Object-Based Attention in Naturalistic Auditory Streams

Giorgio Marinato

2019/2020

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

CIMeC – Center for Mind and Brain Sciences
University of Trento

Supervisor: Prof. Dr. Daniel Baldauf

Table of Contents

1 General Introduction

1.1 Introduction

In many ecological environments, we find ourselves surrounded by complex auditory soundscapes. These auditory scenes are often composed of many concurrent sound sources with both spatially and temporally overlapping spectral details. Imagine for example an underground metro station with people waiting for the train, other people running to reach the next connection, others talking on the phone, a group of teenagers enthusiastically narrating anecdotes to each other, a street performer singing a bit further away, and a loud-speaker voice announcing the imminent arrival of a train. Many of those signals may instantaneously catch our attention. This rather complex auditory scene resembles only one such possible typical situation in everyday life, and humans are consistently able to parse such overlapping signals seemingly effortlessly in order to navigate their surroundings. This capacity was first outlined in the form of a seminal paradigm, the so-called “cocktail party problem” (Cherry, 1953) which resulted in a novel experimental approach that is still one of the most dominant and fruitful approaches in the study of auditory perception. The concept was first introduced to depict the specific situation of a multi-talker environment - like a cocktail party, in which one has to select a specific speech input, suppressing other competing and distracting speech signals. However, under the umbrella of the “cocktail-party” label, the scientific work has actually proceeded along several different lines of research (Bronkhorst, 2000; Bronkhorst, 2015), which - according to Mc Dermott (2009) - can be traced back to two conceptually different, fundamental perspectives: *sound segregation* (or “auditory scene analysis”) and *attentional selection* (first introduced by Cherry). Attention, in every sensory domain, plays a fundamental role to efficiently select the relevant information and ignore the distracting inputs (Posner, 1980). Many seminal studies in vision (e.g., Treisman & Gelade, 1980; Desimone & Duncan, 1995) have shown that attention operates in form of a “biased competition” between

neural representation of perceptual objects, which were depicted as the central "units", on which non-spatial selective attention acts in many natural contexts (see, for example Baldauf & Desimone, 2014; Duncan, 1984; O'Craven, Downing & Kanwisher, 1999).

In the auditory domain, research has focused for a long time on sound segregation mainly applying bottleneck theories of selective attention, which propose that a strict "selective filter" (Broadbent, 1958) limits the amount of sensory information that the system can process. In Broadbent's theory a listener uses physical parameters such as location, pitch, loudness, and timber to intentionally select or filter out information. Studies using the cocktail-party paradigm instead revealed with subsequent observations that specific sound features, meaning (Treisman, 1960), or affective cues, like our own name, (Moray, 1959; Wood & Cowan, 1995) often become inherently salient and effectively capture the focus of attention and cause our attentional to switch without volition. The above observations led to a reconsideration of the Broadbent filter, aliking more a threshold system that enhances the signals of interest (or attenuate the unwanted signals) "rather than acting as an all-or-none barrier" (Treisman, 1960). It was Treisman herself who in the visual domain developed the influential Filter Integration Theory (FIT) (Treisman & Gelade, 1980) shifting the paradigm towards the top-down cognitive process of selective attention considered as the "glue" that binds features together to create perceived objects. In contrast to the Gestalt theory, which claimed that the whole precedes its parts, Treisman and Gelade proposed that "features are registered early, automatically and in parallel across the visual field, while objects are identified separately and only at a later stage, which require attention" (Treisman & Gelade, 1980). The later and recent auditory research heavily draws from these insights from vision sciences in terms of *auditory object formation* and *auditory object selection.* Together the two processes constitute different aspects of how auditory selective attention may operate and contribute to our ability to precisely select the signal of interest at any given time, balancing top-down goals with stimuli salience and previous statistical knowledge of the real word auditory

scenes. However, the interaction between *auditory* selective attention and object formation itself remain a subject of debate.

In this dissertation, I will focus on object-based auditory selective attention, investigating specifically the *object selection* component of auditory attention processes. The aim of the thesis is to provide new insights that support the hypothesis that objects are the “units” of attention, also in auditory domain, therefore contributing to the debate about the cocktail party problem and the mechanisms human beings use to give meaning to sound scenes. The scene processing at object level is indeed of fundamental interest for a higher level comprehension of the soundscapes, because of its complementary inner work involved in the listening to rich complex natural sound environments. The “other half” of the work is thought to be performed by typical “scene analysis” mechanisms such as voluntary and involuntary learned schemas (Woods & McDermott, 2018) or “primitive auditory scene analysis” strategies (Bregman, 1993; McDermott, Wrobleski & Oxenham, 2011).

The strength of the current work is threefold. First, a specific behavioral paradigm – an auditory repetition detection task - has been designed combining a Posner cueing task (Posner, 1980) with an n-back task (Kirchner, 1958) to highlight the object-based aspect of the auditory selective attention. Such a repetition detection task enforces the listeners to fully process the ongoing acoustic stream to an extent of cognitive processing that allows them to recognize a specific acoustic snippet with its low-level properties as an auditory unit (object) and to interpret it as a direct repetition. Crucially, this type of repetition detection task cannot be accomplished by focusing on just one distinct *low-level feature* (e.g., a certain pitch) itself or by making use of spatial information since both streams to the participants from the same external source. Second, carefully crafted stimuli had been employed to resemble as much as possible real-world soundscapes, made of a mixture of environmental sounds created by humans and speech signals (also known as anthropophony). The use of more complex, real-world auditory scenes serves the purpose of bridging the gap between laboratory and life that has been pointed out as one of the

challenges of the current research, for auditory attention in particular (Shinn-Cunningham et al., 2015). Actually, in stark contrast to visual perception, auditory sources are inherently temporal and the cognitive mechanisms underlying the solution of cocktail party tasks that use simple stimuli like pure tones, do not necessarily reflect or guarantee that the same basic mechanism is at work to solve the auditory scene analysis also in real-world settings.
Third, the use of magnetoencephalography (MEG), which directly measures the magnetic field of neurons non-invasively from outside the scalp, guarantee a perfect temporal resolution because the signal is distorted by transitions through tissue with different electric conductivities, like the dura mater, the skull, and the scalp and skin. Thanks to the mathematical advancement of the source reconstruction methods it is also possible to achieve a very good spatial resolution. Therefore MEG is a particularly suited method to study the neural temporal dynamics of the temporal dynamics of auditory selective attention in a cocktail party situation - like also recently collected evidence suggests (e.g., Mesgarani & Chang, 2012; Zion Golumbic et al., 2013a; Lee et al., 2014; Simon, 2015a).

In the subsequent paragraphs of this first introductory chapter, the concepts of auditory object, *auditory object formation*, *auditory object selection* are going to be discussed in greater details, to present at the end of the section the state of the art of the neural signatures of auditory selective attention.
In the second chapter, I will discuss in detail several behavioral tasks that are already available to investigate the object-based attention in auditory domain. The reasons that motivated the design of a novel task and stimulus set, which were specifically designed for this thesis project, will be supported by the presentation of the results of the two behavioral experiments conducted.
The third chapter investigates the neural correlates of auditory object-based attention perspective and is dedicated to the presentation of our MEG study.
Chapter four, finally, summarizes the results, and discusses the overall findings in a broader context.

1.2 Attention, a brief overview

There exists currently a wide range of diverse theories and approaches regarding the concept of *attention*, which developed historically often as metaphors and analogies in visual sciences first. Some of those became the starting blocks also for the latest developments in *auditory attention* research. It is therefore of particular interest for the next sections, to introduce an overview of the most relevant conceptualizations of attention.

As living systems, the humans process signals coming from the external environment as well as from their internal states. As every information processing system, the resources available at any moment, pose a dynamic constrain on the effective capacity of the system. Here, attention plays a crucial role in the orchestrated allocation of limited information processing resources in order to maximize the efficiency of the system. This is achieved in various forms within the different sensory modalities and domains, studied from cognitive psychologists and neuroscientists for many decades.

In the past, attention has been described to work at various levels and in various modes. Whereas Broadbend (1958) proposed attentional selection to occur on an early processing stage - *Early Selection,* (see Broadbent, 1958) - Deutsch and Deutsch provided evidence for a *Late Selection* process (Deutsch & Deutsch, 1963). Treisman prominently expanded Broadbend's ideas to the *Attenuator Theory of Attention* (Treisman, 1960; Treisman, 1964), and later well-known as *Feature Integration Theory* (Treisman & Gelade, 1980).

Shulman and colleagues characterized spatial attention in the form of a

metaphor of a *Spotlight* (Shulman, Remington & Mclean, 1979) or zoom-lense, focusing spatially on important parts of the scene and leaving other, irrelevant parts in the dark or out-of-focus. Also Rizzolatti and colleagues followed this concept closely by integrating attentional selection in their *premotor theory* (Rizzolatti et al., 1987).

Later on, Duncan and Desimone started describing mechanistic understanding of attentional facilitation in their *biased-competition model* (Desimone & Duncan, 1995), postulating for the first time an *object-based nature of visual attention* (Duncan, 1984), or even as a kind of *'shrink-wrap'* type of mechanism (Moran & Desimone, 1985). Also a form of non-spatial attention, feature-based approaches were based on experimental findings of attention to a certain property, e.g., a color, spreading over the whole visual field, even to parts of the scene that were not behavioral relevant, what led to the *feature similarity gain theory* (Treue & Trujillo, 1999). Among all these approaches the query of the underlying mechanisms of visual attention focus on empirical phenomena where visual attention can modulate the *sensitivity* of early perceptual filters (both in space and time), or where attention influences the *selection* of stimuli of interest: How and at which level of the processing hierarchy are representations modulated (in the sense of facilitation and inhibition), and what are the neural computations underlying the selection processes?

1.2.1 Spatial attention: Posner's network of attention

In the context of covertly orienting the focus of processing in space (covert spatial attention) we mainly refer to Posner's terminology (Posner, 1980; Posner & Petersen, 1990) characterize attention as a combination of facilitation and inhibition. From neuroimaging studies we gained knowledge about the holistic structure of the neural networks dedicated to the selective processing of incoming information with the identifications of three specific networks related to

distinct aspects of attention: alerting, orienting, and executive control (Posner & Petersen, 1990; Raz & Buhle, 2006). The alerting network maintains a state of altered responsiveness in preparation to the entering stimuli, is task specific and was experimentally found to be linked to frontal and parietal regions, particularly of the right hemisphere (Assmus et al., 2005; Critchley et al., 2003). It is closely related to the efficiency of a system and usually studied through a paradigm that subtracts conditions that give temporal, but not spatial, information from a neutral condition without any cue presented (Fan et al., 2002).

The orienting network is the capability to shift the inner focus to various parts or aspects of a scene, and it can function either ‘overtly’ or ‘covertly’, and either based on ‘exogenous’ or ‘endogenous’ control signals. It is associated with a widespread dorsal network involving both posterior brain areas, including the superior parietal lobe, the temporal parietal junction and frontal sites, like the frontal eye fields (Kastner et al., 2004; Martínez et al., 1999; Ungerleider, 2000). Traditionally the orienting network has been investigated by measuring the reduction in responses’ reaction time to a target following a cue, which gives valuable information about the likely position in space, but not about the point in time of such a future event. Valid trials show benefits of orienting to a correctly cued location, compared to a no-cue condition, as well as substantial reaction time costs in response to incorrect cues, i.e. responses in the invalid trials compared to and a neutral, ‘no-cue’ condition (Posner, 1980; Treisman & Gelade, 1980).

Endogenous (top-down) and exogenous (bottom-up) orienting of the attentional focus, usually enhance performance by increasing the neural activity of the corresponding sensory system. More specifically, according to the biased competition model (Desimone & Duncan, 1995; Kastner & Ungerleider, 2001) the control system interacts with the sensory information to ensure the faster and more accurate representation of the target.

The executive network, in contrast, is thought to monitor and solve conflict between computations in different neural areas. Therefore, it is conceptualized to involve planning and decision-making stages as well as error detection and

supervision. Imaging studies have identified a prominent role of the anterior cingulate cortex (ACC) in processing of cognitive dissonance. From this literature, it is not yet completely clear if the ACC's role is more in monitoring conflicts or in resolving them, but its function seems to be closer to the response side than the perceptual side of the underlying sensory-motor-transformation (Botvinick, Cohen & Carter, 2004; Bush, Luu & Posner, 2000; Fan et al., 2003; Kerns et al., 2004). Executive attention, finally, is often operationalized by measuring some sort of incompatibility in certain stimulus -response mappings. The most classic examples here are the Stroop task and the Simon tasks (see Eriksen & Eriksen, 1974; Simon, 1969; Stroop, 1935).

1.2.2 Non-spatial attention: feature-based and object-based mechanisms

The domains of attention conceive spatial and non-spatial attention, the latter of which can operate as feature-based and object-based attention. In the case of feature-based attention, resources are selectively deployed to specific visual properties (e.g., a color, a dominant orientation, or a dominant motion direction) present in the environment, independently of their spatial location. Object-based attention, on the other hand, is guided by object structure (Störmer, Cohen & Alvarez, 2019a). Independently from the domain of attention we can also distinguish between exogenous and endogenous modes. Exogenous modes allocate attention in a bottom-up driven fashion from the sensory world, and is mostly a reflexive response based on strong salience or previous experience (eg., Störmer, McDonald & Hillyard, 2019b). Endogenous modes, however, allocate attention based on internal goals, local contingencies and tasks.

The spatial domain of attention was mainly studied in the context of vision and can be considered from two subdomains: overt and covert attention. Overt attention manifests as explicit eye movements that place the visual information at the center of the fovea, where sensitivity is higher. Covert attention, in contrast, manifests as processing of localized stimuli even in absence of eye movements.

A powerful and typical behavioral paradigm (Theeuwes et al., 1998) to study the overt spatial attention use six disks around a fixation point and when five change colors, the participant should saccade to the singleton object. In some trials an extra distracting object is added. Accuracy of eye movements and the speed of saccadic responses are measured in trials without the distractor, as a baseline performance which is compared to the performance of trials with the distractor present. The error rate of around 30% in trials with distractor, represent the measure of the two competing mechanisms of attention (endogenous and exogenous) while performing the task (Trappenberg et al., 2001).

The spatial cueing paradigm developed by Posner has been the elective tool to study the covert spatial attention in which participants respond to targets that are located peripherally from fixation and preceded by cues. When the exogenous mode is the focus of the exploration, the cue appears randomly at any possible target location. However, when the focus is on the endogenous mode cues are presented at fixation and probabilistically predict the location of the behavioral target.

Within the non-spatial domain, it has been demonstrated that attentional shifting and selection can be successfully deployed to specific features of the stimuli of interest (Maunsell & Treue, 2006; Moore & Egeth, 1998) selecting objects on the base of feature, like color or size (Schwedhelm, Baldauf & Treue, 2017; Schwedhelm, Baldauf & Treue, 2020). Such a mechanism has been shown to enhance the neural response also for stimuli that are spatially distant or irrelevant to the task if they share the same feature (Saenz, Buracas & Boynton, 2002). It works both as a bottom-up process with an effect of automatically priming the system to a feature, and as a top-down mechanism (Theeuwes, 2013). Often it is complementary to the spatial-based mechanisms (Liu, Stevens & Carrasco, 2007a) with the distinction that when the location is involved in feature-based attention, the location is also selected by features and not strictly by location properties.

1.3 Auditory attention

1.3.1 Auditory object-based attention

A specific object-based perspective on auditory attention could emerge from the work of Shinn-Cunningham (2008), supporting the idea of extending the theories of visual attention. Shinn-Cunningham defines an auditory object as a perceptual entity coming from one physical source and she argues that the auditory object formation takes place, similarly to vision, at various different analysis scales. These scales are thought to move in a non-hierarchical manner from grouping at the level of local structures (Bregman, 1990) to organization at longer spatial and temporal scales, influenced by top-down attention. In the presence of complex scenes as visual input, it has been shown in visual studies that there is a object-based advantage in processing the various inputs (Desimone & Duncan, 1995; Baldauf & Desimone, 2014). Since attention can work in an object-based mode, several simultaneously present sound sources in a complex auditory scene can compete and perceptually interfere in several different ways. The first one is the energetic masking, which can be understood in a simple way as an overlapping of sounds signals, both in time and frequency (Cooke, 2006). Other forms of masking can be on the level of more high-level information (informational masking) with all sorts of non-energetic masking effects. These two dimensions lead to two types of failure in identifying auditory objects, and consequently parsing effectively the soundscape. The failures of object formation occur whenever the local arrangement is deficient for separating diverse signals from each other (Best et al., 2007). The failures of object selection occur when the listener directs his or her attention on the wrong object either because they don’t know what exactly the target is or because the

bottom-up salience of competing sources is stronger than the top-down goal-oriented attention (Buschman & Miller, 2007).

