Modeling the auditory scene: predictive regularity representations and perceptual objects

István Winkler[1,2], Susan L. Denham[3] and Israel Nelken[4]

[1]Department of General Psychology, Institute for Psychology, Hungarian Academy of Sciences, 1394 Budapest, P.O. Box 398, Hungary

[2]Institute of Psychology, University of Szeged, 6722 Szeged, Petőfi S. sgt. 30-34, Hungary

[3]Centre for Theoretical and Computational Neuroscience, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK

[4]Department of Neurobiology, The Silberman Institute of Life Sciences, and the Interdisciplinary Center for Neural Computation, The Hebrew University, Edmond Safra Campus - Givat Ram, Jerusalem 91904, Israel

Corresponding author: Winkler, I. (iwinkler@cogpsyphy.hu)

Abstract

Predictive processing of information is essential for goal-directed behavior. We offer an account of auditory perception suggesting that representations of predictable patterns, or ‘regularities’, extracted from the incoming sounds serve as auditory perceptual objects. The auditory system continuously searches for regularities within the acoustic signal. Primitive regularities may be encoded by neurons adapting their response to specific sounds. Such neurons have been observed in many parts of the auditory system. Representations of the detected regularities produce predictions of upcoming sounds as well as alternative solutions for parsing the composite input into coherent sequences potentially emitted by putative sound sources. Accuracy of the predictions can be utilized for selecting the most likely interpretation of the auditory input. Thus in our view, perception generates hypotheses about the causal structure of the world.

Prediction underlies adaptive behavior

Achieving one's goals in constantly changing environments requires actions directed at future states of the world. For example, when crossing a street, one has to anticipate the location of cars at the moment when one is likely to intersect their trajectories. Predicting future events is essential for everything we do, from taking into account the immediate sensory consequences of our own actions to signing up to a pension plan. The realization that we constantly interact with the future led to recent theoretical proposals for predictive descriptions of cognitive processes and their implementation in the brain in various domains of cognitive neuroscience. These theories are typically informed by concepts from *Bayesian inference* and consider that the 'purpose' of perception is to generate testable hypotheses about the causal structure of the external world, based both on prior knowledge and the current sensory input[1]. The various theories differ in their emphasis, spanning the range from cognitive, functional approaches[2, 3] through approaches focusing on the two-way transfer of information along *sensory hierarchies*[4] to system approaches specifying details of the architecture and computations involved[5].

In this review, we draw on the notion that prediction underlies perception. We focus on the auditory modality, stressing the importance of the representation of temporal *regularities* as intrinsic to prediction. We argue that regularity representations play an essential role in parsing the complex acoustic input into discrete object representations and in providing continuity for perception by maintaining a cognitive model of the auditory environment. We review evidence showing that some processing of regularities occurs at quite low levels in the auditory system and suggest that auditory perceptual objects are mental constructs based on representations of temporal regularities which are inherently predictive, continuously generating expectations of the future behavior of sound sources. Finally, we examine the role of focused attention in forming auditory object representations.

We conclude that the auditory objects appearing in perception are based on detecting regular features within the acoustic signal. Regularity representations provide alternative interpretations of the acoustic input. Testing the predictions of these representations against incoming sounds guides selection of the dominant (perceived) alternative.

Predictive representations in analyzing the auditory scene

Orderly perception of complex auditory scenes requires them to be broken down into internally coherent constituents. According to Bregman's theory[6] (see Box 1), *auditory scene analysis (ASA)* consists of two phases; the first phase is concerned with the *formation* of alternative sound organizations, while the second is concerned with selecting one of the alternatives to be perceived. Although perceptually it is difficult to separate these processes, the existence of the two phases was demonstrated using *event-related brain potentials (ERPs)*[7, 8]. Winkler and colleagues[8] found two distinct ERP components elicited in sound sequences whose perception spontaneously alternated between two different organizations. The earlier component was elicited when stimulation parameters promoted one organization irrespective of which organization was perceived, whereas the later component only accompanied the actual perception of this organization. The results were interpreted as reflecting the initial formation of alternative interpretations and, separately, the selection of one sound organization.

How does the initial sound organization emerge? In the absence of contextual influences, segregation can be initially based on *simultaneous grouping cues* (see Box 1). For example, Alain and colleagues[9] discovered an event-related brain potential (ERP) component (termed *Object Related Negativity – ORN*), which is elicited when one harmonic of a *complex tone* is sufficiently mistuned, so that it is perceived as separate from the rest of the tone. However, simultaneous cues are insufficient for resolving most natural scenes, and auditory scene analysis also utilizes regularities which link multiple sound events. The key to this process is the formation of a representation which captures the regularities common to a coherent sequence of sounds; a 'model' of a putative sound source. This notion of regularity representation stems from the Gestalt principles of perception[10]. However, in addition to encoding a regularity, this representation is predictive of the sounds that the source is likely to emit and hence can underpin the formation of an identifiable perceptual unit (object) as well as its separation from other units[11]. Direct ERP correlates of stimulus prediction are limited to the initial 80 ms of sound processing[12], suggesting fast generation and processing of the predictions. Although regularity detection is mainly *stimulus-driven*[13], some types of regularities can only be detected by persons with previous specialized training (such as learning to speak a language or playing a musical instrument)[14-16].

Those regularities which are easiest to discover are extracted first and hence determine the organization that is initially perceived. For example, in the *auditory streaming paradigm* (see Box 1), the initial links are most often those between temporally adjacent tones. Later, links are formed between tones sharing some stimulus parameter[17], such as frequency in the example in Box 1. Competition between these links determines the perception of either a single sequence (when the links between temporally adjacent tones are dominant) or the perception of two sequences (when the links between same-feature tones dominate)[18]. Encoding the links has possible neuronal correlates in the responses of auditory neurons to the two different sounds. When many neurons respond to both sounds, the links between temporally adjacent sounds are presumably stronger and a single sequence is perceived, whereas if most neurons respond only to one or to the other, but not to both sounds, two *streams* are formed. *Neural adaptation* to repeating sounds can be stimulus-specific[19-21]. Thus, even neurons that initially respond similarly to both sounds may eventually develop an imbalance, a weakening of the temporally-adjacent links in favor of the repeating-feature ones. Although the location of the neurons encoding these links is debated[19-21], the model accounts well for the effects of the acoustic parameters on the time course of the *build-up of streaming*[6, 22, 23] It predicts faster onset for streaming with larger feature differences and with faster presentation rates, since both lead to faster and stronger adaptation.

