Hearing Research 307 (2014) 98–110

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Review

Functional imaging of auditory scene analysis

Alexander Gutschalk*, Andrew R. Dykstra

Department of Neurology, Ruprecht-Karls-University Heidelberg, Heidelberg, Germany

ARTICLE INFO

Article history:
Received 21 May 2013
Received in revised form
26 July 2013
Accepted 8 August 2013
Available online 19 August 2013

ABSTRACT

Our auditory system is constantly faced with the task of decomposing the complex mixture of sound arriving at the ears into perceptually independent streams constituting accurate representations of individual sound sources. This decomposition, termed auditory scene analysis, is critical for both survival and communication, and is thought to underlie both speech and music perception. The neural underpinnings of auditory scene analysis have been studied utilizing invasive experiments with animal models as well as non-invasive (MEG, EEG, and fMRI) and invasive (intracranial EEG) studies conducted with human listeners. The present article reviews human neurophysiological research investigating the neural basis of auditory scene analysis, with emphasis on two classical paradigms termed streaming and informational masking. Other paradigms – such as the continuity illusion, mistuned harmonics, and multi-speaker environments – are briefly addressed thereafter. We conclude by discussing the emerging evidence for the role of auditory cortex in remapping incoming acoustic signals into a perceptual representation of auditory streams, which are then available for selective attention and further conscious processing.

This article is part of a Special Issue entitled <Human Auditory Neuroimaging>.

1. Introduction

At any given moment, our environment is comprised of multiple sound sources such that the sound arriving at our ear canals is a complex mixture, with acoustic energy from each source overlapping in both time and frequency with other sources. One of the primary functions of the human auditory system is to break down this mixture into individual sound elements that ideally, when grouped together, constitute all the elements produced by an individual source while excluding elements from all other sources. Subsequent sounds that bind together perceptually are referred to as an *auditory stream*, and the process of perceptual organization by *integrating* sound into auditory streams and *segregating* two or more streams from each other has been termed *auditory scene analysis* (Bregman, 1990).

Auditory scene analysis relies on a number of physiological processes, some of which are well studied in other contexts, and others which may be specifically related to perceptual organization. The quest for the neural mechanisms specifically related to auditory scene analysis has received increasing research interest in recent years, utilizing invasive animal models as well as non-invasive functional imaging and electrophysiological studies in human listeners. Previous reviews on auditory scene analysis have discussed in detail the research performed with animal models (Micheyl et al., 2007b) as well as the behavioral and mismatch-negativity (MMN) literature (Snyder and Alain, 2007). The present review will focus on functional imaging studies conducted in human listeners, streaming cues other than pure-tone frequency, and more complex sequence configurations.

The first section focuses on the auditory stream segregation (or streaming) paradigm, a classical paradigm often used to study basic, sequential source segregation. Although the stimuli themselves are quite simple, the streaming paradigm has nonetheless proven particularly fruitful in light of the fact that it produces bistable perception; that is, perception that changes despite identical stimuli. The use of bistability in examining the neural basis of perception, *per se*, is addressed in the following section. We then turn to more complex stimulation paradigms, where multiple tones are presented in variable configurations, focusing in particular on so-called multi-tone informational masking paradigms. We then

Abbreviations: ARN, awareness related negativity; BOLD, blood oxygenation level dependent; ΔF, frequency difference; EEG, electroencephalography; fMRI, functional magnetic resonance imaging; ISI, inter-stimulus interval; ITD, inter-aural time difference; MEG, magnetoencephalography; ROI, region of interest; SSR, steady-state response; STG, superior temporal gyrus

* Corresponding author. Department of Neurology, University of Heidelberg, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany. Tel.: +49 6221 56 36811; fax: +49 6221 56 5258.

E-mail address: Alexander.Gutschalk@med.uni-heidelberg.de (A. Gutschalk).

0378-5955/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.heares.2013.08.003

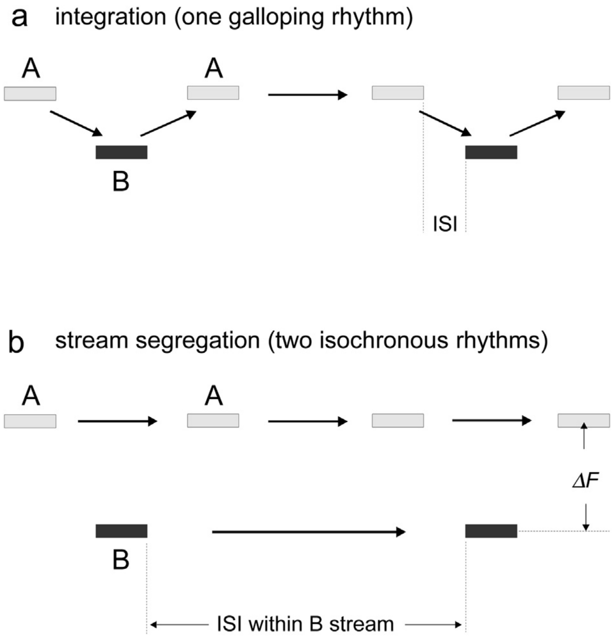
a integration (one galloping rhythm)
A
A
B
ISI
b stream segregation (two isochronous rhythms)
A
A
ΔF
B
ISI within B stream

