Brain & Language 148 (2015) 1–22

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Auditory perceptual objects as generative models: Setting the stage for communication by sound

István Winkler [a,b,*], Erich Schröger [c,*]

[a] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary
[b] Institute of Psychology, University of Szeged, Hungary
[c] Institute for Psychology, University of Leipzig, Germany

ARTICLE INFO

Article history:
Accepted 3 May 2015
Available online 13 July 2015

Keywords:
Audition
Cognition
Auditory object
Auditory scene analysis
Deviance (irregularity) detection
Predictive modeling
Prediction
Speech
Streaming

ABSTRACT

Communication by sounds requires that the communication channels (i.e. speech/speakers and other sound sources) had been established. This allows to separate concurrently active sound sources, to track their identity, to assess the type of message arriving from them, and to decide whether and when to react (e.g., reply to the message). We propose that these functions rely on a common generative model of the auditory environment. This model predicts upcoming sounds on the basis of representations describing temporal/sequential regularities. Predictions help to identify the continuation of the previously discovered sound sources to detect the emergence of new sources as well as changes in the behavior of the known ones. It produces auditory event representations which provide a full sensory description of the sounds, including their relation to the auditory context and the current goals of the organism. Event representations can be consciously perceived and serve as objects in various cognitive operations.

1. Communication channels

Communication requires a channel open between the participants allowing them to exchange information. Communication by sound typically occurs in environments rich in sound sources. In order to listen to someone speaking, we have to be able to create and maintain the channel conveying the information provided by the speaker. This involves separating the speaker's voice from all concurrent streams of sound which themselves are potential alternative channels to choose. For example, while driving a car, we can hear the sound of the car engine, the noise of the tires rolling over the surface, music from the radio while still being able to conduct a conversation with another person. Parsing the mixture of sounds arriving at our ears (termed Auditory Scene Analysis; Bregman, 1990) results in the formation of perceptual units called auditory objects (e.g. the speaker's voice; Griffiths & Warren, 2004; Kubovy & van Valkenburg, 2001; Winkler, Denham, & Nelken, 2009).

Every-day experience tells us that sounds deviating from the acoustic context often break into our conscious experience even if previously we did not attend their source. For example, in the previous mentioned situation (i.e., having a conversation while driving a car), one typically only notices the sound of the car engine, if it starts to cough. Deviance detection has been often studied using electric brain responses elicited by auditory events, termed auditory event-related potentials (ERPs). Sounds violating some regular feature of the preceding sequence have been shown to elicit a specific component within the auditory ERPs, termed the mismatch negativity (MMN; Näätänen, Gaillard, & Mäntysalo, 1978; for reviews, see Kujala, Tervaniemi, & Schröger, 2007; Näätänen, Kujala, & Winkler, 2011). Human and animal research in the past 30 years have revealed many details about how auditory scenes are analyzed, as well as how deviant sounds are detected within the auditory system. However, the two areas of research – auditory scene analysis and auditory deviance detection – have proceeded largely independently from each other. Here, we provide an integrative research review that develops connections between these two areas.

One common thread between the two functions is that they both require some representation of the immediate history of the stimulation. Such a representation allows discrete sounds to be linked together to form an auditory perceptual object, as well as

* Corresponding authors at: Research Centre for Natural Sciences of the Hungarian Academy Sciences, P.O. Box 286, Budapest H-1519, Hungary (I. Winkler). Institut für Psychologie, Universität Leipzig, Neumarkt 9-14, 04109 Leipzig, Germany (E. Schröger).

E-mail addresses: winkler.istvan@ttk.mta.hu (I. Winkler), schroger@uni-leipzig.de (E. Schröger).

http://dx.doi.org/10.1016/j.bandl.2015.05.003
0093-934X/© 2015 The Authors. Published by Elsevier Inc.

to assess whether they carry new information with respect to what we already know about the sound sources in the environment. We will argue that a second common feature is that both auditory scene analysis and auditory deviance detection look into the future. That is, we provide a theoretical framework linking auditory scene analysis and deviance detection via predictive auditory representations.[1]

The idea of human information processing and specifically perception operating in a predictive manner has a long tradition both in psychology and neuroscience. For example, Gregory's (1980) influential contemporary empiricist theory likens perception to scientific hypotheses, which provide the brain's "best guess" of the causes (distal objects) of the stimulation reaching the sensory organs (the proximal stimuli) and can produce extrapolations to parts of the environment, which are currently not accessible to the senses. Recent theories following the empiricist tradition, which started with Helmholtz's (1867) notion of unconscious inference and has been arguably the most influential school for explaining perception (see, e.g., Clark, 2013), posit predictive models integrating perception, attention, learning, and even actions (e.g., Ahissar & Hochstein, 2004; Bar, 2007; Friston, 2010; Hohwy, 2007; Hommel, Musseler, Aschersleben, & Prinz, 2001; Summerfield & Egner, 2009; Tishby & Polani, 2011). In neuroscience, Helmholtz's theory coupled with Bayesian rules for optimal inference generation (Kersten, Mamassian, & Yuille, 2004; Knill & Pouget, 2004) engendered the predictive coding theories appearing first in the 1990s (e.g., Mumford, 1992; Rao & Ballard, 1999). Modern versions of predictive coding assume the existence of a hierarchy of generative models with increasing levels of abstraction (see e.g., the free energy principle of Friston, 2005, 2010). At each level of the hierarchy, predictions from a generative model are compared with the input and the difference is treated as an error signal. The system aims at suppressing (minimizing) the error by adjusting models, with higher levels governing model selection at lower levels.

Effects of stimulus predictability have been shown on auditory scene analysis (e.g., Andreou, Kashino, & Chait, 2011; Bendixen, Denham, Gyimesi, & Winkler, 2010; Rimmele, Schröger, & Bendixen, 2012; initially suggested by Jones, 1976; for a review, see Bendixen, 2014). Regular (predictable) tone patterns embedded separately within two interleaved sequences increased the probability of hearing two concurrent sound streams as opposed to a single streams (Bendixen, Denham, et al., 2010; Bendixen et al., 2013; Szalárdy et al., 2014), while predictable patterns connecting tones across the two interleaved sequences that did not at the same time produce such patterns separately for the two sequences increased the probability of perceiving a single stream over two concurrent ones (Bendixen, Denham, & Winkler, 2014). Further, a predictable pattern (a tune) embedded in one of two interleaved sound sequences made it easier for listeners to follow the other sound sequence (Andreou et al., 2011; Rimmele et al., 2012). Predictive processes probably also play a crucial role in auditory deviance detection (e.g., Bendixen, Schröger, Ritter, & Winkler, 2012; Lieder, Stephan, Daunizeau, Garrido, & Friston, 2013; Paavilainen, Arajärvi, & Takegata, 2007; initially suggested by Winkler, Karmos, & Näätänen, 1996; for a review, see Bendixen, SanMiguel, & Schröger, 2012). Winkler, Karmos, et al. (1996; see also Winkler, 2007) have suggested that deviance is established by comparing incoming sounds against those predicted by the representations of previously detected regularities. For example, when a tone sequence followed the rule "long tones are followed by high ones, whereas short tones by low ones", rare low tones following long ones and high tones following short ones elicited the MMN response signaling that the rule violation was detected (Paavilainen et al., 2007; see also Bendixen, Prinz, Horváth, Trujillo-Barreto, & Schröger, 2008). In this sequence, deviant tones did not contain any rare feature of feature combination, *per se*. Only because the previous tone predicted a different tone to arrive next in the sequence made these tones to violate the acoustic regularity of the sequence, and therefore to be processed as deviants. Bendixen, Schröger, and Winkler (2009) have also found that differences between ERPs elicited by the occasional omission of a predictable vs. an unpredictable tone. These and other evidence reviewed by Bendixen, SanMiguel, et al. (2012) strongly support the notion of the involvement of predictive processes in MMN generation.

Our theoretical framework linking auditory scene analysis and deviance detection is compatible with the general idea of predictive coding. We will argue that regularities detected from the relationship between successive sounds are encoded into generative models of the acoustic environment. Predictions from these models help to construct auditory sensory memory representations and they are compared to the currently dominant interpretation of the auditory input. The outcome of the comparison is used to update the model.

Research on speech processing usually focuses on how the brain decodes spoken messages. The input of most of these models is a stream of speech. That is, they assume that the communication channel is already established. Here we provide a conceptual framework for how the auditory system sets the stage for this. Since using predictions to reduce the amount of computation required to decode messages have also been suggested for language processing (Federmeier, 2007; Hosemann, Herrmann, Steinbach, Bornkessel-Schlesewsky, & Schlesewsky, 2013; van Petten & Luka, 2012), the model proposed here fits seamlessly with such models, specifying some lower levels of the hierarchy.

2. The building bricks: Regularity, deviance, predictive information processing

Deviance can only be defined in relation to something regular. An event is deviant if it does not fit at least one of the relationships connecting the previous events within the environment. That is, a deviant event violates some existing regularity of the context within which it appears. By regularity we mean an implicit sequential rule, which is extracted from the series of sound events by the auditory system. Later, we will specify the types of regularities involved in auditory deviance detection (e.g., concrete and statistical regularities), how they are utilized, and how such regularities are extracted from a sequence of sound. In the auditory modality, deviations range from simple cases, such as breaking the repetition of a discrete sound, to complex ones, such as violating a harmonic or rhythmic rule in music. From the above definition follows that within a sequence of sounds with no regular relationships no sound event can be deviant. Another consequence is that deviance is not equal to physical (acoustic) change. Let us consider a spoken sentence with monotonously falling pitch (such as is typical in statements spoken in Hungarian). Although the pitch of each word is different from the previous one, because it fits the regularity, it is not a pitch deviant. On the other hand, while a word having the same pitch as the previous one represents no pitch change it deviates from

[1] We do not speculate about the neural implementation of this framework or about the neural substrate of the processes being described as part of this framework. We do, however refer to neural markers of these processes, the generators of which have (to some extent) been localized (see respective references). These locations may serve as starting points to determine the neural network underlying the process proposed in our model.

the pitch regularity of the preceding ones (i.e., it is a pitch deviant event).[2]

When describing perception, the above definition of deviance should be further specified in order to take into account the capabilities of the perceiver, that is, the system that would detect deviance. Because the presence of regularity is a prerequisite of deviance detection, the system can only detect deviants that break some regularity that the system "knows about". One can only detect a rhythmic violation in a poem if one remembers the poem or if the word violates some general rhythmic convention the person has experience with. This means that for detecting deviance, the system (in our case the human auditory system) must have access to some representation of the regular relationships applicable to the current environment as well as mechanisms which allow it to determine whether or not a given sound matches these regularity representations.

One should also consider the environment. For the perceiver, the environment is not equal to the physical effects reaching the senses. Our experience (stored representations) of the environment co-determines what we detect as deviant. In our previous example, someone regularly listening to poems in English would detect violations of meter even in English poems he/she never heard before. Thus when we refer to the environment, we mean the combination of two things: the environment and the listeners pre-existing representations of this environment, the context.

What is common between an acoustic regularity established by the recently encountered sounds (such as a sequence of two alternating tones) and rhythmic conventions in poems? They both allow one to predict which sounds are likely to follow the ones just heard. We shall argue that in the human auditory system, regularity representations are used to generate predictions for future events and incoming sounds are checked against these predictions. Consider the situation of crossing a street: We are not only interested where cars are at the moment, but, rather, where they will be when we reach their lane. Recent accounts of perception (Bar, 2007; Enns & Lleras, 2008; Ghahramani & Wolpert, 1997; Gregory, 1980; Schubotz, 2007; Summerfield & Egner, 2009; Winkler et al., 2009) as well as computational models of sensory processes (Friston, 2005; Friston & Kiebel, 2009; Tishby & Polani, 2011) emphasize that information processing is directed towards the future. In the same vein, we term the set of representations of the known regularities of a given environment the (predictive/-generative) model of this environment and we suggest that the auditory system maintains such a predictive model of the acoustic environment.

Why is it advantageous to establish such a model? Living organisms require information for reaching their goals and to successfully adapt their behavior to the environment. Deviance is of special importance as it represents new information that may require some response from the organism. In fact, deviance in the above defined sense is equal to new information for the organism, because sounds conforming to previously detected regularities could be predicted by the organism. Having a model of the environment allows the organism to predict a part of the input and thus prepare to take appropriate action. The larger the part of the input, that the sensory systems can predict, the fewer the information that requires detailed evaluation. As a consequence, fewer resources are needed for processing the actual sensory input. In other words, it is advantageous for the organism to invest into building a good model as the model will permit it to successfully adapt to the environment while conserving resources. The predicted part of the input does not require further processing, unless the event is actively monitored (e.g., one wishes to synchronize an action with an expected event). No information is lost by this type of filtering. At the same time, by identifying deviance, new, possibly important information gets a better chance to receive detailed processing.

Such generative models may perhaps be even more important in the auditory modality than in vision, because the acoustic environment is ephemeral; it lacks elements which can be revisited at will. An important characteristic of sound is that it unfolds in terms of temporally varying signals. Even the most elementary acoustic features, such as pitch or the direction of the sound source require the processing of sound segments of some duration. Moreover, any meaningful analysis of the acoustic environment involves connecting discontinuous segments of the incoming sound flow. How else could we understand prosodic information or tell whether or not the footsteps we hear signal that someone approaches us. Therefore, in order to establish perceptual events, the system must take into account the temporal behavior of the input. The generative model of the environment to be described here serves this purpose. We shall outline the various processes involved in establishing, maintaining, and utilizing a predictive model of the acoustic environment. The information within the model serves multiple purposes, deviance detection being only one of them. We shall argue that the model provides the basis of organizing the acoustic input into perceptual units (objects), which represent the concurrently active independent sources in the environment. That is, the model to be described here is an essential element of auditory scene analysis (Bregman, 1990).

3. An overview of detecting new information in the auditory modality

We regard the set of processes and memory resources involved in detecting new auditory information a system with functional module-like properties.[3] The input to this functional module is sensory data analyzed for basic auditory features. On its output, it delivers a sensory event representation, which, in addition to describing the sensory features of the incoming sound, also specifies the relation of this sound to the auditory context including an evaluation of how well it conforms to the regularities detected from the preceding sounds. Therefore, we term this functional module the Auditory Event Representation System (AERS).[4] AERS is module-like in the sense that it can properly function based on the auditory input alone (i.e., without voluntary effort or focused attention). For this reason, deviance detection has often been described as pre-attentive in the literature. However, the notion of pre-attentiveness assumes a strict serial order between stimulus-driven and attentive processing, which most likely does not hold for auditory deviance detection (see, e.g., Haroush, Hochstein, & Deouell, 2010; Sussman, Winkler, Huotilainen, Ritter, & Näätänen, 2002). These studies have shown that top-down effects, especially those biasing how the auditory input is structured into streams and patterns affect which regularities are extracted and, as a consequence, which sounds are detected as deviants (for a review, see Sussman, 2007). Thus deviance detection is not pre-attentive. However, many deviations are detected even when attention is not focused on the sound sequence (for a review, see Näätänen, 1990). Moreover, even when the auditory input is generally unattended, AERS relies on a memory store, which

[2] Note that although syntactic rule violations and semantic mismatches also fit the above definition, they will not be discussed here as these deviations are not of auditory nature and they are processed at higher levels of the hierarchy; for models of speech perception explaining syntactic and semantic violation phenomena, see e.g., Friederici, 2002; Hagoort, 2008.

[3] Module-like properties refer to the set of functions, as opposed to the neural substrate. In fact, it is quite likely that these functions are implemented by distributed neural networks, which may include several distinct loci in the brain.

[4] Some essentials of AERS and the extraction of computational principles can be found in Schröger et al. (2014).

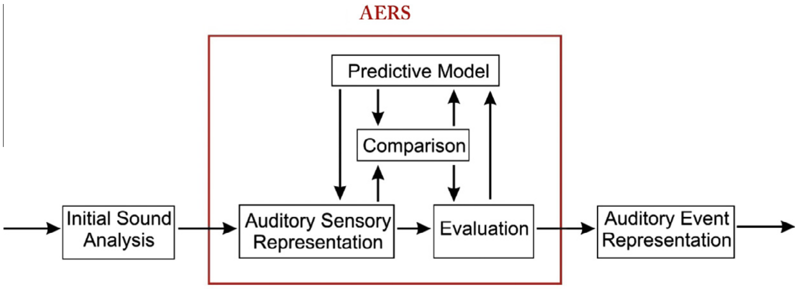

Fig. 1. An overview of the auditory event representation system (AERS). The primary input to AERS is the incoming sound with its basic features established. The main components of the system include a Predictive Model of the auditory environment storing representations of regularities extracted from the preceding sounds. This model generates predictions for upcoming sounds, thus helping to establish Auditory Sensory Representations of the incoming sounds. The resulting representation is compared with the predictions. The outcome of the Comparison is used to (1) update the model and (2) evaluated together with information regarding the current goals of the organism. The result is an Auditory Event Representation (the main output of AERS), which can enter various mental operations and be consciously perceived. Also, the model is influenced by the Evaluation process, which can initiate the building of new or reactivate old but inactive regularity representations.

interacts with other forms of memory, including long-term memory representations (for a review, see, Näätänen, Tervaniemi, Sussman, Paavilainen, & Winkler, 2001). These studies have shown that information learned on-line (such as a difficult discrimination or the structure of short trains; see e.g., Näätänen et al., 2001; Schröger, Tervaniemi, & Huotilainen, 2004; Winkler & Cowan, 2005) or previously (such as representations of the phonemes of a language spoken by the listener; see e.g., Näätänen et al., 1997; Winkler et al., 1999), and even automatized processing strategies (such as those learned by musicians; see e.g., Brattico, Winkler, Näätänen, Paavilainen, & Tervaniemi, 2002; van Zuijen, Sussman, Winkler, Näätänen, & Tervaniemi, 2004, 2005) modulates the processing of sounds. Although the memory representations and processes of AERS are not necessarily consciously experienced (i.e., they are of implicit nature) as there are deviations registered in the brain, which do not appear in conscious perception (see Paavilainen et al., 2007; Sussman, Winkler, Kreuzer, et al., 2002; van Zuijen, Simoens, Paavilainen, Näätänen, & Tervaniemi, 2006), this is a weaker form of modularity compared with that defined by Fodor (1983), because of the possibility of outside access and modification of some internal processes.

