PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 1001–1012
doi:10.1098/rstb.2011.0359

Review

Multistability in auditory stream segregation: a predictive coding view

István Winkler[1,2,*], Susan Denham[3], Robert Mill[3], Tamás M. Bőhm[1,4] and Alexandra Bendixen[5]

[1]*Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, PO Box 398, 1394 Budapest, Hungary*
[2]*Institute of Psychology, University of Szeged, Egyetem u. 2, 6722 Szeged, Hungary*
[3]*School of Psychology, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, UK*
[4]*Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar Tudósok krt. 2, 1117 Budapest, Hungary*
[5]*Institute for Psychology, University of Leipzig, Seeburgstrasse 14-20, 04103 Leipzig, Germany*

Auditory stream segregation involves linking temporally separate acoustic events into one or more coherent sequences. For any non-trivial sequence of sounds, many alternative descriptions can be formed, only one or very few of which emerge in awareness at any time. Evidence from studies showing bi-/multistability in auditory streaming suggest that some, perhaps many of the alternative descriptions are represented in the brain in parallel and that they continuously vie for conscious perception. Here, based on a predictive coding view, we consider the nature of these sound representations and how they compete with each other. Predictive processing helps to maintain perceptual stability by signalling the continuation of previously established patterns as well as the emergence of new sound sources. It also provides a measure of how well each of the competing representations describes the current acoustic scene. This account of auditory stream segregation has been tested on perceptual data obtained in the auditory streaming paradigm.

Keywords: auditory scene analysis; perceptual bistability; predictive processing; auditory object representation; auditory grouping; computational models

1. INTRODUCTION

Our senses provide us with a continuous stream of information from which we have evolved and learnt to extract cues about the objects within our environment. In the auditory modality, the problem of finding objects is challenging, as there may be many sound sources active within our environment at any time, and their acoustic emissions mix together even before reaching the ears. For each incoming sound, it is essential to determine its relationship with previous events; does it 'belong' with (some subset of) other sound events, or does it signify something new that has no recent history within the scene? Even if the incoming sound belongs to a group, there are still questions to be answered: does it tell us anything new about the behaviour of the object, or is it entirely predictable for us and thus simply confirms what we already know?

The process of segregating and grouping sound events is made even more difficult by the need to make grouping decisions 'on the fly'. While the correct decisions may be easy to identify with hindsight, the problem is to try to get them right when the information needed for an optimal decision is incomplete. So how does the auditory system construct, modify and maintain dynamic representations of putative objects within its environment, and how does it minimize incorrect interpretations? Recent studies of perceptual multistability using the auditory streaming paradigm can help us to provide some answers regarding the strategies employed by the auditory system. There is increasing evidence that the auditory system: (i) creates and maintains representations of multiple alternative groupings simultaneously [1] and (ii) evaluates the reliability/usefulness of these representations by means of comparisons between the predictions they generate and the actual sensory input (see the theoretical reviews [2,3]). Here, we provide a theoretically motivated review aimed at showing how auditory perceptual organization can emerge from the trade-off between timely perceptual decisions and perceptual flexibility.

In this account of auditory perceptual organization, we focus on the computational principles and memory resources involved in the sequential grouping of temporally separate sound events, and in particular, on what studies of multistability in auditory streaming can tell us about the way in which these groups feed into perceptual awareness. While undoubtedly important, we do not consider simultaneous (concurrent) grouping of sound events here.

* Author for correspondence (iwinkler@cogpsyphy.hu).

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

2. THE NATURE OF THE REPRESENTATIONS UNDERLYING SEQUENTIAL GROUPING

In everyday environments, sound sources often generate series of discrete sounds, such as a series of footsteps. Some of the potentially important information characterizing the source, such as whether a person is approaching or receding, is not present separately within the individual sounds: it can be extracted only by relating the individual sounds to each other. Sequential grouping—connecting discrete sound events into a perceptual unit—is thus an essential function of auditory perception.

Grouping sounds across time requires some representation of the preceding sounds. The storage of detailed auditory sensory information is probably not based on a single mechanism, although there is no consensus in the literature regarding the number and the assumed functions of the various forms of storage [4,5]. It is, however, clear that at least for a short period of time (estimated to last from a few hundreds of milliseconds to a few seconds), a large number of sounds can be retained in parallel [6]. This characteristic of auditory sensory memory suggests that the auditory system can represent several sets of sounds at any given time. On the other hand, general access to detailed auditory information appears to be limited in time (2–20 s, depending on the testing procedure), although some forms of access are available even for a much longer time [7]. Listeners can, for example, detect sound patterns periodically repeating with a 5–10 s cycle, but not ones exceeding *ca* 10–20 s [8,9]. However, the sensory memory literature reflects the final outcome, rather than the *building* of auditory pattern representations. It is quite probable that connecting successive sounds is limited to even shorter intervals, perhaps below 2 s, as was found in studies of rhythm perception [10,11]. Additional support for this conclusion comes from studies of stimulus-specific adaptation (SSA) [12]. Adaptation to tone sequences is apparent on several time scales, ranging from hundreds of milliseconds to tens of seconds [13]. However, the principal effect of one tone on the next, or the 'one-trial effect', appears to be connected to a time constant of around 1.5 s [13], which is consistent with the fact that the longest stimulus onset asynchrony (SOA) at which SSA is elicited robustly is around 2 s [12]. An intuitive model of SSA [14] suggests that, with increasing temporal separation, the links between discrete events become very weak and short-lived owing to the decay of the traces.

So, how do we store the auditory information required by sequential grouping processes? One possibility is that the large capacity of auditory sensory memory allows us to store each discrete sound separately. Then, when a new sound arrives, our auditory system applies all possible grouping heuristics to determine whether the new sound can be fitted to the stored representations of the preceding sounds. The advantage of this method is maximal flexibility. It includes a full reinterpretation of all (stored) data with each new sound event. The main disadvantages are high pre-processing and decision-making costs. That is, on the arrival of each new sound, all processing occurs as if we had no knowledge of the acoustic configuration of our surroundings. This explanation is clearly not consistent with our intuitive sense of the continuity of perception. More importantly, it is also inconsistent with the observed 'old + new' strategy of the auditory system, whereby continuing previously discovered groupings takes precedence over detecting possible newly emerging ones [15,16]. This is because if groups were formed anew on the arrival of each sound (i.e. preceding sounds are 'remembered', but not how they were grouped previously), there will be no record of previously discovered 'old' groups, and thus the continuation of these 'old' groups cannot take precedence over new ones.

There are also other memory effects that argue against the separate storage of individual sound events. Although there is a lot of evidence suggesting that detailed auditory sensory information becomes unavailable after a relatively short period of time (see above), it appears that this information is still somehow maintained in the brain and can be reactivated by a single reminder, even after longer periods of time (Glenberg [17]; for a review, see Winkler & Cowan [18]). Importantly, the reminder makes accessible not only strictly sensory details of a past sound but also information of how this sound was related to the ones preceding it [19]. Furthermore, even when listeners are instructed to remember individual sounds, they cannot fully separate them from sounds that preceded them within a short period of time [20]. Therefore, as we have previously suggested, the sensory information related to discrete sounds is stored as a part of larger perceptual units. Some theorists argue that these perceptual units can be regarded as auditory object representations [1,21,22].

Storing auditory object representations offers one additional important advantage over the separate storage of individual sound events. Besides reducing processing requirements and increasing perceptual stability, such a system can continuously produce predictions of future states of the acoustic environment, thus helping one to prepare for upcoming events without having to commit higher level cognitive processing resources, whose capacity is possibly more limited. Recent theoretical [1,23,24] and computational modelling approaches to perception [2,25] suggest that sensory systems are inherently predictive.

