Predictive regularity representations in violation detection and auditory stream segregation: from conceptual to computational models

Erich Schröger[1], Alexandra Bendixen[1,2], Susan L. Denham[3], Robert W. Mill[4], Tamás M. Bőhm[5], & István Winkler[5,6]

[1]Institute of Psychology, University of Leipzig, Leipzig, Germany

[2]Auditory Psychophysiology Lab, Department of Psychology, Cluster of Excellence "Hearing4all", European Medical School, Carl von Ossietzky University, Oldenburg, Germany

[3]Cognition Institute and School of Psychology, University of Plymouth, Plymouth, UK

[4]MRC Institute of Hearing Research, Nottingham, United Kingdom

[5]Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

[6]Institute of Psychology, University of Szeged, Szeged, Hungary

This research was supported by the Hungarian Academy of Sciences (Lendület project, LP2012-36/2012 to IW), by the Reinhart Koselleck grant of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, SCH 375/20-1 to ES), by the DFG Cluster of Excellence 1077 "Hearing4all", by the German Academic Exchange Service (Deutscher Akademischer Austauschdient, DAAD, Project 56265741), and by the Hungarian Scholarship Board (Magyar Ösztöndíj Bizottság, MÖB, Project 39589).

Abstract

Predictive accounts of perception have received increasing attention in the past twenty years. Detecting violations of auditory regularities, as reflected by the Mismatch Negativity (MMN) auditory event-related potential, is amongst the phenomena seamlessly fitting this approach. Largely based on the MMN literature, we propose a psychological conceptual framework called the Auditory Event Representation System (AERS), which is based on the assumption that auditory regularity violation detection and the formation of auditory perceptual objects are based on the same predictive regularity representations. Based on this notion, a computational model of auditory stream segregation, called CHAINS, has been developed. In CHAINS, the auditory sensory event representation of each incoming sound is considered for being the continuation of likely combinations of the preceding sounds in the sequence, thus providing alternative interpretations of the auditory input. Detecting repeating patterns allows predicting upcoming sound events, thus providing a test and potential support for the corresponding interpretation. Alternative interpretations continuously compete for perceptual dominance. In this paper, we briefly describe AERS and deduce some general constraints from this conceptual model. We then go on to illustrate how these constraints are computationally specified in CHAINS.

Keywords: audition, cognition, auditory object, auditory scene analysis, deviance detection, predictive modelling, mismatch negativity (MMN)

The processing of auditory information serves to discover the distal sources of sensory input and to detect potentially important events in the environment. To date, these two important functions have been studied relatively independently of each other in the fields of *auditory regularity violation (deviance) detection* (Näätänen 1990) and *auditory scene analysis* (Bregman 1990).

Auditory regularity violation detection (AVD) is concerned with identifying new information in a given context, which is of potential interest to the listener. The basic idea is that new information requires detailed evaluation because we do not know about it yet (as opposed to the redundant repetition of old information). If we learn the regularities inherent in the dynamic sensory input, we can readily know what is "old" and detect what is "new". The classic orienting reaction approach (Sokolov 1963) inspired this extremely fruitful field of irregularity detection, which established a psychophysiological indicator of the successful registration of new information, the Mismatch Negativity (MMN; Näätänen et al. 2011).

Auditory scene analysis (ASA) is concerned with the problem of identifying the concurrently active sound sources (Bregman, 1990). This is a considerable challenge for the information processing system, because the travelling waves emitted by the different sound sources and their echoes mix together before they reach our ears. The crucial task consists in disentangling this mixture by grouping (*integrating*) information that belongs together and separating (*segregating*) information that belongs to different sources. The classic Gestalt principles (Köhler 1947) such as the laws of common fate, similarity, and continuity have been successfully used as a guideline to study the segregation of the auditory input into coherent sound sequences, termed "auditory streams". Auditory streams are important units of auditory perception and, as argued by Winkler and colleagues (Winkler et al, 2009), they also serve as units of information storage possessing the defining characteristics of perceptual object

representations. That is, representations of auditory streams 1) are temporally persistent; 2) encode conjoined auditory features; 3) are separable from other streams; 4) can absorb natural variability of the input; and 5) predict upcoming sound events (Winkler et al., 2009). Note that auditory streams do not always correspond to a single sound source (Bregman, 1990). However, as has often been noted, sound patterns, such as a melody, can also be regarded as perceptual objects (Kubovy and Van Walkenburg, 2001; Griffith and Warren, 2004).

Because new information must either be related to previously detected sound sources, or to decisions on the presence of new sources, in a general theoretical sense one can assume that the detection of new information must be an important factor in auditory scene analysis. A more specific link between the two sets of phenomena can be formed by noticing that both AVD and ASA require knowledge about the history of stimulation. That is, they both utilize some form of representation of the auditory context. The assumption that AVD and ASA utilize at least partly overlapping representations of the auditory context is the first cornerstone of the following description. The second cornerstone comes from the insight that the representations of regularities that define perceptual objects in our environment also allow us to predict their future behavior. This is important because, compared to the time from which our information originated, any interaction with objects can only occur in the future. Therefore, we assume that the representations involved in AVD and ASA are predictive.

In the following, we will briefly describe a conceptual model linking AVD and ASA. In essence, we propose the existence of an Auditory Event Representation System (AERS) that predictively models the acoustic environment and produces representations of auditory objects (Winkler et al. 2009). These representations provide a sensory description of incoming sounds that includes their relationship to the context and the current goals of the listener. Auditory object representations can be consciously perceived; and they serve as units in

various cognitive operations. Although AERS is a "verbal-boxological" model (cf. Jacobs and Grainger 1994) and lacks detailed computational specificity, it organizes a vast amount of literature and yields various useful constraints for computational models. We will demonstrate this in the second half of the paper by describing a specific computational implementation of AERS, the CHAINS model (Mill et al. 2011; Mill et al. 2013). Please see Table 1 for a definition of key terms and concepts in the two models.

The relation between AERS and the predictive-coding approach to perception

In general, the predictive-modeling account of perception explains perception as the result of an interaction between the current sensory information and a model of what we already know about the world. This idea is by no means novel. In fact, it originates from Helmholtz's notion of unconscious inferences (Helmholtz 1867), extended in two important ways:

1) At any given moment of time, there are multiple models potentially applicable to the current sensory input. Therefore, some mechanism for selecting the optimal model should exist.
2) In order to efficiently represent an ever-changing environment and the arrival of unexpected new information, the perceptual system must monitor how well its current representations suit the environment, improving them when possible, initiating new representations if necessary, and adjusting its selection of dominant representations appropriately.

With these two principles in mind, the nature of a predictive (or generative) representation can be summarized as follows: The representation generates predictions that take into account prior experience and provide probabilistic information about what is likely to appear next in a

given context. Consider, for instance, that you have chosen a whistling voice as the ringtone for your mobile phone. Now, while you are in a crowd of people, you suddenly perceive a whistling voice. At least two explanatory models for this sensory input will be formed by your perceptual system: one of them suggesting that your mobile phone is ringing, the other one suggesting that someone is actually whistling in the crowd. Let us assume that the "mobile phone" model wins the competition initially. This will lead to the prediction that you will immediately hear the whistling voice again, which might be so strong that you misperceive a second sound event that is only vaguely similar to a whistling sound. To put this in theoretical terms, the prediction constrains and biases the interpretation of the sensory input. The actual input is compared with the prediction and the difference between the two is computed as prediction error. This prediction error, in turn, can be used for adjusting current representations and/or for selecting a dominant representation (interpretation) from those available (e.g., someone on the street was whistling).