Fig. 1 shows an overview of AERS. For the sake of simplicity, let us first consider the detection of deviance within a single coherent sound sequence (a single auditory stream typically delivered by a single sound source – for a discussion of the relationship between auditory streams and sound sources, see Bregman, 1990). Later we shall consider the case of multiple sound streams (complex auditory scenes; see Section 5). Let the sequence consist mostly of sounds that conform to some regularity: For example, a sequence of sounds with a common timbre, such as would be produced by a person speaking in a neutral voice. Sounds meeting the regularity are termed “standard” sounds, whereas sounds violating the regularity are termed “deviants” (e.g., a high-pitched sound of surprise in the above example). With such a simple acoustic regularity in mind, we now describe the four major constituents of AERS.

3.1. Forming auditory sensory memory representations

Some characteristics of the incoming sound (proto-features: such as, the possible periodicity of the signal, spectral energy maxima, binaural differences, etc., such as the speech landmarks; Stevens, 2002)[5] are extracted early within the afferent pathways of the auditory system. The traditional view suggests that these features are then bound together to form unitary sound representations[6] (the “feature binding problem”: Bertrand & Tallon-Baudry, 2000; Treisman, 1993, 1998; Treisman & Gelade, 1980; Zhuo & Yu, 2011). This feed-forward account is overly simplistic because even discrete sounds can be quite complex and, when two or more sounds overlap each other correctly establishing the features requires that they are first separated from each other. However, unfortunately, the literature currently provides little evidence regarding the details of the interaction between feature extraction and sound separation beyond establishing the ubiquitous presence of descending pathways throughout the auditory system (Schofield, 2010). Therefore, by necessity, we only note the probable existence of such interactions and focus on how auditory features are utilized in deviance detection and sound grouping.

Thus auditory features are bound together to form unitary auditory sensory memory representations (Fig. 1). There is good reason to suggest that the formation of these representations can be affected by the context. An example is the phonemic restoration effect, when the phoneme correctly completing a spoken word is heard even when the actual phoneme sound is omitted and the gap is filled with a sound spanning a broad frequency band (Samuel, 1981; Shinn-Cunningham & Wang, 2008). In the phonemic restoration effect, we hear a sound that is not present as such in the acoustic input. Thus, this percept must be derived from memory, as somebody unfamiliar with the given language will not correctly restore the phoneme. The memory representation does not have to be previously learned as similar restoration effects can be observed for sounds with a predictable pitch contour (see, e.g., the continuity illusion, Riecke, van Opstal, & Formisano, 2008). This suggests that establishing unitary auditory sensory memory representations rests on the construction of predictions from memories that persist over different time scales.

There is no consensus on whether or not feature binding requires focused attention (see Treisman, 1993, 1998; Treisman & Gelade, 1980; vs. Duncan & Humphreys, 1989). Based on the results of deviance detection studies (see, Näätänen, 1992; Näätänen & Winkler, 1999; Sussman, 2007), we assume that such representations are formed even outside the focus of attention. On the other hand, we will note throughout the description of AERS how its output, termed Auditory Event Representation (Fig. 1) can be affected by attention.

[5] One set of candidates which could serve as the input to AERS are spectrotemporal response patterns observed at subcortical and cortical levels (e.g., Schönwiesner & Zatorre, 2009; Versnel, Zwiers, & van Opstal, 2009). For speech, phonetic-level features for vocal tract control are also viable candidates as an input to AERS (Chang, 2014; Leonard & Chang, 2014).

[6] The notion of feature binding and unitary sensory memory representations assumes that there is a period of time during which auditory features are established and that features established earlier are stored in temporary feature traces until the rest of the features become available (c.f. Cowan, 1984; Näätänen & Winkler, 1999).

3.2. Predictive model

The model (“Predictive Model” box in Fig. 1) stores representations of the currently applicable auditory regularities (e.g., in our above example, sounds having a common timbre). The regularity representations produce predictions for upcoming classes of sound (e.g., the next sound should have the same timbre) (Baldeweg, 2006, 2007; Bendixen et al., 2009; Grimm & Schröger, 2007; Winkler, 2007; Winkler, Karmos, et al., 1996; Winkler et al., 2009; and also specifically to speech sounds, see Bendixen, Scharinger, Strauss, & Obleser, 2014). These predictions (a) guide the formation of auditory sensory memory representation of the incoming sound and (b) are compared with the emerging sensory memory representation of the sound (see below).

Existing regularity representations are updated when the incoming sound mismatches the predictions of the model (Schröger, 2007; Winkler, 2007; Winkler, Karmos, et al., 1996) whereas new regularities are extracted once the predictable part of the auditory input is accounted for. Modulating effects on the model reflect structural information encoded in longer-term stores such as long term experience with certain types of sounds as well as explicit knowledge about the current sound sequence. The former has been shown by studies comparing “experts” and naive listeners: players of a given instrument detect smaller pitch deviations for their own instrument compared to listeners playing other instruments (Tervaniemi, Just, Koelsch, Widmann, & Schröger, 2005) and speakers of a language are superior in detecting phoneme category (Näätänen et al., 1997; Winkler et al., 1999) and word changes (Jacobsen et al., 2004; Pulvermüller et al., 2001) relevant in the given language than listeners, who don’t speak that language. The effect of explicit knowledge on deviance detection was demonstrated in an experiment in which listeners did not notice that the sound sequence consisted of a cyclically repeating pattern until they were informed about it; their brain response to rare sounds within the pattern reflected that they structured the sequence differently afterwards (e.g., Sussman, Winkler, Houtilainen, et al., 2002).

3.3. Comparing model predictions with the sensory representation of the current sound

Depicted at the center of the overview of AERS is the comparison between the sensory representation of the incoming sound and the predictions derived from the model of the acoustic environment (Fig. 1). That is, we assume an explicit comparison function (cf. comparator-based deviance-detection; Opitz, Schröger, & von Cramon, 2005; Siddle, 1991), although, previously implicit solutions to the comparison function have also been suggested (Näätänen, 1984). Unlike in laboratory settings, under everyday circumstances, no sound can be fully predicted. Acoustic variability is introduced by (a) variations in the sound source, (b) changes in the relative position of perceiver and the source (i.e., as they move with respect to each other) as well as by (c) various concurrent changes in the physical environment (e.g., movement of sound-reflecting objects). Therefore, (a) the models stores the distributions of feature values and its predictions are adapted to the experienced variability of the preceding sound sequence by referring to ranges of the feature space and classes of sounds, rather than to a specific sound (see, e.g., Winkler et al., 1990, who found that intensity and pitch deviations were detected despite variations in the intensity of the regular sounds); and (b) the comparison output signal (MMN) may reflect the amount of deviation (Näätänen & Alho, 1997; Schröger & Winkler, 1995; Tiitinen, May, Reinikainen, & Näätänen, 1994; see, however Horváth et al., 2008), as the MMN amplitude increases and the peak latency decreases with increasing amounts of deviance (however, due to overlap between MMN and other ERP responses, it is possible that a large part of this effect originates from a different source).

The outcome of the comparison, which describes the relation between the sensory representation of the incoming sound and the regularities stored by the model, is passed on to the evaluation process (see below). If predictions from the model failed, the model needs to be corrected. Thus the outcome of the comparison feeds back to the model via an updating process. The updating process is reflected by the MMN (Winkler, 2007; Winkler & Czigler, 1998; Winkler et al., 2009). This does not, however, rule out that the MMN signal can also serve as an indicator that new information has been encountered by the perceiver (Escera, Alho, Schröger, & Winkler, 2000; Näätänen, 1990; Schröger, 1997). Thus, the MMN can be seen as one (though not the only) indicator of new information, a brain index of prediction error, which drives the updating of the model.

3.4. Evaluation

This is the point, at which the incoming auditory information can be fully assessed and prepared for possible further processing (outside AERS) for attention control (e.g., orienting), determining the real-life event that gave rise to this sound, and assessing its relevance for the current or some pre-planned actions of the organism. Evaluation takes into account the context set up by the current goal-directed processes (top-down effects). Consider the case of timbre constancy as a detected regularity. Usually, small variation in any stimulus parameter is tolerated. If, however, someone were looking for signs of emotional stress in another person’s voice, he/she would very likely notice even very small timbre deviations. Another possible reason for marking a sound for further processing is when it or its relation to the preceding auditory context meets some preset pattern (Formby, 1967; Roye, Jacobsen, & Schröger, 2007) such as hearing one’s own name (Perrin, Garcia-Larrea, Mauguiere, & Bastuji, 1999). On the other hand, even relatively large sound deviations could go unnoticed, when one’s attention is strongly focused somewhere else. Thus the evaluation of a sound takes into account those aspects of the context which are outside the auditory environment. The resulting information package is the primary output of AERS. We term this an auditory event representation, because it describes the sound together with its relation to both the auditory and the general context.

The other main function of the evaluation processes is to initiate the search for new regularities. Successive deviant events may signal a change of the sound source or its behavior. The full description of the acoustic event can be used to find new regularities within the acoustic environment. Furthermore, this is the point at which the unpredicted part of the input (the residue) can be assessed. The residue may reflect the emergence of a new sound source in the environment. For example, typically, no predictions exist for the emergence of a new voice (a new instrument or person) within the auditory input. Such sounds remain unaccounted by the active generative models, forming the residue, after all predictions are checked. Note that although both deviations from the predictions of existing regularity representations and the residue are de facto prediction errors, the former is specific to a given regularity representation, the latter is general to the whole model and they are utilized differently in AERS: updating a regularity representation vs. initiating the formation of a new one.

The P3a ERP response may reflect the assessment of the information value of the incoming sound by the evaluation process. The currently most widely accepted interpretation of P3a is that it reflects a call for further processing of highly deviant or otherwise unexpected stimulus events (e.g., a sound delivered after a long silent interval; Friedman, Cycowicz, & Gaeta, 2001; Polich, 2007). However, this interpretation has been challenged (Horváth, Winkler, & Bendixen, 2008; Rinne, Sarkka, Degerman, Schröger, & Alho, 2006; Wetzel, Schröger, & Widmann, 2013), because some features of P3a suggest that it reflects the assessment of the "significance" of sensory events, combining the information carried by the stimulus with its relevance within a wider context. The latter interpretation is compatible with the assumed function of "evaluation" within AERS.

4. Initial build-up of an auditory model

It has to be asked how a representation of an auditory regularity is established, such as when the sound sources in the environment have been inactive for some time and, therefore, previously detected regularities may not be available, or when a new regularity comes into play. We will illustrate this for simple regularities, before we consider more complex ones.

4.1. Simple regularities

When a sound arriving after a longer silent period repeats a few times, each sound reaching the ears receives initial analysis in the afferent auditory pathway (Carney, 2002). Units of the input are typically separated by abrupt spectrotemporal changes (onsets and offsets), which may be indexed by the elicitation of the onset- and offset-related ERPs, such as the N1 response (Näätänen & Picton, 1987). Markers for the start of sound-units serve as temporal reference points, allowing the auditory system to compare sound patterns with the corresponding representations along the temporal axis, such as distinguishing the A-B tone pair from B-A tone-pair. Evidence shows that the later parts (>350 ms) of long sounds only affect the building of regularity representations when the sound includes abrupt spectrotemporal changes (Schwartze, Tavano, Schröger, & Kotz, 2012; Weise, Grimm, Müller, & Schröger, 2010). This suggests that representations are based on the initial segment of long continuous sounds. However, abrupt spectrotemporal changes within a long sound (such as most consonants) initiate the formation of a new unit, thus enabling a segmented, but precise description of the full sound. Our notion of the basic unit is compatible with that of the literature of auditory sensory memory (see Cowan, 1987; Demany & Semal, 2008) as well as with the notion of parallel analysis of sound on multiple times scales within the human auditory cortex (Nelken, Fishbach, Las, Ulanovsky, & Farkas, 2003; Poeppel, 2003).

Because we assumed that no competing regularity representations preexist, the established sound features are conjoined unless they contain some spectral or temporal cue indicating the presence of multiple concurrent sound sources, such as a mistuned harmonic within a complex tone or asynchronous onset of different spectral components (for reviews of the cues supporting the instantaneous segregation of concurrent sounds, see Alain, 2007; Carlyon, 2004; Ciocca, 2008; de Cheveigné, 2001). Note, that we postpone discussing multi-source sound configuration till later. Here we assume that a unitary representation of a single sound is formed through projecting the different auditory features onto temporal coordinates, thus constructing a description encoding both static and dynamic aspects of the discrete sound (Näätänen & Winkler, 1999). The resulting representation may then be consciously perceived.

Because we consider the (rare) case that no regularity representations were available when the sound was encountered, no predictions could have been formed for this sound. Thus the whole auditory input becomes "residue" within AERS and thus the formation of a new regularity representation is triggered. Because sounds delivered after a long silent period elicit very large responses in auditory cortex (as characterized by the P1 and N1 ERPs), some part of these auditory cortical responses may be related to the process initiating the formation of a new regularity representation (Winkler et al., 2009). One possibility is that the search for a new regularity is initiated when the strength of these responses exceeds some threshold (for a similar idea referring to memory traces, see Näätänen, 1984). However, a single sound is not sufficient for establishing a regularity representation. This has been demonstrated by the lack of any deviance-related ERP response when the second sound of a train differed from the first one while the train was preceded by a long silent period and no compatible regularity had been established in the preceding train (Cowan, Winkler, Teder, & Näätänen, 1993; Winkler et al., 2002).

When the same sound occurs for the second time, it receives the same initial processing than the first sound. However, because the previous sound initiated the formation of a new regularity representation, the relation between this sound and the representation of the first sound is also determined. Their temporal relationship as well as their relationship along the established features is encoded into an episodic representation connecting the two events. Detecting the repetition of a sound could already give rise to the prediction that the next sound will also be the same, as was shown by the elicitation of the MMN response for the 3rd sound of a train that differed from the two previous (identical) sounds (Bendixen, Roeber, & Schröger, 2007). However, in other studies, two repetitions were required before a deviant could trigger a deviance-detection response (Cowan et al., 1993; Winkler, Karmos, et al., 1996). A possible explanation is that when participants attend to the sounds (as in Bendixen et al.'s study), a single repetition can give rise to a regularity representation. This account is compatible with Bayesian inference rules in model selection (Kersten et al., 2004; Knill & Pouget, 2004; Yuille & Kersten, 2006) with priors determined by higher-level models, which may be directly related to behavioral goals (and thus would be regarded as voluntary or attentive in the terminology of cognitive psychology).

For unattended sequences, in which MMN was only elicited by a sound after two repetitions of a different sound, the relationship between the 3rd and the 2nd sound is compared with that between the 2nd and the 1st sound. Note that relationships between successive sounds are compared as opposed to representations of individual sounds (Winkler, 2007). When the two relationships are found to be matching, a regularity representation connecting successive sounds is formed. This representation can now predict future events. Enabling the ability of a regularity representation to predict upcoming sounds can also be regarded as an activation process. Taking this notion, we can distinguish "active" and "dormant" regularity representations. Thus an established regularity representation can be dormant (i.e., does not affect the processing of sounds). It can then be activated either by an additional confirming sound event or by attention directed to the sounds (see above). Thus AERS can learn and react fast to emerging patterns in the environment without forsaking prudence (i.e., wasting processing capacity on chance patterns).

Although auditory regularities are detected even when participants focus their attention on a different modality (for reviews, see Haroush et al., 2010; Sussman, 2007), it is less clear whether strong focusing within the auditory modality can prevent or at least modulate the formation of regularity representations. Modulating effects on deviance detection have been observed by Haroush et al. (2010), whereas Sussman, Winkler, and Wang (2003) showed that the ERP response for deviance in a given feature is suppressed in an unattended sound sequence when participants detect deviants in the same feature in a separate but attended sound stream (for compatible evidence, see Näätänen, Paavilainen, Tiitinen, Jiang, & Alho, 1993; Woldorff, Hackley, & Hillyard, 1991; Woldorff, Hillyard, Gallen, Hampson, & Bloom, 1998). Either way, attention focused strongly on one stream of sound does not in general prevent the detection of deviations in an unattended stream.

