need to perform auditory scene analysis. The benefits of assigning sounds to specific sources accrue to all species that communicate acoustically. In this issue, Itatani and Klump [5] provide an overview of the paradigms applied in the study of auditory scene analysis and streaming of sequential sounds in animal models. They compare psychophysical results from human studies to the evidence for auditory streaming obtained using animal psychophysics. The comparison reveals that similar requirements in the analysis of acoustic scenes have resulted in similar perceptual and neuronal processing mechanisms in the many species that are capable of auditory scene analysis. Again, these processing mechanisms seem to involve both bottom-up and top-down processing.

It has often been stated that the aim of visual analysis is to create an invariant and robust representation of a scene, i.e., of what is 'out there'. However, a natural scene is more than just a collection of objects. In this issue Groen and colleagues [7] propose that we should try to understand how multiple scene properties contribute to scene analysis and what kind of representation is needed to achieve particular higher-level goals, such as recognition and navigation. They stress the importance of the contributions of bottom-up visual analysis to the representation of complex scenes. Such contributions include retinotopic biases and receptive field properties of scene-selective regions in the brain. Moreover, the authors give examples of studies on the temporal dynamics of scene perception demonstrating that low- and mid-level feature representations overlap with those based on higher-level scene categories. Therefore, scene perception is more than just the activation of higher-level scene-selective regions in the brain.

Early theories were based on the assumption that multisensory processing was restricted to higher-level cortical areas, and did not occur in cortical areas concerned with primary sensory analysis. However, the primary sensory cortices receive not only domain-specific information through their primary afferent pathways, but also information from the other senses. For example, high-level auditory information is sent to primary visual cortex via cortical feedback and top-down pathways [26]. These multisensory inputs to the sensory cortices therefore refute the assumption that multisensory processing is limited to higher cortical areas. In this issue, Petro and colleagues [27] discuss the implications of the recent discovery of auditory input to the visual cortex. They propose that auditory input to primary visual cortex could activate visual predictive

codes to facilitate perception. Additionally, Petro et al. [27] suggest that the auditory input may play a role in what they call 'counterfactual processing', by triggering imagery, dreaming and mind wandering, when the visual image is completely different from the visual scene that is actually present. Such processing may be important for allowing people to play out scenarios in their minds to test consequences and make decisions.

It remains unclear how multimodal scenes are represented in the brain as a result of the rapid and complex neural dynamics underlying visual and auditory information processing. In this issue, Cichy and Teng [8] describe three key problems in understanding these dynamics: brain signals measured non-invasively are inherently noisy; the nature of the neural 'code' is currently unknown; and transformations between representations are often non-linear and complex. In their opinion piece, they argue that progress can be made by making use of recent methodological developments such as complex computational modelling (e.g., deep neural networks), in combination with imaging methods (magneto- and electro-encephalography, MEG/EEG, and fMRI) and other types of models (e.g., using representational similarity analysis), and by applying sensitive analysis techniques, such as decoding and cross-classification. The latter is used when assessing the generalisability of a deep neural network. Different conditions are assigned to the training and the testing set. Correct classification of the testing set indicates that the network can correctly classify novel stimuli.

Taking all of this evidence together, it is clear that scene analysis does not involve a simple hierarchical cascade of processing steps, but a complex interplay across modalities, brain regions, and time, depending on the top-down goals of the observer.

The Role of Salience and Attention in Scene Analysis

The dynamic interplay between different levels of processing can be nicely illustrated by the concept of 'salience'. An aspect of a scene can be salient because of its strong physical features (salience driven by bottom-up processing) or because of its top-down relevance (e.g., a goal-directed priority for successful behaviour). As a result, several brain areas can play a role in computation of a 'salience map' of a scene. In this issue, Veale and colleagues [22] review recent advances in the neural and computational basis of visual salience maps. They provide a conceptual framework for how salience maps can be constructed from stimulus features, and assess the progression from feature-specific salience maps to feature-agnostic salience and priority maps. In parallel, the authors evaluate which of these types of maps are represented in various cortical regions and in the superior colliculus. The authors then focus on how salience and priority maps of the superior colliculus are encoded in its superficial and deeper layers, respectively, while providing supporting evidence from slice studies of a rodent model and computational studies that simulate these experimental data.