The most complete variant of predictive coding, Friston's free energy principle (Friston and Kiebel 2009b), suggests that a) model selection is based on Bayesian inference and b) there is a hierarchy of models with increasing generality, with prediction errors from each level propagating upwards in the hierarchy (bottom-up) and models for each level being selected by the next higher level (top-down). The goal of the system is to minimize the overall prediction error formulated as an entropy type of measure, the *free energy*. In this hierarchical predictive coding model, no level within the hierarchy is of special relevance. This approach aims at describing the overall functioning of the perceptual system and can be validated by comparing the behavior of computational implementations of the model with human perceptual decisions and brain activation measures (in other words, with the *outcome* of perceptual processes). In contrast, psychological descriptions of ASA and AVD focus on the *processes* leading to the conscious perception of sounds.

Some recent computational models of MMN (Garrido et al. 2009; Lieder et al. 2013; Wacongne et al. 2012) have provided a link between predictive coding and the brain's response to auditory regularity violations. Winkler and Czigler (2012) suggested that the representations of auditory regularities involved in AVD may be mapped to models of an intermediate level in a predictive coding hierarchy. However, predictive processing has not yet been applied systematically to explain ASA. We will do this by specifying the nature of the predictive regularity representations that compete to explain the auditory sensory input, and by considering how they are initially formed and maintained. This approach fills important holes in both bodies of literature:

1) In general, predictive coding studies have rarely addressed the issue of how models are initially formed or maintained (see, however, Kiebel et al. 2009).
2) Models of ASA have largely disregarded bi-/multi-stable perceptual phenomena (for a general review, see Schwartz et al. 2012), which has led to the underestimation of the role of competition between alternative representations in ASA (see, however, Mill et al. 2013).

In summary, although AERS is generally compatible with predictive coding models (though not necessarily with any particular model), it differs from predictive coding in its roots as well as in its aims. AERS is not an instantiation of the predictive coding principle for auditory deviance detection; instead, AERS describes the common basis for auditory deviance detection and stream segregation in terms of the formation and maintenance of the memory representations underlying these processes. By AERS, we lay the groundwork for computational models describing ASA in terms of continuous competition between regularity-based descriptors of auditory event sequences.

Table 1: Definition of terms

Auditory event representation system (AERS): A cognitive system producing *auditory perceptual event representations* from *auditory stimulus events*.

Auditory perceptual object (or auditory object representation): A member of the set of currently dominant *proto-objects* that occupies perceptual awareness. 'Auditory stream' is a similar term but one which is not explicitly identified in terms of predictive representations.

Auditory predictive regularity representation or ***proto-object***: Terms used to describe the representation of a sequence of sounds linked together by some detected rule or repeating pattern in the form of a generative model. Incoming sounds are checked against the predictions of current proto-objects that compete to 'explain' them. These representations have the potential to become the perceptual objects in conscious awareness, if and when they are dominant (i.e., they are in a highly activated state, assumed to indicate their selection as the most likely description of the current input).

Auditory stimulus event: A discrete sound, localized in time and generated by some source in the external world; i.e. the physical input to our sensory systems (e.g., each of the sounds in the particular sequence of sounds generated by our mobile phone when a text message arrives).

Auditory stimulus event representation: The integrated description of the perceived features of an *auditory stimulus event*; is shaped by the predictions from *AERS*.

Auditory perceptual event representation: The description of an *auditory stimulus event* in the brain; an *auditory sensory stimulus representation*, which is linked to a perceptual object, and expanded with the description of its relationship to the auditory and the general context (e.g. the text-message sound as it appears in our perception). Auditory perceptual event representations are the output of *AERS*.

CHAINS: A computational model (and its implementation into a Matlab/C-based computer program) that incorporates aspects of *AERS*. It can be used to simulate possible *perceptual organizations* for specific parameters, which – in turn – can be tested experimentally.

Chains: A formal description (within *CHAINS*) of a sequence of *auditory stimulus events* including their timing.

Initial sound analysis: The early (bottom-up) activation patterns in the auditory system caused by each *auditory stimulus event*; this analysis is not regarded as part of *AERS*.

Perceptual organization: A complete description of the auditory environment, in terms of auditory object representations, as it appears in perception.

Auditory Event Representation System (AERS)

AERS is characterized by four major constituents, (1) the formation of auditory stimulus event representations, (2) the formation of regularity representations that predict subsequent sensory input, (3) comparison between the predictions and the sensory input, and (4) evaluation of the relevance of the relationship between the incoming sound events and the context (Figure 1). AERS is assumed to receive its input from subcortical and cortical levels, for example, in the form of spectrotemporal response patterns encoding features such as spectral energy maxima. These sound features have to be combined into a unitary auditory stimulus event representation that is held accessible for some time. The auditory N1 ERP response may (partly) reflect processes engaged in this function (Näätänen and Winkler 1999). However, the formation of stimulus event representations does not only rely on input but also on the "bias" exerted by the prior context. This context supports the formation of predictive representations that are used to compute a-priori probabilities for events embedded in the sensory input. These representations capture current auditory regularities such as a pitch alternation regularity of two tones (A and B) differing in frequency presented in a regular pace (ABABABAB…).

Regularity representations are generative models in the sense that they produce predictions for future expected parts of the pattern (i.e. upcoming sounds, such as that the tone following an A will be a B). These predictions strongly guide the formation of auditory stimulus event representations[1] of the incoming sounds. As predictive regularity representations have been normally active before the occurrence of the current sound, the auditory stimulus event representations are always shaped by the a-priori probabilities. Regularity representation is a concept that is well known and frequently used in AVD research. Winkler and colleagues (2009) suggested that the concept of a *stream* essentially corresponds to a regularity

[1] These correspond to auditory sensory memory representations of the classical MMN model (e.g. Näätänen, 1990). We prefer the term auditory stimulus event representations as a stimulus event is represented.

representation, although the notion of a *regularity* is seldom mentioned in this context (instead, the term *coherence* is used to refer to the principles holding together the sounds that form a stream) and streams are primarily regarded as perceptual, not as encoding units (however, these two aspects are obviously not contradictory). Within AERS, streams are regarded as generative models based on detected regularities. Any incoming sound receives an interpretation biased towards being a new token of the currently dominant stream (such as the continuation of the voice of a speaker). So far, few MMN studies have attempted to distinguish predictive processing from alternative explanations (e.g., reevaluating the immediate history of stimulation at the arrival of each new sound event; but see Paavilainen et al., 2007; Bendixen et al., 2008, 2009). Further, if regularity representations lie at the heart of auditory streams, then MMN should only be elicited by sounds belonging to the stream whose regularity is violated. Ritter and colleagues' (2000, 2006) results appear to support this assumption. These assumptions of AERS should be further tested by future research.

Predictions are compared with the emerging auditory stimulus event representations created for the auditory input. This comparison is not a single unitary process. It is computed at multiple anatomical and temporal levels of sound processing. Recent studies have revealed that even subcortical areas of the brain can be involved and that some form of deviance detection can take place as early as 30 ms from the onset of the violation (for a review, see Grimm and Escera 2012). A regularity representation needs to be updated when the incoming sound mismatches its prediction; this updating process is reflected by the MMN (Winkler 2007; Winkler and Czigler 1998; Winkler et al. 2009). According to the proposal of Grimm and Escera (2012) of a hierarchical novelty system (see also Escera et al., this issue), respective updating processes can possibly also be initiated by the precursors of MMN. However, there is evidence that more complex regularities such as feature conjunctions are not encoded at these early levels (Althen et al., 2013).

If the input matches the prediction, no updating is required. Instead, confidence in the model might be strengthened. The repetition positivity (RP) component (Haenschel et al. 2005; Costa-Faidella et al. 2011) and the induced gamma-band response (Herrmann et al. 2004) might be candidates for brain signals reflecting this matching process. So far no study tested whether the early deviance-related responses (as reviewed by Grimm and Escera, 2012) can be regarded as signs of prediction error. Likewise, no study tested whether RP is only elicited by true repetition or also by non-repeating, but predictable sounds (although Bendixen et al., 2008, noted an effect similar to RP in a complex MMN paradigm based on predictive regularities). Future studies may shed light on whether or not these ERP responses reflect processes assumed by AERS.