'Predictive' in the present framework refers to detecting sequential dependencies within a series of sounds and extrapolating these dependencies towards future incoming stimuli. We suggest that such predictions serve more than just preparatory processes: they also provide a way of linking an incoming auditory event to the sound source that correctly predicted the occurrence of this event. For instance, in the typical auditory streaming paradigm introduced by van Noorden [26] ('ABA_ABA_ . . . ', where 'A' and 'B' denote two different tones and '_' stands for a silent period equal to the common duration of the tones), predictive representations such as 'repetition of A with a time constant T' and 'repetition of B with a time constant 2T' can be formed. Any future sound 'A' or 'B' arriving at the predicted time is then automatically grouped with the correct representation, with no need to re-evaluate the auditory

configuration at the arrival of each sound. Note, however, that the representation 'repetition of ABA_ with a time constant of 4T' likewise predicts the occurrence of 'A' and 'B' sounds correctly. Thus, predictive representations are not yet equivalent to auditory objects that are consciously perceived, but they form 'proto-objects' that enter into a competition for perceptual dominance.

Predictive relations, as conceptualized in the present framework, entail repetition of a single sound (e.g. 'A'), a sound pattern (e.g. 'ABA'), or an abstract relation between sounds (e.g. 'the next sound is higher than the current sound'). Evidence from numerous electrophysiological studies (for review, see Bendixen *et al.* [27]) suggests that these types of predictive relations can be extracted from sound sequences outside the focus of attention and that the extracted information is turned into predictions about forthcoming stimuli. For instance, Bendixen *et al.* [28] compared the processing of sound omissions in three different cases: (i) the auditory features of the omitted sound could be predicted based on the preceding sounds, (ii) sound features could be determined only from the sound following the omitted one, and (iii) sound features could neither be predicted before nor determined after the omission. Participants' attention was engaged by watching a movie, and they received no information about the sound sequences. The electrical signals recorded from the brain (event-related potentials, ERPs) for omissions of fully predictable sounds were highly similar to those elicited by actual sounds within the first 50 ms from the expected onset of the omitted sound. In contrast, after *ca* 10 ms from the expected onset of the omitted sound, the ERP responses elicited by sound omissions in the other two conditions already differed from those elicited by actual sounds. These results support the notion of predictive processing in the auditory system, and are compatible with results obtained in the field of auditory deviance detection (for a review, see Winkler [3]), auditory imagery [29] and illusory continuity percepts in animals [30]. Thus, we suggest that (perceptual object) representations of auditory sensory information in the human brain are inherently predictive.

Two types of information are necessary for forming these predictive representations: sensory data (i.e. the actual features of the individual sounds) and link data (i.e. the relations between the individual sounds). These two types of information are, however, redundant. Winkler & Cowan [18] suggested that the amount of sensory information that is needed to be stored can be radically reduced by taking links into account. If full sensory representations are retained only for critical positions of a sequence, then such 'anchors' are sufficient to estimate the sensory information for any position. This type of storage is similar to the way in which movies are encoded for digital computers. Instead of storing each frame, only key frames are stored in full, whereas intermediate frames are represented in terms of changes from the preceding frame, or from the last key frame. By combining the anchors with a representation of the links (the rules or regularities connecting the elements of the group), these representations can predict the continuation of the given sound object in the future. This hypothesis on the storage of auditory objects is compatible with the generative models assumed by predictive coding theories [2,31].

It is important to remember that even when using predictive object representations, perceptual decisions have to be made continuously and without fully reliable knowledge of future sound events. It is not possible to know at any given moment whether the auditory input experienced up to that point will continue unchanged. A new event may be part of a previously detected sound pattern or may represent the onset of a new pattern. We suggest that it is in building and maintaining *multiple alternative* groupings in parallel that the auditory system manages to alleviate this problem, thus ensuring flexibility in perceptual decisions. This notion is consistent with the reactivation of auditory representations, and was also confirmed in studies of auditory deviance detection [32].

In summary, we suggest that when a new sound arrives, the auditory system attempts to connect it to all existing sound representations. Because we assume that these representations are predictive, each incoming sound provides a test of the validity of these representations, while competition between alternatives allows flexible reinterpretation of the acoustic input [1,33]. The large capacity of auditory sensory memory [4] makes it credible to suggest that alternative representations of groupings (proto-objects) can be maintained in parallel, with new ones being initiated all the time. In addition, isolated sounds (i.e. sounds having no relationship to either previous or succeeding sounds) can form representations of their own. However, the number of concurrently active auditory proto-objects is probably limited. It remains to be seen, whether this limit is rather small (e.g. similar, and perhaps even related to, the common limitation found for representations in short-term/working memory [34]) as was suggested by some studies [35] or can be large, provided that the concurrently active proto-objects are sufficiently distinct [36,37]. We discuss the issue of alternative sound representations further in §5.

3. THE FORMATION OF SEQUENTIAL GROUPS

We have seen that perceptual groups are formed by linking individual sounds across time. But what provides the 'glue' that binds sounds together? We suggest that there are two main principles for binding individual elements of a group. The first principle is based on perceptual similarity between the individual sound events: links are more likely to be made between individual sound events that share some or all of their features. The second principle is based on sequential predictability: links are more likely to be formed between sound events predictably following one another. Both types of grouping are governed by temporal aspects of the stimulus configuration; time provides the medium within which links between discrete sound events can be formed and thus determines the cost of maintaining such links (see §4).

The first binding principle, based on perceptual similarity, originates from the law of proximity already expressed by the Gestalt school of psychology [38]. The underlying idea is that the more similar (proximal in feature space) two individual sound events are, the more likely it is that they were emitted by the same sound source, and, therefore, binding them together is more likely to result in veridical perception. Many different acoustic features can influence the sense of similarity; these are described in detail by Moore & Gockel [39] (see also Moore & Gockel [40]).

However, similarity is a relative concept. In terms of a sound sequence, similarity is mediated by time. That is, even small differences may form contrasts when the two sounds are presented within a short period of time, whereas relatively large differences may be tolerated when the two sounds are more removed from each other in time. It is therefore useful to turn from the raw notion of similarity, which may provide a reasonable account for grouping phenomena in static visual displays, to the notion of change in time, which can be characterized as the rate at which sound features change. Indeed, Jones [41] suggested that auditory streaming results from the fact that we are not able to follow fast changes. Thus, highly different sounds presented within a short period of time fall apart (i.e. cannot be effectively bound together; see Shinozaki *et al.* [42]). Although the notion that streaming results from the 'sluggishness' of attention was not confirmed by studies measuring brain responses in the absence of attention (for a review, see Sussman [43]), it is reasonable to assume that in forming links between sounds, feature difference is related to the rate of change, rather than to the raw feature difference. The ecological basis for this assumption is quite obvious: gradual changes are much more likely to characterize a single source than abrupt ones. Thus, when we refer to similarity, we really mean the inverse of the rate of change.

Similarity-based binding plays a crucial role in the initial formation of perceptual groups. Such early binding processes may operate, for instance, based on SSA [12,13] or other refractoriness-related mechanisms [44]. The notion that grouping and selection are based on featural proximity has been introduced to psychology by early filtering theories of selective attention [45] and is also implicitly invoked by the neurophysiological assumption of sensory gating [46].

