Auditory and visual scene analysis: an overview

Hirohito M. Kondo1,†, Anouk M. van Loon2,3,†, Jun-Ichi Kawahara4 and Brian C. J. Moore5

1Human Information Science Laboratory, NTT Communication Science Laboratories, NTT Corporation, Atsugi, Kanagawa 243-0198, Japan
2Department of Experimental and Applied Psychology, and 3Institute of Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, Amsterdam 1081 BT, The Netherlands
4Department of Psychology, Graduate School of Letters, Hokkaido University, Sapporo 060-0810, Japan
5Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

We perceive the world as stable and composed of discrete objects even though auditory and visual inputs are often ambiguous owing 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.

This article is part of the themed issue ‘Auditory and visual scene analysis’.

1. 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 is shouting and the bell of the nearby school is indicating that school just finished. Meanwhile you note a 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 perceptions corresponding to the individual sound sources was termed ‘auditory scene analysis’ by 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 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 multidisciplinary perspective, emphasizing new approaches and developments. While early work on auditory and visual scene analysis focused 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 focused 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.

2. 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 subcortical 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 & 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 psychophysiology. 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 et al. [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 et al. [27] discuss the implications of the recent discovery of auditory input to the visual cortex. They propose that auditory input to the 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 & 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 nonlinear 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.

3. 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 et al. [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 et al. [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 & 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 et al. [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 minimize surprise and uncertainty about the world. The influence of attention is further addressed by Mehta et al. [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.

4. 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 because they can lead to more than one percept although their perceptual input is constant. In this issue, Pelofi et al. [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, whereas 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 et al. [16] report that people diagnosed with 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 et al. [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 the 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 et al. [19] observed a relationship between the concentration of Glx and visual motion perception only 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, whereas 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, whereas 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, because the bottom-up input remains constant while the perception dynamically changes. How and which aspects of neural activity give rise to consciousness are still fundamental questions of cognitive neuroscience. To date, the vast majority of research devoted to this question has come from the field of visual perception. In this issue, Dykstra et al. [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.

5. Conclusion

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 and 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.

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 no. RG78536).

Acknowledgements. 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.
Guest editor biographies

Hirohito Kondo is a Research Scientist of NTT Communication Science Laboratories. He studied Experimental Psychology at Kyoto University and conducted research on the nature of working memory and attentional control in humans while a PhD student. His interest in conscious awareness has led him to investigate where and how meaningful perceptual objects are formed in the brain. He combines psychophysical, neuroimaging (mainly fMRI and MRS), and genotyping methods to better understand the neural mechanisms of transformations from sensory information to conscious perception. His particular focus concerns individual differences in auditory and visual perceptual organization.

Jun Kawahara received his PhD in Experimental Psychology from Hiroshima University. Prior to joining the Department of Psychology of Hokkaido University as an Associate Professor, he conducted human factors and cognitive stress research at the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. His research interests include visual attention, perception of attractiveness, mental workload and metacognition.

Anouk van Loon is a post-doc at the Vrije Universiteit Amsterdam at the Department of Cognitive Psychology. She studied Psychology and Cognitive Neuroscience at the University of Amsterdam. During her PhD, she studied the underlying neural mechanisms of visual awareness. She combines pharmacological interventions, neuroimaging and psychophysical techniques to understand how the human brain integrates the ‘picture in our head’ with incoming sensory information. Currently, she is investigating the relationship between perception, attention and working memory.

Brian Moore is Emeritus Professor of Auditory Perception at the University of Cambridge. His research focuses on hearing and hearing loss, especially the perceptual analysis of complex sounds. He is a Fellow of the Royal Society, the Academy of Medical Sciences, the Acoustical Society of America and the Audio Engineering Society. He has been awarded the Silver and Gold Medals of the Acoustical Society of America, the International Award in Hearing from the American Academy of Audiology, the Award ofMerit from the Association for Research in Otolaryngology, and the Hugh Knowles Prize for Distinguished Achievement from Northwestern University.

References

4. Cherry EC. 1953 Some experiments on the recognition of speech, with one and two ears. /Acoust. Soc. Am. 25, 975 – 979. (doi:10.1121/1.1907229)


Hillstrom AP, Segabini JD, Godwin HJ, Liversedge SP, Benson V. 2017 Cat and mouse search: the influence of scene and object analysis on eye movements when targets change locations during search. Phil. Trans. R. Soc. B 372, 20160106. (doi:10.1098/rstb.2016.0106)