When the model contains at least one (active) regularity representation, predictions for incoming sounds are produced. In our simple case, the model predicts that the next sound is probably identical to the three previous ones. These predictions have an immediate impact on the processing of the input. The sensory memory representations of further sounds are compared with the representation of the sound predicted by the model. In case of a match the regularity representation may be further strengthened. One possible interpretation is that stimuli conforming to the predictions may increase the weight the system attaches to the predictions from the given regularity (i.e., the "confidence" of the system regarding the given prediction). This notion is similar to Friston's predictive coding theory according to which predictions do not only send down contents to the lower level, but also their inferred precision (Feldman & Friston, 2010). Indeed, change-related neural activity has been reported to increase with increasing number of repetitions preceding a change (Javitt, Grochowski, Shelley, & Ritter, 1998). Some studies found a similar effect for the ERP amplitude difference between deviant and standard responses (Bendixen et al., 2007; Haenschel, Vernon, Dwivedi, Gruzelier, & Baldeweg, 2005), although others failed to observe a significant effect (Cowan et al., 1993). On the other hand, Winkler, Karmos, et al.'s (1996) results suggest that stimuli confirming the predictions make the related regularity representations more resistant to elimination. These authors delivered to participants short trains starting with six presentations of tone 'A' followed by 0, 2, 4, or 6 presentations of tone 'B'. The train ended with tone 'C', which differed from both "A' and 'B". Tone 'C' elicited an MMN with respect to tone 'A' even after 4 intervening presentations of tone 'B', showing that the repetition regularity of tone 'A' was not eliminated by repeated presentations of tone 'B'. Although there is no unequivocal proof for the existence of these processes, it stands to reason that regularities, whose predictions are often confirmed, have increased utility for AERS. Haenschel et al. (2005) found an ERP response elicited by regular sounds which increased together with the number of preceding regular sounds (repetition positivity, RP; see also Baldeweg, 2006). The neural process generating RP may be involved in strengthening (sharpening) or making more resistant the corresponding regularity representation(s).

In case a difference is detected between the incoming sound and the prediction from the regularity representation (a prediction error), the stimulus is marked as containing new information. This increases the chance that the stimulus representation receives more detailed processing. The representation of the violated regularity (which predicted the reoccurrence of the same sound in our simple example) is then updated. There is evidence showing that the primary function of the process reflected by the MMN component is related to the regularity representation, as opposed to the deviant sound itself (Winkler & Czigler, 1998). Winkler and Czigler (1998) found that a deviant sound violating two separate regularities within 200 ms elicited two successive MMN responses. In contrast, a deviant sound violating the same regularity two times within 200 ms elicited only a single MMN response. This pattern of results suggests that MMN is primarily related to the regularity violated as opposed to the sound that violates it. We hypothesize that the updating process makes the affected regularity representation (1) carry less weight (confidence) and (2) less resistant to elimination in the future. Several studies (e.g., Winkler, Cowan, Csépe, Czigler, & Näätänen, 1996; Winkler, Karmos, et al., 1996) showed that a single deviant does not prevent a consecutive deviant from eliciting MMN. This means that regularity representations are never eliminated by a single non-conforming auditory event. On the other hand, several deviants in a row or a long silent interval following the last regular sound prevent further deviants from eliciting the deviance-related MMN response. The longest silent interval after which a deviant elicited the MMN was found to be ca. 10–12 s (Sams, Hari, Rif, & Knuutila, 1993; Winkler, Schröger, & Cowan, 2001), thus placing an upper bound on the temporal extent of predictions. However, even in these cases, a single regular sound (termed "reminder) can "reactivate" the regularity representation. That is, a deviant sound following a single regular sound (termed the reminder) will again elicit the MMN (for review of the reactivation phenomenon, see Winkler & Cowan, 2005). Reactivation was observed with the reminder separated from the previous regular sound by 30 s (Winkler et al., 2002). Furthermore, reactivation also occurred when the reminder followed six consecutive different deviant sounds (Winkler, Cowan, et al., 1996). Thus it is yet unknown when or how auditory regularity representations are truly eliminated. They only become dormant (not affecting the processing of sounds) by long silent periods or by repeated failure to correctly predict the incoming sound. A similar "dormant" state can be assumed for regularity representations under construction with the third presentation of the standard activating the regularity representation (i.e., making it produce predictions in the future).

Note that once a regularity representation has been established, it starts producing predictions for upcoming sounds in the sequence – that is, it acts as a (possibly partial) generative model of a putative perceptual object (for a description of how some of these "proto-objects" can emerge as perceptual objects that can be consciously experienced, see Section 6.3). However, the build-up of a regularity representation, as described above, is not a predictive process itself. Predictions are about (possible) objects. Therefore, no prediction can be made before a possible object is detected.

It is easy to see, how the system benefits from this mode of operation. Immediate elimination of the regularity representations would be disadvantageous as random fluctuations, which often occur in everyday acoustic environments, and discrete deviant events (exceptions) would reset the system (returning it to the initial non-predicting state), even when the majority of the stimuli follow the detected rule. The existence of a dormant state for regularities further improves the chances of rapidly finding adequate regularity representations for incoming sounds. Together, these features make AERS a robust system in terms of maximizing its predictive capabilities under natural, noisy and variable circumstances.

4.2. More complex regularities

So far, we focused on how a regularity representation is formed for a sequence of a repeating sound. However, there is evidence that no sound repetition is needed for establishing a regularity representation. For example, if some feature or features are constant within a sequence while other features randomly vary, deviations from the common feature(s) elicit the MMN (Gomes, Ritter, & Vaughan, 1995; Huotilainen et al., 1993; Winkler et al., 1990). From this, we infer that regularity representations have been constructed for the common (invariant) features. Furthermore, when a regularity representation based on feature repetition becomes dormant, it can be reactivated similarly to that described for fully repeating sounds (Ritter, Sussman, Molholm, & Foxe, 2002). This suggests that the relationships between successive sounds are established for each stimulus feature (Nousak, Deacon, Ritter, & Vaughan, 1996; Ritter, Deacon, Gomes, Javitt, & Vaughan, 1995).

Identity is a special case of the possible inter-sound relationships. Therefore, one should expect that regularity representations are also established for regular non-repetitive inter-sound relationships. Indeed, this is the case. For example, the regularity of successive sounds continuously increasing or decreasing in pitch is detected similarly to sound repetition (Tervaniemi, Maury, & Näätänen, 1994). Our description of the processes involved in establishing a new regularity representation takes into account this and similar non-repetitive inter-sound relationships. When any detectable relationship between successive sounds repeats (such as, pitch increases from sound 1 to 2 and from sound 2 to 3, etc.) the corresponding regularity representation is activated. Viewed this way, the rising-pitch regularity is no more complex than feature repetition. Studies showing that sounds violating different feature-regularities of the same sound sequence elicit somewhat different MMN responses (e.g., Deacon, Nousak, Pilotti, Ritter, & Yang, 1998; Giard et al., 1995) suggest that several regularities are maintained in parallel in AERS.

In the above discussed regularities, inter-sound relationships repeated immediately (i.e., the length of the repeating cycle was one: the inter-sound relationship was always the same). However, everyday sound sequences often include repeating cycles consisting of several sounds with a characteristic pattern of inter-sound relationships (such as bird trills). Indeed, for example, exchanging two segments within a repetitive cycle of five tones elicits the MMN response (Winkler & Schröger, 1995). Furthermore, Sussman et al. (Sussman, Ritter, & Vaughan, 1998a; Sussman, Winkler, Houtilainen, et al., 2002) demonstrated that when a sound sequence is perceived in terms of a repeating pattern, the regularity representations underlying deviance detection are also based on the same pattern. These authors presented repeating cycles consisting of five tones, the first four of which was identical and the last different from them (AAAABAAAAB...). When listeners perceived the repeating cycle, no MMN was elicited by the fifth sound, even though MMN was elicited by this sound when the same sounds were presented in a randomized order or when the listener did not detect the repeating cycle. These results suggest that when the repeating cycle was not present or perceived, the listener's brain treated each sound as a separate unit and predicted the repetition of the more frequent (0.8 probability) sound. However, when the cyclic repetition was detected, the pattern of five sounds became the unit and the "rare" sound was part of this predictable standard.

In order to accommodate repeating cycles with >1 length, we need to extend our previous description of establishing a regularity representation. One possible algorithmic approach assumes that when the formation of a new regularity has been initiated the regularity building process opens a chain of inter-sound relationships (i.e., a sequence of relationships between consecutive sounds). The chain is then either completed when two full repetitions are encountered (establishing a new regularity representation) or discarded when no full repetition is reached before exceeding the capacity of the memory involved in building regularity representations. In support of this hypothesis, Sussman, Ritter, and Vaughan (1998b) found that the five-tone repeating cycle described above is processed in terms of the repeating tone pattern when the presentation rate was sufficiently fast so that two full repetitions of the pattern fit into 10 s. In contrast, regularities of the same sequence were processed in terms of individual sounds, when the presentation rate was slower (cf. Scherg, Vajsar, & Picton, 1989; Sussman, Winkler, Houtilainen, et al., 2002). Further evidence for a temporal capacity limitation in finding repeating cycles has been obtained in studies investigating cyclically repeating noise segments (Kaernbach, 2004).[7] In a recent study, Barascud, Pearce, Griffiths, Friston, and Chait (personal communication by Maria Chait, 10.07.2014) revealed that it is not necessarily needed that the full pattern is repeated in order to for the auditory system to predict its reoccurrence. In one condition studied by Chait and colleagues, the pattern consisted of 10 tones and the listener could detect the regularity after the presentation of only 14 tones as was indicated by an enhancement of the magnetic brain response elicited when the continuation of the pattern was terminated and tones of random pitch were presented instead. This result suggests that the threshold for activating a regularity representation (i.e., making it affect the processing of upcoming sounds) is not two full cycles of the pattern; rather, it is possibly a few (perhaps two or three) hits after the initial pattern has been closed. The simplest form of closure (cf. the gestalt term) comes from encountering again the first sound of the pattern whose representation is being constructed.

There are also regularities which cannot be described by a repeating chain of inter-sound relationships. The simplest example of this type is a sequence with two frequent sounds ('A' and 'B'). Such a sequence includes four frequent inter-sound relationships (A -> A, A -> B, B -> A, and B -> B), each of which is represented. Because, as we argued above, the respective representations (memory traces) are not eliminated by the emergence of other relationship, they can be reinforced by conforming evidence. After a few repetitions of the same relationship, a regularity-representation can be formed. In order to account for this types of regularities, the simplified description of building regularity representations described in the previous paragraphs needs to be extended by relaxing the constrain of regularity building requiring immediate repetition of an inter-sound relationship. Instead, we suggest that repetition must come within the life-time of such traces. Based on studies of testing the temporal limits of rhythm perception (Duke, 1989; van Norden, 1975), we tentatively suggest that the life-time of these traces is in the order of 1–2 s. Further, there is no specific evidence regarding how many times such an inter-sound relationship must be encountered for the corresponding regularity representation to be activated. Given that the repetitions are not immediate, we assume that more than two recurrence of the same inter-sound relationship is needed.

The regularity representations based on the various frequently encountered inter-sound relationships are built and become active in parallel. Each covers a certain percentage of the incoming sounds, but as long as no inter-sound relationship disappears for long period from the sequence, they will coexist and, as we will show later (see Section 6) by being compatible with each other, they can form a common sound organization. Results showing that even when two sounds have equally high global probability, either one appearing after a longer micro-sequence of the other elicits the MMN (Sams, Alho, & Näätänen, 1983; Winkler, Paavilainen, & Näätänen, 1992) are compatible with the above description.

The results of many deviance detection studies are compatible with the extended description of forming predictive regularity representations. For example, when pitch increases between the two tones in a tone pair (Ahveninen et al., 2000; Saarinen, Paavilainen, Schröger, Tervaniemi, & Näätänen, 1992) or when tones follow the rule linking different auditory features (Bendixen et al., 2008; Paavilainen et al., 2007) regularity

[7] The temporal characteristics of regularity building appear to be compatible with those obtained for auditory sensory memory (Cowan, 1984). Others, such as those underlying the orienting reflex (Sokolov, 1963) probably fall outside the scope of AERS, possibly being based on some higher hierarchical level in a predictive coding model.

representations can be formed based on a few frequently occurring inter-sound relationships. Indeed, violating these rules results in the elicitation of MMN.

Finally, recent evidence suggests that regularities can also be extracted for non-adjacent sounds, but only when the sounds intervening between successive sounds of one regularity formed a separate regularity themselves (Bendixen, Schröger, et al., 2012). It should be noted that the paradigm did not allow the two sets of sounds to be segregated by any primitive cue. This is important, because after stream segregation, the sounds forming the two regularities would have become separately adjacent to each other (for other effects related to auditory stream segregation, see Section 6). This result suggests that the auditory system not only registers the relationship between adjacent sounds, but possibly also between non-adjacent ones. However, the fact that such regularities are only detected when the intervening sounds give up a separate regularity suggests that such regularity representations are quite weak normally and require to helping each other to become active. Such help from the other regularity is possible, as these regularities are also compatible with each other (see, again Section 6).

So far, we only considered regularities in which auditory features could be predicted with high accuracy. However, this rarely is the case in real-life situations. Dealing with natural variance involves constructing categories and defining regularities by relationships between stimulus categories instead of between concrete stimuli. Such regularities can be considered as abstract rules as opposed to concrete rules, which are based on relationships between concrete sounds. The categories can be pre-existing (i.e., stored in long-term memory), such as the phoneme set of a language, or episodic, such as for example elements of a bird thrill we get adapted to when spending some time in the vicinity of some birds belonging to the same species. In terms of AERS, categories are distributions of feature values and, as was already described, predictions (a) refer to such distributions and (b) are adapted to the experienced variability of the preceding sound sequence.

Indeed, there is evidence that regularities based on categories are processed within AERS. For example, a generalized version of the pitch-alternation regularity can be constructed if every second tone is set higher in pitch than the preceding one whereas every other tone is set lower than the preceding one. Evidence that this extended regularity of pitch alternation is detected and applied by AERS was obtained by Horváth, Czigler, Sussman, and Winkler (2001; see further evidence for other category-based regularities in Paavilainen, Jaramillo, Näätänen, & Winkler, 1999; Phillips et al., 2000; Saarinen et al., 1992; Tervaniemi, Rytkönen, Schröger, Ilmoniemi, & Näätänen, 2001). There is also evidence that, similarly to concrete regularity representations, representations of abstract regularities can be reactivated (Korzyukov, Winkler, Gumenyuk, & Alho, 2003). It is thus highly likely that concrete and abstract regularity representations are one and the same within AERS. That is, the auditory system is always prepared for constructing representations for "abstract" (non-exact) regularities, treating concrete regularities as abstract ones with very small feature variance, because concrete regularities can seldom be found outside the laboratory.

In summary, it is clearly advantageous for the auditory system to establish representations of the acoustic regularities of the sounds encountered within the environment. Such representations can absorb a large part of the incoming sound, acting as filters for new information (Schröger, 1997; Sinkkonen, 1999; Winkler, Reinikainen, & Näätänen, 1993). To this end, AERS registers the inter-sound relationships as well as their sequential order. Once the same order of inter-sound relationships has been detected at least a few times (possibly with other relationships intervening), a regularity representation is formed. Thus regularity representations are formed quite fast. In contrast, elimination of these representations is slow. The latter is important, because, it is clear that in any realistic environment, each of the regularity representations of AERS will often fail to correctly predict upcoming sounds, as even the most complex regularity representations cannot capture alone the complexity of a real-life scene. However, as will be discussed in the next sections, continuously summing the predictions of a large set of such simple and fallible regularity representations can produce a robust and flexible representation system, maximizing the predictive power of the model of the acoustic environment.

5. Auditory regularity representations, auditory streams, auditory perceptual objects

In every-day auditory environments, most of the time, one encounters several concurrent and intermittent sounds originating from different sources. Due to the physical nature of sounds, the acoustic information arriving from various sources interact with each other within the air according to the mathematical laws applying to waves. If we assume that the auditory system has evolved to find out information about distal objects (the sound sources) and events that gave rise to the sounds, then one of the most important functions of the central auditory system is to break down the incoming signal according to their sources. This is not a trivial task, because there are no simple cues separating the contributions of different sound sources that would work in most situations. In fact, an analysis of natural auditory scenes suggest that there is no unique mathematical solution to finding moving sound sources by the information available to the auditory system (Stoffregen & Bardy, 2001). According to the empiricist point of view (Helmholtz, 1867), the auditory system must, therefore use heuristic computational processes, which are based on assumptions regarding the nature of the sound sources to determine the actual source configuration (see, however, the contrasting view of direct perception; Gibson, 1979). This function has been termed the "auditory scene analysis" by Bregman (1990; for recent reviews, see Ciocca, 2008; Denham & Winkler, 2014; Haykin & Chen, 2005; Shinn-Cunningham & Wang, 2008; Snyder & Alain, 2007). Many of these assumptions have been described as the laws of perception by the Gestalt school of psychology (Köhler, 1947). Some of them rely on the spectrotemporal configuration of short auditory segments (e.g., co-occurrence of harmonics of a common base and common onset for sounds produced by the same source), but the majority of these constraining assumptions is concerned with the sequential/temporal relationship between sounds coming from a single source (e.g., smooth continuation, common behavior of the sound components originating from the same source, etc.). In order to utilize these principles, the auditory system must store the recent history of the various active sound sources, representing their characteristic acoustic features as well as their dynamic behavior. We propose that the regularity representations described in the previous sections serve also this purpose and that deviance detection and auditory object formation are two tightly interwoven functions of the auditory system.