The second binding principle, based on predictability, was introduced by Jones [41] and has received further attention recently [1,33]. It rests on the assumption that sound events are bound together across time if the system detects predictive relations between them. Events predictably following one another typically signify a single underlying cause. A natural situation in which predictability is likely to be useful for binding sound events is a repeating cycle of sounds (*pattern*) as emitted, for instance, by a train moving on the rails with a constant speed. As with timbre, sound patterns often distinctively characterize a given source, possibly allowing one to identify it. It seems reasonable to assume that predictability is used as a binding principle; yet, evidence has long been equivocal in this regard ([47]; see discussion in Bregman [15]). However, a recent study based on the bi-stable nature of the auditory streaming paradigm [48] revealed an influence of the predictability of the sound sequence on grouping. Although grouping by prediction appeared to occur at a later stage than grouping based on perceptual similarity, detecting a predictive relationship between sounds substantially increased the stability of groups derived from similarity-based grouping. Consistent evidence has been obtained by Andreou *et al.* [49] with an indirect measure of streaming based on task performance. In accordance with our account of predictive auditory representations, we suggest that proto-objects are characterized by how often they correctly predict incoming sounds, and that higher densities of successful predictions provide advantages for the proto-objects in the competition for perceptual dominance.

The two binding principles—perceptual similarity and sequential predictability—can be conceived as two sets of heuristics by which the auditory system attempts to solve the inverse problem of perception (i.e. determining distal objects from the proximal input). However, none of these heuristics offers a final solution for grouping sounds; rather, they provide a measure of how well individual sound elements in an auditory scene can be connected.

4. PREDICTIVE REPRESENTATIONS AND THE ROLE OF TIME

Under everyday listening conditions, the properties of sounds in a sequence are not constant. Even slight changes in the source, the relative position of the source and the listener, or some acoustic property of the environment will make the sensory input vary. In some cases, the actual acoustic variability can be quite large. If these kinds of variations were not tolerated, auditory perception would become ineffective. Anchors and links must, therefore, refer to *distributions* of values in the auditory parameter space and reflect the previously experienced variance (see Helson's [50] adaptation-level theory). Support for tolerance to stimulus variance in auditory stream segregation has been provided by results showing that jittering stimulus parameters in the auditory streaming paradigm did not prevent the formation of auditory streams [51]. However, when the amount of feature variation is held constant, groups that include predictable variation in some feature are more stable in perception than groups that include random variation in the same feature [48]. This may also apply to temporal predictability [49]. However, although often treated similar to other features in auditory streaming experiments, several studies showed that the processing of the temporal aspects of stimulation differs from that of other acoustic parameters (e.g. [52]).

When describing how sound groups are formed and represented in the brain, one should emphasize the effects of time. We introduce the term 'link strength' as a measure of how easily two sounds can be bound together by similarity. Link strength arises as the product of two factors. The first is a *temporal* component, which decays exponentially during the interval between two sounds with a fairly short time

constant (see §2). This makes it difficult to link sounds that are separated by long intervals. The second is a *rate-of-change* component that is low when two highly different sounds are presented in quick succession, and reaches a maximum only when two sounds are identical. In the simple case of two brief tones separated only in pitch, this quantity can be thought of as being inversely proportional to the (absolute) gradient of a line joining two points in the time-frequency plane. This factor makes it difficult to link sounds with an abrupt change between them. Many of the effects underlying theories of temporal distinctiveness [53] can be modelled as temporal effects on link strength. Further, preparations for processing a stimulus can be focused on a given time range. For example, the phase of slow rhythmic activity in the sensory areas is entrained to the expected time of the predicted stimulus, and the better this synchronization, the more accurate the response [54]. When the system is optimally set for encountering a new stimulus, its processing is enhanced, which can help to link it with other sounds. Thus, our predictive account of auditory grouping considers temporal effects at the level of events as both constraining and modulating the effects of other stimulus parameters.

5. THE TIME COURSE OF SEQUENTIAL GROUP FORMATION

If we store sounds as part of auditory objects, as argued earlier, how then are alternative interpretations of the input formed and compared with each other? This is an important issue, because a large number of studies have reported that when listening to certain types of sound sequences, such as the auditory streaming paradigm [26], perception can switch between two or more alternative organizations [33,48,51,55–57]. Following the principles of grouping described earlier, here we consider, using the auditory streaming paradigm, the process of group formation.

Consider a typical experiment. At the start, no representation can predict the incoming sounds, because previously formed representations, e.g. the one that models the experimenter's voice, have lost their predictive power. Thus, new representations start to be built, and as the various regularities that can be extracted from the sequence are detected and represented, they begin to compete for perceptual dominance. This suggests that perceptual experience during the start up may be rather different from that later in the sequence, and indeed, this is the case. In our experiments, we found evidence for strong differences between the initial phase and subsequent ones [51], as is also the case in visual bi-stability [58]. Denham *et al.* [51] analysed the continuous record of the participants' reports of their perception in terms of 'perceptual phases', the continuous intervals within which the participant marked that he/she perceived the sound sequence in the same way. Here, we illustrate (figure 1) that the duration of the first perceptual phase lasts substantially longer than the subsequent phases (figure 1*b*,*d*; the difference in phase duration was highly significant) and that listeners mostly, but not always, first experience a percept in which all tones are grouped together (the integrated percept; figure 1*a*,*c*).

Based on these observations, we suggest the following account of the course of auditory stream formation. At the onset of the tone sequence, the system starts to build multiple representations in parallel, and the group that is easiest to discover is the one first perceived. Generally, this will be the integrated organization, because forming links between temporally adjacent sounds, i.e. those that follow each other in sequence, is easier than skipping events. Because the abstract links we refer to here manifest as associated patterns of neural activity in the brain, it seems reasonable to suppose that neural activity patterns that follow each other sequentially become associated using neural mechanisms of short-term plasticity [59]. When tones with highly different features follow each other within a very short time (fast presentation rate), then initial links may form between non-adjacent events. In this case, links between same-feature events are established first, because the reduction of link strength between adjacent sounds owing to high feature separation (e.g. by topological distance in a tonotopically organized area of the auditory cortex; [60,61]) exceeds that caused by longer temporal separation between the non-adjacent (but same feature) sounds. In other words, the cost of establishing the link between topologically highly separate focuses of neural activity exceeds that of establishing the link between the neural activity elicited by the incoming sound and a relatively more decayed neural after-effect of the previous (less recent) same-feature sound.

Because no competition can occur until alternative groups have been discovered and a predictive representation of them has emerged, the first percept is prolonged by the time needed to form alternative groups (for a detailed discussion of the duration of the first percept, see Hupé & Pressnitzer [62]). In addition, there is evidence from visual experiments that the currently dominant (perceived) organization is strengthened by correctly describing the stimulus configuration [63]; this also would extend the duration of the first perceptual phase, because the group perceived first gets an extra boost compared with the other groupings. As was suggested in §3, the initial formation of groups may largely (or even exclusively) rely on the feature-proximity principle and not (or not so much) on the predictive principle. Hence, stimulus parameters mainly affect how fast a given group can be discovered and its representation constructed.

In time, the neural associations underlying the alternative sound organizations become stronger and start to vie for dominance. We assume that this competition is between patterns of neural activity far removed from the early sensory processes of stimulus feature extraction (i.e. we posit competition between proto-objects). These patterns of neural activity are much more dependent on the intrinsic brain circuitry and neural mechanisms involved than on the parameters of the stimulation. This implies that the grouping alternatives might compete on more equal terms. Thus, the probabilities of perceiving different organizations tend to become more balanced with time. This also is supported by experimental data [58].