The past ten years or so has seen an increasing number of studies focusing on auditory object formation and object selection, two processes that are thought to happen in an 'heterarchical' way – instead of sequentially 'hierarchical' – influencing each other at various stages both in terms of auditory scene analysis as well in terms of neural processing necessary to successfully parse the soundscape. Regarding the auditory object formation, studies focus both on local spectro-temporal cues where energetic masking has a great impact (e.g., Ihlefeld & Shinn-Cunningham, 2008; Maddox et al., 2015) as well as on higher order features that unfold over time and form the auditory "streams" (Best et al., 2008; Bressler et al., 2014). Regarding the auditory object selection, studies mostly focus on the interaction between the top-down control and bottom-up salience and the way it influences the selective attention at the object level. Top-down attention can be directed to different acoustic dimensions that influence the object formation, e.g. pitch (Maddox & Shinn-Cunningham, 2012) or sound level (Kitterick et al., 2013). Inherent salience of the sounds in terms of features and statistics (Kaya & Elhilali, 2014; Kaya & Elhilali, 2017) have proven to also affect selective attention.

1.3.2 What is an auditory object?

If auditory selective attention is object-based, it is important to first attempt a theoretical and operational explanation of what an auditory object really constitutes - taking into account the consensus which emerges from the literature and the diverse approaches of the sound's segregation and object formation (see following chapter).

In contrast to visual domain, in which the boundaries of an object can be clearly defined in three dimensions, in the auditory modality, the dimensionality of representations is less clear. Both temporal or spectral characteristics allow to describe a sound, and the overall shape of those characteristics (being it in time or frequency domain) in itself could constitute the border of an auditory object

(Griffiths & Warren, 2004). For example, auditory objects are inherently temporal, therefore onset synchronicity is an important parameter but not sufficient to segregate sounds if they both have the same onset.

In analogy to visual objects, an auditory object can be conceived as a specific composition of low-level features (e.g. temporal and frequency continuity, harmonicity), which then becomes a grouped unit. An auditory scene or 'soundscape' could then consist of several such auditory objects combined together, or temporally overlaid, in form of a superordinate entity of individual objects (Bizley & Cohen, 2013), e.g., the characteristic soundscape of a train station, or the inter-mixed conversation at a party. In such complex auditory settings, the individual acoustic objects are temporally confined and bound entities, e.g., the individual words in a conversation or a loud-speaker announcement in the soundscape of a train station. Bregman (1990) with his seminal work on auditory scene analysis already defined some of the rules for the perceptual organization of sound mixtures, that in the past ten years have been subject of an updated conceptualization of attention. To successfully give meaning to an auditory scene, humans bind sounds at a "local" scale as well as at longer time scales (Bizley & Cohen, 2013). The "local" binding happens at a time scale of milliseconds: Here, the sounds are grouped based on the spectro-temporal features like time-frequency proximity and correlated fluctuation in amplitude modulation as these are the strongest cues at this time scale. This stage of object formation has also been named "syllable-level" (Shinn-Cunningham, Best & Lee, 2017). The higher-order grouping happens instead at a longer timescale of possibly up to several seconds, with the formation of longer auditory objects defined as "streams" (Bizley & Cohen, 2013; Bregman, 1990). At this stage the auditory objects formed at the "syllable-level" are grouped into coherent streams mostly influenced by features like location, pitch, and timber.

Importantly, several previous studies have attempted to define what constitutes an auditory object. For example, by describing the rules that guide the formation of an auditory object in front of (and in contrast to) background noise.

However, no clear definitional consensus on how the diverse mechanisms work together has yet been reached (Bregman, 1990; Griffiths & Warren, 2004). A prominent and persuasive operational definition has been recommended by Griffiths and Warren (2004). In their view, an auditory object is defined as (i) a sound structure that has a real correspondence in the sensory world, that (2) 'can be isolated from the rest of the sensory world', and (3) that 'can be recognized or generalized beyond the single particular sensory experience'. the authors also point out that object analysis would also imply some level of generalization across different sensory modalities, such as the correspondence between the auditory and visual domain (Shinn-Cunningham, 2008). Also in cognitive-neuroscience studies investigating neural correlates of object-based processing, this operational definition has been very prominent (Carlyon, 2004; Simon, 2015a).

1.3.3 Auditory object formation

As stated previously object formation works in two rather extended time scales: one "local" that binds together the sound features that are spectro-temporally connected; one "longer" and of higher-order character that groups together sounds that emerge through time and that form what Bregman defined as "streams" (for review see Bizley & Cohen, 2013; Carlyon, 2004; Griffiths & Warren, 2004).

The "local" scale, at which spectro-temporal features are grouped together to form sounds, is also called "syllable-level", because is derived from the slow oscillation rhythm of the language syllable (Shinn-Cunningham et al., 2017). The object binding at this level is generally based on correlated fluctuation in amplitude modulation, a spectro-temporal proximity that is realized if the sounds are continuous in time and/or frequency. Spatial cues at this level of analysis have minor contribution importance to the object formation and are used mainly if the spectro-temporal cues are ambiguous (Ihlefeld & Shinn-Cunningham, 2008;

Schwartz, McDermott & Shinn-Cunningham, 2012). A possible explanation is that separating auditory sources in space requires comparing the inputs of both ears. This implies that it requires more processing time, too - while amplitude cues happen at the periphery of the sound representation. Also harmonic cues have a weak contribution at the syllable level to the object formation (Darwin, 1997). Superfically, these local spectro-temporal grouping cues, both strong and weak, all reflect probabilistic coincidences in acoustic spectro-temporal structure that happen to occur whenever a sound's energy is originating from a defined source in space. Consequently, sound components in a real-world scenario typically have coincidental structures of their envelopes. But even if there are no dominant cues present on how to group objects in a scene, just the repetition of acoustic components can make them to form as objects (McDermott & Simoncelli, 2011).

When grouping at a longer time scale syllable-level features into streams, this requires higher order perceptual features. For example, in a sequence, the continuity or similarity of "syllables" like pitch, timbre, amplitude modulation, and spatial location contribute to the sensation that a sound persists as a single ongoing source (Best et al., 2008; Maddox & Shinn-Cunningham, 2012). The role of selective attention in the object formation is still heavily debated. It has been suggested that auditory objects form only if a stream, i.e., an auditory object that persists over a certain time window, is attended(Alain & Arnott, 2000; Alain & Woods, 1997; Cusack et al., 2004a); others suggest that auditory streams form automatically and pre-attentively (Macken et al., 2003; Sussman et al., 2007). Most likely both factors contribute in an heterarchical way, helping the formation of the auditory objects (Carlyon et al., 2003).

1.3.4 Auditory object selection

The object formation and the grouping of objects in coherent streams are the most important prerequisite of auditory object-based attention. If object formation fails, there is no object to select, therefore the object selection process

cannot take place. However, when the formation of objects and streams succeed accurately, human beings immersed in a complex auditory mixture can profitably select what to attend (both in form of objects and streams), and they will have to do so due to their innate limitation in the capacity of analyzing the entire soundscape and give meaning to it all at once. Moreover, such comprehensive analysis is not necessarily the scope of the everyday communication.

Especially at the level of perceptual objects, in vision it is thought that attention modulates activity in form of a “biased-competition” (Desimone & Duncan, 1995). The biased-competition view argues that an attended object is processed preferentially and in much greater detail than other, competing objects on the display. Both the salience of the various stimulus objects seems to determine the ongoing competition for resources (‘’exogenous attention), and internal goal representations (‘endogenous’ attention). Some experimental evidence of the similar effects in the auditory domain have started to emerge (see, e.g., Mesgarani & Chang, 2012).

Top-down attention can be directed to various dimensions of the acoustic scene, which in turn can influence the object and stream formation. Anybody who listens to such a acoustic scene can deploy his or her top-down attention to a various frequency regions (as shown, e.g., by Cusack et al., 2004), a certain spatial location (Kidd et al., 2005), to a pitch (Maddox & Shinn-Cunningham, 2012), to acoustic level (Kitterick et al., 2013), to timber (Darwin, 1997), or to time . The above studies support the idea that the interaction between formation and selection are two non-hierarchical inextricable processes, however the indications that objects can also be the units of auditory attention come from more recent studies that take into consideration the concept of “streams”. For example, if a listener deploys his attention to a given word, subsequent words which have some perceptual features in common, are statistically more often attended as well (Bressler et al., 2014). Another example, in a task that involves the detection of sequences of digits, the spatial continuity and continuity of target voice led to benefits in the selectivity of attention across time. Such results speak in favor of the fact that attentional selectivity often becomes even stronger over

time, as attention is directed towards an auditory object embedded in a rich and complex scene (Best et al., 2008).

1.3.5 Objects and features in auditory attention

The last two paragraphs outlined the role of selective attention into the interlaced processes of object formation and object selection in order to give meanings to rich natural complex soundscapes. However, as explained, auditory objects are not manifestly physical entities that exist in the physical space independently from our necessity to conceptually organize the world. What exists in the physical world are primitive fundamental properties - features - independent from our perception organization (Bregman, 1993). These properties belong to all the sounds in the natural environments and consist of general acoustic regularities that help decompose the signal with or without attention deployed. These regularities were described extensively (Bregman, 1990) and briefly consist of: a) differences in temporal onset of the sounds; b) slow and smooth variations of the properties of the same sound or of multiple sounds from the same source; c) environmental sounds harmonicity; d) changes on one sound's components that keep all the components of the signal bonded together. Critically, the regularities of the soundscape can be used as bottom-up cues and can be also built in schemas that can be easily learned from the listener (Woods & McDermott, 2018). A particular property that stands apart as grouping cue, not requiring any prior knowledge of the characteristic of the signal, is the inherent repetition of a sound (McDermott et al., 2011).

Therefore there is evidence that specific regularities of complex sound environments can successfully help navigate the physical space preattentively in certain circumstances. Crucially, when a higher-level of understanding - or meaning attribution to the scene - is necessary, humans can recruit the same regularities to process the sound scene at the object level (Best et al., 2008; Shinn-Cunningham, 2008).

1.4 Auditory brain imaging with magnetoencephalography (MEG)

1.4.1 MEG signal

Among the non-invasive tools we have at our disposal to gain insight into neural correlates of the human cognitive processes, magentoencephalography (Cohen, 1972) is one of the most powerful and ideally suited to study the auditory system for its temporal and spatial resolution, and for the diverse types of information that can be extracted from the recorded signal thanks to the different range of analysis and ever evolving innovative computational approaches (Gramfort et al., 2013; Oostenveld et al., 2011; Tadel et al., 2019).

Magentoencephalography systems measure the magnetic fields which directly stem form the electric discharge neurons produce when they communicate with each. These electric currents generate extremely small electromagnetic fields (i.e. in the range of tens to hundreds of femto-Tesla (fT)) that can be measured through highly sensitive detectors – superconducting quantum interference devices (SQUIDs, see Zimmerman, Thiene & Harding, 1970)- when embedded in a magnetically shielded environment.

The structure of the pyramidal neurons consists of cell body, an axon and basal and apical dendrites which branch out in the typical tree-like shape. The long apical dendrites extend perpendicularly to the cortical surface in such a way that their magnetic fields often sum up to magnitudes large enough to be detected (Nagarajan & Sekihara, 2019).

The changes in the membrane potential of the neurons are of two kinds: action potentials and postsynaptic potentials. The latter are considered to be the main

contributors to the measurable MEG signal (Murakami & Okada, 2006) because their currents are much more prolonged (duration ~10 ms) in contrast to action potentials that decay very fast (duration 1 ms). The slower post-synaptic potentials decay therefore facilitate the temporal summation (overlap in time) and consequently the strength of the magnetic field. The measurable MEG signal thus originates from a population of cortical pyramidal neurons (Baillet, 2017).

Since MEG data can be acquired at sub-millisecond timescale, the temporal resolution of MEG imaging is only restrained by the sampling rate, that is typically ~1 kHz. With high enough sampling frequencies, neural oscillations can be registered up to about 500 Hz. The spatial resolution in contrast is more variable and depends on which reconstruction method is used and related parameters (e.g., co-registration errors). In general the reconstruction accuracy of the cortical sources of the magnetic signal can be small as 3 mm with an error of the order of 3 mm (Roberts et al., 2000).

The intrinsic characteristics of the MEG signal explained above are therefore very helpful for the investigation of the temporal dynamics of cortical neural activity and what cognitive functions they relate to, it provides a multidimensional resolution comprising time, space, frequency, as well as power and phase of a given frequency band (Hämäläinen et al., 1993; Lopes da Silva, 2013).

1.4.2 Neural oscillatory activity

Neural oscillations, also described as brain rhythms (Buzsaki, 2006) can only be studied when neural activity is recorded with sufficiently high temporal resolution. From the rhythmic activity of the oscillations emerge an organizational principle of the brain that is thought to be important for inter-neuronal communication and the coordination orderly behavior. Synchronous activation happens continuously within a group of neurons and among various functionally specialized areas of the brain (Varela et al., 2001), giving rise to rich network dynamics with

synchronized activity in many different sites of the human brain.

Importantly, synchronized brain activity is a crucial indicator of information transmission between sites. If two neuronal groups are consistently correlated or co-activated over time we can infer from such neural oscillation that information is exchanged between them (see, e.g., Fries, 2015). Neural oscillations are in general characterized by their amplitude, or power, i.e. the relative signal strength, - and phase, i.e. their rhythmic up and down cycles. Phase estimates can be retrieved from either Fourier or Hilbert transforms (Bertrand, Tallon-Baudry & Pernier, 2000; Cohen, 2011), and are crucial for the computation of neural synchrony in form of phase-locking and neural coherence between groups of neurons (Cohen, 2011). In human neurophysiology certain frequency bands are well described to occur preferentially in certain brain areas, such as the alpha rhythm shows a dominant topography over parietal-occipital areas. Also, different cognitive states are classically linked to specific frequency bands (Wang, 2010), such as alpha (~8-12Hz) or gamma oscillations (~30-120Hz) to cognitive processes involved in perception and attention, or theta oscillations often involved in working memory tasks (Buzsaki, 2006; Fries, 2015). The various frequency bands are further not independent but often linked. In this way alpha and gamma oscillations are often coupled in a reciprocal manner: increases in the alpha band go together with decreases in the gamma band and vice versa. Or frequencies are often coupled, or nested, such that for example high-frequency oscillations occur preferentially at the peak or trough of low-frequency brain waves (phase-amplitude coupling, see Canolty et al., 2006; Fries, 2015; Lakatos et al., 2005).

As stated in the previous paragraph, neural oscillations reflect the excitability of a population of neurons. It has been suggested that information from stimuli inputs arriving at more excitable phases will be processed more efficiently (Womelsdorf et al., 2006). In other words, if a sensory input has a regular temporal structure (i.e. is rhythmic or a real-world quasi-rhythmic stimulus, like speech) the maximal processing efficiency happens when the brain oscillations align with the temporal structure of the sensory input. Important

evidence has been observed in an inter-modal audio-visual selection task in primates (Lakatos et al., 2008). The results confirmed the four predictions: (i) when attention is deployed to one of several rhythmic stimuli oscillations in the sensory cortices tend to entrain (phase-lock) to the stimuli; (ii) high excitability phase tend to coincide with events in the attended stream; (iii) neuronal response amplitude and (iv) behavioral measures like accuracy and reaction time, will be related to the phase of the entrained oscillations. Under the framework of cortical entrainment (Peelle & Davis, 2012; Peelle, Gross & Davis, 2013) entrain to something in the real world means that the quasi-periodic system of the brain oscillations match the phase of an external periodic or quasi-periodic stimulus. This is particularly evident in the field of speech perception, where studies exploit the low frequency oscillation – especially in theta range – of the acoustic amplitude envelope of a speech signal (e.g., Ghitza, 2012; Ghitza, Giraud & Poeppel, 2013; Giraud & Poeppel, 2012).

The relevance of the oscillatory mechanisms and cortical entrainment in a cocktail party situation will be clarified in the next paragraph, which take into account the object-based attention dimension, too.

1.5 Neural signatures of object-based auditory selective attention

1.5.1 Neural correlates of the formation of auditory objects

One of the most explored and accredited theories concerning the neural underpinnings of auditory object formation is the temporal coherence theory (TCT), which assumes that neurons will be coherently activated in their response to sound stimuli, providing a mechanism to bind together what occurs at the

same point in time, and therefore to form perceptual objects (O'Sullivan, Shamma & Lalor, 2015; Shamma, Elhilali & Micheyl, 2011).

However, this framework does not explain in which way an acoustic object can be represented by the firing pattern of a group of neurons. Further, any mechanism of auditory attention requires that the representation of the attended object and the source of attentional facilitation is encoded in separate neural populations. Brain oscillations that are in versus out of phase to each other could provide a mechanism to tell competing neural representations apart (Engel et al., 2009); recent evidence seems to propose that low-frequency oscillations in the auditory system take up the low-frequency structure of syllabic input of attended speech input and become phase-consistent to the rhythmic activation of these external stimuli (Ding & Simon, 2012; Lakatos et al., 2016; Mesgarani & Chang, 2012), and that facilitation and inhibition of auditory input occur phase-dependent (Lakatos et al., 2013; Zion Golumbic et al., 2013a).