Before detailing the role of AERS in auditory object formation, here we argue that the features described for the regularity building functions in the previous sections are fully compatible with known properties of auditory streaming, the most widely studied phenomenon within auditory scene analysis (for a detailed analysis, see, Winkler, 2010; Winkler et al., 2009). The auditory streaming paradigm (van Norden, 1975) consists of a sound sequence mixing together two sets of sounds. The typical sequence takes the form of ABA-ABA-…, where 'A' and 'B' denote two tones differing in frequency and '-' represents a silent period equal to the common duration of the two tones (Fig. 2). Depending on the frequency separation between 'A' and 'B' and the presentation rate (usually characterized by the time between the onsets of consecutive sounds, the stimulus onset asynchrony [SOA]), this sequence is most likely to be experienced either in terms of repeating ABA triplets producing a galloping rhythm in perception (the 'integrated percept') or as two concurrent isochronous streams of sound, a faster paced one consisting of the 'A' and a slower one of the 'B' tones (the 'segregated percept'). Note, however, that other relatively stable percepts are also possible (Denham et al., 2014). With large separation between the two tones and/or fast sound presentation rates, segregation is perceived more commonly, whereas with small frequency separation and/or slow presentation integration is the more common percept (Bregman, 1990; van Norden, 1975). Streams can be segregated by separation in a variety of auditory features (e.g., Akeroyd, Carlyon, & Deeks, 2005; Grimault, Bacon, & Micheyl, 2002; Roberts, Glasberg, & Moore, 2002; Vliegen & Oxenham, 1999) suggesting that auditory streaming is generally based on perceptual dissimilarity (Moore & Gockel, 2002), or rather, taking also into account the effect of presentation rate, auditory streaming is based on rate of perceptual change (Mill, Bőhm, Bendixen, Winkler, & Denham, 2013; Winkler, Denham, Mill, Böhm, & Bendixen, 2012). With longer sound sequences, perception inevitably fluctuates between the possible percepts (Anstis & Saida, 1985; Bendixen, Denham, et al., 2010; Denham, Gyimesi, Stefanics, & Winkler, 2010; Denham & Winkler, 2006; Leopold & Logothetis, 1999; Pressnitzer & Hupe, 2006; Rahne & Sussman, 2009; Roberts et al., 2002; Schadwinkel & Gutschalk, 2011; Wessel, 1979). Thus auditory streaming is a multistable perceptual phenomenon (Winkler et al., 2012). Although perceptual multistability is quite rare under everyday circumstances, this phenomenon is very important for perceptual theories (e.g., Gregory, 1980), as it provides insights into the underlying mechanisms. In short, theories and models of perception need to account for bi-/multistable phenomena (Schwartz, Grimault, Hupe, Moore, & Pressnitzer, 2012). Multistability in auditory streaming suggests that alternative sound organizations are maintained in parallel. Indeed, Horváth et al. (2001) found that in an alternating sequence of two tones, representations for both the regularity of "*A* is followed by *B* and vice versa" and also for the rule of "every second tone is *A*, every other is *B*" have been maintained in parallel. This was shown by MMN being elicited by violating either one of these rules, only. On this basis, Winkler et al. (2009, 2012) suggested that the multistability observed in the auditory streaming paradigm stems from competition between alternative regularity representations describing a sound sequence.

On this hypothesis, the properties of auditory regularity representations should be compatible with the phenomena observed for auditory streaming. One can easily draw parallels between the formation of auditory regularity representations (as described in Section 4) and the temporal course of perception at the beginning of an auditory streaming sequence. On the regularity-representation hypothesis, the initial percept is decided by which regularity is discovered first. In the auditory streaming paradigm, with most, but not all combinations of the parameters, the first percept is the integrated one (Denham, Gyimesi, Stefanics, & Winkler, 2013; Hupe & Pressnitzer, 2012; Winkler et al., 2012). Once a second regularity is discovered, then competition begins. It has been shown that whereas the choice and duration of the first percept reported for an auditory streaming sequence is highly sensitive to stimulus parameters, the effects of these parameters on the probability and duration of the percepts reported after the first perceptual switch is much more modest (Deike, Heil, Böckmann-Barthel, & Brechmann, 2012; Denham et al., 2013). The competition between alternative regularity representations is likely based on adaptation and noise (Mill et al., 2013), as was also suggested for bistable visual phenomena (Shpiro, Moreno-Bote, Rubin, & Rinzel, 2009).

Several studies showed that deviance detection, as indexed by the MMN ERP component, goes hand-in-hand with the segregation of auditory streams. That is, violations of regularities specific to one or another stream only elicit MMN, when separate streams are perceived (e.g., Sussman, Ritter, & Vaughan, 1999; Winkler, van Zuijen, Sussman, Horváth, & Näätänen, 2006; Winkler et al., 1993), whereas violations of regularities specific to the whole sequence only elicit the MMN when listeners experience the integrated percept (e.g., Sussman, 2005; Winkler & Cowan, 2005; Yabe et al., 2001). For example, Winkler et al. (2006) presented tones of intermediate pitch that could join only one of the separate streams formed by the intervening high and low tones, forming different repeating temporal patterns with them. Participants were instructed (and checked on) to voluntarily hold either the high-middle or the low-middle patterns. MMN was elicited when infrequent changes in the timing of the intermediate tones violated the voluntarily held pattern, but not when they violated the possible alternative pattern. Note that deviations only occurred on the tones of intermediate pitch, which were attended all the time. Therefore this effect cannot be explained by attentional filtering. Rather, this is an effect of grouping biased by attention. Furthermore, both bottom-up (Rahne & Sussman, 2009; Winkler, Sussman, et al., 2003; Winkler, Takegata, & Sussman, 2005) and top-down (Sussman, Winkler, Houtilainen, et al., 2002; Winkler

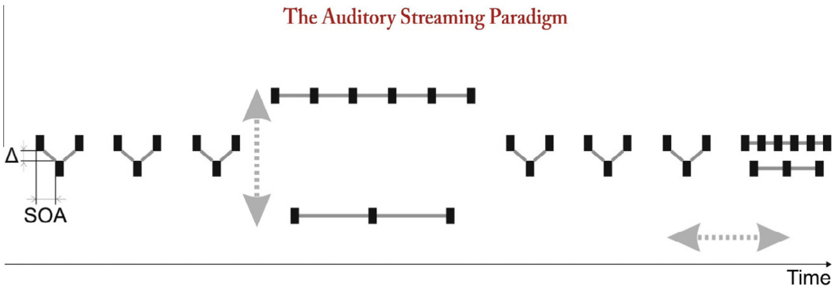

Fig. 2. Schematic diagram of the auditory streaming paradigm (van Norden, 1975). Short sounds (depicted by black rectangles) are presented in a repeating ABA-pattern (the horizontal axis marks the passing of time), where A and B denote two sounds differing in at least one stimulus feature, such as the tone frequency (marked by the vertical position). With small feature separation between the two sounds (marked by Δ on the figure) and slow-to-medium presentation rates (marked with the onset-to-onset interval, the Stimulus Onset Asynchrony on the figure), this sequence of sound is typically experienced as a single coherent stream (marked by connecting adjacent sounds with gray lines on the leftmost and the third segment of the figure). However, when feature separation is increased (second segment, the change marked by the dotted gray vertical arrow before the segment) or the presentation rate is increased (rightmost segment, the change marked by the dotted gray horizontal arrow between the third and the rightmost segments), then listeners tend to perceive the sequence as two separate sound streams, each consisting of similar sounds, only (marked by separately connecting the sounds of each stream by gray lines).

et al., 2006) biasing of the sound organization have parallel effects on deviance detection. The previous example also illustrates a top-down effect on auditory stream segregation. As for a bottom-up effect, Winkler, Sussman, et al. (2003) presented two random tones that intervened between consecutive tones of a repetitive sequence. In one condition, the pitch range of the intervening tones included the pitch of the repeating tone; in the other, the pitch range of the intervening tones was far removed from the pitch of the repeating tone. Occasional intensity deviations of the repeating tone only elicited the MMN in the latter sequences which listeners perceived as segregated into a stream of the repeating tone and a separate stream of the intervening tones. Newborn infants also showed a similar effect, suggesting that this primitive form of stream segregation is already functional at birth (Winkler, Kushnerenko, et al., 2003).

Are auditory streams the true building bricks of sound perception? Cognitive operations are thought to involve objects. Modern theoretical descriptions of auditory objects emphasize similarities between processing principles as opposed to equating features across different modalities (Griffiths & Warren, 2004; Kubovy & van Valkenburg, 2001; Winkler, 2010; Winkler et al., 2009). Winkler et al. (2009; see, also Winkler, 2010) suggested four defining criteria for (auditory) object representations. Object representations (1) bind together auditory features as well as possibly multiple temporally distinct acoustic events; (2) are separable from other (possibly concurrent) objects; (3) generalize across different instances of the same object; (4) can extrapolate to object parts of which no information reached the senses. The first three criteria are probably self-evident. The last one refers to our experience that even when the information from the distal object that reaches our senses does not cover all parts of the object, the representation formed of this object gives us a reasonable assessment of the missing information (Gregory, 1980). Taking one of Gregory's examples, one almost never sees all four legs of a table. Even so, the table we see does not miss the unseen legs. Because the acoustic signal is ephemeral (i.e., there are no still sounds, which could be revisited at will), the missing information typically awaits us in the future. Therefore, for auditory object representations, the criterion of extrapolation primarily means temporal predictions.

Do auditory streams act as sound objects? Bregman (1990) lists plenty of evidence showing that auditory streams meet the first three of the above-listed criteria for auditory object representations (for a point-by-point listing of psychophysiological evidence, see, Winkler et al., 2009). We suggested that predictive auditory regularity provide the basis for auditory streams. Therefore, we regard auditory streams as auditory object representations.[8]

6. How AERS works when the model has been set up

Under everyday circumstances, the model in AERS is almost never "empty"; rather, it contains a mixture of regularity representations that are currently under construction, ones that are active, and ones that are becoming (or already are) dormant, but can still be accessed. The following description of the functioning of AERS discusses how the hypothesized generative models of the auditory environment may be involved in operations necessary for deviance detection as well as in forming auditory streams.

6.1. Simultaneous stream segregation

Fig. 3 illustrates the sequence of processes establishing auditory sensory representations in AERS up to and including deviance detection. As was discussed in Section 3.1, the first estimation of sound features is marked by the box titled "Initial Feature Analysis". This analysis may already separate components of the input based on frequency and ear of origin, since these information are available from the moment the incoming sounds start affecting the receptor surfaces. A first assessing of the sources is denoted by the box titled "Initial Grouping by Simultaneous Cues". This operation can separate sounds by static features, such as outstanding spectral cues, onset relationship, etc. (Bregman, 1990; de Cheveigné, 2001; Micheyl & Oxenham, 2010). Establishing segregation by such cues does not require information about previous sounds (represented in the model) but can rather be performed instantaneously using only information from the current sound. For example, two concurrent sounds having distinct narrow frequency bands or different harmonic structures may be separated from each other (Bregman, 1990). An ERP component reflecting segregation by various instantaneous cues has been discovered by Alain and his colleagues (the "Object-Related Negativity", ORN; Alain, Arnott, & Picton, 2001; Alain, Schuler, & McDonald, 2002; Hautus & Johnson, 2005; Johnson, Hautus, & Clapp, 2003; McDonald & Alain, 2005).

Grouping/segregation by simultaneous cues occurs within a short time from the onset of a sound and it is one source producing candidates for perceptual sound organization. These candidate groups also provide information for sequential/temporal grouping processes (to be discussed in the next subsection). Finally, these grouping processes allow active monitoring of a given feature and thus the detection of deviance within that feature. This operation is depicted by the box titled "Feature-based Target Detection". As was shown by Näätänen and colleagues, a target feature level can be voluntarily maintained in the brain and target sounds can be detected by comparing all incoming sound to this memory trace (for reviews, see e.g., Näätänen, 1990; Näätänen, Alho, & Schröger, 2002; Näätänen et al., 2011). The comparison with a voluntarily maintained memory trace is reflected in an ERP component termed processing negativity (PN). PN is terminated when a sound is found to be different from the target feature or when the target identity is established (reflected by another ERP component, the N2b – see Näätänen, 1990; Näätänen & Gaillard, 1983).

Thus separation by instantaneous cues is done early within AERS. However, sounds separable only by temporal/sequential cues or by a combination of two or more features cannot be segregated by these stream-segregation mechanisms.

6.2. Temporal heuristic analysis

The temporal/sequential heuristic processes hypothesized by Bregman (1990) attempt to account for the contribution of the previously detected active sound sources and identify the onset of new streams from the residual signal. This mode of processing has been termed the old + new strategy. Consequences of the primacy of accounting for the continuation of previously detected streams are demonstrated by the continuity illusion (Riecke et al., 2008) and the phonemic and other auditory restoration effects (Samuel, 1981; Shinn-Cunningham & Wang, 2008). The heuristic processes summarized in the box titled "Temporal Heuristic Analysis" (Fig. 3) receive their input from initial sound analysis (as suggested by, e.g., the continuity illusion) and from the grouping processes evaluating simultaneous cues. Temporal/sequential grouping processes utilize the predictions generated by the representations of previously detected regularities depicted in the "Predictive Model" component. Temporal/sequential processing occurs probably in parallel with the initial grouping by simultaneous cues as was suggested by the results of Bendixen et al. (2009), who found that during the initial 50 ms following the onset of a fully predictable tone, electrical

[8] The notion of perceptual object representation does not include that the object is recognized.

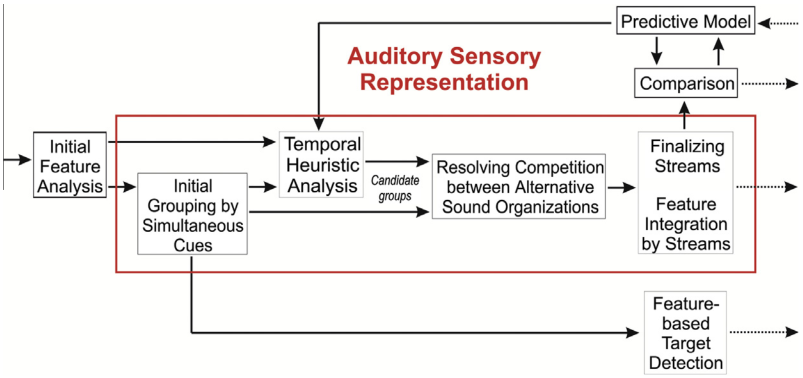
Predictive Model
Auditory Sensory Representation
Comparison
Initial Feature Analysis
Temporal Heuristic Analysis
Candidate groups
Resolving Competition between Alternative Sound Organizations
Finalizing Streams
Feature Integration by Streams
Initial Grouping by Simultaneous Cues
Feature-based Target Detection

Fig. 3. Functional model of the formation of Auditory Sensory Representations. The feature-analyzed sound input enters the processes grouping sounds by (a) Simultaneous Cues and (b) Temporal/Sequential Heuristics. The outcome of initial grouping can serve Target Detection based on a single feature. Also, the same information is passed onto the temporal/sequential grouping process. The two types of grouping processes produce *candidate groups*. Competition between these groups is Resolved, which allows for Finalizing Streams and Integrating Features. The output of Sensory Stimulus Representation is compared with predictions from the Model and Evaluation within the context (see Figs. 1 and 4).

brain potentials were similar irrespective of whether the sound was actually presented or not. In contrast, the omission of an only temporally predictable tone (i.e., one's whose pitch was not predictable) elicited an ERP response that differed from the response elicited by the actual sound. Compatible results were obtained by Friston and Kiebel's (2009) whose hierarchical dynamic Bayesian model yielded a (simulated) percept for an expected but omitted chirp that mimicked the (simulated) percept when the chirp was actually presented to the model. These results indicated the functioning of an early temporal grouping process utilizing predictions for upcoming sounds.

There is also evidence for interaction between the two types of grouping processes (Bendixen, Jones, Klump, & Winkler, 2010; Dyson & Alain, 2008a, 2008b; Dyson, Alain, & He, 2005). For example, Bendixen, Jones, et al. (2010) found that the amplitude and scalp topography of the ORN elicited by complex tones with one mistuned partial was modulated by the probability of mistuned complex tones in the sequence. Interactive processing of simultaneous and temporal cues was also demonstrated by studies using behavioral measures (e.g., Ciocca & Darwin, 1999; Darwin, Hukin, & Alkhatib, 1995; Lee & Shinn-Cunningham, 2008a, 2008b; Steiger & Bregman, 1982; Teki, Chait, Kumar, von Kriegstein, & Griffiths, 2011).

Thus the old + new strategy could be implemented by comparing predictions from the previously detected regularity representations with the incoming sound. Predictions from the model are set on a number of different time scales, in accordance with the diverse temporal basis of the simultaneously active regularity representations. Each of the regularities sets up its own "unit" or temporal chunk of the auditory input. For example, segmental and syllabic units of speech are in the range of 20-80 and 150-300 ms (Poeppel, Idsardi, & van Wassenhove, 2008), whereas the melodic and stress patterns of speech may extend to much longer periods. Thus the analysis of the input must use different temporal chunks or integration periods (Boemio, Fromm, Braun, & Poeppel, 2005; Grimm & Schröger, 2007; Hickok & Poeppel, 2007; Nelken et al., 2003; Poeppel, 2003; Poeppel et al., 2008). One well-known integration period, which is ca. 200 ms long, is termed the temporal window of integration (TWI). A large number of perceptual phenomena have been related to this TWI, such as loudness summation, detection masking, etc. (for a review, see Cowan, 1984; for a quantitative model, Zwislocki, 1969). It is thus no surprise that deviance detection also shows effects related to the TWI, such as the detection of omissions from more or less isochronous sound sequences (Yabe et al., 1998) and the integration/separation of closely spaced deviating events (Czigler & Winkler, 1996; Sussman, Winkler, Ritter, Alho, & Naatanen, 1999). Studies contrasting auditory streaming and temporal integration demonstrated that streaming precedes temporal integration (Sussman, 2005; Yabe et al., 2001). That is, temporal integration occurs within, but not across streams.