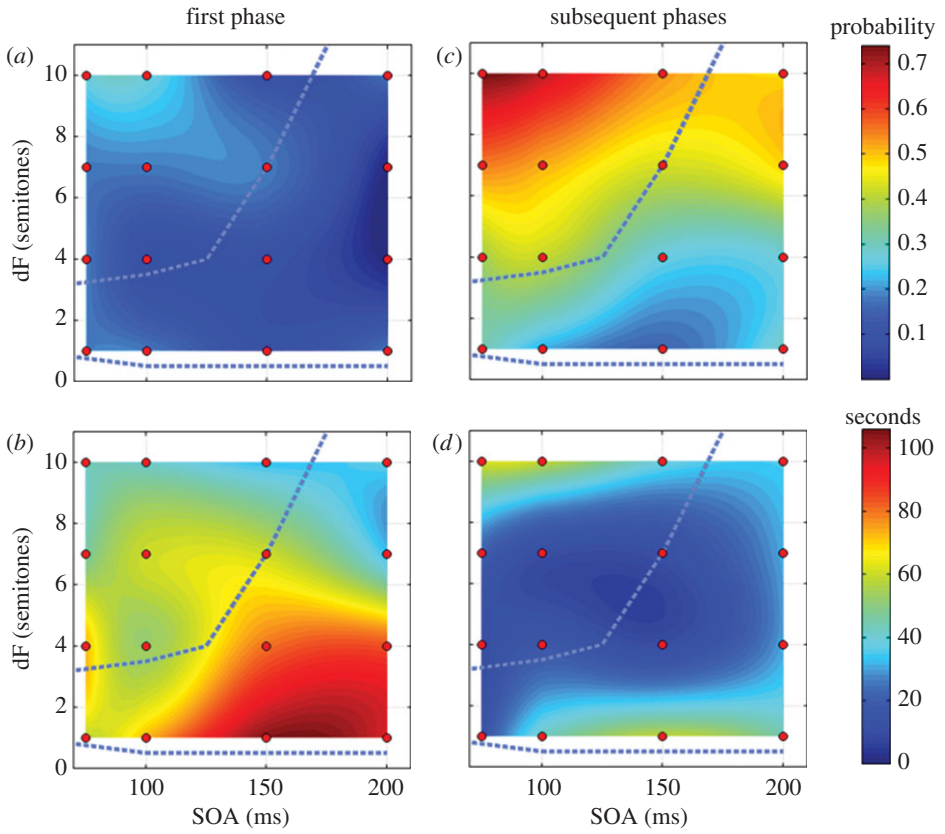

Figure 1. Differences between first and subsequent perceptual phases found in experiment 1 of Denham *et al.* [51]. Listeners were presented with 4-min long trains of 'ABA_' type sequences and were instructed to continuously mark their perception of them. The parameter combinations of frequency difference (dF) and stimulus onset asynchrony (SOA) used are marked by red circles on the figure. Note that SOA was jittered by 20% and dF by 10% centred on the nominal value marked on the figure (see details in Denham *et al.* [51]). Values are interpolated between these points for better visualization. (*a*,*b*) First phase, (*c*,*d*) subsequent phases. (*a*,*c*) Probability of segregation averaged across all participants ($n = 11$). Colour scale to the right, indicating probability of segregation, applies to both plots. (*b*,*d*) Group-mean perceptual phase duration in seconds. Colour scale to the right, indicating duration in seconds, applies to both plots. The perceptual boundaries identified in classical experiments by van Noorden [26] are indicated by blue-dashed lines.

The time course of perceptual organization has also been probed by introducing changes in the stimulus train. Previous studies [36,55,64–66] have consistently reported perceptual reset (i.e. a restart from the state at the outset dominated by the integrated percept, with the necessity to gather evidence for the segregated organization anew) when parameter changes were introduced in the auditory streaming paradigm. This has led to the suggestion that attention is necessary for stream formation, because parameter changes lead to attentional capture, and with an attentional switch, the process of the build up of segregation begins as if from the start. In contrast, we observed a smooth transition towards the segregated organization when the pitch difference between the 'A' and 'B' tones was abruptly increased at a later point during a long stimulus sequence. In this experiment, 4-min sequences of a repeated 'ABA_' pattern were presented with an SOA of 150 ms between consecutive tones. Here, we report data from one condition from the experiment in which, after 2 min, the frequency difference between the 'A' and 'B' sets of tones was suddenly increased from five semitones to seven semitones by lowering the 'A' tones by one semitone and elevating the 'B' tones by one semitone. Both sets of tones contained small, random variations in frequency and intensity identical to Condition 1 reported in Bendixen *et al.* [48]. Participants continuously indicated their perception, and their reports were analysed as described in Bendixen *et al.* [48] and Denham *et al.* [51]. Figure 2 shows the time course of the probability of reporting the segregated percept, averaged across 30 participants. The sudden frequency change is marked by the bold dashed line. No perceptual reset can be observed; instead, perceptual reports of segregation increased.

Two common differences can be found between the paradigm employed in our study and those of the previous ones [36,55,64–66]: (i) most previous studies tested the effects of slightly or substantially larger stimulus changes than we did and (ii) parameters were changed quite early in the sequence (before 20 s), whereas we introduced the changes after 2 min. It is therefore possible that only relatively large stimulus changes trigger a perceptual reset. However, Haywood & Roberts [64] found perceptual reset with a change (deviation) of only three semitones, and Anstis & Saida [55], presenting frequency-modulated tones, observed a reset with a change of two semitones from the centre of adaptation. Thus, the requirement of large stimulus change is doubtful. Regarding the possible effects of stimulus change introduced early

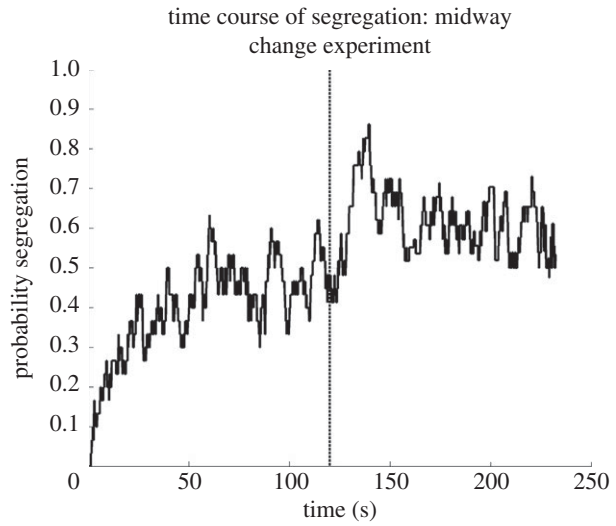

Figure 2. Time course of the probability of reporting segregation throughout a 4-min 'ABA_' sequence with 150-ms stimulus onset asynchrony between consecutive tones. At 120 s (marked by the bold dashed line on the figure), the frequency difference between the 'A' and 'B' sets of tones is suddenly increased from five to seven semitones. Note that no perceptual reset towards the integrated organization is reported by participants.

versus later in the stimulus sequence, in §3, we suggested that proto-objects are evaluated in terms of how often they correctly predict upcoming sounds. Establishing this predictive reliability takes time. Soon after their discovery, proto-objects are weak and stimulus changes resulting in incorrect predictions may eliminate them. In contrast, when a sequence has been presented for a longer period of time, the predictive reliability of the corresponding proto-objects is quite high. Our finding of smooth transitions and no evidence of a return to a first-phase-like state suggests that reliable proto-objects can be modified dynamically to track changes in the ongoing stimulus sequence. Further, rather than causing a complete reset, changes in the stimulation result in modifications in the intrinsic coherence of the proto-objects leading to a shift in their relative strengths within the ongoing competition. These characteristics of the representations of sound groups are important in order to allow the auditory system to track time-varying emissions from the same sound source, thus supporting stable perception with no 'down-time' (i.e. reset time) of the changing sound sources which predominantly inhabit our everyday environment.