1.5.2 Neural mechanisms of auditory object-selection

Although the studies on selective auditory attention encompass different aspects of the attention deployment on an auditory scene (e.g., the auditory spatial attention network, the alerting and orienting network), I will here give a brief overview of the neural code of the top-down auditory object selection in a rich auditory scene, such as a cocktail party situation.

The highly dynamic nature of speech and its eminently characteristic rhythmicity, render it as a complex auditory stimulus, which strikingly fitting to studying the great capacity of attentional selection and the segregation of different streams that compose such a sound-scene. Both its properties lend themselves to investigations of temporal encoding, especially with temporally highly-sensitive brain scanners, such as MEG (Ding et al., 2016; Luo & Poeppel, 2012).

The simplest scenario is the study of speech in noise background (Ding & Simon, 2013) where participants listened to speech masked by spectrally matched stationary noise. The neural representation of the speech stream was measured by the reconstructability of the speech envelope from the low-frequency time-locked MEG responses. The neural representation decreased only mildly with respect to the intelligibility of the sounds, meaning that the neural representation is prelinguistic, as it successfully represents the speech at a noise level sufficiently high that the speech is audible but not intelligible. More interestingly in a complementary study (Ding, Chatterjee & Simon, 2014) the noise was substituted with a vocoded degraded signal which disrupts the temporal and spectral structure without affecting the slow acoustic envelope. The study showed that, while the neural representation of speech remained robust to noise, it was disrupted by the spectral distortions created by vocoding. This shows that the time locking of the neural responses to the speech envelope cannot be explained by mere passive envelope tracking mechanisms, but rather to an attentive access to spectro-temporal structure of the speech signal.

Studies using competing speech streams (Akram et al., 2016; Ding & Simon, 2012; Mesgarani & Chang, 2012; Simon, 2015b; Zion Golumbic et al., 2013a) typically find, for subjects listening to a mixture of two speech streams but attending to only one, that the neural representation of the attended speech is stronger than that of the unattended speech. When the stimuli resemble a real-world auditory scene this cannot be interpreted as simple as an attentional gain (i.e. the neural representation of the object of attention has been amplified) because of the strongly overlapping acoustic properties of the competing signals. It's therefore relevant to try to model the interplay between the neural representation of the object formation and the role of the selective attention in this process (Elhilali & Shamma, 2009; Elhilali et al., 2009b; Kaya & Elhilali, 2017; Shamma et al., 2011).

The study of Mesgarani and Chang recorded with ECoG in humans and analyzed high-gamma local field potential. Using sophisticated speech reconstruction tools they were able to transform the rhythmic neural activity from

auditory cortices back into speech spectrograms. The resulting, reconstructed spectrograms resembled the original spectro-temporal properties remarkably well.

Importantly in terms of attentional modulations of signals, they also tried to reconstruct speech spectrograms in an experimental condition, in which two different speakers were temporally overlaid and mixed together, and the subject was instructed to attend only one of them. In this attentional condition, the reconstruction of the spectrogram of the attended speaker from the neural recordings in the auditory cortex resembled the reconstruction in the single-speaker condition very closely. This is strong support from auditory research for the original biased-competition model. In the biased competition model, visual neurons have been shown to represent stimulus objects in the receptive field under sustained attention as if competing distractors were simply not present ('winner-takes-all', see Desimone & Duncan 1995). And the results also support the "selective entrainment hypothesis" for the auditory modality (Giraud & Poeppel, 2012; Zion-Golumbic & Schroeder, 2012).

In spite of the fact that exploiting the entrainment to the speech envelope with numerous analysis techniques is attracting the focus of current research efforts on auditory object formation and selection by attention mechanism, it is important to highlight that carefully crafted paradigms, and even simpler analysis of neuronal response amplitude still play a crucial role in elucidating the role of attention mechanisms, especially for the selection of auditory objects.

2 Second Chapter:
Behavioral Evidence of Object-Based Attention in Naturalistic Streams

A version of this Chapter was published in January 2019:

Marinato, G., & Baldauf, D. (2019). Object-based attention in complex, naturalistic auditory streams. Scientific reports, 9(1), 2854.

2.1 Abstract

In vision, objects have been described as the 'units' on which non-spatial attention operates in many natural settings. Here, we test the idea of object-based attention in the auditory domain within ecologically valid auditory scenes, composed of two spatially and temporally overlapping sound streams (speech signal vs. environmental soundscapes in Experiment 1 and two speech signals in Experiment 2). Top-down attention was directed to one or the other auditory stream by a non-spatial cue. To test for high-level, object-based attention effects we introduce an auditory repetition detection task in which participants have to detect brief repetitions of auditory objects, ruling out any possible confounds with spatial or feature-based attention. The participants' responses were significantly faster and more accurate in the valid cue condition compared to the invalid cue condition, indicating a robust cue-validity effect of high-level, object-based auditory attention.

2.2 Introduction

In many ecologic environments, the naturalistic auditory scene is composed of several concurrent sounds with their spectral features overlapping both in space and time. Humans can identify and differentiate overlapping auditory

objects surprisingly well (Griffiths & Warren, 2004) . According to McDermott (McDermott, 2009), the identification of different sounds in a complex auditory scene is mainly studied from two conceptually distinct perspectives: sound segregation (or “auditory scene analysis” (Bregman, 1990)) and attentional selection (Cherry, 1953)

As already outlined in the introduction chapter of this dissertation, auditory research has focused primarily on the segregation component (Bronkhorst, 2000; Brungart et al., 2006; Carlyon, 2004; Elhilali & Shamma, 2009), and despite of many efforts to better understand the interaction between auditory attention and segregation processes (Best et al., 2007; Darwin, 1997; Maddox & Shinn-Cunningham, 2012; Shinn-Cunningham et al., 2015; Simon, 2015a; Sussman et al., 2007; Winkler et al., 2003), there is still debate about the mechanisms of auditory object formation and auditory selective attention (Corbetta et al., 1998; Ding & Simon, 2012; Hill & Miller, 2010; Lee et al., 2014; Posner, 1980). However, attentional mechanisms have been described in much detail in other sensory modalities, in particular, in vision. From visual attention research we have learnt how top-down attentional control can operate on visual space (Baldauf & Deubel, 2010; Gregoriou et al., 2009; Mangun & Hillyard, 1991; Moore & Armstrong, 2003; Nobre et al., 2000; Rossi & Paradiso, 1995; Sàenz, Buraĉas & Boynton, 2003; Siegel et al., 2008; Sprague & Serences, 2013) , on low-level perceptual features (Ciaramitaro et al., 2011; Cohen & Tong, 2015; Hopf et al., 2004; Liu, Stevens & Carrasco, 2007b; Maunsell & Treue, 2006; Müller et al., 2006; Störmer & Alvarez, 2014; Treue & Trujillo, 1999; Wegener et al., 2008; Zhang & Luck, 2009), and high-level visual objects (Baldauf & Desimone, 2016; Corbetta et al., 2005; Duncan, 1984; Egly, Driver & Rafal, 1994; O’Craven et al., 1999; Schoenfeld et al., 2014; Scholl, 2001; Shinn-Cunningham, 2008). And especially visual objects have been described as the ‘units’ on which non-spatial attention operates in many natural settings (Falkenberg, Specht & Westerhausen, 2011; O’Craven et al., 1999; Scholl, 2001).

Within the domain of auditory selective attention, early work exploited

mainly the “dichotic listening” paradigm (Cherry, 1953). In this paradigm, participants listen to a different audio stream presented to each ear and are asked to pay attention to either one of them (Alho et al., 2012; Hugdahl et al., 2009; Treisman, 1960), or sometimes to both (Ding & Simon, 2013; Kimura, 1964; Petkov et al., 2004). However, the dichotic listening paradigm always has a spatial component to them and therefore leaves plenty of room for attentional lateralization confounds, which constitute a major shortcoming for using them to investigate non-spatial attention. Later work used paradigms that manipulated specific features of the acoustic stimulus to demonstrate successful tracking of one sound signal over the other. Some studies modulated pitch (Ding et al., 2014; Posner, 1980), others intensity level (Morillon & Schroeder, 2015) or spatial features, such as location (Posner, 1980). More recent studies, focused on the mechanisms of the neural representation of speech signals, using neural recordings for precisely tracking speech signals (Alain & Arnott, 2000; Poeppel, Idsardi & Wassenhove, 2008; Shamma et al., 2011; Zion Golumbic et al., 2013b). Lastly high-level attention modulation in a complex auditory scene was investigated from the neural perspective also with paradigms that involve competing speech streams (Alain & Winkler, 2012; Elhilali et al., 2009b), speech in noise (Morillon et al., 2012), and tone streams (Bizley & Cohen, 2013; Xiang, Simon & Elhilali, 2010).

Here, we introduce a novel stimulus set and task to study object-based attention in the auditory domain. In analogy to visual objects, we defined an auditory object as the aggregation of low-level features into grouped entities. Several auditory objects together can then constitute an auditory scene, or soundscape, e.g. the characteristic soundscape of a railroad station or a multi-talker conversation at a party. In such natural, complex auditory environments, auditory objects are temporally confined and bound entities, e.g., the words constituting a conversation or a train whistle in the soundscape of a railroad station. Notably, there have already previously been various attempts to define what an auditory object is. One influential operational definition was proposed by Griffiths and Warren (Griffiths & Warren, 2004). Here, an auditory

object is defined as something (1) that corresponds to things in the sensory world, (2) that can be isolated from the rest of the sensory world, and (3) that can be recognized or generalized beyond the single particular sensory experience. Further, object analysis may also involve generalization across different sensory modalities, such as the correspondence between the auditory and visual domain. This operational definition has also been used to define the neural representation of auditory objects. Our definition borrows from the previous ones and is in line with the concept of acoustic stream, or 'soundscape', as a superordinate entity of individual objects (Bizley & Cohen, 2013).

Again in analogy to visual paradigms used to study object-based attention (Baldauf & Desimone, 2014), we introduce an auditory repetition detection task, in which participants had to detect brief repetitions of auditory objects within the acoustic stream of a soundscape. The logic behind this new task is that such a repetition detection task requires the participants to fully process the acoustic stream to a cognitive level that allows them to recognize a certain, temporally extended set of low-level features as an object and to understand that this set of features was repeated. Importantly, this attention task cannot be solved by attending to a distinct low-level feature itself (e.g., a certain pitch). To also rule-out spatial attention, we presented two overlapping auditory scenes (e.g., in Experiment 1 a foreign language conversation and a railroad station soundscape) at the same external speaker, attentionally cuing one or the other.

In every trial, a 750 ms long repetition is introduced in one of the two overlapping streams and participants are asked to detect any such repetitions of auditory objects as fast as possible. This task requires the processing of the acoustic stream to the level of auditory objects and is specifically designed to investigate object-based mechanism of selective attention, i.e., whether top-down selective attention can weigh incoming acoustic information on the level of segregated auditory objects by facilitation and/or inhibition processes.

2.3 Experiment 1: Attentional weighting of speech versus environmental sound-scenes

2.3.1 Methods

Participants

Eleven participants (6 females, 5males, mean age 25.7 years, range 23-32 years, all of them right-handed and normal hearing) took part in the behavioral experiment and were paid for their participation. All participants were naïve in respect to the purpose of the study and they were not familiar with any of the languages used to create the speech stimuli. All participants provided written, informed consent in accordance with the University of Trento Ethical Committee on the Use of Humans as Experimental Subjects. One participant had to be excluded from further analyses because he failed to follow the task instructions.

Stimuli

Speech and environmental sound signals:

The experimental stimuli were auditory scenes, consisting of overlapping streams of (a) speech conversations embedded in (b) environmental sounds. All the speech signals were extracted from newscast recordings of various foreign languages: (1) African dialect, (2) Amharic, (3) Armenian, (4) Bihari, (5) Hindi, (6) Japanese, (7) Kurdish, (8) Pashto, (9) Sudanese, (10) Urdu, (11) Basque, (12) Croatian, (13) Estonian, (14) Finnish, (15) Hungarian, (16) Icelandic, (17) Macedonian, (18), Mongolian, and (19) Bulgarian. The environmental sound source signals were selected from soundscapes of public human places,

recorded at (1) airports, (2) canteens, (3) malls, (4) markets, (5) railway stations, (6) restaurants, (7) streets, (8) trains, and (9) subways.

From each recording, we extracted a central part using Audacity software, discarding the very beginning and end of the original signal. All recording segments were processed with Matlab custom functions to cut the sound segments to 5 seconds length, convert them to mono by averaging both channels, and normalize them to -23db. Guided by the Urban Sound Taxonomy 70 and Google's Audio Set 71 we chose the stimuli from high quality YouTube recordings.

Enveloping

After these processing steps, speech signals and environmental signals still differed in their low-frequency rhythmicity and overall signal envelope: the analytical envelopes of the environmental sound epochs were rather stationary whereas speech signals are characterized by prominent quasi-rhythmic envelope modulations in the 4-8Hz range. In order to further equalize the two sound streams and make them as comparable as possible we dynamically modulated the envelope of the environmental sounds using envelopes randomly extracted from the speech signals. To do so envelopes of the speech signals were first extracted using the 'Envelope' functionality in Matlab, which is based on the spline interpolation over local maxima separated by at least 4410 samples, corresponding to 0.1s at a sample rate of 44.1KHz. This relatively large number of samples was chosen in order to keep the environment sound clearly recognizable after applying a quasi-rhythmic temporal.

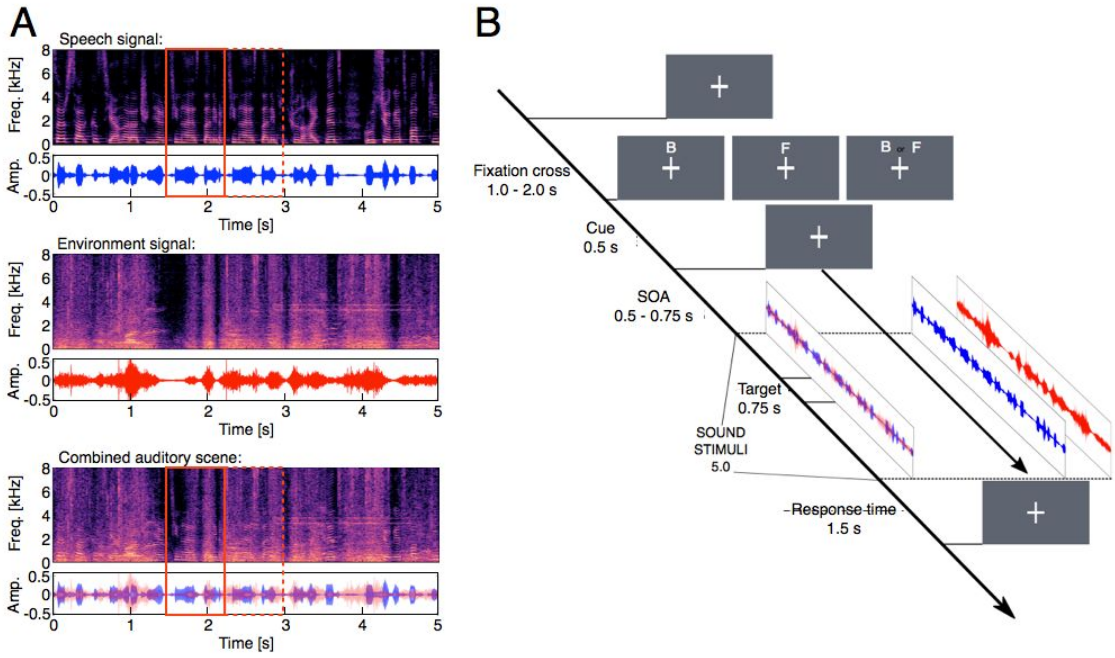

Figure 1. *(A) Experimental acoustic stimuli. In each trial, the acoustic stimuli consisted of two soundscapes, one foreign language speech signal and one environmental sound signal (e.g., the soundscape of a train station), which were temporally and spatially overlaid, presented from the same centrally positioned speaker for a total of 5 seconds. The three subpanels show the time-frequency spectrogram and raw amplitude spectra of examples of the original sound streams (i.e., speech signal and the environmental signal, respectively), as well as for the combined auditory scene that was presented to the participants (lower panel). In one of the two streams (here in the speech signal), a repetition target was embedded by replicating a 750-ms interval (see the solid red box), and repeating it directly after the original segment (see the dashed red box). Linear ramping and cross-fading algorithms were applied to avoid cutting artifacts and to render the transition between segments unnoticeable. The repetition targets had to be detected as fast and as accurate as possible. (B) Sequence of a typical trial. In each trial, a cue was presented indicating either the speech component of the signal ('F') or the environmental component ('B') or both (neutral cueing condition). Subjects were instructed to shift their attention to the cued channel and to detect any repetition targets as fast and as accurate as possible, while keeping central eye fixation throughout the trial. Cue validity was 70%.*

One-back repetition targets and overlay

In a next step, we inserted small segment repetitions to be used as repetition targets in our listening task (see Fig.1A). For this we randomly sampled and extracted a short sound epoch of 750 ms and repeated it immediately after the end of the segment that has been sampled. The length of the repetition targets was chosen to roughly correspond to a functional unit like a typical acoustic event in the environment sounds or a couple of syllables/words in the speech signals. In order to implement the repetition in Matlab, the initial sound signal was cut at a randomly selected sample, then the original beginning, the 750 ms repetition, and the original end to the stream were all concatenated by a linear ramping and cross-fading. The linear ramping is made by a window of 220 samples that corresponds to 5ms at a sample rate of 44.1 KHz. The cross-fading is achieved by simply adding together the ramping down part of the previous segment with the ramping up part of the subsequent segment.