The build-up of streaming has been interpreted as the gathering of evidence in favor of the segregated organization[6]. Within the present framework, we interpret this as competition between alternative sequential associations[18]. In accordance with our view, when listeners are presented with long unchanging sound sequences, such as in the auditory streaming paradigm, their perception fluctuates between the alternative organizations even when the stimulus parameters strongly promote one or the other organization[13, 18, 24]. The neuronal model, described above, while accounting for the build-up, is as yet insufficient to account for the continued perceptual switching. We argue that in addition, it is necessary to assume that competition between alternative sequential associations is a constant feature of ASA[18].

Thus predictive regularity representations provide initial hypotheses for the constituents of the complex auditory input (i.e., they are putative auditory objects). The formation and dynamical behavior of these representations can be related to neural mechanisms observed in several stations of the auditory system.

Maintaining the representation of the auditory scene

Once possible object representations are formed, inconsistencies between them need to be resolved while preferably maintaining the continuity of perception. Figure 1 shows a conceptualization of ASA. First-phase grouping processes are represented on the left with simultaneous and sequential grouping processes separately marked (bottom left box). Sequential grouping is based on predictions produced by representations encoding the previously detected acoustic regularities (upper left box). Competition between alternative sound groupings is resolved in the second phase of ASA (bottom right). Bregman[6] describes this process as “voting” by the grouping processes supporting one or another alternative. Representations reflecting the selected organization are passed onto higher level processes, such as conscious perception. Thus, we always experience sounds as part of some pattern and as belonging to a given stream (lower right arrow).

The various grouping primitives probably have different weights in the voting procedure. Weights reflect confidence in the grouping process. Figure 1 emphasizes the online adjustment of weights according to the reliability of the predictions based on the given regularity representation (Figure 1, upper right). Weights are adjusted after predictions are matched against the parsed input. When a prediction fails, the weight of the corresponding regularity representation is decreased. This process is probably reflected in the *Mismatch Negativity (MMN)* event-related potential[11, 25] (see Box 2). Switching between alternative sound organizations can result from dynamical fluctuations of the weights when both alternatives are strongly supported[18] or from active exploration of alternative interpretations of the input (conveyed by top-down biasing). MMN elicitation has been shown to correspond to the actually perceived sound organization[13].

The auditory system is thought to use an “old+new” strategy in parsing the sound input[6]. Once continuation of the previously detected streams is accounted for, the residue (unexplained input) is regarded as originating from a newly activated source (Figure 1, upper right). Some of the *exogenous ERP responses* (*P1, N1, P2*) may reflect the emergence of new auditory streams. These responses are sensitive to large changes in stimulus energy, which is a prime cue for the activation of a new sound source. Furthermore, they shortly follow the initial 80 ms of the processing of an incoming sound for which direct ERP correlates of prediction were observed[12], and within which the residue is probably estimated. The N1 wave[26] (see Box 2) may be the best candidate, because its frontal subcomponent can be linked to the attentional capture often resulting from the detection of a new object in the environment. In terms of our model of ASA (Figure 1), residue detection feeds into the processes forming new sequential associations (see the previous section).

Our analysis suggests that competition between alternative sound organizations is resolved by taking into account the within-context predictive reliability of the competing regularity representations. New streams are detected by processing the residual acoustic signal, i.e. that which could not be explained by continuation of the previously detected streams.

Neural bases for detecting change and deviance

Possible neural correlates of the processes that are reviewed in the previous sections may be found in various stations of the auditory system. The 'core' auditory pathway (Figure 2) seems to keep a high-fidelity representation of sounds at least up to the level of the primary auditory cortex, although contributions to the buildup of streaming could occur as early as the cochlear nucleus[21]. In the primary auditory cortex itself, a number of response features may already encode information that is related to the formation of auditory objects. For example, the discrete events that are the subject of sequential grouping may be marked by eliciting well-timed onset responses in auditory cortex. These onset responses correspond to the perception of *temporal edges*[27] and can be linked with the N1 wave and, possibly, with ORN (Figure 1).

Recently, *stimulus-specific adaptation (SSA)* has been intensively studied in the ascending auditory pathways. SSA is the reduction in the responses of a neuron to a common sound which does not generalize to other, rare, sounds[28-31]. SSA may be a neural correlate of regularity-based change detection[32]; a process underlying the maintenance and update of auditory representations. In the core ascending pathway of the auditory system, it seems that ubiquitous SSA first appears in *A1*[28, 29]. However, strong SSA is present in *non-lemniscal stations* of the auditory system (Figure 2), starting as early as the external nuclei of the *inferior colliculus*[31]. The properties of SSA (its high sensitivity to small deviations and its fast time course) make it a prime candidate for encoding inter-sound relationships and detecting deviations. SSA has been linked to the ERP components associated with various processes of ASA[25, 29, 33] (N1, ORN, and MMN; see Fig. 1). However, subcortical and cortical SSA activity occurs earlier than any of these ERP responses[32]. Thus, the SSA observed in animals presumably lies upstream of the generation of these ERPs.

As suggested by the short survey above, neural correlates of auditory scene analysis and change detection abound in the auditory system (Figure 2). It may be that they are constructed hierarchically, with the earlier stations using the more obvious stimulus properties and higher stations using derived properties. Alternatively, neural correlates of high-level processes in subcortical stations may be at least partially a reflection of the strong descending system of projections that is present in all sensory systems. These issues will have to be resolved in future experiments.

Predictive regularity representations as perceptual objects

We have argued that auditory regularity representations supported by the SSA mechanism observable in many parts of the auditory system play an essential role in parsing complex auditory scenes. Here we examine whether regularity representations may form the core of auditory object representations. Recent theories of auditory object representation[34, 35] emphasize the requirement of common characteristics for object representations across different modalities. So, what do we expect of perceptual objects? 1) In natural everyday environments, almost no sound occurs in isolation. Therefore, object representations must span multiple acoustic events. 2) An object is described by the combination of its features. 3) An object is a unit which is separable from other objects. Therefore, auditory object representations should specify which parts of the acoustic signal belong to the given object. 4) The actual information arriving from an object to our senses is quite variable in time. Therefore, object representations must generalize across the different ways the same object appears to the senses. 5) Finally, in accord with Gregory's[1] theory of perception, we expect object representations to predict parts of the object for which no input is currently available.

The predictive regularity representations fit all of these criteria.

(1) Auditory regularity representations are temporally persistent; they have been shown to connect sounds separated by up to circa 10 seconds[36] and persist for at least 30 seconds[37].