6. MULTI-STABLE PERCEPTION THROUGH COMPETITION

The multi-stable nature of auditory stream segregation implies competition between alternative percepts. But what actually competes? Intuitively, the answer seems quite straightforward: it is obviously not the individual sound events that are competing, as these are usually part of more than one alternative percept. What must compete, then, is the way in which these individual sound events are organized. In §2, we described the notion of proto-objects, the alternative groupings describing the sound sequence. The question we address here is how these proto-objects enter the competition: that is, through what principles are these proto-objects used in perceptual sound organization.

In the classical streaming paradigm [26], it has been generally assumed that competition takes place between the *Integrated* organization (described by repetition of the perceptual unit 'ABA_') and the *Segregated* organization (described by repetition of the perceptual unit 'A' and, at half the rate, repetition of the perceptual unit 'B') [15]. Note that there may be even more ways of perceiving this stimulus sequence [33,51]. Yet conceiving, the *Segregated* organization as one of the competitors already reveals a complication with this account: it implies that the percepts 'repetition of A' and 'repetition of B' always come as a pair. Allowing them to compete with each other (e.g. one stream having higher saliency than the other, or being favoured by attentional selection) introduces a new level of competition within one of the already defined competitors. Thus, based on this notion, competition occurs hierarchically, as in the hierarchical decomposition model proposed by Cusack *et al.* [36].

Alternatively, one may argue that competition takes place among 'repetition of A', 'repetition of B' and 'repetition of ABA' at the same level, without any hierarchical relations. In the framework set up above, this would be expressed as each proto-object entering the competition on its own. In this account, the joint organization of 'repetition of A' and 'repetition of B' into the *Segregated* percept arises only at a post-competition stage in perception. (Note that in many experiments, listeners were asked about their perception in terms of *Integrated* versus *Segregated*, and it was assumed that those listeners who experience the *Segregated* A percept also experienced *Segregated* B. This assumption may not be fully justified.) In the following, we shall consider three ways of conceptualizing the competition (figure 3) and examine them for their theoretical plausibility and compatibility with experimental evidence, with an eye towards how they could be computationally implemented.

(a) *Hierarchical competition*

In this account, full perceptual organizations compete with each other. We consider an *organization* to be one possible coherent interpretation of the entire auditory scene (e.g. integrating all the sounds that are currently present, or forming two or more subgroups). If there are subgroups within the currently dominant organization, then there is a second level of competition, in which the corresponding proto-objects compete to be in the foreground. There is thus separate competition between and within organizations; a graphical representation of this process is shown in figure 3*a*. This notion is consistent with the qualitatively different character of these two forms of competition. The first-level competition (between organizations) deals with mutually exclusive percepts. The second-level competition (within organizations) deals with compatible percepts in the sense that switching from one to the other does not require a reinterpretation of the auditory input. In this sense, the hierarchical-system

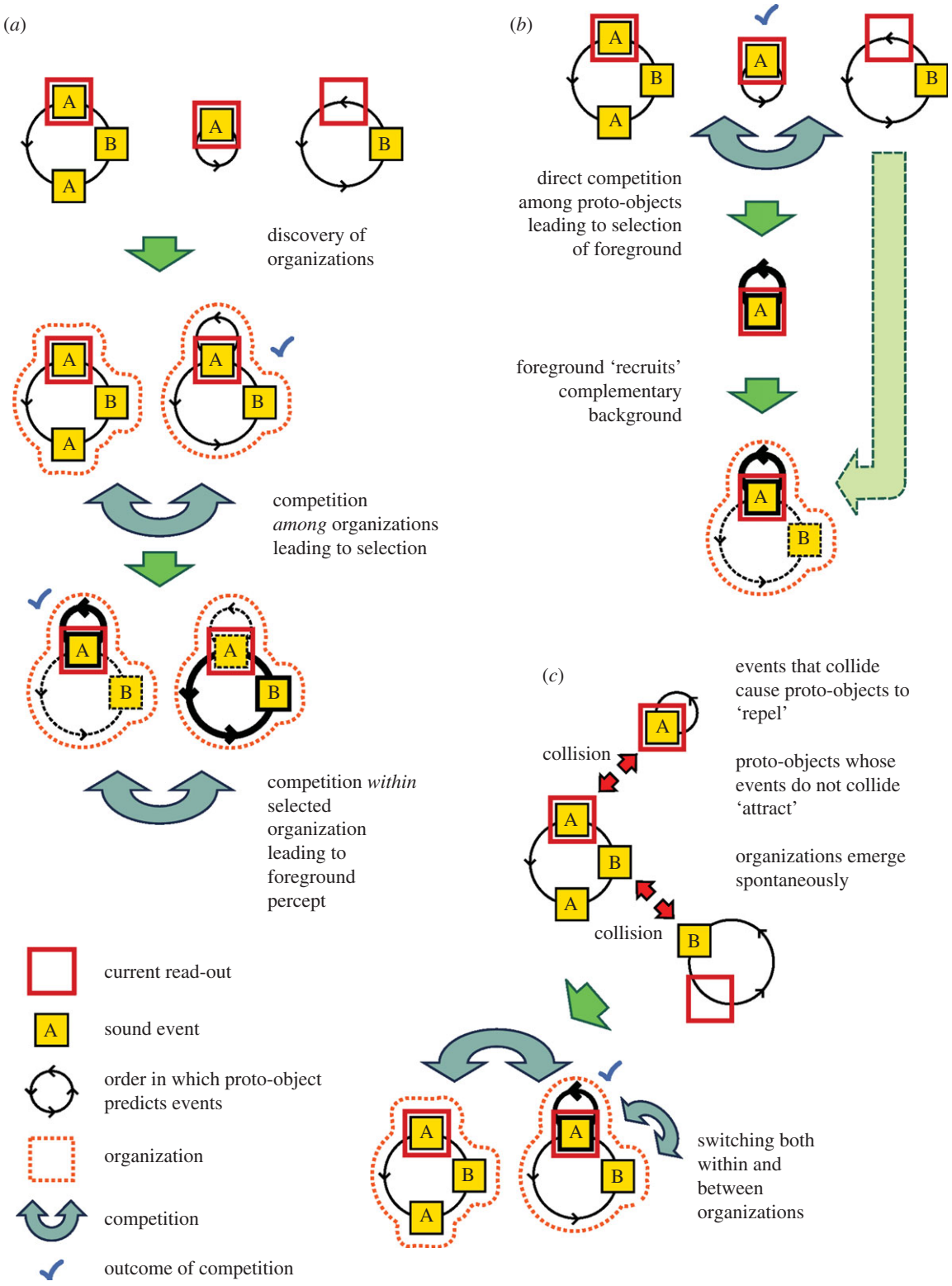

Figure 3. Alternative strategies for explaining perceptual switches in multistability. Sub-panels show, separately for each strategy, how the segregated organization with the A-stream in the foreground would emerge in the ABA_ (auditory streaming) paradigm. Proto-objects are depicted as circular graphs, in which yellow nodes indicate sound events, and the lengths of the arcs connecting the nodes are proportional to the times between the sound events. (Thus, the radius of the circle is proportional to the period of the corresponding sequence.) The sound nodes are to be thought of as moving on 'tracks', in the direction of the chevrons, and are predicted as they pass under the read-out (red square). Proto-objects that are currently perceived in the foreground are highlighted with thicker edges. Green arrows indicate the flow of processing stages; blue arrows indicate a competition at a given stage. (*a*) *Hierarchical competition*: in this scheme, competition takes place in two stages. The first stage selects an organization (an interpretation of the entire auditory scene), and the second stage selects from within this organization a proto-object to appear in the foreground. (*b*) *Foreground–background*: a single competition selects a proto-object to appear in the foreground, after which complementary proto-objects are drawn from those remaining to form an appropriate background. (*c*) *Local collisions*: when two proto-objects predict the same sound event, they mutually inhibit each other. Eventually, the overall balance of cross-inhibition determines which proto-object (or objects) are included in the dominant organization.