Finally, for each trial's audio presentation one resulting speech signal and one environmental sound signal were overlapped to form an auditory scene, consisting of speech conversation embedded in environmental sounds (see Fig. 1A bottom panel). In each trial only one of them could contain a repetition target. A set of the experimental stimuli can be freely downloaded at https://doi.org/10.5281/zenodo.1491058.

Trial Sequence and Experimental Design

All stimuli were presented using Psychophysics Toolbox Version 3 (Brainard, 1997; Kleiner, Brainard & Pelli, 2007). Figure 1B provides an overview of a typical trial sequence. We implemented an attentional cueing paradigm with three cue validity conditions, i.e. valid, neutral, and invalid cues. Cue validity was

70%, 20% of cues were invalid, and 10% neutral. At the beginning of each trial, a fixation-cross appeared and subjects were instructed to keep central eye fixation throughout the trial (see Figure 1B). After an interval of 1.0-2.0s (randomly jittered) a visual cue was presented, directing auditory attention either to the "Speech" signal stream or to the auditory "Environment" stream, or to neither of them in the neutral condition. After another interval of 0.5-0.75s (randomly jittered) the combined audio scene with overlapping speech and environmental sounds started playing for 5.0 s. The participants were instructed to pay attention to the cued stream and to respond with a button press as soon as they recognize any repetition in the sound stimuli. Accuracy and speed were equally emphasized during the instruction.

Before the actual data collection, participants were first familiarized with the sound scenes and had a chance to practice their responses to repetitions for one block of 100 trials. For practice purposes, we initially presented only one of the two sound streams individually so participants had an easier time understanding what repetition signals to watch out for. This training lasted for 17 minutes in total.

Each subsequent testing block consisted of 100 trials but now with overlapping sound scenes consisting of both a speech and an environmental sound stream and with the described attentional cueing paradigm. Each participant performed three experimental blocks, resulting in 300 experimental trials in total. Overall our experimental design had two factors: (1) Cue validity with the conditions valid (70% of trials), neutral (10% of trials) and invalid (20% of trials), and (2) Position of the repetition target in either the speech (50% of trials) or environmental (50% of trials) sound stream. All conditions were trial-wise intermixed.

Data Analysis

All data analyses were performed with custom scripts in MATLAB. A combination of built-in function and custom code was used in order to conduct

descriptive and inferential statistics. For each condition in our 2x3 factor, mean and standard error of the mean (SEM) were calculated both for reaction times and response accuracies. Repeated-measurement analyses of variance were computed on accuracy data, mean reaction times, signal detection sensitivity and response biases. To further investigate systematic differences between individual conditions we computed planned contrasts in form of paired-samples t-tests between the repetition detection rates and reaction times in the valid versus invalid versus neutral cueing condition (both in the speech and environmental sound stream).

However, differences in detection accuracy reaction times can also result from changes in the response bias, for example, by a tendency to reduce the amount of evidence that is required to decide whether a target had occurred. To better understand the stage of selection, i.e., whether increases in detection rate are due to changes in sensitivity or changes in the decision criterion, or both, we further computed signal detection theory (SDT) indices in form of the sensitivity indices (d') and response bias or criterion (c).

2.3.2 Results

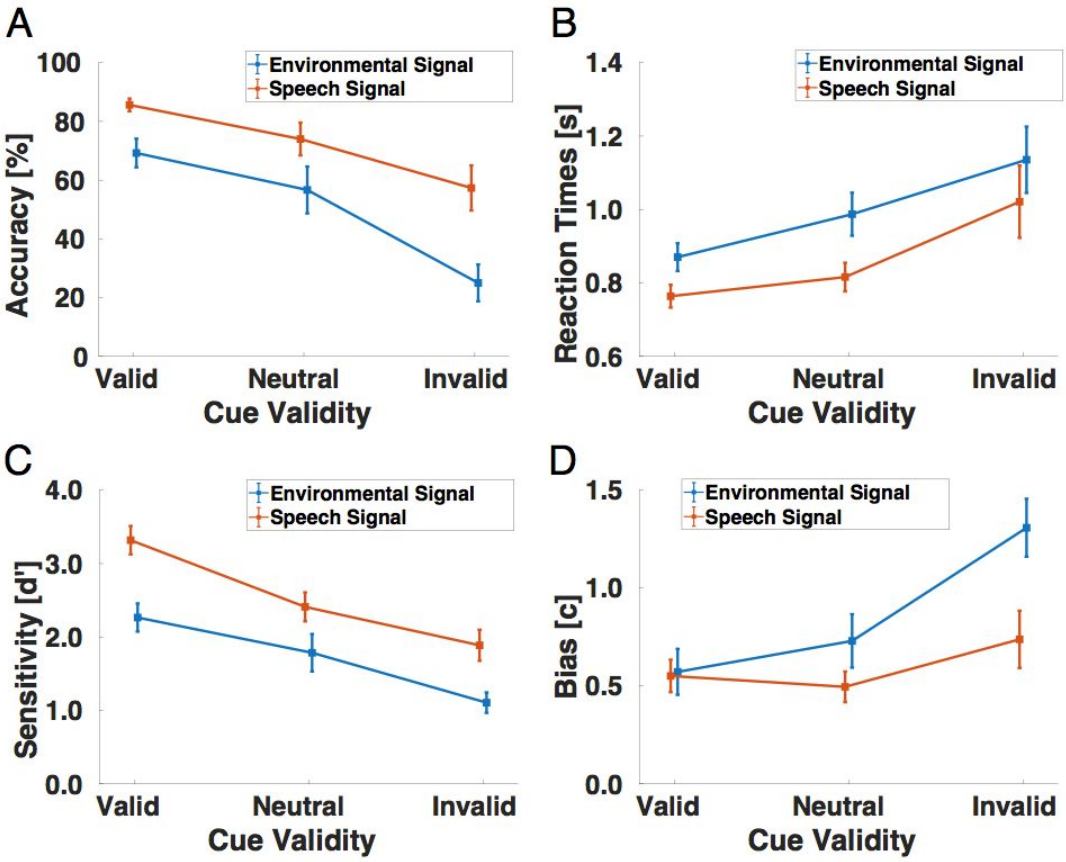

Figure 2. *Experimental results in the repetition detection task of Experiment 1 with overlaid speech and environmental-noise streams. (A) Detection performances (in percent correct) as a function of cue validity (valid, neutral and invalid cueing condition), separately for both components of the acoustic stimuli, i.e., the environmental signal (blue) and the speech signal (red). The data are shown as means and SEM. (B) Reaction times (in seconds) as a function of cue validity (valid, neutral and invalid cueing condition), separately for both components of the acoustic stimuli (blue: environmental signal, red: speech signal). The data are shown as means and SEM.*

(C) Sensitivity scores (d') and decision criteria (D) as a function of cue validity (valid, neutral and invalid cueing condition), separately for both components of the acoustic stimuli, i.e., the environmental signal (blue) and the speech signal (red). The data are shown as means and SEM.

Accuracy

Figure 2A shows the average accuracy with which repetition targets were

detected in both the speech and environmental sound stream, and Figure 2B shows the corresponding reaction times with which the responses were given. Repetition targets were detected well above chance, but performance was clearly not ceiling with up to 85% correct responses in the valid cueing condition. In general, valid cues helped the participants detect repetition targets and also speeded up their responses by about 100 ms in respect to the neutral cueing condition. Invalid cues had the opposite effect. Table 1 provides an overview of the numeric values of mean detection accuracy and reaction times.

A two-way repeated-measures analysis of variance (ANOVA) on the mean detection accuracy statistically confirmed a main effect of the factor Cue validity, with $F(2,18) = 28.36$, $p < 0.001$. There was also a significant main effect of the second factor Position of the repetition target (speech signal vs. environmental signal), with $F(1,9) = 22.53$, $p = 0.001$. Importantly, there was no significant interaction between the two factors, with $F(2,18) = 1.61$, $p = 0.226$, indicating that the attentional modulation by the cue validity worked similarly for both streams. Planned contrasts in form of paired t-tests confirmed the expected direction of the attentional modulation effect: for speech and environmental sound targets combined, participants were significantly better in detecting the repetition targets in the valid then in the invalid cueing condition, $t(9) = 7.5$, $p < 0.001$. Participants responded significantly better also in the valid than in neutral condition, $t(9) = 2.83$, $p = 0.02$ and worse in invalid compared to the neutral condition: $t(9) = -4.38$, $p = 0.002$. Also for the speech and environmental sound stream targets separately, t-tests revealed that valid cues made participants respond faster compared to invalid cues, with $t(9) = 7.13$, $p < 0.001$ in the environmental signal and $t(9) = 3.75$, $p = 0.005$ in the speech signal. A significantly more accurate response was also found between the valid and neutral condition (i.e., facilitation), with $t(9) = 2.32$, $p = 0.045$ for the speech signal and $t(9) = 2.34$, $p = 0.044$ for the environmental signal. However, comparing the invalid versus neutral condition for the two different streams (i.e. suppression effects) gave a significant better response accuracy only for the environment signal, with $t(9) = -4.07$. $p = 0.003$, but not for the speech signal, with $t(9) = -1.83$, $p = 0.1$.

Comparing the detection accuracy under valid cueing conditions for speech signals versus environmental sound signals, a paired t-test revealed that it was a bit harder to detect embedded repetition targets in the environmental signal then in the speech signal, with $t(9) = 3.273$, $p = 0.010$.

Reaction times

A data analysis similar to the one performed for accuracy was also conducted on reaction times, revealing congruent effects. The numeric values of the average reaction time performance in the six experimental conditions are also provided in Table 1.

A two-way repeated-measure analysis of variance (ANOVA) on mean reaction times revealed a main effect of factor Cue validity, $F(2,18) = 8.63$, $p = 0.002$, and a main effect of factor Position of the repetition target (speech signal vs. environmental signal, $F(1,9) = 13.02$, $p = 0.005$). Again, there was no significant interaction between both factors, with $F(2,18) = 0.51$, $p = 0.610$.

To investigate the direction of the observed effects, planned contrasts in the form of paired t-tests were performed between the valid and invalid attention cue for speech and background, combined as well as separately. Combining data from both sound streams, participants were significantly faster identifying targets in the valid compared to the invalid cue condition, $t(9)=-3.218$, $p=0.010$. There were also significant differences between the valid and neutral condition, $t(9) = -2.41$, $p = 0.039$ (i.e. facilitation), and invalid and neutral conditions, $t(9) = 2.44$, $p = 0.037$ (i.e. suppression). Also for the speech and environmental sound stream targets separately, t-tests revealed that valid cues made participants respond faster compared to invalid cues, with $t(9) = -3.85$, $p < 0.004$ for targets in the environmental stream $t(9) = -2.62$, $p = 0.028$ for repetition targets hidden in the speech stream. For the environmental signal, we found evidence for facilitation effects, i.e. faster responses in the valid than in the neutral condition, with $t(9) = -2.48$, $p = 0.035$. In the speech stream, however, we did not find any significant advantage between the valid and neutral cueing condition, with $t(9) = -1.83$, $p =$

0.198. The opposite was true for suppression effects, i.e. comparing the invalid with the neutral cueing condition. Here, for the environmental signal, participants did not show any significant advantage between invalidly and neutrally cued trials, with $t(9) = 1.70$, $p = 0.123$. Instead, participants were faster in the neutral condition if the repetition was in the speech stream, with $t(9) = 2.55$, $p = 0.03$. Finally, comparing the detection accuracy under valid cueing conditions for speech signals versus environmental sound signals, a paired t-test revealed that the repetition targets were detected faster in the speech signal then in the environmental signal, $t(9) = -3.683$, $p = 0.005$. These results of mean reaction times are therefore consistent with the analysis of the detection accuracy data.

Signal-detection theory (SDT) analyses

We also computed sensitivity indices (d') using the method suggested by Macmillan and Creelman (Macmillan & Creelman, 1991). False alarms were detected as responses given before the presentation of the target. We first calculated sensitivity indices separately for each subject and each condition and averaged the computed values separately for each of the six conditions in our 3x2 factorial design (with factors Cue validity and Position of the repetition target).

Figure 2C shows the average sensitivity indices across all participants as a function of cue validity and the relative position of the repetition target. Participants were clearly more sensitive to repetition targets when they were validly cued. In comparison to the neutral cue condition, valid cues made participants more sensitive to repetition targets in both the speech and environmental noise stream. Invalid cues had the opposite effect, hindering subjects' sensitivity to those subtle auditory targets (see also Table 1 for an overview of the numeric values of d' sensitivity and criterion. Therefore, the signal detection sensitivity analysis results were congruent with both the

accuracy and reaction time data.

A two-way repeated-measures analysis of variance (ANOVA) of the sensitivity scores statistically confirmed a main effect of the factor Cue validity, with $F(2,18) = 21.3$, $p < 0.001$. There was also a significant main effect of the factor Position of the repetition target (speech signal vs. environmental signal), with $F(1,9) = 23.73$, $p < 0.001$. There was no significant interaction between the two factors, with $F(2,18) = = 0.78$, $p = 0.47$, indicating that the attentional modulation by the cue validity worked similarly for both streams. Planned contrasts in form of paired t-tests confirmed the expected direction of the attentional modulation effect: for speech and environmental sound targets combined, participants were more sensitive to repetition targets in the valid then in the invalid cueing condition, $t(9) = 7.19$, $p < 0.001$. Comparing the valid and invalid condition with the neutral condition a significant effect of facilitation was detected for the valid condition, with $t(9) = 2.98$, $p = 0.02$ and an suppression effect was found for the invalid condition, with $t(9) = -3.39$, $p = 0.008$. Also for the speech and environmental sound stream targets separately, t-tests revealed that valid cues made participants more sensitive than invalid cues, with $t(9) = 6.19$, $p < 0.001$ and $t(9) = 5.53$, $p < 0.001$ in the environmental signal and in the speech signal, respectively. Regarding the environmental signal, validly cued trials were not significantly different from trial with neutral cues, with $t(9) = 1.83$, $p=0.1$, but there was a facilitation of sensitivity for the speech signal, with $t(9) = 2.94$, $p = 0.02$. An opposite pattern was observed comparing the invalid condition with the neutral one, revealing a significant difference when the target was in the environmental signal, with $t(9) = -2.63$, $p = 0.03$, but no significant difference for targets in the speech stream, with $t(9) = -1.80$, $p = 0.11$. Comparing the sensitivity under valid cueing conditions for speech signals versus environmental sound signals, a paired t-test revealed that sensitivity was in general higher for the speech signals compared to environmental noise signals, with $t(9) = 4.56$, $p = 0.001$.

Figure 2D shows the average criterion (c) indices as a function of the factors Cue validity and Position of the repetition target. Participants have a

similar bias and relatively liberal response criterion in the valid cueing conditions for both the speech and the environmental stream. They become more conservative in the invalid cueing condition especially when the target was embedded in the environmental signal.

A two way repeated-measures analysis of variance of the criterion scores confirmed a main effect of the factor Cue validity, with $F(2,18) = 18.17$, $p < 0.001$, and of the factor Position of the repetition target, with $F(2,18) = 18.09$, $p = 0.002$. There was also a significant interaction between the two factors, with $F(2,18) = 4.48$, $p = 0.03$.

Planned paired t-test were conducted to test the direction of the observed effects. In general there was a significantly more liberal decision criterion in the valid than in the invalid cueing condition, with $t(9) = -5.70$, $p < 0.001$. The difference in response criterion was aslo significant between the invalid and neutral condition, with $t(9) = 4$, $p = 0.003$, but not between the valid and neutral conditions , with $t(9) = -0.804$, $p = 0.44$. Interesting, for the speech and environmental signal stream separately, there was a significant liberalization of the the response criterion for the environmental signal (i.e. contrasting the valid versus invalid cue condition, with $t(9) = -4.86$, $p = 0.001$), and a more conservative answering scheme when comparing the invalid and neutral condition, with $t(9) = 3.87$, p 0.004. In any other contrast no significant differences were observed.