Interestingly, the concept of a 'salience map' topographically encoding stimulus conspicuity over the visual scene has proved to be an efficient predictor of eye movements. Inherent in visual scene analysis is a bottleneck associated with the need to sequentially sample locations with foveating eye movements. In this issue, Hillstrom and colleagues [28] examine the effect of early scene analysis and plausibility of the target location on eye movements when searching for objects in photographs of scenes. A novel feature of their study was that, after the first saccade, the target location was moved either to a position that was equally likely but elsewhere

in the scene, or to an unlikely but possible location, or to a physically improbable location. There was a clear influence of the likelihood of the location on the guidance of visual search. This study offers a nice illustration of the role of top-down factors in goal-directed behaviour during visual scene analysis.

Interest in top-down effects on auditory scene analysis has grown in recent years [24, 25, 29]. There is no doubt that perceptual experience can be modified by attention or intention [23] and by previous experience [17], in a way that is specific for each individual. In this issue, Kaya and Elhilali [9] describe computational models of attention in auditory perception that can incorporate the effects of both bottom-up attention based on perceptual salience and top-down attention. According to these models, attention acts as a selection process or processes that focus both sensory and cognitive resources on the most relevant events in the soundscape. Relevance can be dictated by the stimulus itself (e.g., a loud explosion) or by a task at hand (e.g., listening to announcements in a busy airport). In this issue, Southwell and colleagues [15] explore whether attention can be influenced by the predictability of sounds. In a series of behavioural and EEG experiments they used repeating patterns of sounds to capture attention. Their EEG data demonstrate that the brain rapidly recognizes predictable patterns, as manifested by a rapid increase in responses (the root-mean-square power) to regular relative to random sound sequences. This finding seems contrary to a large body of work showing reduced responses to predictable stimuli. However, Southwell et al. [15] also used two behavioural tasks to reveal that sound regularity did not capture attention. Here, the pattern of results can be interpreted by considering mechanisms that minimise surprise and uncertainty about the world. The influence of attention is further addressed by Mehta and colleagues [30]. They studied the influence of attention and temporal synchrony on the perceptual organization of sound sources using the 'octave illusion'. In their study, they combined behavioural and human EEG data. Based on their results they suggest that the illusion involves an attentional misattribution of time across perceptual streams, rather than an attentional misattribution of location within a stream. Thus, in complex acoustic scenes, attention plays a key role in parsing competing features of the incoming sounds into auditory streams.

Individual Differences in Auditory and Visual Scene Analysis

Perception is a private process for each individual, and perceptual experiences may differ across individuals even when the physical environment is the same [31]. Individual differences in human behaviour and perception are often considered to be "noise" and are therefore discarded through averaging data from a group of participants [32]. However, individual variability can be exploited to better understand the neural computations underlying perceptual experience, cognitive abilities, and motor skills (cf. personalized medicine). Bistable and multistable stimuli are useful tools for investigating where and how perceptual objects are formed in the brain since they can lead to more than one percept although their perceptual input is constant. In this issue, Pelofi and colleagues [17] report experiments comparing musicians and non-musicians responding to sequences of two 'Shepard tones', each of which contains octave-spaced sinusoidal components. For certain sequences of these tones, the direction of the pitch change is ambiguous; either an upward or a downward shift

may be heard. Pelofi et al. obtained both behavioural reports of the direction of the shift and confidence ratings in the reports. No differences were observed between musicians and non-musicians in their judgements of pitch-shift direction. The most ambiguous case resulted in chance performance (50% 'up' judgements) for both groups. However, the non-musicians gave high confidence ratings for the ambiguous case while the musicians gave lower confidence ratings. Pelofi et al. argued that musicians were more likely to hear out components within the complex tones, and hence detected the ambiguity, perhaps unconsciously. In contrast the non-musicians probably heard the complex sound as a whole, and did not detect the ambiguity.

Social deficits and communication difficulties may be partly explained by individual variability in both basic auditory abilities and in scene analysis abilities. In this issue, Lin and colleagues [16] report that people diagnosed with autism spectrum disorder (ASD) are characterized by difficulty in acquiring relevant auditory and visual information in daily environments, despite the fact that people with ASD have normal audiometric thresholds and normal visual acuity. People diagnosed with ASD appear to perceive the world differently than 'normal' people, sometimes having superior abilities in discriminating details of a scene while having difficulties in judging or discriminating more global properties. There may also be substantial and consistent individual differences within those diagnosed with ASD. Interestingly a comparison of the characteristics of scene analysis between auditory and visual modalities in people with ASD reveals some essential commonalities, which could provide clues for the underlying neural mechanisms of ASD.