account is theoretically appealing. Problems arise, however, when trying to set up a hierarchical system starting with the second level, which is what the auditory system is faced with: organizations must be built bottom-up, that is, proto-objects must be formed before an organization can be constructed. For instance, there may be one proto-object containing only As, another containing only Bs, a third containing ABA_ and so on. The system now has to combine these proto-objects into perceptual organizations containing only compatible ones. How does the system find out that the As and Bs are compatible and can form an organization together (the *Segregated* percept), whereas the ABA_ proto-object can form an organization (the *Integrated* percept) on its own? Any such hierarchichal system must include a global compatibility check that governs the formation of organisations. Moreover, this compatibility check must fail occasionally to allow for the explanation of duplex percepts [67]. The picture becomes a lot more complicated when there are several possible proto-objects, which can be combined in multiple ways to form organisations. Checking all possible combinations of perceptual proto-objects for their compatibility requires full synchronization among them, including aligning proto-objects with different temporal cycles, which is computationally highly demanding. Thus, with an increasing number of proto-objects (the case of multi- as opposed to bi-stability in perception; see these auditory [48,51] and visual [68] studies showing multi-stability), forming several organizations and establishing competition between them poses specific implementation problems for the hierarchical competition solution.

(b) *Foreground–background solution*

One solution to the problem described earlier is to deal with proto-objects and organizations in a subtractive rather than an additive manner. First, the system chooses (by means of competition) the proto-object that is currently dominant. Thus, competition has only a single level with each possible proto-object competing with all the others. The dominant proto-object (e.g. a stream of repeating As) is then perceived in the foreground (figure 3*b*). Next, the system groups any sound element that is not part of the currently dominant proto-object (in this simple case, all the Bs) to form the complementary background. This foreground–background approach inherently guarantees compatibility of the current decomposition as well as complete coverage of the auditory scene. (Still, the decomposition mechanism needs to fail occasionally to account for duplex percepts.) The approach is consistent with findings showing that no distinctions are made in the background unless qualitatively very different sound sources are present in the auditory scene [35,36]. However, some results suggest that the background is not always treated as a single unit [37]. Thus, it is unclear whether a pure foreground–background solution would be supported by the currently available data. In the foreground–background account, competition takes place between all of the candidate proto-objects. When there is a change of the foreground, a new background is immediately set up. There is thus only a single level of competition and no need to compute compatibility between alternative proto-objects. This solution, however, entails two qualitatively different types of switches: (i) switching between different proto-objects for the foreground and (ii) switching between the foreground and the background. In the simple ABA_ situation, switching between the repeating As in the foreground and the repeating Bs in the background could occur seamlessly. In other situations, the system may find when it switches from foreground to background that this background cannot be described as a real proto-object and thus it needs to re-examine the whole scene and choose a new foreground. To allow for the different types of switches, the system needs to maintain the current background and all the possible foreground proto-objects in parallel. This makes the representations somewhat heterogeneous. Explaining storage and access to the representation of the 'background', which may be a loose assembly of 'leftovers' from the foreground, could pose the most difficult challenge for implementing the foreground–background account.

(c) *Local collision-based interference between proto-objects*

In the third account, we examine the possibility of whether mutual inhibition between proto-objects can result in the implicit emergence of organizations (i.e. descriptions of the whole scene) without the inhomogeneity of representations introduced by the foreground–background solution. This alternative is also based on single-level competition between the proto-objects. Interactions between proto-objects (local incompatibility) occur when they predict the same sound at the same time (within some tolerance boundary). Thus, unlike the first alternative, incompatibility information is not used for establishing a global picture of the auditory scene. The influence of incompatibility is, instead, entirely local both in time and in terms of its effects on the proto-objects: in proportion to its strength, a proto-object momentarily weakens all proto-objects with whom it collides. No other interactions occur between proto-objects. Organizations are an emergent property of this system, rather than being explicitly formed (figure 3*c*). This is because two or more proto-objects that do not inhibit each other, but which are in inhibitory relationships with some other proto-objects, can become strong or weak in the competition together and will thus implicitly form an organization. The property of being able to jointly strengthen also explains why switching between the proto-objects forming such an (implicit) organization is relatively easy: both proto-objects are free from strong suppression at the same time. Moreover, duplex percepts can be more easily explained in this framework: they are experienced when two mutually exclusive proto-objects happen to be comparably strong, whereas other proto-objects are relatively weak. Finally, unlike the hierarchical competition alternative, this solution handles multistability just as well as bi-stability. That is, the number of proto-objects does not increase the computational demand in an exponential manner, because there is no need for a global checking for

compatibility or for building a large number of possible organizations from compatible subsets of proto-objects. A local collision-based interference system would thus handle many of the competition phenomena considered above, while it is relieved from having to maintain a special background construct or a global compatibility check to form perceptual organizations.

In summary, we tentatively suggest that some form of local competition between colliding proto-objects can help to explain characteristics of perceptual multistability observed in studies of auditory stream segregation and the emergence of alternative perceptual organizations.

Finally, we expect that the ongoing competition between the discovered proto-objects is mediated by their relative strengths. Recent experimental evidence suggests that a combination of noise and adaptation is responsible for the switches in dominance [69,70], which are perceived as the emergence of a new foreground. In accordance with the local collision-based interference hypothesis, proto-objects not in competition with the foreground (if any) form the perceptual background.

7. CONCLUSIONS AND OUTSTANDING ISSUES

In this review, we suggest a framework for describing auditory scene analysis [15] in terms of the formation of multiple alternative representations of sound groups, which then continuously vie for dominance. We argue that these representations of groups are inherently predictive; allowing them to absorb predicted incoming sounds and to detect the emergence of new sound sources. Representations of groups compete based on a measure derived from the strength of the connections between the individual sounds forming the group and modulated, amongst other factors, by the predictive success of the representation. A large part of the relevant evidence originated from studies focusing on bi-/multistability in auditory stream segregation.

Several questions remain for further research. There may be more than one time scale involved in building, maintaining and eliminating the representations of sound groups. Measures of decay in the classical studies of auditory sensory memory [4] appear to match effects on existing group representations, rather than temporal limitations on connections between individual sounds. Thus, studies focusing on establishing the temporal limits of connecting sounds are required. Similarly, it is unclear at this point whether the apparently very large capacity of auditory sensory memory refers to individual sound events or also to representations of groups. Only a handful of studies have been published that systematically test interactions between multiple cues in sequential sound grouping [51]. At this point, it is not clear how the human auditory system uses redundancy or how it deals with conflicting cues. Even less is known about the possible mechanisms mediating the effects of stimulus predictability on auditory stream segregation [48] and how these effects relate to grouping based on similarity cues [49]. The issue of how stimulus parameters govern the duration of the first perceptual phase in the auditory streaming paradigm requires further careful consideration, as it may provide important constraints on modelling the formation of auditory groups and how and when competition commences between the alternatives. Finally, an explanation is required for what makes perception stable in real-life auditory scenes. That is, what is the crucial difference between natural auditory scenes and the auditory streaming paradigm, as the latter appears to be multi-stable irrespective of the stimulus parameters. Further experimental and modelling work will hopefully bring new insights into the structure and temporal dynamics of competition between alternative sound organizations in auditory stream segregation.