The outcome of the comparison describes the relation between the auditory stimulus event representation of the incoming sound and the prediction(s) stemming from previously detected regularities. In fact, several predictive representations may coexist, providing alternative descriptions of the auditory scene. Let us again consider the alternating pitch regularity ABABAB… Alternation is only one of several possible descriptions of the tone sequence. It has been termed the *local* rule in the literature (Horváth et al. 2001) as it makes local predictions regarding the next expected sound (sound *n+1*); that is, after a tone A tone B is predicted, and after B, A is predicted. Another possible regularity that can be derived from this sequence is that every second tone is A, while every other sound is B. This regularity generates the same sound sequence, but makes its predictions with regard to the second upcoming sound (sound *n+2*), thus termed the *global* rule. Horváth and colleagues (2001) showed that auditory predictive regularity representations for both local and global regularities are active in parallel.

The alternation regularity example can also be used to illustrate the conceptual similarity of streams and predictive regularity representations. When presented with an 'ABABAB…' stimulus, participants often report hearing an integrated percept; that is, the perception that one sound source has produced all the tones by regularly alternating between A and B tones. This corresponds to the "n+1" (local) description of the alternation regularity. However, participants may also report hearing two separate sound sources: an A-stream, consisting only of the 'A' sounds, and a B-stream, consisting only of the 'B' sounds. This corresponds to the "n+2" (global) description of the alternation regularity. Few MMN studies addressed the issue whether MMN is only related to the currently dominant (perceived) stream/regularity-representation or also to the currently non-dominant ones (e.g., Szalárdy et al., in press; Winkler et al., 2005, 2006) and the results are somewhat equivocal. This issue requires further research.

Having realized this similarity between regularity representations (in AVD research) and streams (in ASA research), how can the two perspectives inform each other in a fruitful manner? One aspect of the AVD field that can provide new insights for ASA research is the issue of how predictive models are formed. The two research fields have opposite approaches here: while AVD assumes that a mixture of sounds comes as an incoherent series of events in which the regularities must first be discovered, ASA research typically assumes that a mixture is by default interpreted as a series of events belonging together (*integrated*), while the formation of more than one representations to describe the input sequence (i.e., *segregated*) only happens after the accumulation of corresponding evidence. However, this integration-by-default view has recently been challenged by a number of groups (e.g., Deike et al. 2012; Denham et al. 2013) and thus the ASA research field must face the question of how any stream representation initially develops (Winkler et al. 2009, 2012). It may prove highly fruitful to borrow paradigms and findings from AVD research here (e.g., roving standard

paradigms as put forward by Bendixen et al. 2007; Cowan et al. 1993; Haenschel et al. 2005; Sussman et al. 2007; Winkler et al. 1996). At the same time, one aspect in which AVD may benefit from ASA concerns the notion of perceptual organization and the fact that any sequence can have multiple interpretations. This bi- or multistability is rarely considered in AVD research; usually the sequence of sounds is assumed to be processed as one stream throughout. Finally, the notion of a joint representational basis of AVD and ASA has already led to a re-consideration of how the two processes are arranged in time. While previous studies had concluded that ASA precedes AVD (e.g., Müller et al. 2005; Sussman 2005), more recent evidence suggests that this temporal relation is more flexible and depends on the strength of the acoustic and regularity cues that are available to the auditory system (Bendixen et al. 2012).

After a predictive representation (a regularity representation or stream) has been set up, its validity is tested by the occurrence of every new sound event that it predicts. The information about whether the prediction was met – and if not, how far the incoming sound deviated from it – is passed on to the evaluation process. Here, representations of incoming sounds are related to what is currently known about the environment; i.e., the relationship between the incoming sound and the context (including the current goals of the listener) is evaluated. We propose that the P3a ERP response reflects the outcome of this evaluation process and acts as a kind of "significance" marker of sensory events (Horváth et al. 2008; Rinne et al. 2006). The resulting information package is the primary output of AERS that is available for further processing. We term this an auditory perceptual event representation, because it describes the sound event together with its relation to both the auditory and the general context. This is a more realistic conceptualization of the minimal "units" of auditory perception than one restricted to the physical stimulus. In contrast to classical accounts of sensory memory such as echoic memory (Neisser 1967) or (short and long) auditory stores (Cowan 1984; Massaro

1972), AERS emphasizes that auditory perceptual event representations usually are more than a mental "echo" of the auditory stimulus event and incorporate our knowledge regarding the context, our intentions, and even affective affordances ("answer the mobile phone").

Central to AERS is the proposal that the evaluation process also participates in the search for new regularities. If a prediction is not met, this may be due to the fact that the existing predictive model is basically valid but an exception with respect to the regularity occurred (e.g., one footstep of a walking person sounded a bit different from the ones experienced in the past because the person stepped onto some object). In such a case, although the predictive regularity representation may need some modification (updating) it should not be discarded as a valid description (i.e., the model that we are hearing a series of footsteps can be maintained). However, it may also be the case that a new regularity started, that is, a new stream came into play, which does not require the updating of the existing regularity representation but rather the creation of a new one (e.g., another walking person approaches). The former case (i.e. a mismatch between the prediction and the actual continuation of the stream) corresponds to a prediction error in predictive coding models. However, the latter (i.e. the residue that cannot be explained by current predictive representations) corresponds to what Bregman (1990) captured in his "old+new strategy", a heuristic the auditory system utilizes to detect the emergence of new streams. The information that cannot be accounted for by the existing streams (i.e. the residue) can be assessed at this stage and the presence of a new sound source can be considered. As noted earlier, comparisons are only done within streams (see Ritter et al., 2000, 2006). Sounds that do not belong to the given stream are not compared and no deviance (error) signal is generated. This is marked by the "old" (i.e., belonging to one of the detected streams) information entering the comparison, whereas "new" (belonging to no previously detected stream) information initiates the formation of a new regularity representation (see Figure 1).

At this point, the description given by AERS diverges from existing predictive coding models, which lump together deviation from a prediction and the residue (the ‘old’ and the ‘new’) under a single error signal, whereas AERS distinguishes these two prediction errors and assigns different follow-up actions to them (i.e., updating an existing vs. forming a new auditory predictive regularity representation). The distinction between processing prediction errors and the residue may be reflected in ERPs: the former is assumed to elicit the MMN (and possibly earlier deviance-detection-related ERP components), whereas the latter may be reflected by components notably sensitive to large acoustic changes, such as the P1 and N1. Although most known ERP data are compatible with this assumption, it has not been directly tested.

Specifying AERS and extracting some computational principles

Formation of proto-objects

AERS provides a general scheme for forming predictive representations of repeating sound patterns. However, it makes no suggestion about how distinct sounds are linked together into a coherent representation. Thus the first issue to be addressed by a computational model based on AERS is how associations between temporally separate sounds are formed. Intuitively, it should be easier to connect similar than highly dissimilar sounds. This principle has been termed the law of similarity by the Gestalt school of psychology. However, the Gestalt school focused on vision and space, where display items are present side by side. In contrast, in the auditory modality, similarity is mediated primarily by time. Thus the principle of similarity translates to a temporal version of smooth continuation. That is, similarity is better expressed in terms of temporal rate of feature change (Jones 1976; Winkler et al. 2012). This modified definition of similarity receives support from numerous studies of auditory stream segregation

(for reviews see Moore and Gockel 2002; Moore and Gockel 2012). These results show that sounds with even moderate frequency separation may segregate when presented in close succession (high rate of change due to the short inter-stimulus interval), whereas sounds with much higher frequency separation may be perceptually grouped together if there are longer time intervals between them (low rate of change due to the longer inter-stimulus interval). The interplay between featural and temporal separation is limited by the temporal constraints of the underlying memory processes as well as by the organization of feature spaces in the auditory system. Thus a computational implementation must choose parameters in accordance with the known perceptual and neurophysiological properties. Recent studies suggest that the initial formation of predictive representations (termed "proto-objects" in Mill et al. 2011; cf. also Rensink 2000) may primarily rely on the above notion of similarity. Similarity (rate of feature change) may also determine the time needed to establish a proto-object, if we assume that links are formed with a probability related to their similarity; i.e., more similar sounds are more likely to be associated and, therefore, such groups of sounds (and the corresponding regularities) are found more quickly. In response to a sequence of sounds, the proto-object discovered first will emerge first in perception, and will remain there without competition until at least one more alternative proto-object is discovered (cf. Winkler et al. 2012).