Individual differences in perceptual organization may result from genetic, neurochemical, and anatomical factors. An early study revealed large genetic effects on the perception of illusory movement [33], which occurs when observers view shaded stripes peripherally. Binocular rivalry occurs when different images are presented to each eye; the percept tends to switch irregularly from the image in one eye to the image in the other eye. A large-sample twin heritability study demonstrated that additive genetic factors account for approximately 50% of the variance in the spontaneous switching rate during binocular rivalry [34]. Brain measures, such as regional volumes [35] and interregional connections [36], are associated with perceptual switching in visual rivalry. Moreover, the inhibitory neurotransmitter γ-aminobutyric acid (GABA) has been linked to the perceptual dynamics of a range of different visual bistable illusions [37]. In addition, the number of perceptual switches for auditory and visual stimuli differs between genotype groups related to the dopaminergic and serotonergic systems, respectively [38, 39]. Thus, neurochemical modulations underlie individual differences in the temporal dynamics of the perceptual organization of scenes.

In this issue, Kondo and colleagues [18] used auditory multistability to examine to what extent neurochemical and cognitive factors influence the observed idiosyncratic patterns of switching between percepts. The concentrations of glutamate-glutamine (Glx) and GABA in different brain regions were measured by magnetic resonance spectroscopy (MRS), and personality traits and executive functions were assessed using questionnaires and response inhibition tasks. Intriguingly, although switching patterns within each individual differed between auditory streaming and verbal transformations (where a syllable or word is repeated many times, and the perceived speech sounds change over time), similar dimensions were extracted separately from the two datasets. Individual switching patterns were significantly correlated with Glx and GABA concentrations

in auditory cortex and inferior frontal cortex but not with the personality traits and executive functions. The results suggest that auditory perceptual organization depends on the balance between neural excitation and inhibition in different brain regions.

In contrast, in this issue Takeuchi and colleagues [19] only observed a relationship between the concentration of Glx and visual motion perception in the prefrontal cortex and not in visual areas. They examined two types of motion phenomena – motion assimilation and contrast – and found that, following the presentation of the same stimulus, some participants perceived motion contrast, while others perceived motion assimilation. The tendency of participants to perceive motion assimilation over motion contrast was positively correlated with the concentration of Glx in the prefrontal cortex, while GABA had only a weak effect.

Apart from these examples of applying multistable stimuli to assess individual differences in perception, multistable stimuli offer a powerful tool for studying subjective perceptual experience and conscious perception, since the bottom-up input remains constant while the perception dynamically changes. How and which aspects of neural activity give rise to consciousness are still a fundamental questions of cognitive neuroscience. To date, the vast majority of research devoted this question has come from the field of visual perception. In this issue Dykstra and colleagues [20] discuss the recent literature concerning candidate neural correlates of conscious auditory perception. They consider the phenomena that need to be incorporated into a theory of conscious auditory perception and consider the implications for a general theory of conscious perception, encompassing all of the senses. Additionally, Dykstra et al. [20] suggest the approaches and techniques that can best be applied to investigate this.

Conclusions

The topics described above have been investigated using a wide range of methods and analyses, including studies of different species and of individual differences in humans. Especially, the integration across psychophysics, imaging methods (MRS, MEG/EEG & fMRI) and computational models, has yielded valuable insights and may be increasingly used in the future. Additionally, we would like to stress the importance of applying multimodal stimuli [40, 41] and neurophysiological studies in animals [42-44] for identifying the neural mechanisms of scene analysis. The papers in this special issue have covered both auditory and visual scene analysis, helping to bridge the gap between sensory modalities. We hope that the complementary contributions in this issue will stimulate new lines of research and promote fruitful collaborations across disciplines.

Additional Information

Acknowledgments

We thank Helen Eaton, Commissioning Editor of the journal, for her crucial support and advice in our guest editor work for this special issue. We also thank the authors who have contributed to this issue and all the referees for their important remarks, comments, and suggestions.

Ethics

N/A

Data Accessibility

N/A

Authors' Contributions

H.M.K., A.M.V.L., J.I.K., and B.C.J.M. conceived and wrote the manuscript.

Competing Interests

We have no competing interests.