The theoretical framework proposed within this review was as much aimed at providing a novel explanation for a wide range of multistability phenomena observed in auditory stream segregation as to bring to light the underlying conceptual issues. Coupled with our modelling efforts [71], we hope that it will inspire further research and ultimately lead to a better understanding of the time-honoured question of how we as humans experience the auditory 'world as we find it, naively and uncritically' [38].

This research was supported by the European Community's Seventh Framework Programme FP7/2007-2013—Challenge 2: Cognitive Systems, Interaction, Robotics—under grant agreement no. 231168 acoustic scene analysis for detecting living entities (SCANDLE).

REFERENCES

1 Winkler, I., Denham, S. L. & Nelken, I. 2009 Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn. Sci.* **13**, 532–540. (doi:10.1016/j.tics.2009.09.003)
2 Friston, K. & Kiebel, S. 2009 Cortical circuits for perceptual inference. *Neural Netw.* **22**, 1093–1104. (doi:10.1016/j.neunet.2009.07.023)
3 Winkler, I. 2007 Interpreting the mismatch negativity. *J. Psychophysiol.* **21**, 147–163. (doi:10.1027/0269-8803.21.34.147)
4 Cowan, N. 1984 On short and long auditory stores. *Psychol. Bull.* **96**, 341–370. (doi:10.1037/0033-2909.96.2.341)
5 Demany, L. & Semal, C. 2007 The role of memory in auditory perception. In *Auditory perception of sound sources. Springer handbook of auditory research* (eds W. A. Yost, A. N. Popper & R. A. Fay), pp. 77–113. New York, NY: Springer.
6 Cowan, N. 1988 Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychol. Bull.* **104**, 163–191. (doi:10.1037/0033-2909.104.2.163)
7 Agus, T. R., Thorpe, S. J. & Pressnitzer, D. 2010 Rapid formation of robust auditory memories: insights from noise. *Neuron* **66**, 610–618. (doi:10.1016/j.neuron.2010.04.014)
8 Kaernbach, C. 2004 The memory of noise. *Exp. Psychol.* **51**, 240–248. (doi:10.1027/1618-3169.51.4.240)
9 Warren, R. M., Bashford, J. A., Cooley, J. M. & Brubaker, B. S. 2001 Detection of acoustic repetition for very long stochastic patterns. *Percept. Psychophys.* **63**, 175–182. (doi:10.3758/BF03200511)

10 Duke, R. A. 1989 Musicians' perception of beat in monotonic stimuli. *J. Res. Music. Educ.* **37**, 61–71. (doi:10.2307/3344953)
11 van Noorden, L. & Moelants, D. 1999 Resonance in the perception of musical pulse. *J. New Music Res.* **28**, 43–66. (doi:10.1076/jnmr.28.1.43.3122)
12 Ulanovsky, N., Las, L. & Nelken, I. 2003 Processing of low-probability sounds by cortical neurons. *Nat. Neurosci.* **6**, 391–398. (doi:10.1038/Nn1032)
13 Ulanovsky, N., Las, L., Farkas, D. & Nelken, I. 2004 Multiple time scales of adaptation in auditory cortex neurons. *J. Neurosci.* **24**, 10 440–10 453. (doi:10.1523/Jneurosci.1905-04.2004)
14 Mill, R., Coath, M., Wennekers, T. & Denham, S. L. 2011 Abstract stimulus-specific adaptation models. *Neural Comput.* **23**, 435–476. (doi:10.1162/NECO_a_00077)
15 Bregman, A. S. 1990 *Auditory scene analysis. The perceptual organization of sound*. Cambridge, MA: MIT Press.
16 Darwin, C. J. 1995 Perceiving vowels in the presence of another sound: a quantitative test of the 'old-plus-new' heuristic. In *Levels in speech communication: relations and interactions: a tribute to Max Wajskop* (eds C. Sorin, J. Mariani, H. Méloni & J. Schoentgen), pp. 1–12. Amsterdam, The Netherlands: Elsevier.
17 Glenberg, A. M. 1984 A retrieval account of the long-term modality effect. *J. Exp. Psychol. Learn. Mem. Cogn.* **10**, 16–31. (doi:10.1037/0278-7393.10.1.16)
18 Winkler, I. & Cowan, N. 2005 From sensory to long-term memory: evidence from auditory memory reactivation studies. *Exp. Psychol.* **52**, 3–20. (doi:10.1027/1618-3169.52.1.3)
19 Korzyukov, O. A., Winkler, I., Gumenyuk, V. I. & Alho, K. 2003 Processing abstract auditory features in the human auditory cortex. *Neuroimage* **20**, 2245–2258. (doi:10.1016/j.neuroimage.2003.08.014)
20 Cowan, N., Saults, S. & Nugent, L. 2001 The ravages of absolute and relative amounts of time on memory. In *The nature of remembering: essays in honor of Robert G. Crowder* (eds H. L. Roediger III, J. S. Nairne III, I. Neath & A. Surprenant), pp. 315–330. Washington, DC: American Psychological Association.
21 Griffiths, T. D. & Warren, J. D. 2004 What is an auditory object? *Nat. Rev. Neurosci.* **5**, 887–892. (doi:10.1038/nrn1538)
22 Kubovy, M. & Van Valkenburg, D. 2001 Auditory and visual objects. *Cognition* **80**, 97–126. (doi:10.1016/S0010-0277(00)00155-4)
23 Bar, M. 2007 The proactive brain: using analogies and associations to generate predictions. *Trends Cogn. Sci.* **11**, 280–289. (doi:10.1016/j.tics.2007.05.005)
24 Summerfield, C. & Egner, T. 2009 Expectation (and attention) in visual cognition. *Trends Cogn. Sci.* **13**, 403–409. (doi:10.1016/j.tics.2009.06.003)
25 Creutzig, F., Globerson, A. & Tishby, N. 2009 Past–future information bottleneck in dynamical systems. *Phys. Rev. E* **79**, 041925. (doi:10.1103/PhysRevE.79.041925)
26 van Noorden, L. P. A. S. 1975 Temporal coherence in the perception of tone sequences. Doctoral dissertation, Technical University Eindhoven, Eindhoven, The Netherlands.
27 Bendixen, A., SanMiguel, I. & Schröger, E. In press. Early electrophysiological indicators for predictive processing in audition: a review. *Int. J. Psychophysiol.* (doi:10.1016/j.ijpsycho.2011.08.003)
28 Bendixen, A., Schröger, E. & Winkler, I. 2009 I heard that coming: event-related potential evidence for stimulus-driven prediction in the auditory system. *J. Neurosci.* **29**, 8447–8451. (doi:10.1523/jneurosci.1493-09.2009)
29 Kraemer, D. J. M., Macrae, C. N., Green, A. E. & Kelley, W. M. 2005 Musical imagery: sound of silence activates auditory cortex. *Nature* **434**, 158. (doi:10.1038/434158a)
30 Petkov, C. I., O'Connor, K. N. & Sutter, M. L. 2007 Encoding of illusory continuity in primary auditory cortex. *Neuron* **54**, 153–165. (doi:10.1016/j.neuron.2007.02.031)
31 Mumford, D. 1992 On the computational architecture of the neocortex. II. The role of corticocortical loops. *Biol. Cybern.* **66**, 241–251. (doi:10.1007/BF00202389)
32 Horváth, J., Czigler, I., Sussman, E. & Winkler, I. 2001 Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Cogn. Brain Res.* **12**, 131–144. (doi:10.1016/S0926-6410(01)00038-6)
33 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)
34 Cowan, N. 2001 The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114. (doi:10.1017/S0140525X01003922)
35 Brochard, R., Drake, C., Botte, M.-C. & McAdams, S. 1999 Perceptual organization of complex auditory sequences: effect of number of simultaneous subsequences and frequency separation. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 1742–1759. (doi:10.1037/0096-1523.25.6.1742)
36 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)
37 Winkler, I., Teder-Sälejärvi, W. A., Horváth, J., Näätänen, R. & Sussman, E. 2003 Human auditory cortex tracks task-irrelevant sound sources. *Neuroreport* **14**, 2053–2056. (doi:10.1097/01.wnr.0000095496.09138.6d)
38 Köhler, W. 1947 *Gestalt psychology: an introduction to new concepts in modern psychology*. New York, NY: Liveright Publishing Corporation.
39 Moore, B. C. J. & Gockel, H. E. 2012 Properties of perceptual stream formation. *Phil. Trans. R. Soc. B* **367**, 919–931. (doi:10.1098/rstb.2011.0355)
40 Moore, B. C. J. & Gockel, H. 2002 Factors influencing sequential stream segregation. *Acta Acust. United Acust.* **88**, 320–333.
41 Jones, M. R. 1976 Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychol. Rev.* **83**, 323–355. (doi:10.1037/0033-295X.83.5.323)
42 Shinozaki, N., Yabe, H., Sato, Y., Hiruma, T., Sutoh, T., Matsuoka, T. & Kaneko, S. 2003 Spectrotemporal window of integration of auditory information in the human brain. *Cogn. Brain Res.* **17**, 563–571. (doi:10.1016/S0926-6410(03)00170-8)
43 Sussman, E. S. 2007 A new view on the MMN and attention debate: the role of context in processing auditory events. *J. Psychophysiol.* **21**, 164–175. (doi:10.1027/0269-8803.21.34.164)
44 May, P. J. C. & Tiitinen, H. 2010 Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology* **47**, 66–122. (doi:10.1111/j.1469-8986.2009.00856.x)
45 Broadbent, D. E. 1958 *Perception and communication*. Oxford, UK: Pergamon.
46 Cromwell, H. C., Mears, R. P., Wan, L. & Boutros, N. N. 2008 Sensory gating: a translational effort from basic to clinical science. *Clin. EEG Neurosci.* **39**, 69–72.