The formation of a proto-object, however, needs an additional step beyond establishing links between sounds. Sounds linked together by similarity can only affect the processing of upcoming sounds when one can draw predictions from them. That is, the building of a proto-object is only complete once it has shown the potential to predict upcoming sound events (because only then can new events be "absorbed" by this proto-object). The simplest way this can happen is when a repeating pattern is detected. AERS suggests a more general formulation: repetition of an inter-sound relationship, with predictions taking the form of value distributions in the parameter space. In implementations, one needs to carefully choose

realistic parameters both for what kind of inter-sound relations are handled by the model and for limiting the possible length of proto-objects. The auditory system appears to show quite surprising constraints in terms of the length of the patterns (number of items within and/or duration spanned by the pattern) that can be extracted (e.g., Sussman and Gumenyuk 2005; Boh et al. 2011). These constraints need to be (even) more systematically investigated within the field of AVD to permit their inclusion within computational models of ASA.

Finally, it is important to note that temporal adjacency is not a necessary prerequisite for linking sounds together, as was shown by Bendixen and colleagues (2012b) for AVD and by numerous streaming experiments for ASA (e.g., Müller et al. 2005). This is important as it allows the auditory system to form parallel representations of alternative proto-objects for the same sequence of sounds. Evidence for the auditory system maintaining alternative regularity representations for describing the same sound sequence has been obtained in MMN studies (Horváth et al. 2001).

Maintenance of proto-objects

Predictive processes may have a dual role in maintaining proto-object representations. Firstly, they may help to improve internal cohesion, by strengthening links between the elements of a proto-object. This has been suggested by the results of studies showing that proto-objects within which individual sounds or sound features can be predicted by some simple rule (e.g. a regularly repeating pattern) are more likely to emerge in perception for longer periods of time compared with proto-objects within which sounds are only linked by more diffuse rules (e.g. a predictable feature distribution) (Bendixen et al. 2010, 2013). A proto-object with an internal predictive rule can be regarded as a proto-object with an internal structure, which allows it to provide more precise predictions compared to similar proto-objects with no internal structure.

Second, successful predictions should help to maintain a proto-object, whereas failures should decrease its chances of survival. AERS suggests that predictive failures reduce the effectiveness of the given proto-object in the competition for perceptual dominance. Further, it suggests that below some threshold, proto-objects become ineffective and stop affecting the processing of incoming sounds. However, they remain in an inactive state, from which they can be reactivated for rather long periods of time (Winkler et al. 2002).

Internal cohesion (the strength of associations between its elements) and predictive success determine the competitiveness of a proto-object, i.e., its effect on other proto-objects within the competition. The moment-to-moment activation levels of the competing proto-objects (resulting from these factors) determine which of them (possibly more than one) are part of conscious awareness at any given time.

Competition and the emergence of perceptual organizations

A perceptual organization is a complete description of the auditory environment as it appears in perception. For example, the repetitive 'ABA_' sequence (van Noorden, 1975) is most commonly heard either as a repeating three-tone pattern (i.e., all sounds appearing as a single integrated unit) or as two parallel streams of sound, one consisting of the A, the other of the B sounds, with one of them appearing in the foreground, the other in the background. Whereas in the first case, perceptual organization consists of a single sound object, in the second case, perceptual organization consists of two sound objects and the assignment of the foreground. (Note that in real-life situations, there are almost always multiple sound objects with some of them falling to the background.) As these are alternative perceptual organizations of the same sequence, only one of them can appear in perception in any given time. The questions to be addressed by a computational model implementing the AERS principles are:

1) How are perceptual organizations formed from proto-objects?

2) What is the unit that enters the competition: the proto-objects or the alternative full perceptual organizations?

3) What form does the competition take? How do the competitors affect each other?

Compatibility/incompatibility between proto-objects has been discussed by Winkler et al. (2012; see also Mill et al. 2013), who defined any pair of proto-objects as incompatible if they predicted the same sound event (termed collision). This definition is based on the principle of exclusive allocation (i.e., that any given sound can be part of only one percept at a time). Exclusive allocation mostly holds for auditory perception (see, e.g., Winkler et al. 2006), although there are examples of duplex perception (e.g., Fowler and Rosenblum 1990). An analysis of the possible forms of competition led to the suggestion of collisions as the basic building block of competition between proto-objects: two proto-objects compete with each other, when, and only when, they collide. It can be shown that competition based on this simple principle implicitly leads to the emergence of perceptual organizations as reported by human participants (Mill et al., 2013).

These are then the computational principles extracted from AERS, which underlie the development of the computational model called CHAINS (Mill et al. 2011). We now go on to describe the principles of CHAINS (for detailed accounts, see Mill et al. 2011, 2013). We would like to note that many other computational models of ASA have been formulated (Beauvois and Meddis 1991; McCabe and Denham 1997; Wang and Chang 2008), some of them also incorporating the concept of predictive processing (e.g., Elhilali and Shamma 2008; Grossberg et al. 2004). Nevertheless, we limit ourselves to the description of CHAINS here because it is intimately connected to the MMN-based AERS framework.

Competition and Cooperation between Proto-Objects in a Model of Auditory Scene Analysis (CHAINS)

Formation of proto-objects

CHAINS is a computational model that allows one to flexibly implement aspects of the AERS conceptual framework already outlined. In keeping with the AERS schematic, CHAINS receives a temporal pattern of pre-analyzed sound events as input, and explores ways to form representations of the embedded regularities. Sounds are encoded throughout CHAINS as discrete tokens, which represent a single point in feature space at a specific instant in time. In our terminology, tokens within CHAIN relate to auditory stimulus events within AERS. The CHAINS algorithm does not access the absolute features of a token. Instead, it measures the distance between a token and an incoming stimulus event, and in this way links together "constellations" of stimulus events in an unfolding time-feature space to form *chains* (Figure 2); i.e. its representations are based on relative representations of feature distributions.

The likelihood of a pattern of stimulus events coalescing into a chain in the first place is determined probabilistically according to the interaction of two functions that serve complementary roles. One function specifies the probability that an incoming event is added to a chain; the other specifies the probability that an incoming event is left out, i.e., skipped over. The CHAINS model does not specify what form these functions are to take: this decision is deferred to the modeler. Nevertheless, their general influence will be heavily informed by the empirical data considered earlier, namely, that it is difficult to link events whose features change abruptly; and, conversely, it is difficult *not* to link events whose features change gradually (Figure 2A,B). It is important to emphasize that these two outcomes are not mutually exclusive. On the contrary, it is an essential feature of CHAINS that, when presented with an input event, each chain has the possibility of splitting into *two* parallel

variants, one that includes the event, and one that excludes it. This principle leads to an exponentially-growing set of chains, the proliferation of which is to some extent constrained by the low probabilities of perceptually unreasonable linkages.