Fig. 1. Example of the classical *ABA*_streaming paradigm introduced by Van Noorden (1975), where *A* and *B* are pure tones with a frequency difference Δ*F* and "_" is a silent pause. (a) The sequence is perceived as one stream with a characteristic, galloping rhythm when the Δ*F* is small. (b) At larger Δ*F*, the pattern is usually heard as two segregated, isochronous streams. In the latter case, the predominant perceptual interstimulus interval (ISI) within the *B*-tone stream is prominently longer than the time interval between the *B* tones and the leading *A* tones when both are integrated into one stream.

briefly address a number of other paradigms from the scene-analysis literature, before attempting a synthesis of the various findings in the final section. Based on the studies reviewed, we argue that auditory cortex may represent the major interface between faithful representations of acoustic stimuli and perceptual representations of auditory streams.

2. Auditory stream segregation

Stream segregation is a now-classical paradigm with which to study the segregation of temporally interleaved tone sequences. In the simplest case, two tones, *A* and *B*, continuously alternate in a regular *ABAB*… pattern. When the frequency difference (Δ*F*) between the tones is small and the presentation rate slow, the sequence is perceived as a single stream of alternating tones (a 'trill') (Miller and Heise, 1950). Conversely, when the rate is sufficiently high and the Δ*F* sufficiently large, two separate streams — one of *A* tones and another of *B* tones — are perceived. The latter phenomenon has been referred to as the streaming effect (Bregman, 1990).[1] Several variants of such patterns have been used in the auditory scene analysis literature, with the most ubiquitous being the *ABA_ABA_*… pattern (Fig. 1) introduced by Van Noorden (1975). This pattern produces a characteristic change of pattern perception that is well suited to instruct experimental listeners, because the rhythmic perception is less abstract than the theoretical explanation of what are one or two streams. When the *ABA_* pattern is perceived as one stream, it produces a distinct, galloping rhythm. When the pattern splits, two isochronous streams are perceived, one with double the rate (*A*) of the other (*B*). Apart from the modification of rate or rhythmic percept that can be effected by streaming, there are other, objective effects as well. For example, the separation of streams makes it more difficult to estimate the temporal relationship between two sound elements, even if they are adjacent in time, if they do not belong to the same stream (Bregman and Campbell, 1971; Vliegen et al., 1999b).

2.1. Computational models for stream segregation

The earliest neuronal models purporting to explain stream segregation suggested that segregation of pure tone sequences can be explained based on neuronal representation distance along the frequency axis of the cochlea — the so-called peripheral channeling hypothesis (Hartmann and Johnson, 1991) — along with an additional temporal integrator in the central nervous system (Beauvois and Meddis, 1996; McCabe and Denham, 1997). Multi-unit recordings in macaque monkeys suggested that frequency separation in the auditory cortex is modulated by forward suppression (Brosch and Schreiner, 1997), such that the separation of the individual neuronal representations of the different stimuli is enhanced at shorter inter-stimulus intervals (ISI) (Bee and Klump, 2004; Fishman et al., 2004, 2001), potentially explaining why stream segregation can also be observed with smaller Δ*F* at faster rates and shorter ISIs (Bregman et al., 2000; Van Noorden, 1975). Furthermore, streaming often is not an instantaneous percept, but may build-up over the course of seconds (Anstis and Saida, 1985; Bregman, 1978). Adaptation processes with longer time constants than forward suppression have been proposed to explain this gradual buildup of streaming-related activity in auditory cortex that is often observed at intermediate Δ*F* (Micheyl et al., 2005). Similar multi-second adaptation has later been observed as early in the auditory pathway as the cochlear nucleus (Pressnitzer et al., 2008). The models used to explain streaming based on a separation of streams into distinct neuronal representations are generally summarized as the *population-separation model* of auditory stream segregation (Fishman et al., 2012; Micheyl et al., 2007b). The population-separation model of stream segregation goes beyond the previous peripheral channeling model in also considering neuronal representations of other feature representations than ear and tone frequency, which supposedly emerge in the central auditory system. The adaptation phenomena described above are an additional component of the model to explain temporal phenomena such as buildup and rate dependency.

While the population-separation model of stream segregation can explain the classical streaming effect introduced above, it may not be universal enough to explain why stream segregation does or does not occur with other stimulus configurations. For example, it has been pointed out that the separation of two streams of tones along the tonotopic axis in auditory cortex was similar for alternating and synchronous pure-tone sequences, but that the synchronous sequences are generally perceived as a single coherent stream of chords (Elhilali et al., 2009b). In the framework of Bregman (1990), one would argue that the common onsets of the synchronous sequence are a stronger cue for integration than frequency separation is for segregation. An alternative model that accounts for synchronicity cues and additionally for other temporal characteristics of auditory objects was introduced as the *temporal coherence model* of auditory stream segregation (Elhilali et al., 2009b, Shamma et al., 2011). This model adds a module subsequent to the separation of sounds into different neuronal populations (i.e. feature extraction), which computes the coherence between stimulus-locked activity in all neural channels in a time interval of up to 500 ms. Sound elements with high coherence are

[1] Note that the term *streaming* has alternatively been used to characterize any kind of sequential grouping in auditory perception. Following Bregman (1990, page 47), we will only use streaming in the context of the classical, alternating stream segregation paradigm in this paper.