6.3. Competition and establishing the perceived sound organization

Bregman (1990) likened the process of selecting one of the alternatives to "voting", where the decomposition of the input that receives the most support from the grouping processes (i.e., many of the grouping processes lead to computing this solution) becomes dominant and it is embraced by the system. This process is marked in Fig. 3 as "Resolving Competition between Alternative Sound Organizations". Considering the structure of this competition, Winkler et al. (2012) suggested that regularity representations (termed proto-objects by Winkler et al., 2012) compete when they predict the same sound at the same time (termed collision). This local form of competition allow the emergence of full sound organizations (coalitions of proto-objects), which are compatible with each other in the sense Bregman (1990) suggested. Assuming that when two proto-objects collide, they mutually inhibit each other, Mill et al. (2013) showed that compatible proto-objects become "weak/strong" (see below) together in the auditory streaming sequences. That is, whereas the proto-object describing the integrated percept (A-B-A) collides with both of the proto-objects describing the two segregated streams (A---A and B-------B), the latter never collide with each other. Thus, when the integrated proto-object is dominant (i.e., it is perceived), it suppresses both segregated proto-objects; when it is weakened (by adaptation and noise), the two segregated proto-objects become stronger and together they suppress the integrated proto-object (as suggested by Bregman, 1990). One or the other of them becomes dominant (perceived in the foreground) while the other is perceived in the background (i.e., it is not suppressed, as opposed to when the integrated proto-object is dominant). These features of Mill et al. (2013) computational model fully match the perception of the auditory streaming sequences. Further, this notion is also compatible with results suggesting that both redundant and contradictory predictions can be generated for some stimulus sequences (Horváth et al., 2001; Pieszek, Widmann, Gruber, & Schröger, 2013; Widmann, Kujala, Tervaniemi, Kujala, & Schröger, 2004).

The voting process (competition) is based upon some "strength" (termed activation by Mill et al., 2013) measure of the alternative groupings (proto-objects). Strength is provided by the regularity representation supporting the given alternative. It may depend on the type of regularity (representing learning through evolution and individual experience) and also on the "reliability" of each solution: solutions based on regularity representations whose predictions have often been met in the recent past are stronger than alternatives based on regularity representations whose predictions have not always been confirmed by the incoming sounds (for an analysis of the MMN literature on this issue, see Winkler, 2007). Thus prediction error influences the selection between alternative groupings/organizations by decreasing the strength of the related alternatives in the competition. This notion is also compatible with the "model optimization by minimizing prediction errors" principle of predictive coding theories, specifically with a Bayes-optimal (Robert, 2007) variant of model selection (i.e., the winning model provides the greatest evidence or minimum surprise, such as that described by Friston & Kiebel, 2009).[9]

In AERS, the output of the "Finalizing Streams and Feature Integration by Streams" box can be consciously perceived (Fig. 3). In many cases, the solution is (almost) unequivocal (unambiguous auditory scenes). That is, one of the alternatives receives far greater support than any of the others. However, it is also possible that two or more alternative solutions get substantial support from the grouping processes (ambiguous auditory scenes). In this case, perception will fluctuate between the alternatives and, unlike in unambiguous cases, one may voluntarily choose one perception over the other. Thus voting can be biased by top-down effects, but only to a certain degree; that is, one cannot choose an arbitrary solution against an existing dominant stimulus-driven one. This is supported both by results of a large number of behavioral (e.g., van Norden, 1975) and electrophysiological studies (e.g., Sussman, Winkler, Huotilainen, et al., 2002; Winkler et al., 2006).

6.4. Finalizing feature integration

Once the dominant sound organization is selected, the feature-combinations making up the sounds appearing in the dominant organization are bound together, separately for each of the concurrent sounds, thus creating sound representations, which are inherently linked to auditory streams. Although some influential theories based on visual experiments suggest that feature integration requires focused attention (e.g., Treisman, 1998; see, however, e.g., Duncan & Humphreys, 1989; Winkler, Takegata, & Sussman, 2005), several studies investigating auditory feature binding found that it can occur even in the absence of focused attention (Gomes, Bernstein, Ritter, Vaughan, & Miller, 1997; Sussman, Gomes, Nousak, Ritter, & Vaughan, 1998; Takegata, Huotilainen, Rinne, Näätänen, & Winkler, 2001; Takegata, Paavilainen, Näätänen, & Winkler, 1999; Takegata et al., 2005; Winkler et al., 2005). However, there is also evidence showing that under some circumstances, the integration of auditory features may not work correctly and illusory feature conjunctions emerge (Hall, Pastore, Acker, & Huang, 2000; Thompson, Hall, & Pressing, 2001). For example, when two or more sounds differing both in pitch and timbre are presented in a concurrent array, listeners may identify sounds as being part of the array that have the pitch of one sound and the timbre of another sound from the array. Takegata et al.' (2005) results suggest that correct automatic integration of features occurs also in such cases. Therefore, miscombination of the features may occur during task-related processes. One possible explanation is that listeners use strategies relying on the processes of "feature-based target detection". That is, when the sounds can be segregated by one of the two features, this feature may become the primary cue for the listener in performing a conjunction-search task. A comparison between the results of Thompson et al. (2001) and those of Woods and colleagues (Woods & Alain, 2001; Woods, Alain, & Ogawa, 1998) supports this interpretation.

At this point in auditory processing, the sensory representation of the incoming sound(s) is complete and it now enters the processes establishing which regularity representations accurately predicted the behavior of the auditory environment ("Comparison" in Figs. 1, 3 and 4). Furthermore, the sensory description of the auditory input can now be evaluated with respect to the current goals of the organism (see "Evaluation" in Figs. 1 and 4) and those parts of the acoustic input for which no prediction existed (residue) can be identified.

6.5. Deviance detection

The output of the feature-integration operation is compared with the predictions produced by the active regularity representations. The outcome of this comparison is used (1) for adjusting the reliability value of the regularity representations (thus affecting their weight in the competition between alternative sound organizations) and (2) for providing the sensory representation of the incoming sound with information about its relationship with the auditory context (i.e., how well it fit the various regularities detected for the acoustic environment). Although the two functions are compatible, as was noted earlier, Winkler and Czigler (1998) showed that the deviance detection signal reflected by MMN is primarily related to adjusting the regularity representations.

In terms of information, deviation from what has been inferred from the generative model represents the new information content of the acoustic event. Thus the sound representations produced by AERS carry with them a description of how much acoustic information they contain. Of crucial importance to the above description of

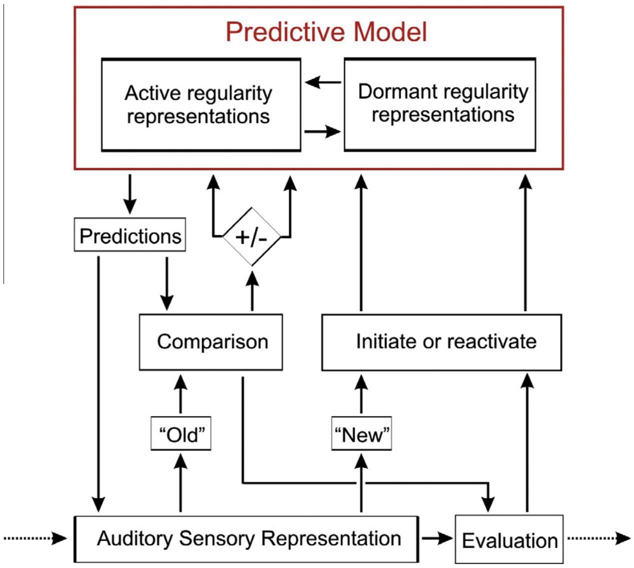
Predictive Model
Active regularity representations
Dormant regularity representations
Predictions
+/-
Comparison
Initiate or reactivate
"Old"
"New"
Auditory Sensory Representation
Evaluation

Fig. 4. Functional model of sound evaluation and model maintenance. Continuations of streams found during the formation of Auditory Sensory Representations are compared with predictions from the currently Active Regularity Representations of the Model. The outcome of the Comparison process updates the corresponding regularity representations as well as informing the Evaluation process. Sounds for which no predictions existed Initiate the formation of a new regularity representation, which is initially in a dormant state, until its predictive value is confirmed). Dormant regularity representations can also be (re)activated by their behavioral relevance through the evaluation process. In turn, active regularity representation become dormant through updating, if their prediction is violated several times or if no corresponding auditory input is received for a longer period of time (relative to the predicted timing).

[9] Note that Bayes-optimal model selection alone does not explain the bi-/multistable phenomena found in auditory streaming. Adaptation and noise are also required for dominance switches (Mill et al., 2013; Shpiro et al., 2009).

deviance detection is that the sound events separated in the auditory sensory representation are only compared with the regularities that apply to the stream they belong to. This was confirmed by Ritter and colleagues (Ritter, de Sanctis, Molholm, Javitt, & Foxe, 2006; Ritter, Sussman, & Molholm, 2000), who presented a sound sequence made up from two distinct sets of tones differing from each other on several features, among them tone duration. The two sets of sounds segregated in perception. Occasional sounds with a duration differing from the regular sound durations in either stream only elicited MMN with respect to the stream with which they shared all other auditory features. In contrast, a tone differing in duration from both of two frequently presented tones within the same stream elicits two separate MMNs, one for each of the regular tones (Sussman, Sheridan, Kreuzer, & Winkler, 2003; Winkler, Karmos, et al., 1996). Studies delivering sound sequences with some regularity that could only be discovered when the sound sequence was organized in a specific way found that MMN was only elicited, when participants perceived the sound sequence in the given way (Sussman, Horváth, Winkler, & Orr, 2007; Sussman, Ritter, et al., 1999; Sussman, Winkler, Huotilainen, et al., 2002; Winkler, Kushnerenko, et al., 2003; Winkler, Sussman, et al., 2003; Winkler et al., 2006). These results strongly support the hypothesis that deviance detection occurs separately within each stream. Furthermore, Bendixen, Denham, et al. (2010), Bendixen et al. (2013) and Bendixen, Denham, et al. (2014) found that when a regularity is only discovered when the sounds are organized in a given way, but not within the alternative sound organizations, the presence of this regularity extends the dominance periods of the given organization. Bendixen, Denham, et al. (2010) interleaved two tone sequences separated in pitch. Listeners reported perceiving segregation for longer periods of time when separately, each sequence consisted of repeating tone patterns as compared with when the order of the tones was randomized separately in each sequence. Together, the MMN and perceptual data suggest that repeating patterns are only extracted for streams segregated by other regularities. But once such regularities are discovered, they further support stream segregation and deviance detection.

In summary, the output of AERS not only provides a finely resolved description of the sound event, but also places this event into the context of currently known auditory streams and marks how well it fits the sounds preceding it. That is, the deviance detection system functions as a filter, flagging each sound that carries new information about its source.

6.6. Maintenance of the model

The acoustical environment is in constant flux. New sources become active, whereas active sources may discontinue or change their emission; sources may synchronize or fall out of synchrony. Thus the regular characteristics of the input change all the time. Therefore, the model of the auditory environment requires constant maintenance.

Fig. 4 depicts the functions involved in maintaining and updating the model of the auditory environment. Of crucial importance is the separation of the continuation of streams for which the model already contains representations and the residue, the sounds which did not fit into any of the previously detected streams. When streams are finalized, after competition between the alternative groupings has been resolved, the two types of information are separated. Fig. 4 depicts the two types of information as "Old" (continuation of a known stream) and "New" (residue). The continuation of previously detected streams is used to adjust the regularity representations of the corresponding stream. Just because the given sound has been found to continue the stream, it does not mean that all the regularity representations of the given stream correctly predicted this sound. Therefore predictions for each regularity representation belonging to the stream are compared with the sound continuing the stream in parallel and adjustments are made accordingly. The residue requires the formation of new regularity representations or the reactivation of an existing but dormant one (Fig. 4). This process is modulated by contextual information through the "Evaluation" function.

As was already mentioned in Section 4, the representations of those regularities whose predictions are met by the incoming stimulus acquire additional weight as their validity extended in time. In contrast, when a given prediction is not met, then the generative model needs to be updated, as its value for successfully anticipating auditory events has been reduced (Fig. 4). Regularity representations referring to the same sound stream may be fully independent of each other; that is, they describe different aspects of the sound stream. The updating of such regularity representations takes place independently. For example, when two tones with different pitch but uniform duration are alternated, a tone may fit both regularities, violate the constancy of duration, but correctly continue the pitch alternation and vice versa or, violate both regularities. In such cases, multiple regularity violations have been shown to elicit additive MMN components (Alain, Achim, & Woods, 1999; Levänen, Hari, McEvoy, & Sams, 1993; Schröger, 1995, 1996; Takegata, Huotilainen, et al., 2001; Takegata, Paavilainen, Näätänen, & Winkler, 2001; Takegata et al., 1999; Winkler & Czigler, 1998; Winkler, Karmos, et al., 1996). In contrast, violating regularities based on related auditory features, such as multiple temporal or spectral regularities, interfere with each other, typically resulting in subadditivity between the MMN components (Alain, Cortese, & Picton, 1999; Czigler & Winkler, 1996; Takegata, Paavilainen, et al., 2001).

New regularity-representations are created for the newly emerging sound sources, the emission of which shows up as residue after the continuation of the known regularities have been accounted for (Fig. 4). Bendixen et al.'s (2009) results shed some light on the possible timing of the onset of this process. These authors found that when an exact prediction is available for the incoming sound, the ERP response elicited by omitting this sound is not significantly different from that elicited by the sound itself within ca. 50–80 ms from the (expected) onset of the sound. This suggests that the continuation of known regular streams is assessed within this period of time and the residue becomes available by the end of this period. Starting at about this time, a series of auditory cortical ERP components (P1, N1, and P2) can be recorded from the scalp, all of which are sensitive to large acoustic changes in a sequence or sounds presented after long silent periods (Näätänen, 1992; Näätänen & Picton, 1987). We tentatively suggest that these ERP components may reflect processes involved in building a representation for new auditory objects.[10] In agreement with this notion, a part of the N1 wave is known to be elicited with quite large amplitude by a stimulus delivered after a long silent period then sharply decreasing within the first few sounds of a new train and reaching an asymptotic level after 4–5 sounds (Cowan et al., 1993; Näätänen & Picton, 1987). This sequence of events is compatible with the assumed phases of establishing a new regularity representation (see Section 4). Furthermore, stimuli eliciting an N1 of high amplitude are likely to capture attention (Näätänen, 1990; Näätänen et al., 2011), which agrees with one's everyday experience of noticing new sound sources. Also, N1 is increased for familiar sounds (Kirmse, Jacobsen, & Schröger, 2009). Familiar sounds are more likely to enter consciousness (see e.g., hearing one's own name

[10] Note, however, that as of yet no evidence relates these ERP components directly to the formation of new regularity representations. Furthermore, P1, N1, and P2 are elicited by all sounds with a sufficiently abrupt onset; although as was noted, their amplitude is higher, when the incoming sound largely differs from previous sounds.

in an unattended channel; Cherry, 1953). Finally, the N1 response is highly sensitive to the direction of focused attention, which is compatible with the assumed top-down influence on detecting new sound sources (Hillyard, Hink, Schwent, & Picton, 1973).

As was described in Section 4, regularity representations can be in a dormant state either because the number of sounds conforming to the given regularity has not yet reached the required level or, because no sounds meeting the regularity have been encountered for some time. The residue may, however, contain a sound conforming to a dormant regularity. In this case, the regularity is reactivated (Winkler & Cowan, 2005; Fig. 4).

Both the formation of new regularity representations and the reactivation of a dormant one can be influenced by top-down processes, including attentional effects (for the interaction of prediction and attention in audition, see Schröger, Marzecová, & SanMiguel, 2015). The relation of the incoming sound to the context is part of the auditory event representation, the outcome of AERS. This includes how well it fit the existing regularities (possible prediction errors) thus allowing evaluation of the sound information with respect to the larger context (including behavioral goals). In response, higher levels of the perceptual/cognitive system can adjust the functioning of AERS by forcing it to look for certain regularities or to reactivate ones, which have become dormant based on stimulus (bottom-up) information alone. Note that the output of the "Comparison" function as specified here is not identical to prediction error in predictive coding models. We separated prediction error into two parts: mismatch between the prediction and the actual continuation of the stream (the "-" branch in the "Old" route; MMN ERP response) and the residue triggering the formation of a new or reactivating a dormant regularity (the "New" route; possibly related to some subcomponent of the N1 ERP response). Within a hierarchical predictive coding model, the former can be addressed locally (within the same level in the hierarchy), whereas the latter requires intervention from higher levels.

Overall, the maintenance of the generative model must ensure that the functional module can quickly adapt to changes in the acoustic environment while keeping its predictive value high all the time. Redundancies in the model, the possibility to reuse outdated regularities, as well as maintaining each of the regularity representations in parallel allows AERS to mark new information for subsequent processing.

7. Comparison with existing models of predictive processing in perception

The assumption of predictive processing is not unique to AERS. For example, predictive modeling theories are based on the same assumption. Garrido and colleagues' (Garrido, Kilner, Stephan, & Friston, 2009; Lieder et al., 2013) and Wacongne and colleagues model (Wacongne, Changeux, & Dehaene, 2012) of ERP responses elicited in the auditory oddball paradigm come closest to the current description. These models have been created to explain the observable MMN response in some specific cases of deviance detection basing on the known neurophysiological properties of the auditory system. However, while these models might be relatively easily extended to cover different auditory features and some additional regularities, as of yet, no attempt has been made to generalize them to the large variety of regularities, whose violation elicits the MMN response. Further, neither model addresses most of the other issues covered by AERS (model build-up, reactivation, separating the updating of existing models from the formation of a new one, or auditory stream segregation, in general) and does not provide a psychological interpretation of the assumed processes. In contrast, AERS covers all of the known regularities, at the cost of relinquishing the neural specificity of the models mentioned above. Thus the two approaches are complementary and may provide synergy in the future. One possible link has been suggested by Winkler and Czigler (2012), who suggested that deviance detection, as reflected by MMN may fit as an intermediate level into a hierarchical predictive coding model. Pre-MMN ERP responses elicited by simpler forms of deviations (for reviews, see Malmierca, Sanchez-Vives, Escera, & Bendixen, 2014, and Grimm & Escera, 2012) may reflect lower levels of the hierarchy.

