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/

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.