The simple chain-building scheme we have described lays a concrete groundwork for two key principles that appear in the AERS framework, namely, predictive regularity and residual input. A predictive regularity is established when a repeating pattern appears in a chain. Specifically, if a chain is found to consist of a repeating sequence, it closes to form a loop, and thereupon ceases to grow by incorporating incoming events. This is when it becomes a proto-object and starts to predict events according to the regularity it encodes (Figure 2C). The chain will persist in some form as long as its predictions are correct. At the same time, correctly predicting a given input event has immediate implications for the formation of further alternative chains: The probability of adding an event to a chain is reduced in accordance with how many times this event has been predicted by already existing chains. This can be implemented, for instance, by modifying the link probability functions described above. The probability reduction naturally gives rise to a graded interpretation of residual input: events that are predicted by fewer chains are more likely to be built into existing chains, or seed new ones. On the other hand, one can tailor the exclusion probability function to make it easier for a chain, when building, to skip over an event that has been predicted (i.e., accounted for) by many other chains.

Maintenance of proto-objects

Not all features of AERS related to the maintenance of proto-objects have been implemented in CHAINS up to now. For example, in the current state of CHAINS an incoming sound violating a predictive regularity prediction erases the respective chain. In AERS, however, it is assumed that it takes some time before a proto-object becomes ineffective. This is

suggested from AVD research, where it has been shown that it takes several repetitions of a violation in order to abolish the MMN elicited by the violations (Winkler et al. 1996). Moreover, AERS claims that regularity representations that no longer affect the processing of the incoming sounds can remain in an inactive, "dormant" state, and can be reactivated by a single "reminder" (Cowan et al. 1993). This dormant state does not (yet) have an explicit computational analogue in CHAINS; it may, however, map onto the chain's continuously varying level of excitation described in the next section. Further aspects that are not yet incorporated in the CHAINS model despite existing evidence from MMN/AVD research are summarized in the computational description of CHAINS (Mill et al. 2013). Notwithstanding these future challenges, we now go on to describe the already implemented aspects of the CHAINS model.

The level of excitation of each chain depends on its predictive success and the presence of collisions with concurrent chains (proto-objects). CHAINS simulates the dynamics of the changing and inter-related excitation levels of all existing chains, which determines its predictions of the perceptual organization of the scene.

Competition and the emergence of perceptual organizations

From the fact that CHAINS is permitted to skip over input events when building chains, it follows that most individual chains do not describe the input in its entirety. Moreover, the lack of a strict principle of mutual exclusion during the building process implies that the same input event can be incorporated into many chains, and the population of chains at a given moment will contain a degree of redundancy. A single chain, considered in isolation, therefore makes a good candidate for a *proto*-object, in that it may predict only a fragment of the auditory scene, with incomplete coverage but no internal inconsistencies. At the same time, the population of chains, considered as a whole, provides an overly exhaustive

predictive account of a single scene, with complete coverage but many colliding predictions. The ideal circumstances lie between these two extremes: insofar as a scene is predictable, a subpopulation of the chains should interleave their predictions so as to account for every event exactly once (and the remainder should acquiesce). We refer to these subpopulations as *perceptual organizations*. Of course, for any given scene, there may be more than one organization latent within a population. This provides a natural basis for reasoning about perceptual multi-stability: if the elements within a population compete directly with each other to predict events, then subsets of chains that together predict disjoint aspects of a sensory scene will (implicitly) cooperate to form organizations. This is the process by means of which CHAINS discovers and maintains organizations. We now examine the details of this process.

At the outset, we introduce the notion of a chain's *excitation*, a quantity denoted E_i, which can fall between zero (not excited) and one (fully excited). In a sense, all chains in the population are predictive, but it is the excited chains' predictions that are taken to account for the events in a scene. There are many conceivable contributions to the excitation of chain *i*, three of which are essential. Firstly, the predictive success of the chain, defined as the rate at which it makes successful predictions (S_i), increases its excitability. Secondly, collisions with any other chain, *j*, defined as the rate at which predictions of chain *i* collide with those of chain *j* (C_{ij}) multiplied by the latter chain's excitation (E_j), reduces the chain's excitability. Thirdly, there is a persistent noise term (U_i), which perturbs the chain's excitability over time.
This behaviour described above can be captured in a simple system of first-order non-linear equations. For each *i*,

$$\tau_m \frac{dE_i}{dt} = -E_i + \varphi(\alpha_S S_i - \alpha_C \sum_{j \neq i} C_{ij} E_j + \alpha_U U_i + \text{const.})$$

where $\varphi(x) = (1 + e^{-x})^{-1}$ is a sigmoid function. Alternatively, the effect of collisions can be

mediated indirectly via an *inhibitory* variable, I_i:

$$\tau_m \frac{dE_i}{dt} = -E_i + \varphi(\alpha_S S_i - \alpha_{IE} I_i + \alpha_U U_i + \text{const.})$$

$$\tau_m \frac{dI_i}{dt} = -I_i + \varphi(\alpha_C \sum_{j \neq i} C_{ij} E_j + \text{const.})$$

The benefit of the latter scheme is that it limits the effect of collisions with many chains by introducing saturation. In either case, the α parameters control how successes, collisions and noise contribute to the chain's excitation, and τ_m is a time constant that controls how rapidly excitation evolves. (The terms S_i and C_{ij} are dynamic state variables found by leaky integration, though for a repeating stationary sequence such as 'ABA_', we can treat them as though they are constant.)

Consider now how CHAINS might respond to the repeating 'ABA_' sequence. Firstly, we assume that the building process outlined earlier has assembled three chains: one that predicts all three tones (the 'ABA' chain), a second that predicts only the As (the 'A' chain), and a third that predicts only the Bs (the 'B' chain). Principally, there are two ways into which these chains can organise themselves, given the dynamical system mentioned.

In the first scenario, the excitation of chain 'ABA', E_{ABA}, is initially high and the excitation of the other two chains is low. Chain 'ABA' issues three correct predictions per stimulus cycle, whereas chains 'A' and 'B' issue only two and one, respectively. Consequently, chain 'ABA' will be more excited due to more successful predictions than 'A', and 'A' more excited than 'B' in turn. In addition, the predictions of 'ABA' will regularly collide with those of 'A' and 'B', tending to reduce the excitation of the latter even further. This process will stabilise, with E_{ABA} near to one, and E_{A} and E_{B} near to zero. In this *integrated* organization, the dominant chain 'ABA' predicts the input events by itself, and chains 'A' and 'B' are non-dominant.

In the second scenario, the excitation of chains ‘A’ and ‘B’ (E_A and E_B, respectively) are initially high, and that of chain ‘ABA’ is low. Here, the contributions due to successful predictions are the same as those in the integrated scenario. However, the excitation of chain ‘ABA’ is substantially reduced by its collisions with chains ‘A’ and ‘B’, whereas chains ‘A’ and ‘B’ are relatively uninhibited: their own predictions do not collide with each other at all, and E_{ABA} is low, so the impact of collisions with chain ‘ABA’ is small. This process will also stabilise, with E_{ABA} nearer zero, and E_A and E_B nearer one. In this *segregated* organization, the dominant chains, ‘A’ and ‘B’, alternately predict the stimulus events, and chain ‘ABA’ is non-dominant.