(2) Auditory regularity representations encode all sound features with a resolution comparable to perception, since perceptually discriminable deviations elicit MMN (for a review, see[38]). Importantly, MMN is also elicited by rare sounds differing from two or more frequent sounds only in the *combination* of two auditory features[39, 40]. Thus, auditory regularity representations describe sounds by the combination of their features.

(3) When two sound streams are perceptually separated, MMN reflects the perceived sound organization[11], its elicitation dynamically follows perceptual fluctuations between two alternative sound organizations and the effects of priming sequences on perception[13]. Critically, if two concurrent auditory streams are characterized by separate regularities, then deviant sounds only elicit an MMN with respect to the stream to which they belong perceptually[41, 42]. Thus regularity representations correspond to the perceptually separable units of the auditory input.

(4) Regularities are extracted from acoustically widely different exemplars in a sequence[43-45], including the natural variation of environmental sounds[46]. Moreover, regularities governing the variation of sounds are also extracted from a sound sequence (e.g., "the higher the pitch the softer the tones in the sequence"; see[47]). Thus auditory regularity representations generalize across different instances of the same object.

(5) Violations of predictive rules have been shown to elicit the MMN (for recent reviews, see[11, 48, 49]). For example, delivering a low tone after a short one elicited the MMN, when for most tones the rule "short tones are followed by high-pitched tones, long tones by low-pitched tones" held[50, 51]. Direct evidence for the generation of predictions was obtained by Bendixen and colleagues[12], who observed short-latency ERP correlates of auditory anticipation. Compatible results were obtained with a wide variety of stimulus paradigms[52-56]. Thus it appears that auditory regularity representations provide predictions of future sound events.

We therefore suggest that representations of auditory regularities serve as perceptual objects. That is, auditory objects are described in the brain by predictive rules linking together coherent sequences of sounds. Although there are obvious modality-specific phenomena, the notion of describing objects by the rules binding them into a unit could also be applicable in other modalities. Many Gestalt principles appear to work similarly in different modalities and the requirement for object representations to interpolate and extrapolate from the available data was initially conceived largely on the basis of visual evidence[1]. Violating visual and somatosensory temporal regularities elicits visual and somatosensory analogues of the auditory MMN, respectively[57, 58]. Very recently, an MMN-like component has been observed in response to violating an audiovisual regularity[59, 60]. Thus it appears that regularity representations are formed and utilized even in cross-modal integration.

Auditory object representations and attention

The hypothesis that auditory object representations are representations of the regularities linking together sounds forming a coherent sequence allows us to reexamine the long-standing debate in psychology regarding whether object formation requires focused attention[61, 62]. Within the present framework, we should ask whether forming regularity representations requires attention. Several studies suggest that deviations from auditory regularities are detected even when attention is not focused on the sounds[38, 63], including regularities based on the conjunction of auditory features[39, 40], a focal point of the debate about the role of attention in object formation. Furthermore, auditory streams may also be formed outside the focus of attention[64]. Most convincingly, acoustic regularities are detected in comatose patients[65] and in sleeping newborns[66]. For example, neonates detect violations of the beat in a rhythm with natural variations[67] and the ratio of different constituent sounds within sound patterns[68]. Stream-formation dependent regularity detection was also observed in newborns[69]. Thus it appears that in the auditory modality, forming predictive regularity representations does not require focused attention. This may also be true for vision. Summerfield and Egner[70] argue that expectation and attention have complementary functions in visual perception and that they are produced by separate neural mechanisms[71].

However, it is unknown whether sleeping newborns or comatose patients form perceptual object representations. Furthermore, attention can affect auditory deviance detection[72], *feature binding*[39], resetting of stream segregation[23], and determining which streams are segregated within a complex auditory scene[73]. Thus it seems plausible that although object representations can be formed outside the focus of attention, attentive processes have a strong modulating effect.

Conclusions

We have argued that predictive representations of temporal regularities constitute the core of auditory objects in the brain. This notion of auditory object formation is compatible with recent accounts of perception in other modalities[3, 70], with theories of motor control[74], and the interaction between motor control and perception[75]. Although there are several outstanding questions regarding the mechanisms underlying the proposed model (Box 3), it appears that predictive processing occurs at all levels of cognitive function in the human brain[5]. We therefore hypothesize that auditory sensory memory and predictions are but the two sides of the same coin.

Acknowledgements

Supported by the European Community's Seventh Framework Programme (grant n° 231168 – SCANDLE; I.W. and S.D.) and by a grant of the Israeli Science Foundation (ISF) to I.N.

Box 1: Auditory scene analysis and the auditory streaming paradigm

The pressure waves which we experience as sounds are a combination of all the sounds present in the environment at any time. If we are to make sense of the auditory world and interact with it effectively, it is necessary for the brain to isolate the information relating to different sound sources. The phrase 'auditory scene analysis' was coined by Bregman[6] to describe this basic problem, and processing strategies which allow the brain to segregate sounds have been extensively investigated (for recent reviews, see[22, 76, 77]).

Essentially, grouping strategies fall into two classes, *simultaneous* (used to assign concurrently active features to one or more objects) and *sequential* (used to form associations between discrete sound events). Spectral regularity, *harmonicity* and common onset are primary simultaneous grouping cues. However, sequential grouping actually turns out to be the more important, in that it can override the organisations formed by simultaneous grouping cues. Ecologically this makes sense as most informative sounds, especially communication sounds, are intermittent, and it is necessary to form associations between events which may be separated in time by fairly long intervals; i.e. there is a trade-off between global and local decisions, and the global context constrains local decisions.

Sequential grouping has often been investigated using the auditory streaming paradigm (see Figure I below) to determine the physical parameters which govern the associations formed between alternating sounds. The importance of this approach is that the same sequence of sounds can be perceived in (typically two) different ways depending on the sequential grouping decision, and there are salient perceptual differences between the different groupings. For example, if all sounds illustrated in the figure below are considered to belong to the same group (*integration*), then subjects perceive and report a galloping rhythm; however, if the sounds marked red form a separate group from the sounds marked green, then the galloping rhythm is no longer heard, and one sound sequence pops into the perceptual foreground (*streaming* or *segregation*), while the other falls into the background. It turns out that although differences in frequency are probably the most important factor, virtually any type of detectable difference can trigger streaming[17]. There is also a trade-off between featural differences and the time intervals between successive sounds, with shorter intervals increasing the tendency to report streaming.