	Valid	Neutral	Invalid
Accuracy [%]			
Speech	85.62 (2.17)	74.0 (5.57)	57.33 (7.66)
Environment	69.24 (4.87)	56.67 (7.97)	25.0 (6.27)
Reaction time [ms]			
Speech	764 (31)	816 (39)	1021 (98)
Environment	870 (38)	987 (59)	1136 (90)
Sensitivity index (d')			
Speech	3.32 (0.19)	2.41 (0.20)	1.89 (0.21)
Environment	2.27 (0.19)	1.79 (0.25)	1.11 (0.14)
Decision criterion (c)			
Speech	0.55 (0.08)	0.50 (0.08)	0.74 (0.15)
Environment	0.57 (0.12)	0.73 (0.14)	1.31 (0.15)

Table 1. *Experiment 1 with overlaid speech and environmental sound streams. Numeric values of the detection accuracy (in percent correct), reaction time (in ms), sensitivity indices (d'), and decision criteria (c), all across all ten participants for all three cueing conditions (valid, neutral and invalid cues), separately for the speech and environmental component of the signal. Values represent the means and standard errors of the mean.*

2.4 Experiment 2: Attentional weighting of two competing speech streams

In Experiment 1, we used an ecologically valid scenario of a speech signal being overlaid with environmental noise and asked participants to tune their attention to track one or the other input stream. Importantly, we equaled the low-level rhythmicity and the signal envelope, however, there exists the

possibility that some low-level differences remained between the two types of stimuli and that any attentional weighting was based only on such subtle differences alone. Maybe participants could have done the task in Experiment 1 by focusing on lower-level features instead.

Therefore, we address the question of object-based attention in a second experiment in which we present two overlaid sound streams from only one category (speech) that largely match in all low-level properties and thus require participants to fully attend to the higher-level properties. In Experiment 2, we therefore employ the same object-based repetition detection task as in Experiment 1, but have people attend one voice among other voices (both streams again overlaid spatially and temporally congruent), i.e., a listening scenario that is more similar to the classic cocktail party problem but without spatial separability of the signal sources.

2.4.1 Methods

Participants

Ten participants (6 females, 4 males, mean age 27.5 years, range 25-33 years, all of them right-handed and with normal hearing) took part in Experiment 2. They all were naïve in respect to the purpose of the study, and none of them had participated in Experiment 1. They were not familiar with any of the languages used to create the speech stimuli. All participants provided written, informed consent in accordance with the University of Trento Ethical Committee on the Use of Humans as Experimental Subjects.

Stimuli

Speech sound signals and overlay:

In Experiment 2 we presented auditory scenes that consisted of two

overlapping streams of speech conversation. There were no further embedded environmental sounds. The speech signals overlaid here were the same speech signals used also in Experiment 1. Again, a repetition segment of 750 ms was randomly embedded in either one of them, serving as a repetition target that had to be detected as fast and as accurately as possible. Both speech signals were presented from the same central position, making it impossible to use spatial information to solve the task. A set of the experimental stimuli can be freely downloaded at https://doi.org/10.5281/zenodo.1491058.

Trial Sequence and Experimental Design

As in the previous experiment we had three cueing conditions, i.e. valid, neutral, and invalid cues. Cue validity was 70%, 20% of cues were invalid, and another 10% neutral. To direct the participants' selective attention towards one or the other speech stream, we used an auditory cue, which consisted of the first 1.0s segment of the isolated speech signal of one of the two speakers.

A typical trial sequence in Experiment 2 is very similar to the first experiment, but now the attention was cued to one of two speech streams by a short acoustic cue, which consisted of a short pre-play segment of one of the voices. At the beginning of each trial, a fixation-cross appeared and subjects were instructed to keep central eye fixation throughout the trial. After a random interval of 1.0-1.5 s, the auditory cue was presented, directing auditory attention to one of the two speakers. In trials with neutral cue condition, no cue was given at all. After another jittered interval of 1.0-1.5s, the combined audio scene with both overlapping speech streams started playing and continued for 5.0 s. The participants were instructed to pay attention to the cued stream and to respond with a button press as fast and as accurately as they recognized any repetition segments. Accuracy and speed were equally emphasized during the instruction.

Before the actual data collection, participants were first familiarized with the repetition segments by listening to ten individual example presentations and then performing one short sample block of 20 overlaid sound scenes in order to

practice their responses to repetitions. Each subsequent testing block consisted of 60 trials. Each participant performed five experimental blocks, resulting in 300 experimental trials in total. In this second experiment, only the factor Cue validity (with the three conditions valid, neutral, and invalid) was relevant for the behavioral analyses. All conditions were trial-wise intermixed.

Data Analysis

For each condition of the factor Cue validity, the mean and standard error of the mean (SEM) were calculated both for reaction times, response accuracies, and sensitivity indices. For the purpose of inferential statistics, repeated-measurement analysis of variance were computed on all those three behavioral measures. To further investigate systematic differences between individual conditions we computed planned contrasts in form of paired-samples t-tests between the repetition detection rates, reaction times and sensitivity scores in the valid versus invalid cueing condition.

2.4.2 Results

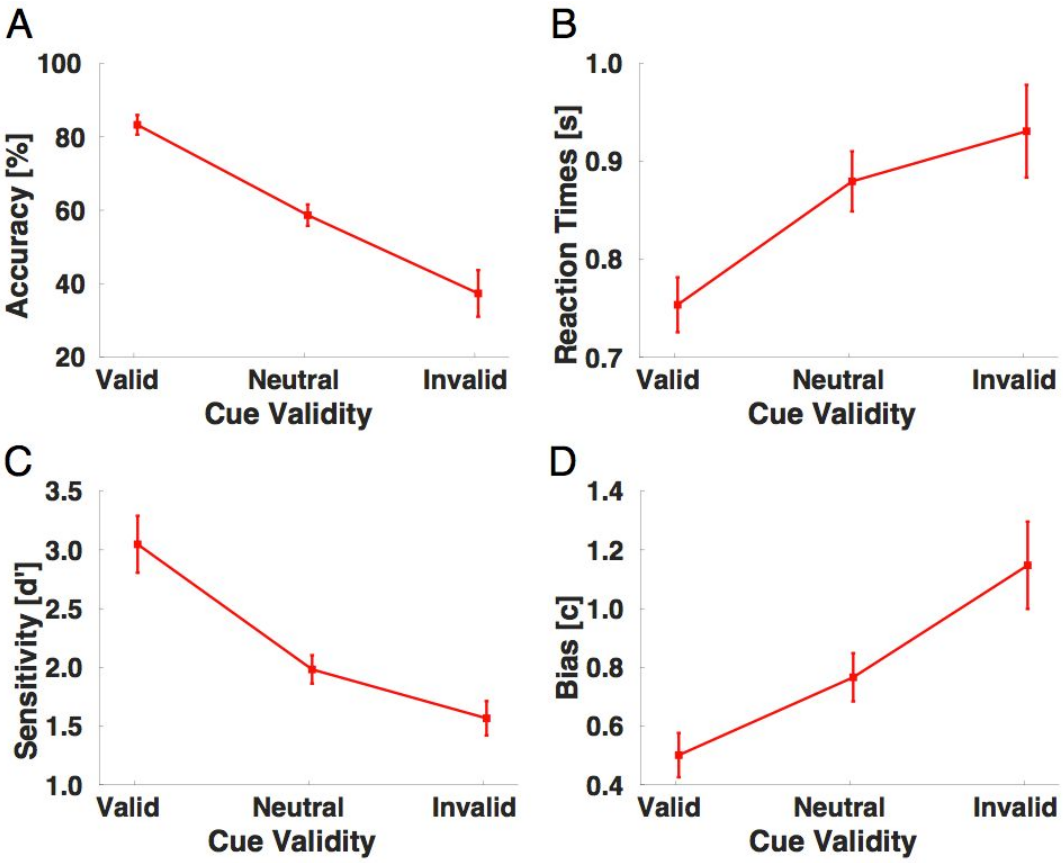

Figure 3. *Experimental results in the repetition detection task of Experiment 2 with two overlaid speech streams. (A) Detection performances (in percent correct) as a function of cue validity (valid, neutral and invalid cueing condition), for the second experiment with two speech sounds. The data are shown as means and SEM. (B) Reaction times as a function of cue validity (valid, neutral and invalid cueing condition), for the second experiment with two speech sounds. The data are shown as means and SEM. (C) Sensitivity scores (d') and decision criteria (D) as a function of cue validity (valid, neutral and invalid cueing condition) for the second experiment with two speech sounds. The data are shown as means and SEM.*

Accuracy

Also in Experiment 2, with two competing speech signals, the repetition

targets were detected well above chance, Figure 3A shows the average detection accuracy, and Figure 3B shows the corresponding reaction times with which the responses were given. There was a cue-validity effect in the sense that valid cues helped the participants in better detecting repetition targets and also speeded up their responses. Invalid cues, however, had a hindering effect compared to neutral cues (see Table 2 for all the numeric values of mean detection accuracy and reaction times).

A one-way repeated-measures analysis of variance (ANOVA) on the mean detection accuracy statistically confirmed a main effect of the factor Cue validity, with $F(2,18) = 27.27$, $p < 0.001$. We also calculated planned contrasts in form of paired t-tests to confirm the direction of the attentional modulation effect: participants were significantly better in detecting repetition targets in the valid then in the invalid cueing condition, $t(9) = 5.44$, $p < 0.001$ and then in the neutral cueing condition, with $t(9) = 5.18$, $p < 0.001$. Also a paired t-test comparison between invalid cueing condition and neutral cueing condition revealed a significant better accuracy in detecting the target in the neutral condition, with $t(9) = -4.48$ $p = 0.001$.

Reaction times

An analysis of the response times revealed congruent cue-validity effects. The numeric values of the average reaction time performance in the six experimental conditions are provided in Table 2. A one-way repeated-measure analysis of variance (ANOVA) on mean reaction times revealed a main effect of factor Cue validity, $F(2,18) = 19.25$, $p < 0.001$. To investigate the direction of the observed effects, planned contrasts in the form of paired t-tests were performed between the valid, invalid and neutral cueing conditions. Participants were significantly faster identifying targets in the valid compared to the invalid cue condition, with $t(9) = -5.25$, $p = 0.001$, and compared to the neutral cue, with $t(9) = -5.34$, $p = 0.001$. A paired t-test between invalid and neutral condition revealed no significant effects, with $t(9) = 1.7$ $p = 0.12$. These cue-validity effects on mean

reaction times are therefore consistent with the analysis of the detection accuracy data.

Signal-detection theory (SDT) analyses

False alarms were detected as responses given before the presentation of the target. We first calculated sensitivity indices separately for each subject and each condition and only then averaged the computed values in each of the three cueing conditions. Figure 3C shows the sensitivity scores (d'). Participants became more sensitive to the subtle repetition targets when they were validly cued. Invalid cues, however, were distracting attention and decreased sensibility to repetition targets (see Table 2 for an overview of the numeric values of sensitivity indices and criteria. Overall, the sensitivity analyses revealed congruent effects with the accuracy and reaction time data.

A one-way repeated-measures analysis of variance (ANOVA) on the sensitivity scores statistically confirmed a main effect of the factor Cue validity, with $F(2,18) = 16.06$, $p < 0.001$. Planned contrasts in form of paired t-tests confirmed the expected direction of the attentional modulation effect: participants were significantly more sensitive to repetition targets in the valid then in the invalid cueing condition, with $t(9) = 4.56$, $p = 0.002$, and also compared to the neutral cueing condition, with $t(9) = 4.04$, $p = 0.003$. No significant effect was found in a paired t-test between invalid and neutral condition: $t(9) = -2.05$, $p = 0.07$.

Again, we also calculated measures of the response criterion (c) to better characterize the response bias used by the participants between the conditions. Figure 3D shows the change in the response criterion between the three conditions, with a more liberal criterion in the validly cued trials and a more conservative response bias for the invalidly cued trials (both in respect to the neutral condition, which is in the middle).

A one-way repeated-measure analysis of variance (ANOVA) on the

response bias scores revealed a statistically significant effect of the factor Cue validity, with $F(2,18) = 16.30$, $p < 0.001$. Here, planned contrast in the form of paired t-test confirmed a significantly more liberal bias in the valid cueing condition compared to the invalid cueing condition, with $t(9) = -4.71$, $p=0.001$ but also when comparing the valid cue condition with the neutral cueing condition, $t(9) = -2.84$, $p = 0.02$. Response biases were significantly more conservative in the invalid cueing condition compared to the neutral cueing condition, with $t(9) = 3.62$, $p = 0.006$.

	Valid	Neutral	Invalid
Accuracy [%]			
Speech	83.29 (2.67)	58.67 (2.91)	37.33 (6.34)
Reaction time [ms]			
Speech	754 (28)	880 (31)	931 (47)
Sensitivity index (d')			
Speech	3.05 (0.24)	1.98 (0.12)	1.57 (0.15)
Decision criterion (c)			
Speech	0.50 (0.08)	0.77 (0.08)	1.15 (0.15)

Table 2. *Experiment 2 with two overlaid speech streams. Numeric values of the detection accuracy (in percent correct), reaction time (in ms), sensitivity indices (d'), and decision criteria (c), all across all ten participants for all three cueing conditions (valid, neutral and invalid cues), separately for the speech and environmental component of the signal. Values represent the means of each score.*

2.4.3 Discussion

For the present study we used novel sets of stimuli and a new repetition detection task to study object-based attention in the auditory domain. Our

paradigm and stimuli were specifically conceived to tackle high-level, object-based mechanisms of selective voluntary attention, in analogy to attentional weighting paradigms used in the visual domain (Baldauf & Desimone, 2014; Desimone & Duncan, 1995). By presenting two spatially and temporally overlapping auditory scenes we were able to overcome some shortcomings of previously used dichotic listening paradigms (Cherry, 1953; Ding & Simon, 2012a) regarding the role of spatial information. In classical dichotic listening tasks, participants often listen to two temporally overlapping soundscapes, attending one or the other, and it has been shown that the ability to focus attention to one particular stream depends on certain acoustic factors such as space separation, frequency distance, or semantic level of representation (Alho et al., 2014). However, in classical dichotic listening experiments the two streams are often spatially separable from each other because they are typically presented to the left vs. right ear, respectively. This introduces potential confounds between high-level, e.g., object-based or semantic processes and spatial attention processes. Notably, other recent studies have also addressed the problem of object formation and selective attention without using dichotic stimulation paradigms (Alain & Arnott, 2000; Alho et al., 2014; Best et al., 2007; Bressler et al., 2014; Degerman et al., 2008; Ihlefeld & Shinn-Cunningham, 2008; Lee et al., 2013; Maddox et al., 2015).

Some of the more recent neuroimaging studies also made use of modified dichotic paradigms (i.e. binaural listening) in which the same signal is presented to both ears. In these studies participants selectively listened to one of the superimposed speech streams forming a multi-talker auditory scene (Ding & Simon, 2012; Ghinst et al., 2016; Zion Golumbic et al., 2013a) or speech in synthetic noise (Ding et al., 2014; Ding & Simon, 2013), or tone rhythms (Xiang et al., 2010).

An important difference to these previous studies is that we combined in Experiment 1 two acoustic streams, a speech- and non-speech signal, in an ecologically valid way, as it is a typical scenario in many everyday situations. Combining these different types of streams also brings advantages for the

parsing of the auditory scene in the sense that both streams are less likely to be confused. In order to make the two overlaid acoustic streams comparable we introduce a procedure that allows us to equalize the envelope modulation, i.e. their coarse temporal dynamics, between them by extracting the analytic envelope from one type of signal and (across different trials) re-applying the extracted envelopes to the other type of signals. By this envelope equalization process, the two signals became very comparative in their overall temporal structure, which allowed us to directly compare them within the same attentional weighting experiment. Although we made the two auditory streams in Experiment 1 as comparable as possible, e.g. by adjusting their respective rhythmicity and their signal envelopes, some differences in difficulty remained between the speech versus environmental noise signal. This is most likely due to the fact that the human auditory system is very well tuned to processing human speech signals, resulting in inherent behavioral advantages for identifying targets in the speech stream (Belin et al., 2000; Vouloumanos & Werker, 2004; Zatorre, Belin & Penhune, 2002). Importantly, however, Experiment 2 demonstrated that the object-based attention effects could also be observed in a listening scenario in which two very similar speech streams are overlaid. In this way the second experiment controls for both spatial and low-level feature-based attention (for features such as pitch or frequency), which both cannot be helpful in this specific task. Therefore, while the overlay of two different types of auditory streams clearly has advantages for the parsing of the scene, this is not a prerequisite for object-based attention to work.