Funding

B.C.J.M. was supported by the Engineering and Physical Sciences Research Council (UK, grant number RG78536).

References

[1] Bizley, J.K. & Cohen, Y.E. 2013 The what, where and how of auditory-object perception. *Nat. Rev. Neurosci.* **14**, 693-707. (doi:10.1038/nrn3565).

[2] Epstein, R.A. 2014 Neural systems for visual scene recognition. In *Scene Vision: Making Sense of What We See.* (eds. K. Kveraga & M. Bar), pp. 105-134. Cambridge, MA, MIT Press.

[3] Bregman, A.S. 1990 *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA, MIT Press.

[4] Cherry, E.C. 1953 Some experiments on the recognition of speech, with one and two ears. *J. Acoust. Soc. Am.* **25**, 975-979.

[5] Itatani, N. & Klump, G.M. this issue Animal models for auditory streaming. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0112).

[6] Marr, D. 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York, Freeman.

[7] Groen, I.I.A., Silson, E.H. & Baker, C.I. this issue Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0102).

[8] Cichy, R.M. & Teng, S. this issue Resolving the neural dynamics of visual and auditory scene processing in the human brain: A methodological approach. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0108).

[9] Kaya, E.M. & Elhilali, M. this issue Focusing on the clutter in auditory scenes: Perspecive from modeling auditory attention. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0101).

[10] Kondo, H.M., Pressnitzer, D., Toshima, I. & Kashino, M. 2012 Effects of self-motion on auditory scene analysis. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6775-6780. (doi:10.1073/pnas.1112852109).

[11] Carlin, M.A. & Elhilali, M. 2013 Sustained firing of model central auditory neurons yields a discriminative spectro-temporal representation for natural sounds. *PLoS Comput. Biol.* **9**, e1002982. (doi:10.1371/journal.pcbi.1002982).

[12] Muckli, L., De Martino, F., Vizioli, L., Petro, L.S., Smith, F.W., Ugurbil, K., Goebel, R. & Yacoub, E. 2015 Contextual feedback to superficial layers of V1. *Curr. Biol.* **25**, 2690-2695. (doi:10.1016/j.cub.2015.08.057).

[13] Kravitz, D.J., Peng, C.S. & Baker, C.I. 2011 Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci.* **31**, 7322-7333. (doi:10.1523/JNEUROSCI.4588-10.2011).

[14] van Loon, A.M., Fahrenfort, J.J., van der Velde, B., Lirk, P.B., Vulink, N.C.C., Hollmann, M.W., Scholte, H.S. & Lamme, V.A.F. 2016 NMDA receptor antagonist Ketamine distorts object recognition by reducing feedback to early visual cortex. *Cereb. Cortex* **26**, 1986-1996. (doi:10.1093/cercor/bhv018).

[15] Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K. & Chait, M. this issue Is predictability salient? A study of attentional capture by auditory patterns. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0105).

[16] Lin, I.-F., Shirama, A., Kato, N. & Kashino, M. this issue Specificity of auditory and visual scene analysis in autism. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0115).

[17] Pelofi, C., de Gardelle, V., Egré, P. & Pressnitzer, D. this issue Inter-individual variability in auditory scene analysis revealed by confidence judgments. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0107).

[18] Kondo, H.M., Farkas, D., Denham, S.L., Asai, T. & Winkler, I. this issue Auditory multistability and neurotransmitter concentrations in the human brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0110).

[19] Takeuchi, T., Yoshimoto, S., Shimada, Y., Kochiyama, T. & Kondo, H.M. this issue Individual differences in visual motion perception and neurotransmitter concentrations in the human brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0111).

[20] Dykstra, A.R., Cariani, P.A. & Gutschalk, A. this issue A roadmap for the study of conscious audition and its neural basis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0103).

[21] Kondo, H.M. & Kashino, M. 2009 Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* **29**, 12695-12701. (doi:10.1523/JNEUROSCI.1549-09.2009).

[22] Veale, R., Hafed, Z.M. & Yoshida, M. this issue How is visual saliency computed in the brain? Insights from behaviour, neurobiology, and modeling. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0113).

[23] 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).

[24] Snyder, J.S., Gregg, M.K., Weintraub, D.M. & Alain, C. 2012 Attention, awareness, and the perception of auditory scenes. *Front. Psychol.* **3**, 15. (doi:10.3389/fpsyg.2012.00015).