Box 1 Figure Legend

Figure I. The auditory streaming paradigm[78]. The same sequence of alternating sounds can be perceived as belonging to a single perceptual object (top) or to two separate objects (bottom), one occupying the foreground and the other the background.

Box 2: The auditory N1 and the mismatch negativity (MMN) event-related brain potentials

Event-related brain potentials (ERPs) are usually analyzed in terms of components, i.e. "the contribution to the recorded waveform of a particular generator process" (p. 376 in ref[26]). The auditory N1 deflection appears with negative polarity over the frontocentral scalp, typically peaking between 100 and 120 ms from stimulus onset (Figure I). N1 is elicited by sudden changes in sound energy, such as sound onset or an abrupt change in the spectral make-up of a continuous sound. In short, the auditory N1 is elicited by acoustic change. A large part of the auditory N1 is generated bilaterally within primary auditory cortical areas. However, the auditory N1 is not a single component as it has multiple generators both within and outside the auditory cortex, which are differentially affected by stimulus parameters[26]. Increasing the inter-stimulus interval increases the N1 amplitude up to at least 10 s and the auditory N1 is sensitive to most sound features. These findings suggest that the neuronal generators of N1 are involved in the temporary storage of auditory information. However, the N1 is not sensitive to combinations of auditory stimulus features. Therefore, the neural generators of auditory N1 cannot implement a full memory representation of a sound[36].

The *scalp topography* of the mismatch negativity (MMN) ERP component (Figure II; for a recent review, see[79]) is similar to that of the auditory N1, although the generator locations of the two ERP responses can be distinguished from each other[80]. MMN is elicited by violating some regular feature of a sound sequence and it typically peaks ca. 100-140 ms from the onset of the deviation. Violations of both simple and complex regularities elicit the MMN, whereas MMN is not elicited by isolated sounds or a sound change occurring in the beginning of a sequence. In short, the MMN is elicited by sounds deviating from a detected regularity. The current interpretation of MMN suggests that MMN reflects the detection of failed auditory predictions[11]. There has been a debate in the literature as to whether or not the auditory N1 and MMN are based on separate neural processes[33, 80, 81]. Converging evidence suggests that the two ERP responses are partly but not fully based on common neural mechanisms[25, 82].

Box 2 Figure Legend

Figure II. The auditory N1 and MMN responses elicited in an oddball paradigm. Sequences composed of frequent (90% probability; "standard") low-pitched (300 Hz fundamental frequency) and infrequent (10%; "deviant") high-pitched (600 Hz) missing-fundamental complex tones of 500 ms duration were presented in a random order and with a 400 ms constant inter-stimulus interval to 12 young healthy participants. Participants were reading a book during the stimulus presentation. Group-average frontal (Fz) ERP responses are shown separately for the standard (thin line) and deviant (thick line) tones. The latency of the N1 deflection was significantly modulated by the spectral make-up of the tones (shorter peak latency for the higher-pitched tone); the difference is marked in yellow. Deviant tones elicited a negative-going second peak in the 200-260 ms interval from stimulus onset, which was not present in the standard-tone responses. Although this latency range is later than that typical for MMN (due to the specific make-up of the tones), the differential response (marked in light blue) was identified as MMN. (Figure adapted from[83]).

Box 3: Outstanding questions

- What are the neural processes that are involved in forming sequential associations and extracting regularities?
- Are regularities explicitly represented in neural activity, or implicitly in the pattern of synaptic connections that is plastically adapted to each situation?
- What kind of regularities can be detected without attention being focused on the sounds?
- Do representations of complex sequential rules help in segregating auditory streams or are they only involved in stabilizing and maintaining streams separated by simple feature cues?
- How many auditory objects can be concurrently represented? Is the limit related to the “capacity” of short-term or working memory?
- Are the neural substrates of auditory sensory memory and predictive processes separate?
- Can we find a causal link between the neurons showing SSA and the encoding of regularities (especially complex ones)?

Glossary

Auditory Scene Analysis (ASA)

The process of analyzing a complex mixture of sounds to isolate the information relating to different sound sources.

Auditory streaming

A perceptual phenomenon in which a sequence of sounds is perceived as consisting of two or more auditory streams. When streaming occurs, perceivers experience difficulty in extracting inter-sound relationships across streams, such as the order between two sounds belonging to different streams.

Build-up of auditory streams

The perception of segregated auditory streams (see Box 1) takes some time to develop. The *buildup* of streaming refers to the tendency for the probability of subjects reporting streaming to increase from the onset of the sound sequence for 4-8 s depending on the stimulus parameters.

Complex tone

A tone that contains multiple frequency components (in contrast to a simple or pure tone, which is a sine wave with a single frequency).

Feature binding

Linking together the features of a perceptual unit; e.g., the color, shape, etc. of an object seen.

Harmonicity

The property of a sound composed of harmonics (pure tone components whose frequencies are integer multiples of a greatest common divisor frequency, called the fundamental frequency, commonly within the pitch existence region of 30 – 4000 Hz).

Mismatch Negativity (MMN)

A frontally negative going component human auditory *ERP* elicited by sounds violating some of the detected regularities of the preceding sound sequence (see Box 2).

Missing fundamental complex tone

A harmonic complex tone which does no contain its own fundamental frequency (see harmonicity).

N1

A frontally negative-going exogenous wave of the human *ERP.* The auditory N1 is elicited by sudden changes in sound energy or spectral make-up (see Box 2).

Neural adaptation

The reduction in neural responses following the repetition of a stimulus

Object Related Negativity (ORN)

A component of the event-related potentials that is elicited when two concurrent sounds are separated by simultaneous cues, such as detecting a non-harmonic frequency alongside with a complex harmonic tone.

P1

A frontally positive-going exogenous human *ERP* component elicited by sound onsets. The auditory P1 is generated in primary auditory cortex and in adults, it usually peaks between 40 and 80 ms from stimulus onset.

P2

A frontally positive-going human exogenous ERP component following the N1 wave by 20 to 60 ms. The main neural generators of P2 are located in auditory cortex.

Regularity (auditory)

A repeating property of a sound sequence. Regularities can be as simple as the cyclical repetition of a sound or as complex as the rule that “short tones are followed by high-pitched tones, long tones by low-pitched tones”. In terms of auditory processing, only those regularities, which can be detected by the brain, matter (e.g., setting the frequencies of consecutive sounds in a sequence according to some arbitrary mathematical formula would not necessarily result in the brain detecting any regularity in the sequence). Detection of a regularity requires that 1) the given feature is analyzed and encoded and 2) further occurrences of the feature are matched with the retained code. Thus regularity detection involves memory and (possibly implicit) learning.