Kiebel, von Kriegstein, Daunizeau, and Friston (2009) developed a predictive-coding based computational model online recognizing tokens in a hierarchically structured continuous sound. This model has some of the capabilities we assume for the initial build-up of an auditory model including segmentation of continuous sounds, a feature not addressed in (but obviously necessary for) AERS. Thus we regard this model as a possible implementation of some of the functions of AERS. However, again, this model does not consider auditory stream segregation. In contrast, no previous theory or model of auditory stream segregation (Anstis & Saida, 1985; Bregman, 1990; Carlyon, 2004; Jones & Boltz, 1989; Schwartz et al., 2012; Shamma, Elhilali, & Micheyl, 2011; Snyder & Alain, 2007) allocates a role for predictive processing. Recently, Mill et al. (2013) based their computational model on the ideas described here; we have already referred to their work in previous sections. Finally, the notion that predictions are an essential aspect of information processing has been considered by modern theorists, such as Bar (2004, 2007), Summerfield and Egner (2009), and elements of it also appear in Gregory's (1980) and Ahissar and Hochstein's (2004) work. Our description is compatible with many of these ideas. Unfortunately, these general theories are largely based on visual perception, providing little guidance for solving the special problems of auditory perception.

Results of a series of studies suggest that predictive confidence (our term) or precision (Feldman & Friston, 2010) do not fully describe the stimulus-driven determinants of the MMN amplitude. Todd and colleagues (Mullens et al., 2014; Todd, Heathcote, Mullens, et al., 2014; Todd, Heathcote, Whitson, et al., 2014; Todd, Provost, & Cooper, 2011; Todd, Provost, Whitson, Cooper, & Heathcote, 2013) have repeatedly observed that when the roles of two sounds as frequent standard and rare deviant are periodically exchanged, only the configuration encountered first follows the principle of improved predictive confidence/precision with greater stability of the configuration (i.e., longer periods within which the same standard-deviant configuration remains the same) but not the reversed role configuration, which initially appears after the first role change between the two sounds. The authors refer to this asymmetry as "primacy bias". It probably stems from the different relevance attached to the repetitive and the rare sound by the brain, as the bias can also be manipulated by assigning behavioral relevance to one or the other sound, even when they are first encountered by the listener in a sequence in which the two sounds appear with equal probability (Mullens et al., 2014). If this was the case, then within AERS, primacy bias appears through the evaluation function, which takes into account the contextual relevance attached to each sound by higher-level functions.

8. AERS issues related to communication by sound

We started our review by pointing out that communication requires the maintenance of an open channel between the parties. Here we break down this general function for the auditory modality and describe the role of AERS in implementing them in humans.

First, one should consider that often multiple sources are active concurrently and the listener needs to distinguish them in order to being able to follow one (or a few) of these communication channels. Segregating sound sources, including speakers, is the primary function of AERS. As was suggested before, we regard the regularity representations as forming the core of auditory perceptual objects (Winkler et al., 2009). They encode the characteristic auditory features detected for the sound source, allowing the system to determine which part of the incoming sound was likely generated by this source. The output of AERS is the earliest internal sound representation that can be identified, monitored, processed as a speech stream, etc.

Once the sources are distinguished, they can be identified. Identification of sound sources naturally lies outside AERS. However, it affects the AERs as it allows learned information about the given type of source to fine-tune the predictive model. For example, if the source is a bird, we expect chirping sounds (experience with the exact species may allow even more specific predictions). This knowledge can affect AERS via the Evaluation function, biasing the build-up and reactivation of models. Source (speaker) identity is even more crucial for speech perception. Typically, coherent messages come from a single speaker (or, possibly, several speakers speaking in concert – such as a choir; which can be regarded as single sound source). The acoustic features identifying the speaker are encoded in the regularity representations of AERS. This information is relied on by the semantic processes of speech perception. In turn, syntactic and semantic information may allow much sharper predictions for upcoming sounds. Although predictive processes have been hypothesized for language processing, most studies and models consider reading and possibly sign language, but not speech *per se*. However a predictive framework for speech processing has been developed by Kotz and colleagues (Kotz & Schwartze, 2010; Kotz, Schwartze, & Schmidt-Kassow, 2009; Rothermich & Kotz, 2013; Sammler, Kotz, Eckstein, Ott, & Friederici, 2010) that is compatible with the idea that AERS regularity representations may be specified by predictions based on syntactic or semantic predictability.

The next issues to be solved are whether the message is directed to us and if so, does it require a response. Although the first information is typically resolved by the general context, at least in infants, the mode of speech (infant vs. adult directed speech) has a significant effect on whether the infant regards him/herself as the addressee of the message (Senju & Csibra, 2008). Prosody generally tells whether the message contains a question. Prosodic regularities are encoded in AERS, as was shown by studies testing prosodic violations (e.g., Honbolygó, Csépe, & Ragó, 2004; Leitman, Sehatpour, Shpaner, Foxe, & Javitt, 2009; Tong et al., 2014). Thus AERS serves as a source of prosodic information to speech processing and, in return, prosodic regularity representations in AERS may be modulated by syntactic information.

Finally, conversations require mutual adaptation from the participants. This includes turn-taking as well as co-adapting their rhythms of speech (Jaffe, Beebe, Feldstein, Crown, & Jasnow, 2001). The temporal aspects of sound sequences, including stimulus rate and rhythmic structures are also represented in AERS. Changes in a regular inter-stimulus interval (e.g., Nordby, Roth, & Pfefferbaum, 1988), train offsets (e.g., Bendixen, Scharinger, et al., 2014; Horváth, Müller, Weise, & Schröger, 2010; Yabe et al., 1998), and violations of higher-order rhythmic (Ladinig, Honing, Háden, & Winkler, 2009) elicit MMN responses. Thus, AERS can play an important role in providing information about response timing.

From our point of view, speech is only one of many possible communication channels. Once the channel is established, the predictions are assumed to foster the processing of the information delivered via the channel. In the case of speech, one can hardly imagine that production and perception can work without generative models; consider, e.g., the German sports reporter Heribert Fassbender, who could articulate up to 26 phonemes per second, and listeners were still being able to comprehend it (N. Blotzki, unpublished Master Thesis, Bonn University). Another benefit of predictions is that they help to detect new acoustic information communicated by the channel: through deviance (irregularity) detection, an inherent property of AERS, we may quickly learn about a change of the speaker or a change in the speaker's state (by detecting changes in voice characteristics).

9. Limitations of AERS and future directions

The goal of this review has been to outline a common theoretical framework for conceptualizing phenomena observed in studies of auditory scene analysis and deviance detection. As most theoretical frameworks, the current one is also bound in two different ways: (1) self-imposed limitations regarding the width and depth of the discussion and (2) limitations imposed by the experimental data considered. We already referred to the first in Introduction. We did not review the extensive literature in psychology, acoustics, and neuroscience on early auditory processing. We assume that preprocessed auditory information is available at the input of AERS. Similarly, we did not attempt to outline higher cognitive systems which utilize the information from AERS and can adjust its operation.

It is perhaps more important to consider, how the empirical evidence forming the basis of the current review may limit its scope. Most studies referred in the current review (and also in the literature in general) deal with stimulus configurations, which are far simpler than almost any real-life auditory scene. Specifically, streams in these studies differ from each other by one (in a very few cases two or three) primary auditory features. The same is true for the separation between standard and deviant sounds in most deviance-detection studies. Furthermore, the test sounds are usually short and they seldom fully overlap each other in time; thus promoting discrete sounds as easily discernible units of the incoming sound. Does this mean that the functions and principles described in AERS are only valid for such simplified stimulus configurations? Perhaps not. There is no reason to assume that either auditory streaming or deviance detection would work differently when the sounds are separated by complex spectro-temporal characteristics. For example, Winkler, Teder-Sälejärvi, Horvath, Näätänen, and Sussman (2003) delivered to participants sequences composed of 11 different natural footstep sounds. Ten of these sounds were similar and presented in an approximately even rhythm, thus giving the impression of someone walking. The remaining one sounded as if someone stepped onto a different surface. Although there were no easy-to-define spectral or temporal differences between the two types of footstep sounds, when the different-surface footstep was presented in the 10th position in the footstep sequence, it elicited the MMN. MMN was elicited despite the fact that the participant watched a movie with sounds and street noise was also continuously delivered to the room through loudspeakers. This result demonstrates that the deviance detection system can also utilize complex auditory features within a natural sound environment as for detecting the deviation, the sequence of footsteps had to be segregated from two other continuous streams of sound (street noise and the sound of the movie), both of which covered a wide spectral range, fully overlapping that of the footstep sounds. Thus auditory streams fully overlapping in time and in the spectrum were segregated from each other (for streaming by complex feature differences, see also, Iverson, 1995; for extracting sounds fully embedded in other sounds, see e.g., Chait, 2014; McDermott, Wrobleski, & Oxenham, 2011; Teki, Chait, Kumar, Shamma, & Griffiths, 2013; Teki et al., 2011).

Deviance detection for speech sounds works similarly to other types of sounds with phonetic features showing categorical effects with respect to the languages spoken by the listener (for reviews see, Bishop, 2007; Näätänen, 2001; Pulvermüller & Shtyrov, 2006; Rimmele, Sussman, & Poeppel, 2015). Speaker and speech segregation from noise and from other speakers has been extensively studied in the literature (e.g., Culling & Summerfield, 1995; de Cheveigné, Kawahara, Tsuzaki, & Aikawa, 1997; de Cheveigné, McAdams, & Marin, 1997; for the engineering point of view see, e.g., Loizou, 2007). However, speech/speaker segregation is typically not an auditory-only function, as understanding speech allows one to sharpen predictions for upcoming sounds. As was already noted in Introduction, such effects are fully compatible with AERS. They are conceptualized as higher-level models in a possible predictive coding hierarchy affecting model selection and parameters in AERS through the "Evaluation" function.

Forming and maintaining regularity representations in AERS do not depend on the feature underlying the regularity. However, AERS does not include mechanisms promoting the emergence of context-specific features (e.g., by plastic changes in the spectro-temporal receptive fields of groups of afferent neurons). It is possible that long-term learning effects shape what features are picked up by our auditory system. Further, although our description focuses on the temporal/sequential cues of auditory stream segregation, we also considered stream segregation by spectral/concurrent cues. As for finding sound units within a realistic auditory scene, together with Nelken et al. (2003), we maintain that sound is analyzed on multiple time scales in parallel, thus allowing parallel formation of regularities based on different units. There exist some computational models capable of segmenting continuous sounds (Coath, Brader, Fusi, & Denham, 2005; Kiebel et al., 2009). One exciting future direction will be to connect them with computational models based on AERS (Mill et al., 2013). Larger units can then be built from smaller ones, as was reviewed in Section 4.2. Thus the simplified stimulation employed in most studies of auditory stream segregation and deviance detection does not appear to limit the generality of the functions assumed for AERS. On the other hand, we acknowledge that the current description did not explore the effects of task- and knowledge-based strategies on auditory streaming and deviance detection. AERS includes several different ways in which top-down influence can affect its operation. Specifying these effects is an exciting direction for further research (see, e.g., Niessen, van Maanen, & Andringa, 2008).

In summary, AERS is aimed at capturing the intelligent dynamic aspects of auditory perceptual processing, which allows – paraphrasing Köhler's famous sentence – naïve and uncritical listeners to effortlessly experience the auditory world as organized and to select from it meaningful, identifiable objects at will. So, finally, one can say that we listen to sounds through our AERS.

Acknowledgments

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), and by the Leibniz-Professor Award from the University of Leipzig to IW during the winter term 2003/2004. We thank Johanna Steinberg for proofreading the manuscript.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457–464.

Ahveninen, J., Jaaskelainen, I. P., Pekkonen, E., Hallberg, A., Hietanen, M., Näätänen, R., et al. (2000). Increased distractibility by task-irrelevant sound changes in abstinent alcoholics. *Alcoholism-Clinical and Experimental Research, 24*(12), 1850–1854.

Akeroyd, M. A., Carlyon, R. P., & Deeks, J. M. (2005). Can dichotic pitches form two streams? *Journal of the Acoustical Society of America, 118*(2), 977–981.

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research, 229*(1–2), 225–236.

Alain, C., Achim, A., & Woods, D. L. (1999). Separate memory-related processing for auditory frequency and patterns. *Psychophysiology, 36*(6), 737–744.

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance, 27*(5), 1072–1089.

Alain, C., Cortese, F., & Picton, T. W. (1999). Event-related brain activity associated with auditory pattern processing. *NeuroReport, 10*(11), 2429–2434.

Alain, C., Schuler, B. M., & McDonald, K. L. (2002). Neural activity associated with distinguishing concurrent auditory objects. *Journal of the Acoustical Society of America, 111*(2), 990–995.

Andreou, L. V., Kashino, M., & Chait, M. (2011). The role of temporal regularity in auditory segregation. *Hearing Research, 280*(1–2), 228–235.

Anstis, S., & Saida, S. (1985). Adaptation to auditory streaming of frequency modulated tones. *Journal of Experimental Psychology: Human Perception and Performance, 11*(3), 257–271.

Baldeweg, T. (2006). Repetition effects to sounds: Evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences, 10*(3), 93–94.

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

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617–629.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*(7), 280–289.

Bendixen, A. (2014). Predictability effects in auditory scene analysis: A review. *Frontiers in Neuroscience, 8*, 60.

Bendixen, A., Bőhm, T. M., Szalárdy, O., Mill, R., Denham, S. L., & Winkler, I. (2013). Different roles of similarity and predictability in auditory stream segregation. *Learning and Perception, 5*, 37–54.

Bendixen, A., Denham, S. L., Gyimesi, K., & Winkler, I. (2010). Regular patterns stabilize auditory streams. *Journal of the Acoustical Society of America, 128*(6), 3658–3666.

Bendixen, A., Denham, S. L., & Winkler, I. (2014). Feature predictability flexibly supports auditory stream segregation or integration. *Acta Acustica United with Acustica, 100*(5), 888–899.

Bendixen, A., Jones, S. J., Klump, G., & Winkler, I. (2010). Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components. *Neuroimage, 50*(1), 285–290.

Bendixen, A., Prinz, W. G., Horváth, J., Trujillo-Barreto, N. J., & 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. (2012). Early electrophysiological indicators for predictive processing in audition: A review. *International Journal of Psychophysiology, 83*(2), 120–131.

Bendixen, A., Scharinger, M., Strauss, A., & Obleser, J. (2014). Prediction in the service of comprehension: Modulated early brain responses to omitted speech segments. *Cortex, 53*, 9–26.

Bendixen, A., Schröger, E., Ritter, W., & Winkler, I. (2012). 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. *Journal of Neuroscience, 29*(26), 8447–8451.

Bertrand, O., & Tallon-Baudry, C. (2000). Oscillatory gamma activity in humans: A possible role for object representation. *International Journal of Psychophysiology, 38*(3), 211–223.

Bishop, D. V. M. (2007). Using mismatch negativity to study central auditory processing in developmental language and literacy impairments: Where are we, and where should we be going? *Psychological Bulletin, 133*(4), 651–672.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience, 8*(3), 389–395.

Brattico, E., Winkler, I., Näätänen, R., Paavilainen, P., & Tervaniemi, M. (2002). Simultaneous storage of two complex temporal sound patterns in auditory sensory memory. *NeuroReport, 13*(14), 1747–1751.

Bregman, A. S. (1990). *Auditory scene analysis. The perceptual organization of sound*. Cambridge, MA: MIT Press.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences, 8*(10), 465–471.

Carney, L. H. (2002). Neural basis of audition. In H. Pashler & S. Yantis (Eds.). *Stevens' handbook of experimental psychology. Sensation and perception* (Vol. 1, pp. 341–396). New York: John Wiley & Sons.

Chait, M. (2014). Change detection in complex acoustic scenes. *The Journal of the Acoustical Society of America, 135*(4), 2171.

Chang, E. F. (2014). Feature representation in human speech cortex during perception and production. In E. Budinger (Ed.), *Proceedings of the 15th international conference on auditory cortex* (pp. 14). Magdeburg: University of Magdeburg.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25*(5), 975–979.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience, 13*, 148–169.

Ciocca, V., & Darwin, C. J. (1999). The integration of nonsimultaneous frequency components into a single virtual pitch. *Journal of the Acoustical Society of America, 105*(4), 2421–2430.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Coath, M., Brader, J. M., Fusi, S., & Denham, S. L. (2005). Multiple views of the response of an ensemble of spectro-temporal features support concurrent classification of utterance, prosody, sex and speaker identity. *Network-Computation in Neural Systems, 16*(2–3), 285–300.

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin, 96*(2), 341–370.

Cowan, N. (1987). Auditory sensory storage in relation to the growth of sensation and acoustic information extraction. *Journal of Experimental Psychology: Human Perception and Performance, 13*(2), 204–215.

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.

Culling, J. F., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds – Absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America, 98*(2), 785–797.

Czigler, I., & Winkler, I. (1996). Preattentive auditory change detection relies on unitary sensory memory representations. *NeuroReport, 7*(15–17), 2413–2417.