The *integrated* and *segregated* organizations of chains are both stable with respect to the CHAINS dynamics. If it were not for the noise terms, U_i, the competition would settle into one of these two states and remain there. However, the addition of a moderate level of noise leads to transitions back and forth between one organization and the other. To adequately model perceptual multi-stability, one must choose the α and τ parameters to ensure an appropriate balance in the contribution of success, collisions and noise. There is a broad range of parameter sets that lead to multi-stability, and we can briefly summarise their respective influences as follows. In general increasing α_S (the effect of successes) promotes integration, increasing α_C (or α_{IE}, the effect of collisions) promotes segregation, and increasing α_U (the effect of noise) increases the rate of switching and promotes segregation to a small extent. Figure 2D presents example time courses of the excitations of the ‘ABA’, ‘A’, and ‘B’ chains, as they compete with each other over a 240 second period to explain an ‘ABA_’ input sequence. The ABA chain is discovered initially, and the ‘A’ and ‘B’ chains are discovered somewhat later (~25, 52 sec, respectively). The emergence of perceptual organizations is evident in these series: either the ‘ABA’ chain inhibits the ‘A’ and ‘B’ chains to produce an

integrated phase (e.g., 90–170 sec), or the ‘A’ and ‘B’ chains together inhibit the ‘ABA’ chain to produce a segregated phase (e.g., 60–90 sec). In this example, the noise contribution to each chain (U) suffices to produce phase durations on the order of tens of seconds.

The basic chains dynamics set out above can be augmented in a straightforward manner by adding additional terms inside the sigmoid functions of the excitation and inhibition equations, $\varphi(\cdot)$. For example, a feedback term that causes a chain to excite itself promotes the stability of organizations. (Adding adaptation to this feedback term promotes the stability of organizations for only a limited period after they become dominant.) In addition, one can add a rediscovery term, which excites a chain every time another version of it is rebuilt from the input events. For example, if the parameters of an ‘ABA_’ sequence favor integration (i.e., smooth changes; see Figure 2A), then the ‘ABA’ chain will be rebuilt or rediscovered quite often. If the chains required to form the segregated organization have already been built (‘A’ and ‘B’), the frequent rediscovery of the ‘ABA’ chain will promote integration during the competition. The converse applies if the stimulus parameters promote segregation (Figure 2B). Other terms are also conceivable, for instance, those which encode the effort made to attend to a particular sound event or organization.

The most important feature of the CHAINS dynamics is that they arise naturally from the predictive successes of chains and the collisions between predictions—there is no special effort to predefine *integrated* or *segregated* percepts, as they exist with respect to an ‘ABA_’ sequence. Consequently, CHAINS is also bistable when presented with an alternating tone (‘ABAB…’), or a sequence containing three tones. Furthermore, CHAINS exhibits *multistability* when more than two possible stable states exist. For example, organizations can arise in which all three events in an ABA triplet break into three separate chains, or the first two tones in a triplet form one chain, and the third tone breaks off into its own chain.

Ultimately, which chains participate in the competition will depend on the parameters that govern the probability of their formation in the first place, and which groups of chains subsequently coalesce to form an organization will depend on their compatibility (collisions). Because CHAINS makes no assumptions concerning the form the perceptual organizations should ultimately take, it provides a flexible starting point from which to explore multistable perception driven by ambiguous sequences more complex than the classical alternating and galloping tones. As denoted above, the strength of this approach can be further increased by shaping the probability of chain formation in accordance with the vast body of AVD results based on the MMN, resulting in a fruitful integrated AERS-CHAINS framework.

Summary

We started from the Auditory Event Representation System (AERS), a conceptual framework linking auditory regularity violation detection and auditory scene analysis largely based on the MMN (Näätänen et al. 1978) research (Winkler and Schröger submitted). The notion of predictive processes underlying the elicitation of MMN (Winkler et al. 1996) has gained momentum in the last couple of years (e.g., Baldeweg 2007; Bendixen et al. 2012a; Garrido et al. 2009; Näätänen et al. 2011; Schröger 2007; Wacongne et al. 2011; Winkler 2007), partly because of its compatibility with predictive coding theories (e.g., Friston and Kiebel 2009a; Mumford 1992; Rao and Ballard 1999). AERS takes the next step by linking auditory regularity violation detection and auditory scene analysis through predictive representations of the regularities detected from the sound input, which then serve as proto-objects continuously vying for the possibility of appearing in conscious perception. From AERS, we extracted some theoretical requirements for computational models of auditory stream segregation, many of which have been implemented in the CHAINS model (Mill et al., 2013). CHAINS further specifies the formation, and competition between proto-objects. Applying

CHAINS to the auditory streaming paradigm (Van Noorden, 1975) the time-course of the excitation of the three typical proto-objects has been shown, demonstrating that CHAINS can model the dynamics of the competition and the emergence of perceptual organizations in multistable auditory stimulus configurations in a way that closely resembles perceptual reports of human listeners. Thus CHAINS demonstrates that the principles of AERS provide a viable basis for computational models of auditory stream segregation.

References

Althen H, Grimm S, Escera C (2013) Simple and complex acoustic regularities are encoded at different levels of the auditory hierarchy. European Journal of Neuroscience. doi: 10.1111/ejn.12346. [Epub ahead of print]

Baldeweg T (2007) ERP repetition effects and mismatch negativity generation - A predictive coding perspective. Journal of Psychophysiology 21 (3-4):204-213

Beauvois MW, Meddis R (1991) A computer-model of auditory stream segregation. Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology 43 (3):517-541

Bendixen A, Bőhm TM, Szalárdy O, Mill R, Denham SL, Winkler I (2013) Different roles of similarity and predictability in auditory stream segregation. Learning and Perception 5:37-54

Bendixen A, Denham SL, Gyimesi K, Winkler I (2010) Regular patterns stabilize auditory streams. Journal of the Acoustical Society of America 128 (6):3658-3666

Bendixen A, Prinz WG, Horváth J, Trujillo-Barreto NJ, Schröger E (2008) Rapid extraction of auditory feature contingencies. Neuroimage, 41(3): 1111-1119.

Bendixen A, Roeber U, Schröger E (2007) Regularity extraction and application in dynamic auditory stimulus sequences. Journal of Cognitive Neuroscience 19 (10):1664-1677

Bendixen A, SanMiguel I, Schröger E (2012a) Early electrophysiological indicators for predictive processing in audition: A review. International Journal of Psychophysiology 83 (2):120-131.

Bendixen A, Schröger E, Ritter W, Winkler I (2012b) Regularity extraction from non-adjacent sounds. Frontiers in Psychology 3:143.

Bendixen A, Schröger E, Winkler I (2009) I heard that coming: event-related potential evidence for stimulus-driven prediction in the auditory system. The Journal of Neuroscience, 29 (26): 8447-8451.

Boh B, Herholz SC, Lappe C, Pantev C (2011) Processing of complex auditory patterns in musicians and nonmusicians. PLoS One, 6 (7): e21458.

Bregman AS (1990) Auditory scene analysis. The perceptual organization of sound. MIT Press, Cambridge, MA

Costa-Faidella J, Grimm S, Slabu L, Diaz-Santaella F, Escera C (2011) Multiple time scales of adaptation in the auditory system as revealed by human evoked potentials. Psychophysiology 48 (6):774-783

Cowan N (1984) On short and long auditory stores. Psychological Bulletin 96 (2):341-370.

Cowan N, Winkler I, Teder W, Näätänen R (1993) Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). Journal of Experimental Psychology-Learning Memory and Cognition 19 (4):909-921

Deike S, Heil P, Böckmann-Barthel M, Brechmann A (2012) The build-up of auditory stream segregation: a different perspective. Frontiers in Psychology 3.

Denham SL, Gyimesi K, Stefanics G, Winkler I (2013) Perceptual bistability in auditory streaming: How much do stimulus features matter? Learning & Perception 5 (2):73-100.

Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. Journal of the Acoustical Society of America 124 (6):3751-3771.

Escera C, Leung S, Grimm S (in press) Deviance detection based on regularity encoding

along the auditory hierarchy: electrophysiological evidence in humans. Brain Topography, this issue.

Fowler CA, Rosenblum LD (1990) Duplex perception - a comparison of monosyllables and slamming doors. Journal of Experimental Psychology-Human Perception and Performance 16 (4):742-754

Friston K, Kiebel S (2009a) Cortical circuits for perceptual inference. Neural Networks 22 (8):1093-1104

Friston K, Kiebel S (2009b) Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364 (1521):1211-1221.