Our approach of using a repetition detection task adds an important new behavioral variant to the set of dichotic tasks to study selective auditory attention. Similar repetition detection tasks are often found in working memory studies (Jaeggi et al., 2010; Sussman et al., 2007). Here the repetition detection task is implemented to study high-level object-based attention and was therefore based on a rather long integration window of 750 ms segments. In order to identify the repeating pattern in the auditory stream both segments have to be processed to a relatively deep stage, presumably to a level at which auditory objects are

formed and recognized and at least to some degree attributed some semantic interpretation. This is analogous to recent studies in the visual domain (Baldauf & Desimone, 2014; Kanwisher, McDermott & Chun, 1997) where object-based attention was studied by having participants attend to either a visual stream of spatially overlapping face and house stimuli. In this visual version of object-based attention task, subjects, too, had to identify 1-back repeats in the respectively attended stream, i.e. the re-occurrence of the same face token or house token in two successive presentation cycles. Similar to our present stimuli in the auditory domain, the argument has been made that such a repetition task is logically only possible if the face stimuli have been analyzed at least to the level of face identification processes, which are known to involve comparably late stages of the visual hierarchy in high-level visual areas concerned with object recognition. Consequently, also the attentional modulation by the task was strongest in high-level visual areas in IT cortex. Similar here in the current design, the two segments, i.e. the original sound segment and its repetition about 1s later have to be processed to a comparatively high level of sound recognition, at which at least some meaning or interpretation has been computed from the segments, in order to successfully compare them. Similar to the visual variant of the task, low-level features like the pitch or spectra characteristics of an individual sound, will not allow for a successful comparison and render the detection of a segment repetition very difficult. To accomplish this also from a technical point of view, we put special care in the cutting and clipping process involved in designing the stimulus material for this repetition recognition task: In order not to leave any clipping artifacts or other detectable low-level features in acoustic sequence that could be exploited as low-level, acoustic cues for the to-be-detected repetition targets, we employed special amplitude cross-fading techniques that render the original cutting positions and transition between subsequent segments unnoticeable.

With these carefully designed stimuli and our repetition detection task, we tested mechanisms of selective attentional modulation and the effects of auditory attention on concurrent auditory streams. Following the biased competition

theory (Desimone & Duncan, 1995), selective attention is the central mechanism that biases processing of perceptual stimuli by facilitating the processing of important information and - at the same time - filtering out irrelevant information. In the present study, top-down attention to any of the two acoustic streams (i.e., the speech stream versus the environmental stream in Experiment 1 or either one of the speech streams in Experiment 2) was hypothesized to facilitate the behavioral performance in a high-level, object-based target detection task.

In both experiments, our results clearly showed the hypothesized cue validity effect: the faster and more accurate responses that were given to targets after a valid cue indicate a significant facilitation effect by top-down auditory attention. At the same time, we were also able to see significant inhibition of the respectively non-attended stream (invalid cueing condition) in comparison to a third, neutral attentional condition, in which no cue was presented at all. This replicates many previous cue-validity results and is indicative for the notion that attentional weighting works in a very similar way also on high-level, object-based auditory stimuli, presumably relying on the very same mechanisms as in other modalities or stimulus domains. Both the reaction time data and detection accuracy showed the very same pattern of results and complemented each other. Moreover, analyses of signal detection sensitivity revealed congruent effects with the previous two measures: significantly higher d' sensitivity indices can be observed in the valid cueing condition compared to both the neutral and the invalid cueing condition. There was also a tendency to adjust the decision criterion for the cued versus un-cued auditory stream. Decision criteria were more liberal for the cued and more conservative for the un-cued auditory stream, as it is typical also in classic Posner-type cueing paradigms, for example in vision these tendencies were present in both experiments, accompanying the attentional effects on perceptual sensitivity (d'). They can be explained by the fact that the experimental manipulation of the cue validity requires an unequal number of trials in the valid versus invalid (versus neutral) condition, which changes the a-priori probabilities, with which the target occurs in either stream. Apparently, participants can adopt independent decision criteria (i.e., adjust their

response thresholds) for parts of the auditory stream that are more or less likely to contain the target. Since we designed the experiment with a cue validity of 70%, participants may have also adopted a response strategy of being more liberal in identifying a repetition target in the cued stream, and thus also producing more false alarms than in the invalid cueing condition. The lower the probability with which a repetition target can occur in each of the competing sound streams, the more sensory evidence is required for a decision to report that repetition (and vice versa) (Müller & Findlay, 1987).

Comparable top-down cueing effects to ours were observed behaviorally in tasks based on the Posner-cueing paradigm (Bagherzadeh et al., 2017; Baldauf, 2015; Baldauf & Desimone, 2016; Baldauf & Deubel, 2010; Baldauf et al., 2016; Mangun & Hillyard, 1991; Moore & Zirnsak, 2017; Posner, 1980; Voytek et al., 2017). In a prototypical Posner-cueing paradigm, participants have to fixate a central point on the screen and to attend covertly to either side of the fixation point in order to detect the temporal onset of a brief target stimulus. Such Posner-cueing paradigms also exist for other, non-spatial attentional scenarios such as visual features (Andersen, Fuchs & Müller, 2011; Hopf et al., 2004; Liu et al., 2007b; Maunsell & Treue, 2006; Müller et al., 2006; Sàenz et al., 2003; Störmer & Alvarez, 2014; Treue & Trujillo, 1999), auditory features (Andersen et al., 2011; Costa et al., 2013; Elhilali et al., 2009b; Krumbholz, Eickhoff & Fink, 2007; Shamma et al., 2011; Woods & Alain, 1993; Woods & McDermott, 2015; Zotkin et al., 2003) and visual objects (Baldauf & Desimone, 2014; Kim et al., 2017; Liu, 2016; Zhang et al., 2017), all of which exhibit reliable attentional facilitation effects. The robust finding of such ‘cue-validity effects’ in our study proofs that the concept of attentional weighting and biased competition also hold for high-level attention sets in the auditory domain and that the cueing paradigm in combination with a high-level repetition detection task can be used to study attentional facilitation on an object-based level of the auditory processing hierarchy.

It is hard to provide an exhaustive explanation of how auditory objects are constructed within the auditory processing hierarchy, but clearly the formation of

auditory objects has an inherent temporal dimension, which visual objects don't have necessarily: in audition, we store representation of certain spectro-temporal regularities only with the unfolding of the sounds over time, and on that basis we can then parse the complex auditory scene into discrete object representations. Up to now it is still not fully understood to what extent attention is required to identify irregularities in a sound stream (Cusack et al., 2004b; Ruusuvirta, Huotilainen & Näätänen, 2007; Sussman et al., 2005), in our task the unexpected event was the repetition of a fairly large temporal segment. This could have been identified as such only if the participants have previously directed their attention to a stream, building and recognizing the auditory objects that were forming that stream and recognize the same segment of objects being played again.

Of course, working memory plays a crucial role in solving the repetition detection task. As Conway and colleagues pointed out, auditory working memory poses important constraints on the process of object formation and the involved high-level selection processes (Colflesh & Conway, 2007; Conway, Cowan & Bunting, 2001). Given that the temporal dimension of auditory signals is so inherently important for the parsing of object information, working memory is needed as key component. In our task for example, in order to detect repetitions in one stream, the parsed high-level object information needs to be stored and continuously updated in a working memory buffer so that any new incoming information can sequentially be matched against these stored templates.

In conclusion, our present study complements previous research that used behavioral paradigms to investigate high-level auditory attention, e.g., in a multi-talker, cocktail-party sound scenes, offering two novel aspects compared to the previous literature. First, we combined a modified Posner-paradigm and a repetition detection task in order to study the high-level, object-based aspects of selective attention in acoustic scenes. This attention task has the advantage that it cannot be solved based on the detection of simple low-level features, but instead it strictly requires a deep, object-level or semantic-level processing of the auditory stream, allowing for investigation of the attentional weighting at higher levels of the auditory processing hierarchy. Second, we used speech streams in

combination with field-recordings of environmental sounds as competing sound objects, allowing us to study a particularly ecologically valid situation of competing, spatially overlapping soundscapes. Our results show robust cue-validity effects of object-based auditory attention.

3 Third Chapter:
Neural Correlates of Auditory Object-Based Attention studied with Magnetoencephalography and Naturalistic Soundscapes

3.1 Introduction

The ability of the brain to group together sounds originating from the same source, while simultaneously segregating sounds originating from different sources – “auditory scene analysis” (Bregman, 1990) – and the deployment of attention to select relevant sounds in the soundscape – the “cocktail party problem” (Cherry, 1953) – has been studied mainly by human psychophysical studies. The underlying neuroscience in humans and functional neuroimaging of the topic has instead seen only a relatively recent development. The common experimental methodologies employed vary from invasive neurophysiology with intracranial electrodes, like the electrocorticography (ECoG) in patients that require such methodology for clinical treatment, to whole brain non-invasive techniques with EEG, MEG, and fMRI. To investigate the neural basis that leads to the parsing of relevant sounds, most of the auditory neuroimaging research focuses on tasks with the spatial cues component in the cocktail party phenomenon and experimental paradigms that focus on non-spatial components to solve the same problem. In real-world settings, however, both components contribute to the formation and selection of auditory objects that ultimately enable a coherent perception of the soundscape. The field of auditory research that focuses on spatial information is mostly based on main acoustic cues that allow the separation between target and masker sounds. To segregate sound sources humans leverage the difference in time (Interaural Time Difference, ITD) and the difference in the sound levels (Interaural Level Difference, ILD) at which the acoustic waves reach the two ears (for a comprehensive review see Ahveninen, Kopčo & Jääskeläinen, 2014).

Spatial cues facilitate the sound segregation but are not strictly required to segregate the sound elements in a scene (Hawley, Litovsky & Culling, 2004), therefore the use of experimental stimuli lacking spatial information empathizes other functional aspects of the neural mechanisms involved in the resolution of

an auditory scene. Previous studies have shown that a broad variety of approaches has been applied, with stimuli that range from simple tones to fully naturalistic speech streams, with and without attentional manipulation paradigms.

Experiments that do not direct attention and use simple tones (Gutschalk et al., 2005) like the classic ABA pip-tone triplets and more complex patterns (Gutschalk, Micheyl & Oxenham, 2008; Gutschalk et al., 2007) have revealed an increase in the MEG response similar to the corresponding N1-component when the respective tone or pattern was detected, and no or a much weaker MEG response when the same tones went undetected. These results seem to represent the perceptual segregation of the auditory scene in a possible foreground-background manner.

When controlled manipulation of attention to a rhythmic pattern similar to the one applied by Gutschalk et al. (2008) is integrated in the experimental paradigm (Elhilali et al., 2009b), the event related field (ERF) in the MEG signal - again consistent with the N1-component - is stronger when the listeners' attention was focused on the rhythmic pattern compared to conditions with unattended background random sound patterns. Similar results (Xiang et al., 2010) were found also in case of competing simple patterns with different amplitude modulation rates, which has the advantage of creating a simple acoustic scene with two perceptually different streams. The participants are instructed to attend one or the other stream and consequently perceive the attended acoustic signal as foreground, suppressing the unattended one in background. The evoked MEG response (ERF), as in the previous case, was significantly stronger when the sound pattern was attended, and therefore perceived as 'foreground', than when the same pattern was the unattended stream.

Recent developments in the field have been focusing on employing natural speech stimuli to tackle the neural mechanisms underneath a typical cocktail party problem. The inherent quasi-rhythmic properties of the speech and the innate tuning towards speech signals in humans, make this category of stimuli particularly suitable for the investigation of temporal encoding, especially in high

temporal resolution imaging modalities like MEG (and to some extent also EEG) (Biesmans et al., 2017; Biesmans et al., 2015; Di Liberto, O'Sullivan & Lalor, 2015; Ding et al., 2014; Ding et al., 2016; Luo & Poeppel, 2007; Luo & Poeppel, 2012).

In this kind of studies it has been proven fruitful to apply linear system analysis methods on the acoustic and neural envelope to predict the neural responses to a given stimulus and also to reconstruct a stimulus feature from the neural responses (Ding & Simon, 2012).

The neural mechanisms that direct attention to one speech stream in a soundscape has been tested by Ding and colleagues in its simplest form with one stream masked by an artificial noise background (Ding et al., 2014; Ding & Simon, 2013). The first study revealed that the neural representation of the speech envelope is largely unaffected by stationary noise for moderate and moderately poor signal-to-noise (SNR) ratio, while it suddenly floors close to zero when the SNR is poor. The latter study introduced a spectral distortion by frequency band-vocoding the stimuli, a manipulation that alters the fine temporal and spectral structure while leaving the acoustic envelope unaltered. Even though the neural representation computed with linear system methods mirror just the acoustic envelope, this research demonstrates that the access to spectro-temporal fine structure of speech stream is necessary for neurons to efficiently separate the speech signal from noise. The approach of using a speech stream in stationary noise has potential to highlight the mechanisms of the neural representation of one stream in a sound scene, however these same mechanisms are probably more related to the low-level features of the stream formation in the auditory cortex than to the attention manipulation and its neural underpinning.

A better implementation of the cocktail party problem entails competitive speech streams in which a listener has to segregate one stream from the other. The problem has been investigated in different modalities. In EEG recordings by Power and colleagues (Power et al., 2012), MEG (Akram et al., 2016; Ding & Simon, 2012), and in ECoG by Mesgarani and colleagues (Mesgarani & Chang,

2012; Zion Golumbic et al., 2013a) both showing results in strong agreement with each other that demonstrate that the neural representation of the attended speech is stronger than that of the unattended speech. This effect can be interpreted as “attentional gain”, and this is especially true when the experiment contains a dichotic listening task (Ding & Simon, 2012a) or when low-level spectral features are manipulated (Ahveninen et al., 2011; Elhilali et al., 2009b), in a way that the two streams become more easily separable. Top-down attention indeed seems to enhance the perception of a stream, partly by suppressing the other one (Bidet-Caulet et al., 2007). However it has been argued that attention plays also a role in auditory object formation, but how and when the two mechanism integrate to form a neural representation of auditory objects is not yet clear (Bizley & Cohen, 2013; Elhilali & Shamma, 2009; Shinn-Cunningham, 2008; Shinn-Cunningham et al., 2017).

Expanding from our behavioral work (see Chapter 2 of the present thesis) we designed an MEG study using a similar set of stimuli that has been proven effective to study attentional facilitation and inhibition on an object-based level of the auditory processing hierarchy and that showed robust ‘cue-validity effects’.

In brief, in every trial, a 750 ms long segment is repeated in one of the two overlapping streams (in the speech stream or the environment background) and participants are asked to detect any such repetitions of auditory objects as fast as possible. The logic behind this paradigm is that such a repetition detection task requires the participants to fully process the acoustic stream to a cognitive level that allows them to recognize a certain, temporally extended set of low-level features as an object and to understand that this set of features was repeated.

Most importantly our paradigm in MEG stands out in several aspects from previous neuroimaging paradigms presented above. In fact, it exploits the detection task from studies that usually employ simple stimuli (with and without attention manipulation), and embeds the attentional modulation in a ecologic scenario with naturalistic streams, similar to the ones used by studies that tackle the neural representation of speech streams in cocktail party acoustic scenes.

Since we demonstrated in the behavioral study (see Chapter 2) that the cueing paradigm in combination with a high-level repetition detection task can be used to study attentional effects on an object-based level in auditory domain (i.e. whether top-down selective attention can weigh incoming acoustic information at the level of segregated auditory objects by facilitation and/or inhibition processes), we hypothesized that the same behavioral effect is represented on a cortical level with an enhanced neural activity for the attended stream and at same time an active suppression of the unattended stream, timely related to the identification of the repeated segment. Going beyond the previous behavioral studies, we therefore computed evoked field potentials (ERFs) at the source level to test whether the attentional weighting effect is indeed mapped as early as in auditory cortex as well as in higher level areas involved in the attentional network, supporting the hypothesis that auditory attention is involved in auditory object selection as well as in auditory object formation. We also computed the time-frequency representation for the entire duration of the trial to map the ongoing brain oscillatory activity at the source level as a signature of sustained attention.

We acknowledge that, beyond designing a paradigm to tackle specifically the object level of auditory scene processing, we are not able to rule out completely the possibility that the task can be solved just at a sound feature level. However in this case we expect signatures of the ERFs, timelocked to the repetition onset, more similar to the MEG response corresponding to the N1 component, especially in time, like in Gutschalk's studies (Gutschalk et al., 2005; Gutschalk et al., 2008). Moreover, accordingly to temporal coherence models (Elhilali et al., 2009a; Gutschalk & Dykstra, 2014; Shamma et al., 2011) attending to a particular sound characteristic tunes the neural spectro-temporal receptive fields (STRFs) and boosts the neural signal at times of attended feature. Therefore with simple stimuli it is reasonable to expect a correlation between the spectro-temporal features of the stimuli and their spectro-temporal neural representation (e.g. a variation in tone is evident both in the stimulus and neural spectrogram). Within our paradigm we expect instead that the time-frequency

representation reveals more the attention oscillatory activity rather than representing the tracking of specific sound features.