Sequential grouping of sounds

Linking together sounds, whose onsets are separated in time. These processes require memory of the history of auditory stimuli.

Simultaneous grouping of sounds

Linking together concurrent sounds by common properties, such as harmonicity or common onset. In contrast to sequential grouping, these processes do not require memory of the history of auditory stimuli.

Stimulus-driven processing

Information processing in the brain, which is determined by the incoming stimuli irrespective of the mental state or current goals of the organism.

Stimulus-specific adaptation (SSA)

The reduction in neural responses to a repetitive sound, which does not generalize to other (rare) sounds.

Temporal edge

The onset time of an auditory event

Figure Legends

Figure 1. Box model of Auditory Scene Analysis (ASA). *First phase of ASA (left; magenta):* Auditory information (lower left) enters initial grouping (lower left box). Predictive regularity representations (upper left box) support sequential grouping, whereas segregation by simultaneous cues does not require memory resources. *Second phase of ASA (right; orange):* Competition between candidate groupings is resolved by selecting the alternative supported by grouping processes carrying the highest confidence (lower right box). Confidence in those regularity representations whose predictions failed is reduced and the unpredicted part of the auditory input (residue) is parsed for new regularities (upper right boxes). *ERP components associated with some of the ASA functions (light blue circles linked to the corresponding function by "≈" signs):* ORN reflects the detection mistuned partials, which is an important spectral cue for segregating sound sources. N1 (see Box 2) is marked by an asterisk, because it stands for the exogenous components possibly reflecting the detection of a new stream. MMN (see Box 2) is assumed to reflect the process of adjusting the confidence weight of those regularity representations, whose predictions were not met by the actual input. *Top-down effects modulating ASA (marked violet at the affected processes):* Training and contextual information (i.e., previous experience or knowledge regarding the given context, such as identifying a given sequence as speech) allow one to detect some complex acoustic regularities (such as speech- and music-specific regularities). Actively searching for the emergence of some new or a specific expected object increases the sensitivity of detecting the corresponding regularity. When multiple alternative organizations receive approximately equal support (ambiguous stimulus configurations), selecting the dominant organization can be voluntarily biased. (Figure adapted from[11]).

Figure 2. Schematic representation of the ascending auditory pathways. Auditory nerve fibers from the cochlea terminate in the cochlear nucleus, the first central station of the auditory pathways. Some neurons in the cochlear nucleus already show correlates of the buildup of streaming. A complex set of stations in the brainstem, including the nuclei of the superior olivary complex (which are the first locus of binaural integration) and the nuclei of the lateral lemniscus (who are involved in high-resolution encoding of stimulus onsets and in binaural processing) projects to the inferior colliculus, the major midbrain auditory center (which doesn't have homologues in other sensory systems). Brainstem connectivity is only partially displayed, to make the figure easier to read. Collicular neurons project to the auditory station in the thalamus, the medial geniculate body, which in turn projects to auditory cortex. Binaural interactions occur in the superior olive, but in addition there are substantial connections between the ICs of both sides and between auditory cortical fields on both sides of the brain (marked by thick black arrows). The inferior colliculus, medial geniculate body and auditory cortex are complexes containing multiple subdivisions. Each has a 'core' division (the central nucleus of the inferior colliculus, ICc, the ventral division of the medial geniculate body, vMGB, and primary auditory cortex, all marked in dark blue). ICc projects heavily to vMGB which is the major auditory input to primary auditory cortex, forming the core (or lemniscal) pathway. Many neurons along the core pathway have short latency and narrow V-shaped tuning curves. Surrounding the core subdivisions, the belt or non-lemniscal stations, include the external nuclei of the inferior colliculus, the dorsal and medial divisions of the MGB, and some non-primary fields (marked in light blue). Red arrows indicate stations in which strong SSA has been documented. These include primarily the extralemniscal divisions of the IC and MGB (although weak forms of SSA may be found in the core stations as well) and primary auditory cortex.

References

1. Gregory, R.L. (1980) Perceptions as hypotheses. *Philos Trans R Soc Lond B Biol Sci* 290, 181-197
2. Bar, M. (2004) Visual objects in context. *Nat Rev Neurosci* 5, 617-629
3. Bar, M. (2007) The proactive brain: using analogies and associations to generate predictions. *Trends Cogn Sci* 11, 280-289
4. Ahissar, M., and Hochstein, S. (2004) The reverse hierarchy theory of visual perceptual learning. *Trends Cogn Sci* 8, 457-464
5. Friston, K. (2005) A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* 360, 815-836
6. Bregman, A.S. (1990) *Auditory Scene Analysis*. MIT Press
7. Snyder, J.S.*, et al.* (2006) Effects of attention on neuroelectric correlates of auditory stream segregation. *J Cogn Neurosci* 18, 1-13
8. Winkler, I.*, et al.* (2005) Event-related brain potentials reveal multiple stages in the perceptual organization of sound. *Brain Res Cogn Brain Res* 25, 291-299
9. Alain, C.*, et al.* (2002) Neural activity associated with distinguishing concurrent auditory objects. *J Acoust Soc Am* 111, 990-995
10. Köhler, W. (1947) *Gestalt Psychology*. Liveright
11. Winkler, I. (2007) Interpreting the mismatch negativity (MMN). *J. Psychophysiol* 21, 147-163
12. Bendixen, A.*, et al.* (2009) I heard that coming: ERP evidence for stimulus driven prediction in the auditory system. *J Neurosci* 29, 8447-8451
13. Rahne, T., and Sussman, E. (2009) Neural representations of auditory input accommodate to the context in a dynamically changing acoustic environment. *Eur J Neurosci* 29, 205-211
14. van Zuijen, T.L.*, et al.* (2005) Auditory organization of sound sequences by a temporal or numerical regularity: a mismatch negativity study comparing musicians and non-musicians. *Cogn Brain Res* 23, 270-276
15. Näätänen, R.*, et al.* (1993) Development of a memory trace for a complex sound in the human brain. *NeuroReport* 4, 503-506
16. Winkler, I.*, et al.* (1999) Brain responses reveal the learning of foreign language phonemes. *Psychophysiol* 36, 638-642
17. Moore, B.C.J., and Gockel, H. (2002) Factors influencing sequential stream segregation. *Acta Acust - Acust* 88, 320-333
18. Denham, S.L., and Winkler, I. (2006) The role of predictive models in the formation of auditory streams. *J Physiol Paris* 100, 154-170
19. Fishman, Y.I.*, et al.* (2004) Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J Acoust Soc Am* 116, 1656-1670
20. Micheyl, C.*, et al.* (2005) Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* 48, 139-148
21. Pressnitzer, D.*, et al.* (2008) Perceptual organization of sound begins in the auditory periphery. *Curr Biol* 18, 1124-1128
22. Snyder, J.S., and Alain, C. (2007) Toward a neurophysiological theory of auditory stream segregation. *Psychol Bull* 133, 780-799
23. Cusack, R.*, et al.* (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
24. Pressnitzer, D., and Hupe, J.M. (2006) Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr Biol* 16, 1351-1357
25. Garrido, M.I.*, et al.* (2009) The mismatch negativity: a review of underlying mechanisms. *Clin Neurophysiol* 120, 453-463
26. Näätänen, R., and Picton, T.W. (1987) The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiol* 24, 375-425