Garrido MI, Kilner JM, Stephan KE, Friston KJ (2009) The mismatch negativity: A review of underlying mechanisms. Clinical Neurophysiology 120 (3):453-463.

Griffiths TD, Warren, JD (2004) Opinion: What is an auditory object? Nature Review Neuroscience 5: 887-892

Grimm S, Escera C (2012) Auditory deviance detection revisited: Evidence for a hierarchical novelty system. International Journal of Psychophysiology 85 (1):88-92.

Grossberg S, Govindarajan KK, Wyse LL, Cohen MA (2004) ARTSTREAM: a neural network model of auditory scene analysis and source segregation. Neural Networks 17 (4):511-536.

Haenschel C, Vernon DJ, Dwivedi P, Gruzelier JH, Baldeweg T (2005) Event-related brain potential correlates of human auditory sensory memory-trace formation. Journal of Neuroscience 25 (45):10494-10501

Helmholtz Hv (1867) Handbuch der physiologischen Optik. Allgemeine Encyklopädie der Physik,, vol Bd 9. Voss, Leipzig

Herrmann CS, Munk MHJ, Engel AK (2004) Cognitive functions of gamma-band activity: memory match and utilization. Trends in Cognitive Sciences 8 (8):347-355.

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. Cognitive Brain Research 12 (1):131-144

Horváth J, Winkler I, Bendixen A (2008) Do N1/MMN, P3a, and RON form a strongly coupled chain reflecting the three stages of auditory distraction? Biological Psychology 79 (2):139-147

Jacobs AM, Grainger J (1994) Models of visual word recognition - sampling the state-of-the-art. Journal of Experimental Psychology-Human Perception and Performance 20 (6):1311-1334.

Jones MR (1976) Time, our lost dimension: toward a new theory of perception, attention, and memory. Psychological Review 83 (5):323-355.

Kiebel SJ, von Kriegstein K, Daunizeau J, Friston KJ (2009) Recognizing sequences of sequences. Plos Computational Biology 5 (8):e1000464

Köhler W (1947) Gestalt Psychology: An introduction to new concepts in modern psychology. Liveright Publishing, New York

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

Lieder F, Daunizeau J, Garrido MI, Friston KJ, Stephan KE (2013) Modelling Trial-by-Trial Changes in the Mismatch Negativity. Plos Computational Biology 9 (2).

Massaro DW (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychological Review 79:124-145

McCabe SL, Denham MJ (1997) A model of auditory streaming. Journal of the Acoustical Society of America 101 (3):1611-1621.

Mill R, Bőhm T, Bendixen A, Winkler I, Denham SL CHAINS - Competition and cooperation between fragmentary event predictors in a model of auditory scene

analysis. In: Information Sciences and Systems (CISS), 2011 45th Annual Conference on, Baltimore, 2011. IEEE, pp 1-6.

Mill RW, Bőhm TM, Bendixen A, Winkler I, Denham SL (2013) Modelling the Emergence and Dynamics of Perceptual Organisation in Auditory Streaming. Plos Computational Biology 9 (3).

Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. Acta Acust United Acust 88 (3):320-333

Moore BCJ, Gockel HE (2012) Properties of auditory stream formation. Philosophical Transactions of the Royal Society B-Biological Sciences 367 (1591):919-931.

Müller D, Widmann A, Schröger E (2005). Deviance-repetition effects as a function of stimulus feature, feature value variation, and timing: a mismatch negativity study. Biological Psychology, 68(1): 1-14.

Mumford D (1992) On the computational architecture of the neocortex II. The role of cortico-cortical loops. Biol Cybern 66 (3):241-251.

Näätänen R (1990) The Role of attention in auditory information-processing as revealed by event-related potentials and other brain measures of cognitive function. Behavioral and Brain Sciences 13 (2):201-232

Näätänen R, Gaillard A, Mäntysalo S (1978) Early selective-attention effect on evoked potential reinterpreted. Acta Psychologica 42:313-329

Näätänen R, Kujala T, Winkler I (2011) Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses. Psychophysiology 48 (1):4-22.

Näätänen R, Winkler I (1999) The concept of auditory stimulus representation in cognitive neuroscience. Psychological Bulletin 125 (6):826-859

Neisser U (1967) Cognitive Psychology. Appleton-Century-Crofts, New York

Paavilainen P, Arajärvi P, Takegata R (2007). Preattentive detection of nonsalient contingencies between auditory features. Neuroreport 18(2): 159-163.

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2 (1):79-87.

Rensink RA (2000) Seeing, sensing, and scrutinizing. Vision Research 40 (10-12):1469-1487.

Rinne T, Särkkä A, Degerman A, Schröger E, Alho K (2006) Two separate mechanisms underlie auditory change detection and involuntary control of attention. Brain Research 1077:135-143

Ritter W, De Sanctis P, Molholm S, Javitt DC, Foxe JJ (2006) Preattentively grouped tones do not elicit MMN with respect to each other. Psychophysiology 43(5): 423-430.

Ritter W, Sussman E, Molholm S (2000) Evidence that the mismatch negativity system works on the basis of objects. Neuroreport 11(1): 61-63.

Schröger E (2007) Mismatch negativity - A microphone into auditory memory. Journal of Psychophysiology 21 (3-4):138-146

Schwartz JL, Grimault N, Hupé J-M, Moore BC, Pressnitzer D (2012) Multistability in perception: binding sensory modalities, an overview. Philos Trans R Soc Lond B Biol Sci 367 (1591):896-905

Sokolov EN (1963) Higher nervous functions: The orienting reflex. Annual Review of Physiology 25:545-580.

Sussman ES (2005) Integration and segregation in auditory scene analysis. Journal of the Acoustical Society of America 117(3 Pt 1):1285-98.

Sussman ES, Gumenyuk V (2005) Organization of sequential sounds in auditory memory. Neuroreport 16 (13):1519-1523

Sussman ES, Horváth J, Winkler I, Orr M (2007) The role of attention in the formation of auditory streams. Perception & Psychophysics 69 (1):136-152

Szalárdy O, Winkler I, Schröger E, Widmann A, Bendixen A (in press) Foreground-background discrimination indicated by event-related brain potentials in a new auditory multistability paradigm. Psychophysiology. doi: 10.1111/psyp.12139. [Epub ahead of print]

van Noorden LPAS (1975) Temporal Coherence in the Perception of Tone Sequences. Technical University, Eindhoven

Wacongne C, Changeux JP, Dehaene S (2012) A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. Journal of Neuroscience 32 (11):3665-3678.

Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. Proceedings of the National Academy of Sciences of the United States of America 108 (51):20754-20759.

Wang DL, Chang P (2008) An oscillatory correlation model of auditory streaming. Cognitive Neurodynamics 2 (1):7-19.

Winkler I (2007) Interpreting the mismatch negativity. Journal of Psychophysiology 21 (3-4):147-163

Winkler I, Czigler I (1998) Mismatch negativity: deviance detection or the maintenance of the 'standard'. Neuroreport 9 (17):3809-3813

Winkler I, Czigler I (2012) Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. International Journal of Psychophysiology 83 (2):132-143.

Winkler I, Denham S, Mill R, Bőhm TM, Bendixen A (2012) Multistability in auditory stream segregation: a predictive coding view. Philosophical Transactions of the Royal Society B-Biological Sciences 367 (1591):1001-1012.

Winkler I, Denham SL, Nelken I (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends in Cognitive Sciences 13 (12):532-540

Winkler I, Karmos G, Näätänen R (1996) Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. Brain Research 742 (1-2):239-252

Winkler I, Korzyukov O, Gumenyuk V, Cowan N, Linkenkaer-Hansen K, Ilmoniemi RJ, Alho K, Näätänen R (2002) Temporary and longer term retention of acoustic information. Psychophysiology 39 (4):530-534

Winkler I, Takegata R, Sussman E (2005). Event-related brain potentials reveal multiple stages in the perceptual organization of sound. Cognitive Brain Research, 25 (1), 291-299.