Darwin, C. J., Hukin, R. W., & Alkhatib, B. Y. (1995). Grouping in pitch perception – Evidence for sequential constraints. *Journal of the Acoustical Society of America, 98*(2), 880–885.

de Cheveigné, A. (2001). The auditory system as a separation machine. In A. J. M. Houtsma, A. Kohlrausch, V. F. Prijs, & R. Schoonhoven (Eds.), *Physiological and psychological bases on auditory function* (pp. 453–460). Maastricht, The Netherlands: Shaker.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., & Aikawa, K. (1997). Concurrent vowel identification. 1. Effects of relative amplitude and F0 difference. *Journal of the Acoustical Society of America, 101*(5), 2839–2847.

de Cheveigné, A., McAdams, S., & Marin, C. M. H. (1997). Concurrent vowel identification. 2. Effects of phase, harmonicity, and task. *Journal of the Acoustical Society of America, 101*(5), 2848–2856.

Deacon, D., Nousak, J. M., Pilotti, M., Ritter, W., & Yang, C. M. (1998). Automatic change detection: Does the auditory system use representations of individual stimulus features or gestalts? *Psychophysiology, 35*(4), 413–419.

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

Demany, L., & Semal, C. (2008). The role of memory in auditory perception. In W. A. Yost, A. N. Popper, & R. A. Fay (Eds.), *Auditory perception of sound sources. Springer handbook of auditory research* (pp. 77–113). New York: Springer.

Denham, S. L., & Winkler, I. (2014). Auditory perceptual organization. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization*. http://dx.doi.org/10.1093/oxfordhb/9780199686858.013.001 (online publication).

Denham, S. L., Bőhm, T. M., Bendixen, A., Szalardy, O., Kocsis, Z., Mill, R., et al. (2014). Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. *Frontiers in Neuroscience, 8*, 25.

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

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

Denham, S. L., & Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *Journal of Physiology – Paris, 100*(1–3), 154–170.

Duke, R. A. (1989). Musicians perception of beat in monotonic stimuli. *Journal of Research in Music Education, 37*(1), 61–71.

Duncan, J., & Humphreys, G. W. (1989). Visual-search and stimulus similarity. *Psychological Review, 96*(3), 433–458.

Dyson, B. J., & Alain, C. (2008a). Is a change as good with a rest? Task-dependent effects of inter-trial contingency on concurrent sound segregation. *Brain Research, 1189*, 135–144.

Dyson, B. J., & Alain, C. (2008b). It all sounds the same to me: Sequential ERP and behavioral effects during pitch and harmonicity judgments. *Cognitive Affective & Behavioral Neuroscience, 8*(3), 329–343.

Dyson, B. J., Alain, C., & He, Y. (2005). Effects of visual attentional load on low-level auditory scene analysis. *Cognitive Affective & Behavioral Neuroscience, 5*(3), 319–338.

Enns, J. T., & Lleras, A. (2008). What's next? New evidence for prediction in human vision. *Trends in Cognitive Sciences, 12*(9), 327–333.

Escera, C., Alho, K., Schröger, E., & Winkler, I. (2000). Involuntary attention and distractibility as evaluated with event-related brain potentials. *Audiology and Neuro-Otology, 5*(3–4), 151–166.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491–505.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Formby, D. (1967). Maternal recognition of infants cry. *Developmental Medicine and Child Neurology, 9*(3), 293. 293-&.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*(2), 78–84.

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: An event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience and Biobehavioral Reviews, 25*(4), 355–373.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B – Biological Sciences, 360*(1456), 815–836.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Friston, K. J., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks, 22*(8), 1093–1104.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology, 120*(3), 453–463.

Ghahramani, Z., & Wolpert, D. M. (1997). Modular decomposition in visuomotor learning. *Nature, 386*(6623), 392–395.

Giard, M. H., Lavikainen, J., Reinikainen, K., Perrin, F., Bertrand, O., Pernier, J., et al. (1995). Separate representation of stimulus frequency, intensity, and duration in auditory sensory memory – An event-related potential and dipole-model analysis. *Journal of Cognitive Neuroscience, 7*(2), 133–143.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gomes, H., Bernstein, R., Ritter, W., Vaughan, H. G., & Miller, J. (1997). Storage of feature conjunctions in transient auditory memory. *Psychophysiology, 34*(6), 712–716.

Gomes, H., Ritter, W., & Vaughan, H. G. (1995). The nature of preattentive storage in the auditory-system. *Journal of Cognitive Neuroscience, 7*(1), 81–94.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 290*(1038), 181–197.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*(11), 887–892.

Grimault, N., Bacon, S. P., & Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *Journal of the Acoustical Society of America, 111*(3), 1340–1348.

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

Grimm, S., & Schröger, E. (2007). The processing of frequency deviations within sounds: Evidence for the predictive nature of the mismatch negativity (MMN) system. *Restorative Neurology and Neuroscience, 25*(3–4), 241–249.

Haenschel, C., Vernon, D. J., Dwivedi, P., Gruzelier, J. H., & Baldeweg, T. (2005). Event-related brain potential correlates of human auditory sensory memory-trace formation. *Journal of Neuroscience, 25*(45), 10494–10501.

Hagoort, P. (2008). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society B – Biological Sciences, 363*(1493), 1055–1069.

Hall, M. D., Pastore, R. E., Acker, B. E., & Huang, W. Y. (2000). Evidence for auditory feature integration with spatially distributed items. *Perception & Psychophysics, 62*(6), 1243–1257.

Haroush, K., Hochstein, S., & Deouell, L. Y. (2010). Momentary fluctuations in allocation of attention: Cross-modal effects of visual task load on auditory discrimination. *Journal of Cognitive Neuroscience, 22*(7), 1440–1451.

Hautus, M. J., & Johnson, B. W. (2005). Object-related brain potentials associated with the perceptual segregation of a dichotically embedded pitch. *Journal of the Acoustical Society of America, 117*(1), 275–280.

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation, 17*(9), 1875–1902.

Helmholtz, H. v. (1867). *Handbuch der physiologischen Optik*. Leipzig: Voss.

Hickok, G., & Poeppel, D. (2007). Opinion – The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in human brain. *Science, 182*(4108), 177–180.

Hohwy, J. (2007). Functional integration and the mind. *Synthese, 159*(3), 315–328.

Hommel, B., Musseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*(5), 849–878.

Honbolygó, F., Csépe, V., & Ragó, A. (2004). Suprasegmental speech cues are automatically processed by the human brain: A mismatch negativity study. *Neuroscience Letters, 363*(1), 84–88.

Horváth, J., Czigler, I., Jacobsen, T., Maeß, B., Schröger, E., & Winkler, I. (2008). MMN or no MMN: No magnitude of deviance effect on the MMN amplitude. *Psychophysiology, 45*(1), 60–69.

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., Müller, D., Weise, A., & Schröger, E. (2010). Omission mismatch negativity builds up late. *NeuroReport, 21*(7), 537–541.

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.

Hosemann, J., Herrmann, A., Steinbach, M., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Lexical prediction via forward models: N400 evidence from German Sign Language. *Neuropsychologia, 51*(11), 2224–2237.

Huotilainen, M., Ilmoniemi, R. J., Lavikainen, J., Tiitinen, H., Alho, K., Sinkkonen, J., et al. (1993). Interaction between representations of different features of auditory sensory memory. *NeuroReport, 4*(11), 1279–1281.

Hupe, J. M., & Pressnitzer, D. (2012). The initial phase of auditory and visual scene analysis. *Philosophical Transactions of the Royal Society B – Biological Sciences, 367*(1591), 942–953.

Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance, 21*(4), 751–763.

Jacobsen, T., Horvath, J., Schröger, E., Lattner, S., Widmann, A., & Winkler, I. (2004). Pre-attentive auditory processing of lexicality. *Brain and Language, 88*(1), 54–67.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., & Jasnow, M. D. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development, 66*(2), i–viii, 1–132.

Javitt, D. C., Grochowski, S., Shelley, A. M., & Ritter, W. (1998). Impaired mismatch negativity (MMN) generation in schizophrenia as a function of stimulus deviance, probability, and interstimulus/interdeviant interval. *Evoked Potentials-Electroencephalography and Clinical Neurophysiology, 108*(2), 143–153.

Johnson, B. W., Hautus, M., & Clapp, W. C. (2003). Neural activity associated with binaural processes for the perceptual segregation of pitch. *Clinical Neurophysiology, 114*(12), 2245–2250.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83*(5), 323–355.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96*(3), 459–491.

Kaernbach, C. (2004). The memory of noise. *Experimental Psychology, 51*(4), 240–248.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.

Kiebel, S. J., von Kriegstein, K., Daunizeau, J., & Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Computational Biology, 5*(8), e1000464.

Kirmse, U., Jacobsen, T., & Schröger, E. (2009). Familiarity affects environmental sound processing outside the focus of attention: An event-related potential study. *Clinical Neurophysiology, 120*(5), 887–896.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.

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

Korzyukov, O. A., Winkler, I., Gumenyuk, V. I., & Alho, K. (2003). Processing abstract auditory features in the human auditory cortex. *Neuroimage, 20*(4), 2245–2258.

Kotz, S. A., & Schwartze, M. (2010). Cortical speech processing unplugged: A timely subcortico-cortical framework. *Trends in Cognitive Sciences, 14*(9), 392–399.

Kotz, S. A., Schwartze, M., & Schmidt-Kassow, M. (2009). Non-motor basal ganglia functions: A review and proposal for a model of sensory predictability in auditory language perception. *Cortex, 45*(8), 982–990.

Kubovy, M., & van Valkenburg, D. (2001). Auditory and visual objects. *Cognition, 80*(1–2), 97–126.

Kujala, T., Tervaniemi, M., & Schröger, E. (2007). The mismatch negativity in cognitive and clinical neuroscience: Theoretical and methodological considerations. *Biological Psychology, 74*(1), 1–19.

Ladinig, O., Honing, H., Háden, G., & Winkler, I. (2009). Probing attentive and preattentive emergent meter in adult listeners without extensive music training. *Music Perception, 26*(4), 377–386.

Lee, A. K. C., & Shinn-Cunningham, B. G. (2008a). Effects of frequency disparities on trading of an ambiguous tone between two competing auditory objects. *Journal of the Acoustical Society of America, 123*(6), 4340–4351.

Lee, A. K. C., & Shinn-Cunningham, B. G. (2008b). Effects of reverberant spatial cues on attention-dependent object formation. *JARO – Journal of the Association for Research in Otolaryngology, 9*(1), 150–160.

Leitman, D. I., Sehatpour, P., Shpaner, M., Foxe, J. J., & Javitt, D. C. (2009). Mismatch negativity to tonal contours suggests preattentive perception of prosodic content. *Brain Imaging and Behavior, 3*(3), 284–291.

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences, 18*(9), 472–479.

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences, 3*(7), 254–264.

Levänen, S., Hari, R., McEvoy, L., & Sams, M. (1993). Responses of the human auditory-cortex to changes in one versus two stimulus features. *Experimental Brain Research, 97*(1), 177–183.

Lieder, F., Stephan, K. E., Daunizeau, J., Garrido, M. I., & Friston, K. J. (2013). A neurocomputational model of the mismatch negativity. *PLoS Computational Biology, 9*(11).

Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. Boca Raton, FL: CRC Press.

Malmierca, M. S., Sanchez-Vives, M. V., Escera, C., & Bendixen, A. (2014). Neuronal adaptation, novelty detection and regularity encoding in audition. *Frontiers in Systems Neuroscience, 8*.

McDermott, J. H., Wrobleski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National academy of Sciences of the United States of America, 108*(3), 1188–1193.

McDonald, K. L., & Alain, C. (2005). Contribution of harmonicity and location to auditory object formation in free field: Evidence from event-related brain potentials. *Journal of the Acoustical Society of America, 118*(3), 1593–1604.

Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research, 266*(1–2), 36–51.

Mill, R. W., Bőhm, T. M., Bendixen, A., Winkler, I., & Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Computational Biology, 9*(3).

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica United with Acustica, 88*(3), 320–333.

Mullens, D., Woodley, J., Whitson, L., Provost, A., Heathcote, A., Winkler, I., et al. (2014). Altering the primacy bias – How does a prior task affect mismatch negativity? *Psychophysiology, 51*(5), 437–445.

Mumford, D. (1992). On the computational architecture of the neocortex II. The role of cortico-cortical loops. *Biological Cybernetics, 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. (1992). *Attention and brain function*. Hillsdale, NJ: Erlbaum.

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology, 38*(1), 1–21.

Näätänen, R., & Alho, K. (1997). Mismatch negativity – The measure for central sound representation accuracy. *Audiology and Neuro-Otology, 2*(5), 341–353.

Näätänen, R., Alho, K., & Schröger, E. (2002). Electrophysiology of attention. In H. Pashler & J. Wixted (Eds.). *Stevens' handbook of experimental psychology. Methodology in experimental psychology* (Vol. 4, 3rd ed., pp. 601–653). New York: John Wiley.

Näätänen, R., & Gaillard, A. W. K. (1983). The orienting reflex and the N2 deflection of the event-related potential (ERP). In A. W. K. Gaillard & W. Ritter (Eds.), *Tutorials in event related potential research: Endogenous components* (pp. 119–141). Amsterdam: Elsevier Science Ltd.

Näätänen, R., Gaillard, A. W. K., & 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., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*(6615), 432–434.

Näätänen, R., Paavilainen, P., Tiitinen, H., Jiang, D., & Alho, K. (1993). Attention and mismatch negativity. *Psychophysiology, 30*(5), 436–450.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound – A review and an analysis of the component structure. *Psychophysiology, 24*(4), 375–425.

Näätänen, R. (1984). In search of a short duration memory trace of a stimulus in the human brain. In L. Pulkkinen & P. Lyytinen (Eds.), *Human action and personality. Essays in honor of Martti Takala. Jyväskylä studies in education, psychology and social research 54* (pp. 29–43). Jyväskylä: University of Jyväskylä.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). 'Primitive intelligence' in the auditory cortex. *Trends in Neurosciences, 24*(5), 283–288.

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

Nelken, I., Fishbach, A., Las, L., Ulanovsky, N., & Farkas, D. (2003). Primary auditory cortex of cats: Feature detection or something else? *Biological Cybernetics, 89*(5), 397–406.

Niessen, M. E., van Maanen, L., & Andringa, T. C. (2008). Disambiguating sound through context. *International Journal of Semantic Computing, 2*(3), 327–341.

Nordby, H., Roth, W. T., & Pfefferbaum, A. (1988). Event-related potentials to time-deviant and pitch-deviant tones. *Psychophysiology, 25*(3), 249–261.

Nousak, J. M. K., Deacon, D., Ritter, W., & Vaughan, H. G. (1996). Storage of information in transient auditory memory. *Cognitive Brain Research, 4*(4), 305–317.

Opitz, B., Schröger, E., & von Cramon, D. Y. (2005). Sensory and cognitive mechanisms for preattentive change detection in auditory cortex. *European Journal of Neuroscience, 21*(2), 531–535.

Paavilainen, P., Arajärvi, P., & Takegata, R. (2007). Preattentive detection of nonsalient contingencies between auditory features. *NeuroReport, 18*(2), 159–163.

Paavilainen, P., Jaramillo, M., Näätänen, R., & Winkler, I. (1999). Neuronal populations in the human brain extracting invariant relationships from acoustic variance. *Neuroscience Letters, 265*(3), 179–182.

Perrin, F., Garcia-Larrea, L., Mauguiere, F., & Bastuji, H. (1999). A differential brain response to the subject's own name persists during sleep. *Clinical Neurophysiology, 110*(12), 2153–2164.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al. (2000). Auditory cortex accesses phonological categories: An MEG mismatch study. *Journal of Cognitive Neuroscience, 12*(6), 1038–1055.

Pieszek, M., Widmann, A., Gruber, T., & Schröger, E. (2013). The human brain maintains contradictory and redundant auditory sensory predictions. *PLoS One, 8*(1).

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245–255.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B – Biological Sciences, 363*(1493), 1071–1086.

Polich, J. (2007). Updating p300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148.

Pressnitzer, D., & Hupe, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16*(13), 1351–1357.

Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., et al. (2001). Memory traces for words as revealed by the mismatch negativity. *Neuroimage, 14*(3), 607–616.

Pulvermüller, F., & Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Progress in Neurobiology, 79*(1), 49–71.

Rahne, T., & Sussman, E. (2009). Neural representations of auditory input accommodate to the context in a dynamically changing acoustic environment. *European Journal of Neuroscience, 29*(1), 205–211.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Riecke, L., van Opstal, A. J., & Formisano, E. (2008). The auditory continuity illusion: A parametric investigation and filter model. *Perception & Psychophysics, 70*(1), 1–12.

Rimmele, J., Schröger, E., & Bendixen, A. (2012). Age-related changes in the use of regular patterns for auditory scene analysis. *Hearing Research, 289*(1–2), 98–107.

Rimmele, J., Sussman, E., & Poeppel, D. (2015). The role of temporal structure in the investigation of sensory memory, auditory scene analysis, and speech perception: A healthy-aging perspective. *International Journal of Psychophysiology, 95*(2), 175–183.

Rinne, T., Sarkka, 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, D. C., & Foxe, J. J. (2006). Preattentively grouped tones do not elicit MMN with respect to each other. *Psychophysiology, 43*(5), 423–430.

Ritter, W., Deacon, D., Gomes, H., Javitt, D. C., & Vaughan, H. G. (1995). The mismatch negativity of event-related potentials as a probe of transient auditory memory – A review. *Ear and Hearing, 16*(1), 52–67.

Ritter, W., Sussman, E., & Molholm, S. (2000). Evidence that the mismatch negativity system works on the basis of objects. *NeuroReport, 11*(1), 61–63.