27. Fishbach, A., *et al.* (2001) Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. *J Neurophysiol* 85, 2303-2323

28. Ulanovsky, N., *et al.* (2004) Multiple Time Scales of Adaptation in Auditory Cortex Neurons. *J. Neurosci.* 24, 10440-10453

29. Ulanovsky, N., *et al.* (2003) Processing of low-probability sounds by cortical neurons. *Nat Neurosci* 6, 391-398

30. Perez-Gonzalez, D., *et al.* (2005) Novelty detector neurons in the mammalian auditory midbrain. *Eur J Neurosci* 22, 2879-2885

31. Malmierca, M.S., *et al.* (2009) Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *J Neurosci* 29, 5483-5493

32. Nelken, I., and Ulanovsky, N. (2007) Mismatch negativity and stimulus-specific adaptation in animal models. *J Psychophysiol* 21, 214-223

33. Jääskeläinen, I.P., *et al.* (2004) Human posterior auditory cortex gates novel sounds to consciousness. *Proc Natl Acad Sci U S A* 101, 6809-6814

34. Kubovy, M., and Van Valkenburg, D. (2001) Auditory and visual objects. *Cognition* 80, 97-126

35. Griffiths, T.D., and Warren, J.D. (2004) Opinion: What is an auditory object? *Nat Rev Neurosci* 5, 887-892

36. Näätänen, R., and Winkler, I. (1999) The concept of auditory stimulus representation in cognitive neuroscience. *Psychol Bull* 125, 826-859

37. Winkler, I., and Cowan, N. (2005) From sensory to long-term memory: evidence from auditory memory reactivation studies. *Exp Psychol* 52, 3-20

38. Näätänen, R., *et al.* (2007) The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clin Neurophysiol* 118, 2544-2590

39. Takegata, R., *et al.* (2005) Pre-attentive representation of feature conjunctions for simultaneous, spatially distributed auditory objects. *Brain Res Cogn Brain Res* 25, 169-179

40. Winkler, I., *et al.* (2005) Preattentive binding of auditory and visual stimulus features. *J Cogn Neurosci* 17, 320-339

41. Winkler, I., *et al.* (2006) Object representation in the human auditory system. *Eur J Neurosci* 24, 625-634

42. Ritter, W., *et al.* (2000) Evidence that the mismatch negativity system works on the basis of objects. *NeuroReport* 11, 61-63

43. Korzyukov, O.A., *et al.* (2003) Processing abstract auditory features in the human auditory cortex. *NeuroImage* 20, 2245-2258

44. Näätänen, R., *et al.* (2001) "Primitive intelligence" in the auditory cortex. *Trends Neurosci* 24, 283-288

45. Pakarinen, S., *et al.* (2007) Measurement of extensive auditory discrimination profiles using mismatch negativity (MMN) of the auditory event-related potential. *Clin Neurophysiol* 118, 177-185

46. Winkler, I., *et al.* (2003) Human auditory cortex tracks task-irrelevant sound sources. *NeuroReport* 14, 2053-2056

47. Paavilainen, P., *et al.* (2001) Preattentive extraction of abstract feature conjunctions from auditory stimulation as reflected by the mismatch negativity (MMN). *Psychophysiol* 38, 359-365

48. Baldeweg, T. (2006) Repetition effects to sounds: Evidence for predictive coding in the auditory system. *Trends Cogn Sci* 10, 93-94

49. Baldeweg, T. (2007) ERP repetition effects and Mismatch Negativity generation. A predictive coding perspective. *J Psychophysiol* 21, 204–213

50. Bendixen, A., *et al.* (2008) Rapid extraction of auditory feature contingencies. *NeuroImage* 41, 1111-1119

51. Paavilainen, P., *et al.* (2007) Preattentive detection of nonsalient contingencies between auditory features. *NeuroReport* 18, 159-163

52. Grimm, S., and Schröger, E. (2007) The processing of frequency deviations within sounds: Evidence for the predictive nature of the Mismatch Negativity (MMN) system. *Restor Neurol Neurosci* 25, 241–249

53. Haenschel, C.*, et al.* (2005) Event-related brain potential correlates of human auditory sensory memory-trace formation. *J Neurosci* 25, 10494-10501

54. Kraemer, D.J.*, et al.* (2005) Musical imagery: sound of silence activates auditory cortex. *Nature* 434, 158

55. Leaver, A.M.*, et al.* (2009) Brain activation during anticipation of sound sequences. *J Neurosci* 29, 2477-2485

56. Pariyadath, V., and Eagleman, D. (2007) The effect of predictability on subjective duration. *PLoS One* 2, e1264

57. Czigler, I. (2007) Visual mismatch negativity: Violating of nonattended environmental regularities. *J Psychophysiol* 21, 224-230

58. Akatsuka, K.*, et al.* (2007) Objective examination for two-point stimulation using a somatosensory oddball paradigm: an MEG study. *Clin Neurophysiol* 118, 403-411

59. Winkler, I.*, et al.* (2009) Deviance detection in congruent audiovisual speech: Evidence for implicit integrated audiovisual memory representations. *Biol Psychol* in press, doi:10.1016/j.biopsycho.2009.08.011

60. Widmann, A. *, et al.* (2004) From symbols to sounds: visual symbolic information activates sound representations. *Psychophysiol* 41, 709-715

61. Duncan, J., and Humphreys, G.W. (1989) Visual search and stimulus similarity. *Psychol Rev* 96, 433-458

62. Treisman, A. (1998) Feature binding, attention and object perception. *Philos Trans R Soc Lond B Biol Sci* 353, 1295-1306