47 French-St.George, M. & Bregman, A. S. 1989 Role of predictability of sequence in auditory stream segregation. *Percept. Psychophys.* **46**, 384–386. (doi:10.3758/BF03204992)

48 Bendixen, A., Denham, S. L., Gyimesi, K. & Winkler, I. 2010 Regular patterns stabilize auditory streams. *J. Acoust. Soc. Am.* **128**, 3658–3666. (doi:10.1121/1.3500695)

49 Andreou, L.-V., Kashino, M. & Chait, M. 2011 The role of temporal regularity in auditory segregation. *Hearing Res.* **280**, 228–235. (doi:10.1016/j.heares.2011.06.001)

50 Helson, H. 1964 *Adaptation-level theory: an experimental and systematic approach to behavior*. Oxford, UK: Harper & Row.

51 Denham, S. L., Gyimesi, K., Stefanics, G. & Winkler, I. 2010 Stability of perceptual organisation in auditory streaming. In *The neurophysiological bases of auditory perception* (eds E. A. Lopez-Poveda, A. R. Palmer & R. Meddis), pp. 477–488. New York, NY: Springer.

52 Sperduti, M., Tallon-Baudry, C., Hugueville, L. & Pouthas, V. 2011 Time is more than a sensory feature: attending to duration triggers specific anticipatory activity. *Cogn. Neurosci.* **2**, 11–18. (doi:10.1080/17588928.2010.513433)

53 Glenberg, A. M. & Swanson, N. G. 1986 A temporal distinctiveness theory of recency and modality effects. *J. Exp. Psychol. Learn. Mem. Cogn.* **12**, 3–15. (doi:10.1037/0278-7393.12.1.3)

54 Stefanics, G., Hangya, B., Hernádi, I., Winkler, I., Lakatos, P. & Ulbert, I. 2010 Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *J. Neurosci.* **30**, 13 578–13 585. (doi:10.1523/Jneurosci.0703-10.2010)

55 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)

56 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)

57 Roberts, B., Glasberg, B. R. & Moore, B. C. J. 2002 Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *J. Acoust. Soc. Am.* **112**, 2074–2085. (doi:10.1121/1.1508784)

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

59 Bourjaily, M. A. & Miller, P. 2011 Synaptic plasticity and connectivity requirements to produce stimulus-pair specific responses in recurrent networks of spiking neurons. *PLoS Comput. Biol.* **7**, e1001091. (doi:10.1371/journal.pcbi.1001091)

60 Bee, M. A. & Klump, G. M. 2004 Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. *J. Neurophysiol.* **92**, 1088–1104. (doi:10.1152/jn.00884.2003)

61 Fishman, Y. I., Arezzo, J. C. & Steinschneider, M. 2004 Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J. Acoust. Soc. Am.* **116**, 1656–1670. (doi:10.1121/1.1778903)

62 Hupé, J.-M. & Pressnitzer, D. 2012 The initial phase of auditory and visual scene analysis. *Phil. Trans. R. Soc. B* **367**, 942–953. (doi:10.1098/rstb.2011.0368)

63 Pastukhov, A. & Braun, J. 2008 A short-term memory of multi-stable perception. *J. Vis.* **8**, 7–14. (doi:10.1167/8.13.7)

64 Haywood, N. R. & Roberts, B. 2010 Build-up of the tendency to segregate auditory streams: resetting effects evoked by a single deviant tone. *J. Acoust. Soc. Am.* **128**, 3019–3031. (doi:10.1121/1.3488675)

65 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)

66 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)

67 Rand, T. C. 1974 Dichotic release from masking for speech. *J. Acoust. Soc. Am.* **55**, 678–680. (doi:10.1121/1.1914584)

68 Kovács, I., Papathomas, T. V., Yang, M. & Fehér, A. 1996 When the brain changes its mind: interocular grouping during binocular rivalry. *Proc. Natl Acad. Sci. USA* **93**, 15 508–15 511. (doi:10.1073/pnas.93.26.15508)

69 Kang, M. S. & Blake, R. 2010 What causes alternations in dominance during binocular rivalry? *Atten. Percept. Psychophys.* **72**, 179–186. (doi:10.3758/App.72.1.179)

70 van Ee, R. 2009 Stochastic variations in sensory awareness are driven by noisy neuronal adaptation: evidence from serial correlations in perceptual bistability. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **26**, 2612–2622. (doi:10.1364/JOSAA.26.002612)

71 Mill, R., Böhm, T., Bendixen, A., Winkler, I. & Denham, S. L. 2011 CHAINS: competition and cooperation between fragmentary event predictors in a model of auditory scene analysis. In *45th Annual Conference on Information Sciences and Systems* (*CISS*). *Baltimore, MD.* (eds M. Elhilali & H. Weinert), Institute of Electrical and Electronics Engineers. (doi:10.1109/CISS.2011.5766095)