3.2 Methods

3.2.1 Participants

15 healthy participants (Mean = 28,26; SD = 3.23) took part in the study. All had normal or correct-to-normal vision and briefly tested for a balanced left-right hearing perception with a sample of the same stimuli employed in the experiment. All participants were naive in respect to the purpose of the study and they were not familiar with any of the languages used to create the speech stimuli. All participants provided written, informed consent in accordance with the University of Trento Ethical Committee on the Use of Humans as Experimental Subjects. The entire session lasted approximately 2 hours including preparation time (1.2 hours in MEG).

3.2.2 Task and design

To study the neural correlates of object-based attention in a mixed sounds scene, the same task and stimuli of Experiment 1 (Marinato & Baldauf, 2019) have been employed and will be only briefly recapitulate here (for a detailed description of stimuli preparation and task design see Chapter 2).

The experimental stimuli were auditory scenes, consisting of overlapping streams of (a) speech conversations embedded in (b) environmental sounds. All the speech signals were extracted from newscast recordings of various foreign languages from which a segment of 5 seconds was extracted. The environment sounds consisted of field recordings of public human places like airports, streets,

restaurants. We dynamically modulated the envelope of the environmental sounds using envelopes randomly extracted from the speech signals to make the two streams as comparable as possible at the low-level feature of envelope tracking brain mechanism. Moreover, we controlled for spatial confounds converting the two streams in mono by averaging the stereo channels together and presenting them dichotically.

To create the target streams, we inserted small segment repetitions to be used as repetition targets in our listening task (see Fig.1A in Chapter 2). For this we randomly sampled and extracted a short sound epoch of 750 ms and repeated it immediately after the end of the segment that had been sampled. The length of the repetition targets was chosen to roughly correspond to a functional unit like a typical acoustic event in the environment sounds or a couple of syllables/words in the speech signals.

The main experiment was presented using Psychophysics Toolbox Version 3.72 and DataPixxToolbox functions connected to the real-time DataPixx hardware to deliver visual cues and sounds stimuli in a critical real-time manner. Figure 1B in Chapter 2 provides an overview of a typical trial sequence.

At the beginning of each trial, a fixation-cross appeared and participants were instructed to keep central eye fixation throughout the trial (see Figure 1B). After an interval of 1.0-2.0s (randomly jittered) a visual cue was presented, directing auditory attention either to the “Speech” signal stream or to the auditory “Environment” stream, or to neither of them in the neutral condition. After another interval of 0.5-0.75s (randomly jittered) the combined audio scene with overlapping speech and environmental sounds started playing for 5.0 s. Before starting the experiment, the participants were told to pay attention to the cued stream and to respond with a button press as soon as they recognize any repetition in the sound stimuli. Accuracy and speed were equally emphasized during the instruction. Participants responded with their right hand index fingers, using MEG compatible response buttons (DataPixx system).
Overall there were 5 runs of 60 trials each for a total of 300 trials, each trial

starting with a random jitter of 1.5s to 2.5s.

Before the actual data collection, participants were first familiarized with the sound scenes and had a chance to practice their responses to repetitions for a trial block of 15 trials. Overall our experimental design had two factors: (1) Cue validity with the conditions valid (70% of trials), neutral (10% of trials) and invalid (20% of trials), and (2) Position of the repetition target in either the speech (50% of trials) or environmental (50% of trials) sound stream. All conditions were trial-wise intermixed.

3.2.3 Behavioral data analyses

An analysis of the behavioral responses has been conducted following the method used also in Chapter 2: reaction times, accuracy, sensitivity and criterion were first calculated. It is important to check if the participants show the same attentional weighting effects of the previous studies also with the pneumatic MEG audio stimulation equipment instead of standard high-quality headphones that deliver sounds through a moving magnet, not suitable in a magnetically shielded room.

All data analyses were performed with custom scripts in MATLAB. A combination of built-in function and custom code was used in order to conduct descriptive and inferential statistics. For each condition in our 2x3 factor, mean and standard error of the mean (SEM) were calculated both for reaction times and response accuracies.

The accuracy responses were calculated based on the reaction times measured from the participants: trials with early reaction times (responses before the actual repetition onset) were considered wrong answers as well trials with no response at all; all trials with outliers (exceeding 2.5 standard deviations based on each individual subject's mean) were excluded.

Repeated-measurement analyses of variance were computed on accuracy data, mean reaction times, signal detection sensitivity and response biases. To further investigate systematic differences between individual conditions we

computed planned contrasts in form of paired-samples t-tests between the repetition detection rates and reaction times in the valid versus invalid versus neutral cueing condition (both in the speech and environmental sound stream).

3.2.4 MEG data acquisition

Whole-head MEG recordings were obtained at a sampling rate of 1000 Hz with a low pass antialiasing filter at 330Hz and a high pass filter at 0.1Hz.using a 306-channel (204 first-order planar gradiometers, 102 magnetometers) VectorView MEG system (Neuromag, Elekta Inc., Helsinki Finland) in a magnetically shielded room (AK3B, Vacuum Schmelze). For each participant, the individual head shape was digitized with a Polhemus Fastrak digitizer (Polhemus), including fiducial landmarks (nasion, preauricular points) and about 300 additional points on the scalp, all evenly spread out over the participant's head. Landmarks and head-position induction coils were digitized twice to ensure that their spatial accuracy was less than 1 mm. Head movements were monitored by passing small currents through these coils before each run.

From most participants, an anatomical 3D structural image was obtained using a 4T magnetic resonance imaging (MRI) scanner (Bruker Biospin, Ettlingen Germany). with an 8-channel birdcage head coil (magnetization prepared rapid gradient echo, 1 × 1 × 1 mm). The anatomical scans were then 3-D reconstructed using FreeSurfer software (Dale, Fischl & Sereno, 1999; Fischl, Sereno & Dale, 1999) and used in the 3-D forward models of the MEG analyses.

3.2.5 MEG data preprocessing and ERF source space analysis

The data were analyzed with a combination of Brainstorm (Tadel et al., 2011) and FieldTrip (Oostenveld et al., 2011) MATLAB toolboxes as well as custom scripts, following the general standards (Tadel et al., 2019) whenever this

was fitting the goal of our analysis workflow.

The continuous FIF files from each MEG acquisition run were visually inspected for exclusion of noisy recording channels, and system related artifacts (e.g. SQUID jumps) and a maximum of 12 sensors channels per experimental run were removed and interpolated and then external noise was removed offline from the MEG recordings using MaxFilter software (Taulu & Simola, 2006). The continuous data were then linked for each participant and for each run to the Brainstorm Toolbox database and further cardiac and other artifacts were removed using independent components obtained from an extended Infomax ICA (Independent Component Analysis) decomposition algorithm (Lee, Girolami & Sejnowski, 1999). The maximum number of ICA components were extracted as the residual degrees of freedom after signal space separation performed by MaxFilter algorithm.

The continuous data were then segmented in two different sets of epochs of total length 1200ms, with a 200ms baseline period. One set of epochs were time-locked to the stimulus onset and grouped by condition “attend speech” and by condition “attend environment”, serving for the ROIs selection process detailed in the next section. The second set of epochs extracted from the continuous files were time-locked to the repetition onset and grouped by conditions “attended” stream, “not attended” stream and “detected repetition”, “not detected repetition”, serving for the ERF analysis for the object-based attentional effect. Both sets of epochs were processed for an ERF source analysis as follows.

Each epoch was visually inspected, and those containing physiological artifacts or other artifacts not cleaned by previous ICA process were discarded from further analyses, resulting in an average of 19% of trials per participant discarded from further analysis. The epochs were firstly averaged per condition at the sensor level for each run within subjects.

For each participant head-models were computed by co-registering the participant head shape with the reconstructed MRI brain volumes using

FreeSurfer standard pipeline (Dale et al., 1999; Fischl, 2012; Fischl et al., 1999) or, when no individual anatomy was available (six participants) with the standard brain from FreeSurfer warped to the participant brain volume. The source reconstruction was performed at each run level using the minimum-norm estimates (Hämäläinen & Ilmoniemi, 1994) with overlapping spheres implemented in Brainstorm toolbox, for a resulting source space of 15000 vertices. To allow for inter participant comparisons, the averaged source maps were normalized with respect to 200ms baseline (z-scores). Once the source activity was estimated for each individual run, the epochs were averaged across runs and within participants. To obtain brain maps of the neural activity across all participants - group analysis -, each individual source space was projected to a standard FreeSurfer brain as a default anatomy parcellated according to a multi-modal brain atlas from the Human Connectome Project (Glasser et al., 2016). This group analysis served also to the identification and selection of ROIs explained in detail in the next section.

The mean of the vertices from each ROI for each participant was exported from Brainstorm database to MATLAB for an ERF analysis with custom scripts of the “repetition epochs” set. Planned contrast in form of t-test at every time point between two source-reconstructed time courses of each ROI were performed between epochs in which the repetition was embedded in the attended versus not attended stream and between epochs in which the repetition was detected versus not detected trials. The t-tests were corrected for false discovery rate.

For the time-frequency analyses, we used the Fieldtrip toolbox and custom-made code. The signal was Fourier-transformed on a single-trial level and power was estimated for a frequency range from 1-15Hz using a 300ms sliding window with multitapers methods. The resulting power spectra were then also normalized in respect to the baseline period before cue-onset.

3.2.6 ROIs selection

Following the analysis practices established in the lab, we used a recently developed cortical parcellation that provides the most precise insights into the structural and functional organization of the human brain to date (Tabarelli et al., 2020). This parcellation is based on a multi-modal atlas of the human brain developed under the Human Connectome Project and obtained by combining structural, diffusion, functional and resting state MRI data from 210 healthy young individuals to identify 180 regions of interest (ROIs), per hemisphere (Glasser et al., 2016).

The process of ROIs selection unfolds from the mapped cortical activity at source level of the epochs time-locked to the stimulus onset for both conditions, attend speech and attend environment. Overlapping the Glasser's parcellation to the cortical activity maps, we selected a number of areas ranging to early peripheral auditory processing, such as A1, to more high-level processing areas typically involved in the object-based attention networks, such as IFJ (Baldauf & Desimone, 2014). A complete list of this initial selection for attend environment condition is provided in Table 3 as an example. Since we did not find any relevant difference between attend speech and attend environment cortical spatial distribution activity, we averaged the signal between the two hemispheres just for this step. For each area we selected the maximum peak within the epochs length and extracted a 200ms window around that peak. The average activity of this window was tested for significance against the 200ms baseline window with alpha 0.05 and the p-values Bonferroni-corrected for all 360 parcellation of the whole brain. Since the activity of every area was significantly different from the baseline, we used this initial list to form the definitive ROIs selection by grouping the previous selection, this time separately for each hemisphere, in cortical regions as organized in the original Glasser's parcellation. The final ROIs selection is represented with Glasser's regions color code in Figure 4 and comprise the following cortical sectors: 1) Early Auditory Cortex; 2) Auditory Association Cortex; 3) Superior Parietal Cortex; 4) Posterior Opercular Cortex; 5) Insular Frontal Opercular Cortex; 6) Inferior Frontal Cortex; 7)

Temporo-parieto-occipital Junction.

The neural signal extracted from these grouped regions was tested for significant difference against the baseline window with the same method applied for the original ROI parcels, resulting again all significant.

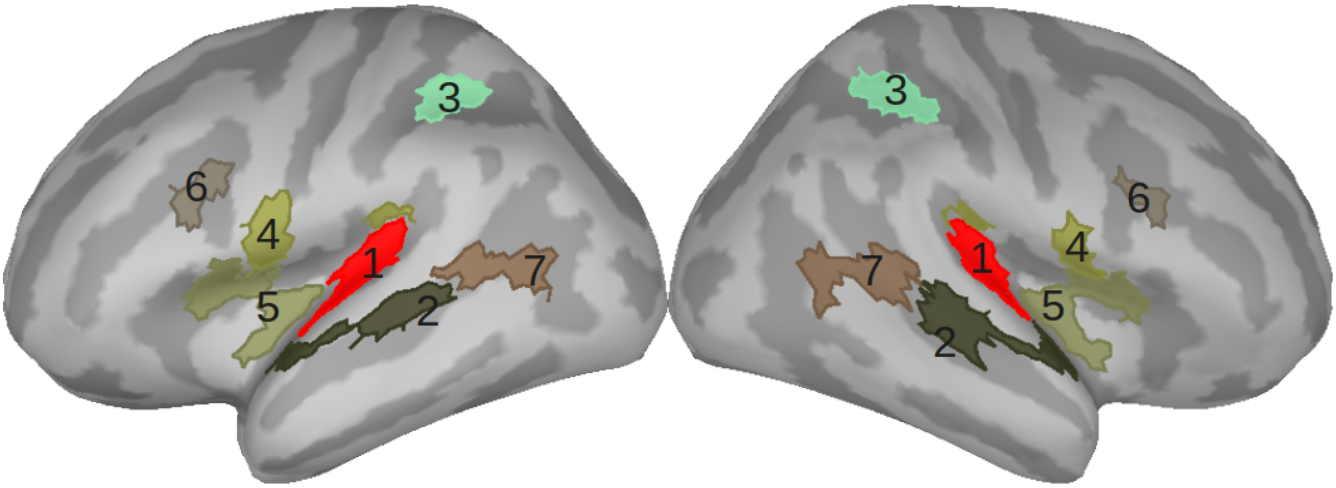

Figure 4: *ROI Selection. Single ROIs from Glasser et al. (2016) were grouped by "Regions" following the same criteria explained in the supplementary material of the same paper. The ROIs correspond in Glasser's terminology to: 1) Early Auditory Cortex; 2) Auditory Association Cortex; 3) Superior Parietal Cortex; 4) Posterior Opercular Cortex; 5) Insular Frontal Opercular Cortex; 6) Inferior Frontal Cortex; 7) Temporo-parieto-occipital Junction.*

Table 3: Significant ROIs for "attend environment" condition for averaged hemispheres

Note: the initial p-values were Bonferroni corrected with total 360 comparison for whole brain areas

Area Name	Area Description	Glasser's Regions	p-Value	Significance at 0.05*
43	Area 43	Posterior Opercular Cortex	0.000006	1
A1	Primary Auditory Cortex	Early Auditory Cortex	0.000000	1
A5	Auditory 5 Complex	Auditory Association	0.000000	1

		Cortex		
AIP	Anterior IntraParietal Area	Superior Parietal Cortex	0.000001	1
FOP1	Frontal Opercular Area 1	Posterior Opercular Cortex	0.000002	1
FOP2	Frontal Opercular Area 2	Insular and Frontal Opercular Cortex	0.000005	1
FOP3	Frontal Opercular Area 3	Insular and Frontal Opercular Cortex	0.000004	1
FOP4	Frontal Opercular Area 4	Insular and Frontal Opercular Cortex	0.000000	1
IFJa	Area IFJa	Inferior Frontal Cortex	0.000002	1
IFJp	Area IFJp	Inferior Frontal Cortex	0.000001	1
LBelt	Lateral Belt Complex	Early Auditory Cortex	0.000000	1
MBelt	Medial Belt Complex	Early Auditory Cortex	0.000002	1
PBelt	ParaBelt Complex	Early Auditory Cortex	0.000000	1
PFcm	Area PFcm	Posterior Opercular Cortex	0.000000	1
PoI1	Area Posterior Insular 1	Insular and Frontal Opercular Cortex	0.000000	1
PoI2	Area Posterior Insular 2	Insular and Frontal Opercular Cortex	0.000000	1
RI	RetroInsular Cortex	Early Auditory Cortex	0.000000	1
STSdp	Area STSd posterior	Auditory Association Cortex	0.000000	1
TA2	Area TA2	Auditory Association Cortex	0.000001	1
TPOJ1	Temporo-parieto-occipital Junction 1	Temporo-parieto-occipital Junction	0.000000	1
TPOJ2	Temporo-parieto-occipital Junction 2	Temporo-parieto-occipital Junction	0.000000	1

3.3 Results

The computation of behavioral measures combined with event related fields analysis at the source level allowed us to better characterize the object-based auditory attention processes across time and space at the cortical level.

3.3.1 Behavioral results

The behavioral results obtained in MEG substantially replicate the results of Experiment 1 detailed in the second chapter, except for the reaction times in the invalid condition of the speech stream which show slower reaction times – although not significantly slower - in respect to the environment signal. One possible explanation could be that the listening fidelity with the pneumatic MEG audio set-up is less optimal than with a high-fidelity pair of headphones, making the task a bit more difficult. Therefore, in the invalid condition, participants were able to catch the repetition segment within the speech stream with more accuracy and sensitivity than within the environmental stream, but it required more time. Instead the response accuracy and sensitivity in the invalid condition for environment signal was so low that the participants were able to catch just the trials in which the repetition segment was more easily detectable, and therefore they were a bit faster than in the speech signal.