63. 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

64. Sussman, E.S.*, et al.* (2007) The role of attention in the formation of auditory streams. *Percept Psychophys* 69, 136-152

65. Fischer, C.*, et al.* (2006) Improved prediction of awakening or nonawakening from severe anoxic coma using tree-based classification analysis. *Crit Care Med* 34, 1520-1524

66. Kushnerenko, E.*, et al.* (2007) Processing acoustic change and novelty in newborn infants. *Eur J Neurosci* 26, 265-274

67. Winkler, I.*, et al.* (2009) Newborn infants detect the beat in music. *Proc Natl Acad Sci USA* 106, 2468-2471

68. Ruusuvirta, T.*, et al.* (2007) Preperceptual human number sense for sequential sounds, as revealed by mismatch negativity brain response? *Cereb Cortex* 17, 2777-2779

69. Winkler, I.*, et al.* (2003) Newborn infants can organize the auditory world. *Proc Natl Acad Sci USA* 100, 1182-1185

70. Summerfield, C., and Egner, T. (2009) Expectation (and attention) in visual cognition. *Trends Cogn Sci* 13, 403-409

71. Bubic, A.*, et al.* (2008) Violation of expectation: neural correlates reflect bases of prediction. *J Cogn Neurosci* 21, 155–168

72. Haroush, K., *et al.* (2009) Momentary fluctuations in allocation of attention: Cross-modal effects of visual task load on auditory discrimination. *J Cogn Neurosci* in press, doi: 10.1162/jocn.2009.21284

73. Sussman, E.S.*, et al.* (2005) Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cogn Affect Behav Neurosci* 5, 93-110

74. Kawato, M. (1999) Internal models for motor control and trajectory planning. *Curr Op Neurobiol* 9, 718-727

75. Bäss, P., *et al.* (2008) Suppression of the auditory N1 event-related potential component with unpredictable self-initiated tones: evidence for internal forward models with dynamic stimulation. *Int J Psychophysiol* 70, 137-143

76. Carlyon, R.P. (2004) How the brain separates sounds. *Trends Cogn Sci* 8, 465-471

77. Ciocca, V. (2008) The auditory organization of complex sounds. *Front Biosci* 13, 148-169

78. van Noorden, L.P.A.S. (1975) Temporal coherence in the perception of tone sequences. Institute for Perception Research, Eindhoven

79. Kujala, T., *et al.* (2007) The mismatch negativity in cognitive and clinical neuroscience: theoretical and methodological considerations. *Biol Psychol* 74, 1-19

80. Näätänen, R., *et al.* (2005) Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. *Psychophysiol* 42, 25-32

81. May, P.J.C., and Tiitinen, H. (2009) Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiol* in press, doi:10.1111/j.1469-8986.2009.00856.x

82. Friston, K. and Kiebel, S. (2009) Cortical circuits for perceptual inference. *Neural Networks* in press, doi:10.1016/j.neunet.2009.07.023

83. Winkler, I., *et al.* (1997) Two separate codes for missing-fundamental pitch in the human auditory cortex. *J Acoust Soc Am* 102, 1072-1082