Ritter, W., Sussman, E., Molholm, S., & Foxe, J. J. (2002). Memory reactivation or reinstatement and the mismatch negativity. *Psychophysiology, 39*(2), 158–165.

Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. New York: Springer.

Roberts, B., Glasberg, B. R., & Moore, B. C. J. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *Journal of the Acoustical Society of America, 112*(5), 2074–2085.

Rothermich, K., & Kotz, S. A. (2013). Predictions in speech comprehension: fMRI evidence on the meter-semantic interface. *Neuroimage, 70*, 89–100.

Roye, A., Jacobsen, T., & Schröger, E. (2007). Personal significance is encoded automatically by the human brain: An event-related potential study with ringtones. *European Journal of Neuroscience, 26*(3), 784–790.

Saarinen, J., Paavilainen, P., Schröger, E., Tervaniemi, M., & Näätänen, R. (1992). Representation of abstract attributes of auditory-stimuli in the human brain. *NeuroReport, 3*(12), 1149–1151.

Sammler, D., Kotz, S. A., Eckstein, K., Ott, D. V. M., & Friederici, A. D. (2010). Prosody meets syntax: The role of the corpus callosum. *Brain, 133*, 2643–2655.

Sams, M., Alho, K., & Näätänen, R. (1983). Sequential effects in the ERP in discriminating two stimuli. *Biological Psychology, 17*(1), 41–58.

Sams, M., Hari, R., Rif, J., & Knuutila, J. (1993). The human auditory sensory memory trace persists about 10 sec – Neuromagnetic evidence. *Journal of Cognitive Neuroscience, 5*(3), 363–370.

Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1124–1131.

Schadwinkel, S., & Gutschalk, A. (2011). Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus. *Journal of Neurophysiology, 105*(5), 1977–1983.

Scherg, M., Vajsar, J., & Picton, T. W. (1989). A source analysis of the late human auditory evoked potentials. *Journal of Cognitive Neuroscience, 1*, 336–355.

Schofield, B. R. (2010). Structural organization of the descending auditory pathway. In A. IRees & A. R. Palmer (Eds.), *The Oxford handbook of auditory science: The auditory brain* (pp. 43–64). Oxford: Oxford University Press.

Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National academy of Sciences of the United States of America, 106*(34), 14611–14616.

Schröger, E. (1995). Processing of auditory deviants with changes in one-stimulus versus two-stimulus dimensions. *Psychophysiology, 32*(1), 55–65.

Schröger, E. (1996). Interaural time and level differences: Integrated or separated processing? *Hearing Research, 96*(1–2), 191–198.

Schröger, E. (1997). On the detection of auditory deviations: A pre-attentive activation model. *Psychophysiology, 34*(3), 245–257.

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

Schröger, E., Bendixen, A., Denham, S. L., Mill, R. W., Böhm, T. M., & Winkler, I. (2014). Predictive regularity representations in violation detection and auditory stream segregation: From conceptual to computational models. *Brain Topography, 27*, 565–577.

Schröger, E., Marzecová, A., & SanMiguel, I. (2015). Attention and prediction in human audition: A lesson from cognitive psychophysiology. *European Journal of Neuroscience, 41*(5), 641–664.

Schröger, E., Tervaniemi, M., & Huotilainen, M. (2004). Bottom-up and top-down flows of information within auditory memory: Electrophysiological evidence. In C. Kaernbach, E. Schröger, & H. J. Müller (Eds.), *Psychophysics beyond sensation: Laws and invariants of human cognition. Scientific psychology series* (pp. 389–407). Mahwah, NJ: Lawrence Erlbaum Associates.

Schröger, E., & Winkler, I. (1995). Presentation rate and magnitude of stimulus deviance effects on human pre-attentive change detection. *Neuroscience Letters, 193*(3), 185–188.

Schubotz, R. I. (2007). Prediction of external events with our motor system: Towards a new framework. *Trends in Cognitive Sciences, 11*(5), 211–218.

Schwartz, J. L., Grimault, N., Hupe, J. M., Moore, B. C. J., & Pressnitzer, D. (2012). Multistability in perception: Binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society B – Biological Sciences, 367*(1591), 896–905.

Schwartze, M., Tavano, A., Schröger, E., & Kotz, S. A. (2012). Temporal aspects of prediction in audition: Cortical and subcortical neural mechanisms. *International Journal of Psychophysiology, 83*(2), 200–207.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology, 18*(9), 668–671.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences, 34*(3), 114–123.

Shinn-Cunningham, B. G., & Wang, D. (2008). Influences of auditory object formation on phonemic restoration. *Journal of the Acoustical Society of America, 123*(1), 295–301.

Shpiro, A., Moreno-Bote, R., Rubin, N., & Rinzel, J. (2009). Balance between noise and adaptation in competition models of perceptual bistability. *Journal of Computational Neuroscience, 27*(1), 37–54.

Siddle, D. A. T. (1991). Orienting, habituation, and resource-allocation – An associative analysis. *Psychophysiology, 28*(3), 245–259.

Sinkkonen, J. (1999). Information and resource allocation. In R. Baddeley, P. Hancock, & P. Foldiak (Eds.), *Information theory and the brain* (pp. 241–254). Cambridge: Cambridge University Press.

Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin, 133*(5), 780–799.

Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology, 25*, 545–580.

Steiger, H., & Bregman, A. S. (1982). Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics, 32*(2), 153–162.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111*(4), 1872–1891.

Stoffregen, T. A., & Bardy, B. G. (2001). On specification and the senses. *Behavioral and Brain Sciences, 24*(2), 195–213.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences, 13*(9), 403–409.

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *Journal of the Acoustical Society of America, 117*(3), 1285–1298.

Sussman, E. S. (2007). A new view on the MMN and attention debate – The role of context in processing auditory events. *Journal of Psychophysiology, 21*(3–4), 164–175.

Sussman, E. S., Gomes, H., Nousak, J. M. K., Ritter, W., & Vaughan, H. G. (1998). Feature conjunctions and auditory sensory memory. *Brain Research, 793*(1–2), 95–102.

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

Sussman, E. S., Ritter, W., & Vaughan, H. G. (1998a). Predictability of stimulus deviance and the mismatch negativity. *NeuroReport, 9*(18), 4167–4170.

Sussman, E. S., Ritter, W., & Vaughan, H. G. (1998b). Attention affects the organization of auditory input associated with the mismatch negativity system. *Brain Research, 789*(1), 130–138.

Sussman, E. S., Ritter, W., & Vaughan, H. G. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology, 36*(1), 22–34.

Sussman, E. S., Sheridan, K., Kreuzer, J., & Winkler, I. (2003). Representation of the standard: Stimulus context effects on the process generating the mismatch negativity component of event-related brain potentials. *Psychophysiology, 40*(3), 465–471.

Sussman, E. S., Winkler, I., Huotilainen, M., Ritter, W., & Näätänen, R. (2002). Top-down effects can modify the initially stimulus-driven auditory organization. *Cognitive Brain Research, 13*(3), 393–405.

Sussman, E. S., Winkler, I., Kreuzer, J., Saher, M., Näätänen, R., & Ritter, W. (2002). Temporal integration: Intentional sound discrimination does not modulate stimulus-driven processes in auditory event synthesis. *Clinical Neurophysiology, 113*(12), 1909–1920.

Sussman, E., Winkler, I., Ritter, W., Alho, K., & Naatanen, R. (1999). Temporal integration of auditory stimulus deviance as reflected by the mismatch negativity. *Neuroscience Letters, 264*(1–3), 161–164.

Sussman, E. S., Winkler, I., & Wang, W. J. (2003). MMN and attention: Competition for deviance detection. *Psychophysiology, 40*(3), 430–435.

Szalárdy, O., Bendixen, A., Böhm, T. M., Davies, L. A., Denham, S. L., & Winkler, I. (2014). The effects of rhythm and melody on auditory stream segregation. *Journal of the Acoustical Society of America, 135*(3), 1392–1405.

Takegata, R., Brattico, E., Tervaniemi, M., Varyagina, O., Näätänen, R., & Winkler, I. (2005). Preattentive representation of feature conjunctions for concurrent spatially distributed auditory objects. *Cognitive Brain Research, 25*(1), 169–179.

Takegata, R., Huotilainen, M., Rinne, T., Näätänen, R., & Winkler, I. (2001). Changes in acoustic features and their conjunctions are processed by separate neuronal populations. *NeuroReport, 12*(3), 525–529.

Takegata, R., Paavilainen, P., Näätänen, R., & Winkler, I. (1999). Independent processing of changes in auditory single features and feature conjunctions in humans as indexed by the mismatch negativity. *Neuroscience Letters, 266*(2), 109–112.

Takegata, R., Paavilainen, P., Näätänen, R., & Winkler, I. (2001). Preattentive processing of spectral, temporal, and structural characteristics of acoustic regularities: A mismatch negativity study. *Psychophysiology, 38*(1), 92–98.

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *Elife, 2*.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. *Journal of Neuroscience, 31*(1), 164–171.

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: An event-related potential and behavioral study. *Experimental Brain Research, 161*(1), 1–10.

Tervaniemi, M., Maury, S., & Näätänen, R. (1994). Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. *NeuroReport, 5*(7), 844–846.

Tervaniemi, M., Rytkönen, M., Schröger, E., Ilmoniemi, R. J., & Näätänen, R. (2001). Superior formation of cortical memory traces for melodic patterns in musicians. *Learning & Memory, 8*(5), 295–300.

Thompson, W. F., Hall, M. D., & Pressing, J. (2001). Illusory conjunctions of pitch and duration in unfamiliar tone sequences. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 128–140.

Tiitinen, H., May, P., Reinikainen, K., & Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature, 372*(6501), 90–92.

Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In V. Cutsuridis, A. Hussain, & J. G. Taylor (Eds.), *Perception-action cycle: Models, architectures, and hardware* (pp. 601–636). New York, Dordrecht, Heidelberg, London: Springer.

Todd, J., Heathcote, A., Mullens, D., Whitson, L. R., Provost, A., & Winkler, I. (2014). What controls gain in gain control? Mismatch negativity (MMN), priors and system biases. *Brain Topography, 27*(4), 578–589.

Todd, J., Heathcote, A., Whitson, L. R., Mullens, D., Provost, A., & Winkler, I. (2014). Mismatch negativity (MMN) to pitch change is susceptible to order-dependent bias. *Frontiers in Neuroscience, 8*, 180.

Todd, J., Provost, A., & Cooper, G. (2011). Lasting first impressions: A conservative bias in automatic filters of the acoustic environment. *Neuropsychologia, 49*(12), 3399–3405.

Todd, J., Provost, A., Whitson, L. R., Cooper, G., & Heathcote, A. (2013). Not so primitive: Context-sensitive meta-learning about unattended sound sequences. *Journal of Neurophysiology, 109*(1), 99–105.

Tong, X., McBride, C., Zhang, J., Chung, K. K. H., Lee, C.-Y., Shuai, L., et al. (2014). Neural correlates of acoustic cues of English lexical stress in Cantonese-speaking children. *Brain and Language, 138*, 61–70.

Treisman, A. M. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 353*(1373), 1295–1306.

Treisman, A. M. (1993). The perception of features and objects. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, & control. A tribute to Donald Broadbent*. Oxford: Clarendon Press.

Treisman, A. M., & Gelade, G. (1980). Feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

van Norden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. Eindhoven: Technical University.

van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190.

van Zuijen, T. L., Simoens, V. L., Paavilainen, P., Näätänen, R., & Tervaniemi, M. (2006). Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: An event-related brain potential study. *Journal of Cognitive Neuroscience, 18*(8), 1292–1303.

van Zuijen, T. L., Sussman, E., Winkler, I., Näätänen, R., & Tervaniemi, M. (2004). Grouping of sequential sounds – An event-related potential study comparing musicians and nonmusicians. *Journal of Cognitive Neuroscience, 16*(2), 331–338.

van Zuijen, T. L., Sussman, E., Winkler, I., Näätänen, R., & Tervaniemi, M. (2005). Auditory organization of sound sequences by a temporal or numerical regularity – A mismatch negativity study comparing musicians and non-musicians. *Cognitive Brain Research, 23*(2–3), 270–276.

Versnel, H., Zwiers, M. P., & van Opstal, A. J. (2009). Spectrotemporal response properties of inferior colliculus neurons in alert monkey. *Journal of Neuroscience, 29*(31), 9725–9739.

Vliegen, J., & Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *Journal of the Acoustical Society of America, 105*(1), 339–346.

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience, 32*(11), 3665–3678.

Weise, A., Grimm, S., Müller, D., & Schröger, E. (2010). A temporal constraint for automatic deviance detection and object formation: A mismatch negativity study. *Brain Research, 1331*, 88–95.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal, 3*(7), 45–52.

Wetzel, N., Schröger, E., & Widmann, A. (2013). The dissociation between the P3a event-related potential and behavioral distraction. *Psychophysiology, 50*, 920–930.

Widmann, A., Kujala, T., Tervaniemi, M., Kujala, A., & Schröger, E. (2004). From symbols to sounds: Visual symbolic information activates sound representations. *Psychophysiology, 41*(5), 709–715.

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology, 21*(3–4), 147–163.

Winkler, I., & Cowan, N. (2005). From sensory to long-term memory – Evidence from auditory memory reactivation studies. *Experimental Psychology, 52*(1), 3–20.

Winkler, I., Cowan, N., Csépe, V., Czigler, I., & Näätänen, R. (1996). Interactions between transient and long-term auditory memory as reflected by the mismatch negativity. *Journal of Cognitive Neuroscience, 8*(5), 403–415.

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. (2010). In search for auditory object representations. In I. Czigler & I. Winkler (Eds.), *Unconscious memory representations in perception: Processes and mechanisms in the brain* (pp. 71–106). Amsterdam and Philadelphia: John Benjamins.

Winkler, I., Denham, S. L., Mill, R., Böhm, T. M., & 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, S. L., & 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, R. J., et al. (2002). Temporary and longer term retention of acoustic information. *Psychophysiology, 39*(4), 530–534.

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., et al. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology, 36*(5), 638–642.

Winkler, I., Kushnerenko, E., Horváth, J., Ceponiene, R., Fellman, V., Huotilainen, M., et al. (2003). Newborn infants can organize the auditory world. *Proceedings of the National academy of Sciences of the United States of America, 100*(20), 11812–11815.

Winkler, I., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., & Näätänen, R. (1990). The effect of small variation of the frequent auditory stimulus on the event-related brain potential to the infrequent stimulus. *Psychophysiology, 27*(2), 228–235.

Winkler, I., Paavilainen, P., & Näätänen, R. (1992). Can echoic memory store two traces simultaneously? A study of event-related brain potentials. *Psychophysiology, 29*(3), 337–349.

Winkler, I., Reinikainen, K., & Näätänen, R. (1993). Event-related brain potentials reflect traces of echoic memory in humans. *Perception & Psychophysics, 53*(4), 443–449.

Winkler, I., & Schröger, E. (1995). Neural representation for the temporal structure of sound patterns. *NeuroReport, 6*(4), 690–694.

Winkler, I., Schröger, E., & Cowan, N. (2001). The role of large-scale memory organization in the mismatch negativity event-related brain potential. *Journal of Cognitive Neuroscience, 13*(1), 59–71.

Winkler, I., Sussman, E., Tervaniemi, M., Horváth, J., Ritter, W., & Näätänen, R. (2003). Preattentive auditory context effects. *Cognitive Affective & Behavioral Neuroscience, 3*(1), 57–77.

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., Teder-Sälejärvi, W. A., Horvath, J., Näätänen, R., & Sussman, E. (2003). Human auditory cortex tracks task-irrelevant sound sources. *NeuroReport, 14*(16), 2053–2056.

Winkler, I., van Zuijen, T. L., Sussman, E., Horváth, J., & Näätänen, R. (2006). Object representation in the human auditory system. *European Journal of Neuroscience, 24*(2), 625–634.

Woldorff, M. G., Hackley, S. A., & Hillyard, S. A. (1991). The effects of channel-selective attention on the mismatch negativity wave elicited by deviant tones. *Psychophysiology, 28*(1), 30–42.

Woldorff, M. G., Hillyard, S. A., Gallen, C. C., Hampson, S. R., & Bloom, F. E. (1998). Magnetoencephalographic recordings demonstrate attentional modulation of mismatch-related neural activity in human auditory cortex. *Psychophysiology, 35*(3), 283–292.

Woods, D. L., & Alain, C. (2001). Conjoining three auditory features: An event-related brain potential study. *Journal of Cognitive Neuroscience, 13*(4), 492–509.

Woods, D. L., Alain, C., & Ogawa, K. H. (1998). Conjoining auditory and visual features during high-rate serial presentation: Processing and conjoining two features can be faster than processing one. *Perception & Psychophysics, 60*(2), 239–249.

Yabe, H., Tervaniemi, M., Sinkkonen, J., Huotilainen, M., Ilmoniemi, R. J., & Näätänen, R. (1998). Temporal window of integration of auditory information in the human brain. *Psychophysiology, 35*(5), 615–619.

Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., et al. (2001). Organizing sound sequences in the human brain: The interplay of auditory streaming and temporal integration. *Brain Research, 897*(1–2), 222–227.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.

Zhuo, G., & Yu, X. (2011). Auditory feature binding and its hierarchical computational model. In H. Deng, D. Miao, J. Lei, & F. Wang (Eds.), *Artificial intelligence and computational intelligence* (pp. 332–338). Berlin, Heidelberg: Springer.

Zwislocki, J. J. (1969). Temporal summation of loudness – An analysis. *Journal of the Acoustical Society of America, 46*(2P2), 431–441.