Winkler I, van Zuijen TL, Sussman E, Horváth J, Näätänen R (2006) Object representation in the human auditory system. European Journal of Neuroscience 24 (2):625-634.

Figures

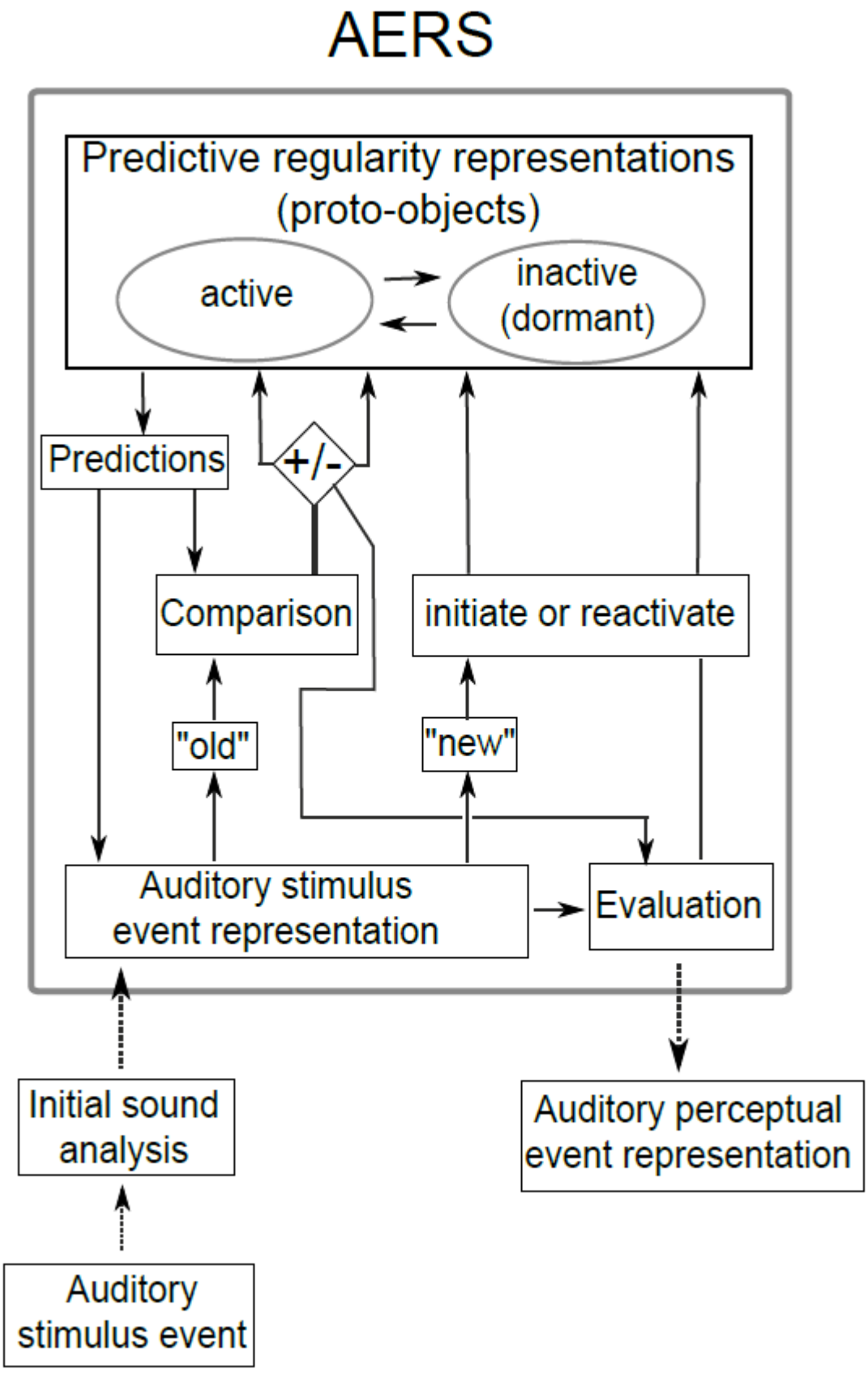

Figure 1. AERS model. The primary input to AERS are the incoming auditory stimulus

events with their basic features established by processes concerned with the initial analysis of the sound. Predictive regularity representations encode detected regularities and predict the upcoming sounds. The established auditory stimulus event representations (which, in turn, are biased by the predictions) are compared with the predictions. The outcome of this comparison is used for updating the predictive regularity representations and for the subsequent evaluation process. There, the auditory stimulus event representations are related to auditory context and to the current goals of the organism. The output of AERS is an auditory perceptual event representation (e.g. a particular tone of a flute). They can enter various mental operations and be consciously perceived. Please note, that this event representation is linked to the respective auditory object representation (e.g. the flute). In addition, the evaluation process can initiate the building of new or reactivate old but inactive regularity representations; for more details see text, for the definition of terms see Table 1.

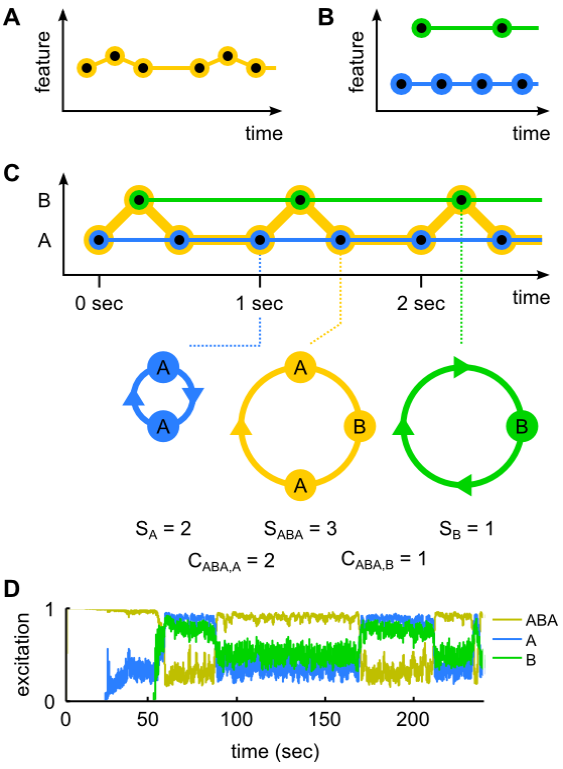

Figure 2. CHAINS model. A) Schematic illustration of circumstances that favor integration. Events plotted in a time-feature space (solid black markers) vary gradually in feature distance and consequently form into a single chain ('ABA', yellow). Links from A to A are difficult to form, however, because it is difficult to skip over B when it forms a smooth continuation with the As, and the same is true for linking the Bs. B) Schematic illustration of circumstances that favor segregation. Events vary abruptly in feature distance. It is therefore easy to form a chain consisting solely of As (green), because it is probable that 'B' will be successfully skipped. For the same reason, a complementary chain consisting solely of Bs is probable (blue). However, the 'ABA' chain is difficult to form, owing to the improbability of building many abrupt links. C) Illustration showing how various aspects of a single 'ABA_' sequence are explained by three chains, with some overlap. (The example assumes that all links that can form, do so with certainty.) At the point where a chain contains a repeating subsequence, it closes to form a predictive loop (shown below). D) Time-varying excitations of the 'ABA', 'A' and 'B' chains (E_{ABA}, E_A and E_B) in competition. Successes and collisions between the chains define a competition that gives rise to organizations that are stable for short periods.

Either ‘ABA’ dominates alone (integration), or ‘A’ and ‘B’ dominate together.