Figure 1

Click here to download high resolution image

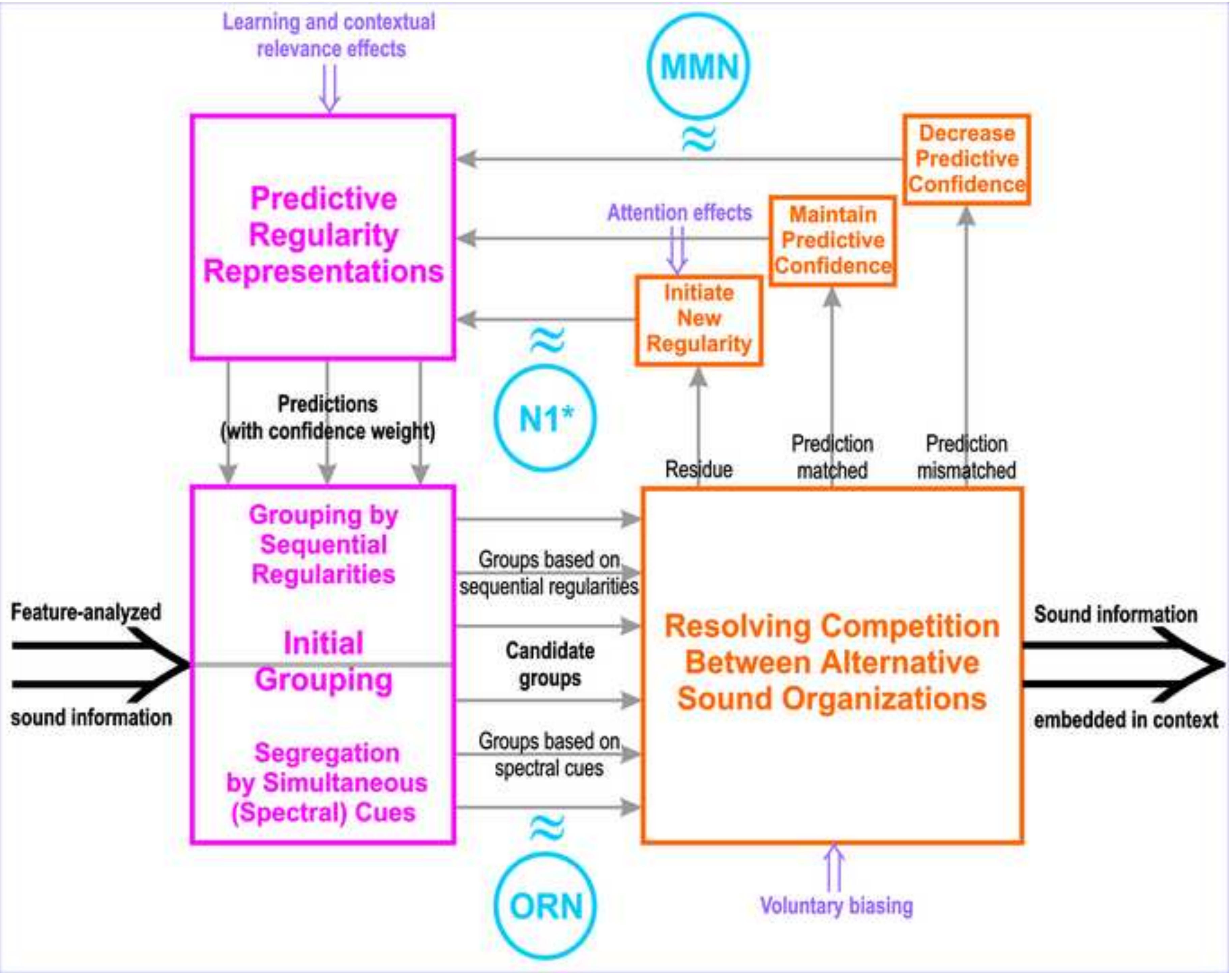

Figure 2

Click here to download high resolution image

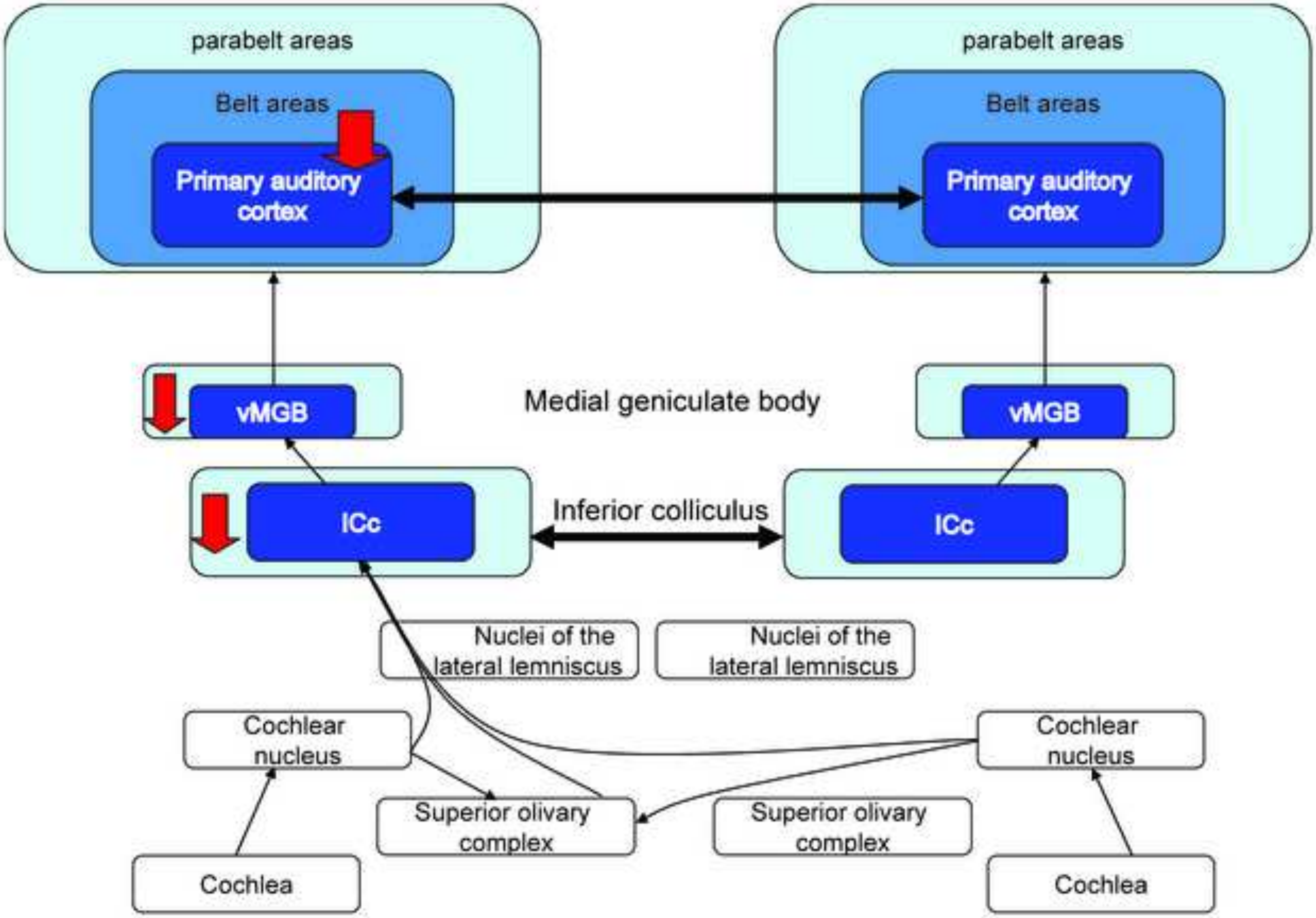

Figure I (Box 1)

Click here to download high resolution image

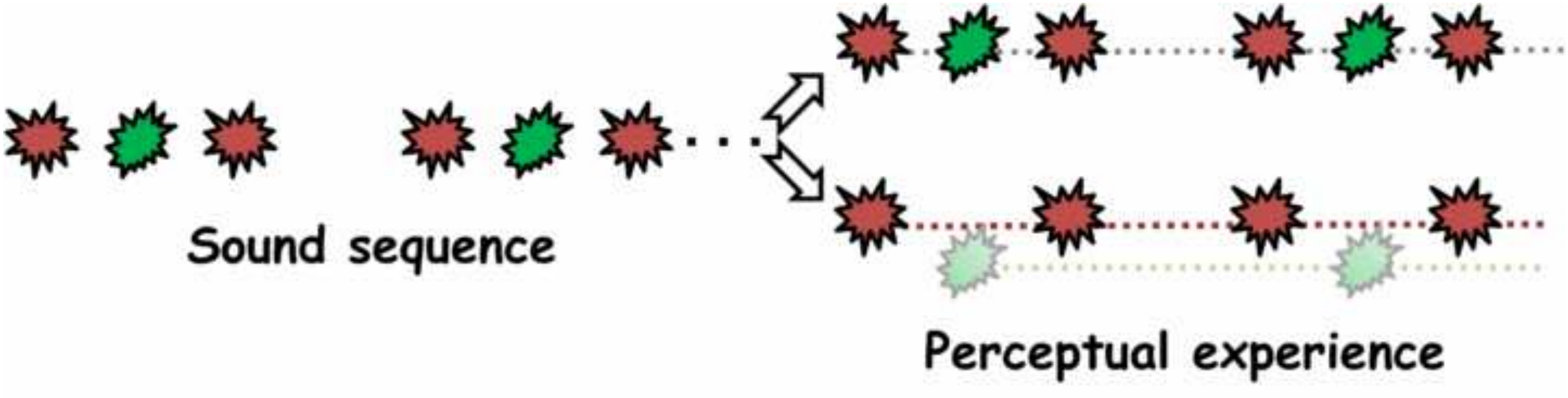

Figure II (Box 2)