[25] Moore, B.C.J. & Gockel, H.E. 2012 Properties of auditory stream formation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 919-931. (doi:10.1098/rstb.2011.0355).

[26] Ghazanfar, A.A. & Schroeder, C.E. 2006 Is neocortex essentially multisensory? *Trends Cogn. Sci.* **10**, 278-285. (doi:10.1016/j.tics.2006.04.008).

[27] Petro, L.S., Paton, A.T. & Muckli, L. this issue Contextual modulation of primary visual cortex by auditory signals. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0104).

[28] Hillstrom, A.P., Segabinazi, J.D., Godwin, H.J., Liversedge, S.P. & Benson, V. this issue Cat and mouse search: The influence of scene and object analysis on eye movements when targets change locations during search. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0106).

[29] Shinn-Cunningham, B.G. 2008 Object-based auditory and visual attention. *Trends Cogn. Sci.* **12**, 182-186. (doi:10.1016/j.tics.2008.02.003).

[30] Mehta, A.H., Jacoby, N., Yasin, I., Oxenham, A.J. & Shamma, S. this issue An auditory illusion reveals the role of streaming in the temporal misallocation of perceptual objects. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (doi:10.1098/rstb.2016.0114).

[31] Wang, D. 2007 Computational scene analysis. In *Challenges for Computational Intelligence.* (eds. W. Duch & J. Mandziuk), pp. 163-191. Berlin, Springer.

[32] Kanai, R. & Rees, G. 2011 The structural basis of inter-individual differences in human behaviour and cognition. *Nat. Rev. Neurosci.* **12**, 231-242. (doi:10.1038/nrn3000).

[33] Fraser, A. & Wilcox, K.J. 1979 Perception of illusory movement. *Nature* **281**, 565-566.

[34] Miller, S.M., Hansell, N.K., Ngo, T.T., Liu, G.B., Pettigrew, J.D., Martin, N.G. & Wright, M.J. 2010 Genetic contribution to individual variation in binocular rivalry rate. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2664-2668. (doi:10.1073/pnas.0912149107).

[35] Kanai, R., Bahrami, B. & Rees, G. 2010 Human parietal cortex structure predicts individual differences in perceptual rivalry. *Curr. Biol.* **20**, 1626-1630. (doi:10.1016/j.cub.2010.07.027).

[36] Genç, E., Bergmann, J., Singer, W. & Kohler, A. 2011 Interhemispheric connections shape subjective experience of bistable motion. *Curr. Biol.* **21**, 1494-1499. (doi:10.1016/j.cub.2011.08.003).

[37] van Loon, A.M., Knapen, T., Scholte, H.S., St. John-Saaltink, E., Donner, T.H. & Lamme, V.A.F. 2013 GABA shapes the dynamics of bistable perception. *Curr. Biol.* **23**, 823-827. (doi:10.1016/j.cub.2013.03.067).

[38] Kondo, H.M., Kitagawa, N., Kitamura, M.S., Koizumi, A., Nomura, M. & Kashino, M. 2012 Separability and commonality of auditory and visual bistable perception. *Cereb. Cortex* **22**, 1915-1922. (doi:10.1093/cercor/bhr266).

[39] Kashino, M. & Kondo, H.M. 2012 Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 977-987. (doi:10.1098/rstb.2011.0370).

[40] van Ee, R., van Boxtel, J.J.A., Parker, A.L. & Alais, D. 2009 Multisensory congruency as a mechanism for attentional control over perceptual selection. *J. Neurosci.* **29**, 11641-11649. (doi:10.1523/JNEUROSCI.0873-09.2009).

[41] Deroy, O., Chen, Y.-C. & Spence, C. 2014 Multisensory constraints on awareness. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **369**, 20130207. (doi:10.1098/rstb.2013.0207).

[42] Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J. & Shamma, S.A. 2009 Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* **61**, 317-329.

[43] Itatani, N. & Klump, G.M. 2014 Neural correlates of auditory streaming in an objective behavioral task. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10738-10743.

[44] Yoshida, M., Itti, L., Berg, D.J., Ikeda, T., Kato, R., Takaura, K., White, B.J., Munoz, D.P. & Isa, T. 2012 Residual attention guidance in blindsight monkeys watching complex natural scenes. *Curr. Biol.* **22**, 1429-1434. (doi:10.1016/j.cub.2012.05.046).