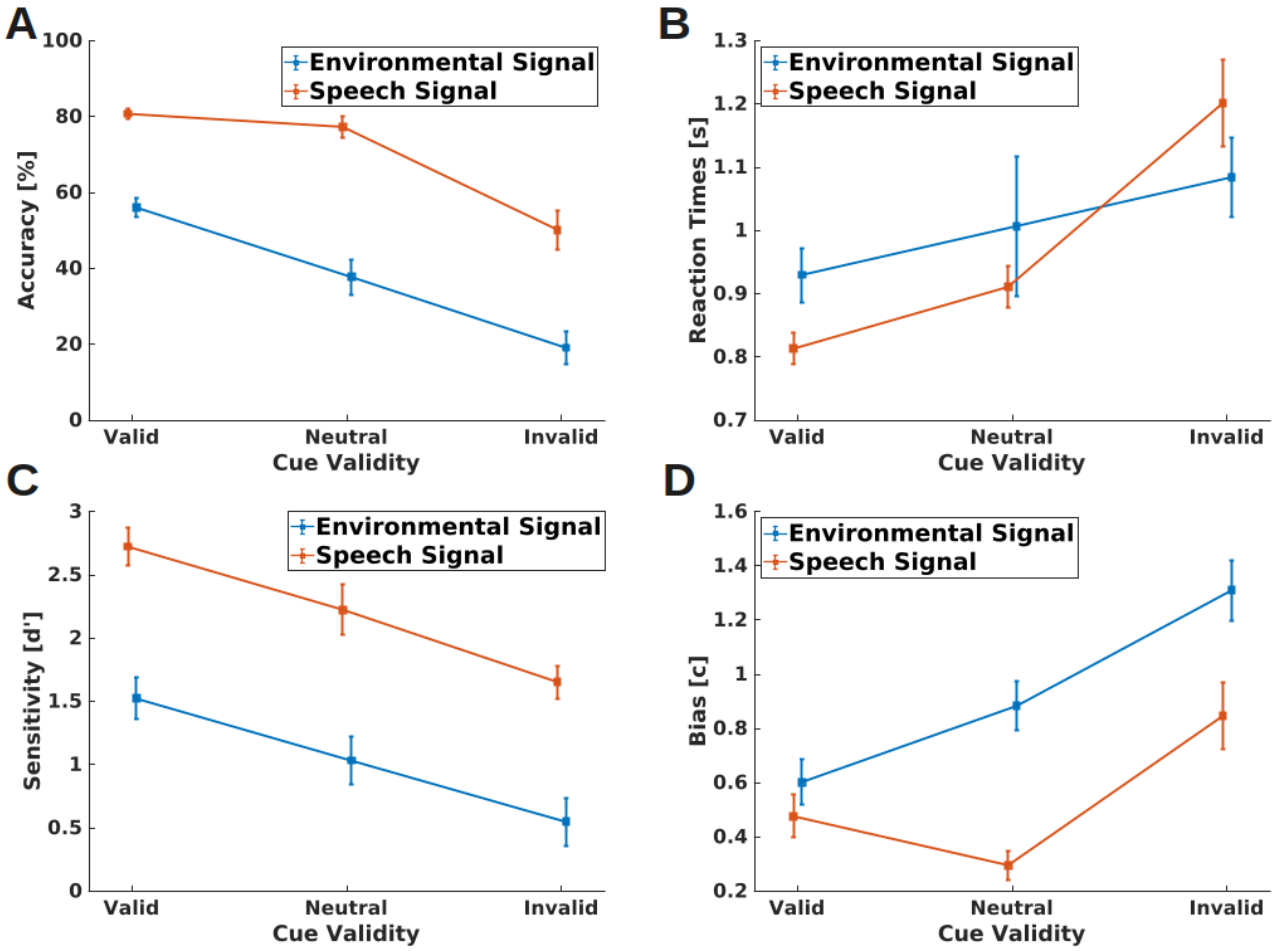

Figure 5 *Experimental results in the repetition detection task with overlaid speech and environmental-noise streams. (A) Detection performances (in percent correct) as a function of cue validity (valid, neutral and invalid cueing condition), separately for both components of the acoustic stimuli, i.e., the environmental signal (blue) and the speech signal (red). The data are shown as means and SEM. (B) Reaction times (in seconds) as a function of cue validity (valid, neutral and invalid cueing condition), separately for both components of the acoustic stimuli (blue: environmental signal, red: speech signal). The data are shown as means and SEM.*

(C) Sensitivity scores (d') and decision criteria (D) as a function of cue validity (valid, neutral and invalid cueing condition), separately for both components of the acoustic stimuli, i.e., the environmental signal (blue) and the speech signal (red). The data are shown as means and SEM.

3.3.2 Sound onset ERF results

The whole brain activity of each participant, time-locked to the sound onset, was projected to a default anatomy template and group averaged to obtain a spatio-temporal map for attend speech and attend environment conditions (see Figure 6, panel A and B respectively). The maps show for both conditions an early activity at 90 ms localized in the early auditory cortex region that later, at about 200ms, also spread to higher-level processing regions like inferior frontal cortex and temporo-parietal occipital junction (see also Figure 4) among others. The event-related field trace for the three regions for the right hemisphere, taken as an example, shows a first prominent peak at 90ms and a more prominent and sustained peak at around 200ms in the early auditory cortex of both conditions. Higher-level processing areas such as insular frontal cortex and temporo-parietal junction show mainly a single peak at a 200ms latency from the sound stimuli presentation, suggesting a hierarchical cascade of information flow. Panels C and D of Figure 6 show three example ROIs of the speech signal onset and environments signal onset ERF analysis. For each area we selected the maximum peak within the epochs length and extracted a fixed 200ms window around that peak (i.e. from -100ms from the selected peak to +100ms from the selected peak), similarly to what we did to identify the ROIs in the previous section. The mean of the 200ms window around the most prominent peak of each ROI was computed and tested for significance difference against the mean of the 200ms baseline window where no stimulation, except for a fixation cross on screen, was presented to the participant. The red triangles in panel C and D of Figure 6, mark the peak and the statistical significance of the window marked by the two red lines, define also a region as selected for the ERF analysis of the subsequent repetition detection epochs.

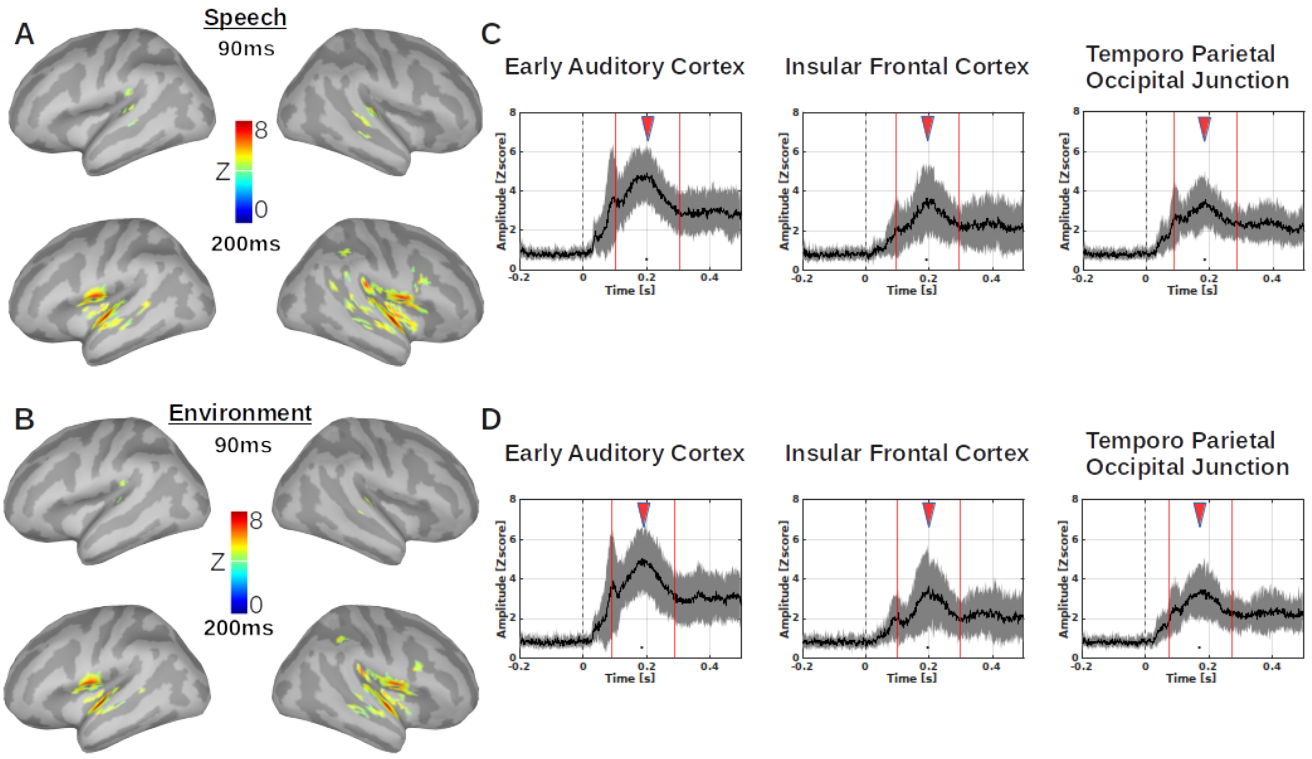

Figure 6. *Cortical activity for condition "attend Speech" and "attend Environment". Panel A and B shows the source reconstructed spatial distribution of the cortical activity for all participants, projected to a default anatomy for "attend speech" and "attend environment" conditions respectively. Panel C and D show the ERF trace for selected ROIs in the right hemisphere for conditions "attend speech" and "attend environment" respectively.*

3.3.3 Repetition onset ERF results

Source-level ERFs of repetition detection epochs were conducted contrasting epochs in which the repetition segment was embedded in the attended stream against epochs in which the target was embedded in the unattended stream, subsequently between epochs in which the repetition was detected against epochs in which the target went undetected.

First in both contrasts the source-level ERFs clearly show a significant pronounced neural activity when the repetition was in the attended stream whereas the repetition in the unattended stream did not (see Figure 7 A-B). Furthermore, the same pronounced effect was shown when the repetition was detected while the neural activity for the undetected repetition is almost completely flat (Figure 8 A-B). Both effects are in accordance with previous

findings in the literature (Gutschalk et al., 2005).

Second, the neural activity in both contrasts and each ROI exhibit a similar flat pattern for about 200ms after the presentation of the repetition which is consistent with the time necessary to recognize that an auditory object was repeated. After this latency period the two signals differentiate from one another.

Third, the time period at which the neural signal unveils a sustained significant difference between the contrasted conditions, largely vary based on the cortical spatial distribution activity which suggests a hierarchy of information flow from lower level processing ROIs to higher-level processing ROIs. In the attended versus not attended contrast (Figure 7 A-B) the auditory associative cortex starts to display a significant distinct pattern at about 350ms followed by insular operational cortex at 400ms, with temporo-parietal occipital junction and superior parietal cortex after 600ms. In the detected versus undetected contrast the auditory associative cortex shows an earlier significantly different neural activity as early as 270ms, the insular frontal operational cortex after 300ms with temporo-parietal junction and superior parietal cortex after 450ms.

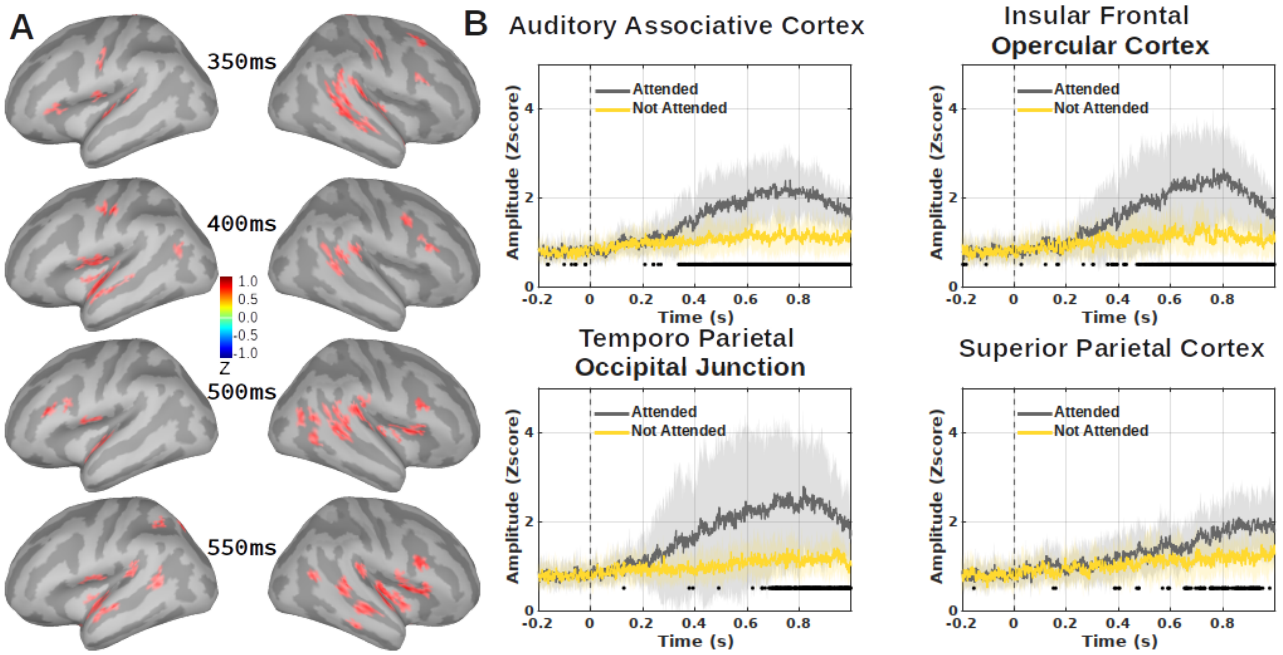

Figure 7. *"Attended" versus "Not Attended" contrast. Panel A show the spatial distribution where the variability of the neural activity between the two condition is greater. Panel B shows the time course for selected ROIs ordered*

following the time point of the significant divergence of the neural activity.

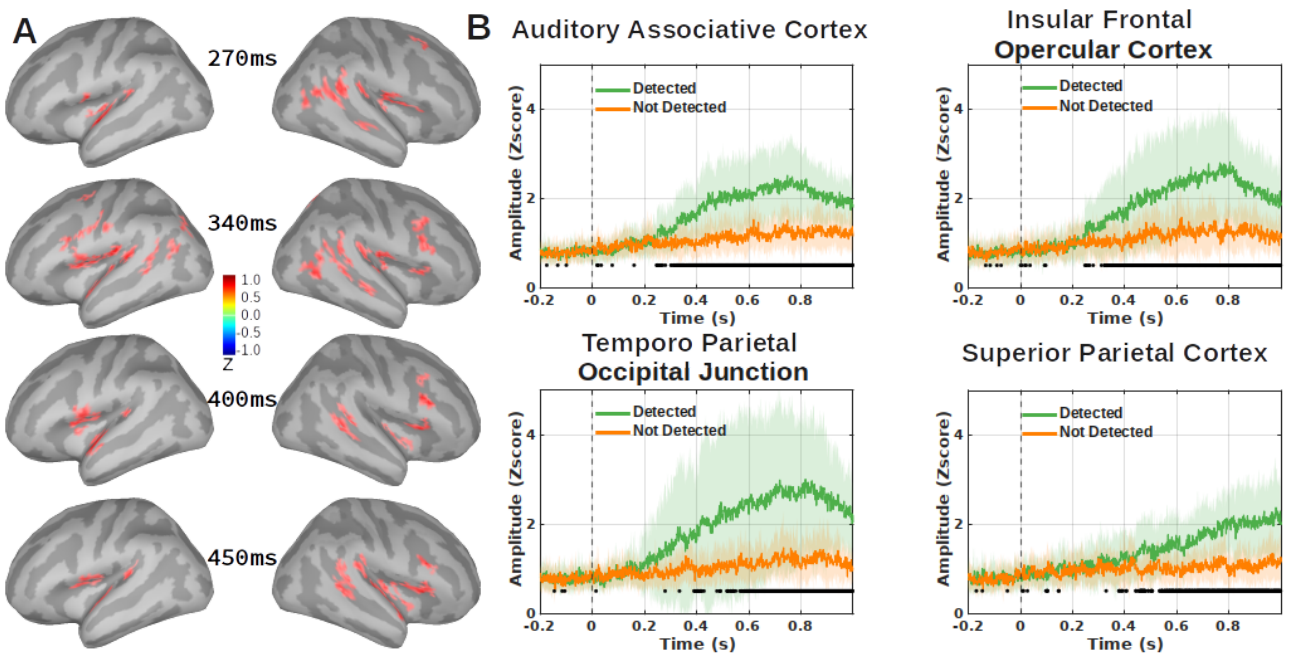

Figure 8. *"Detected" versus "Not Detected" contrast. Panel A show the spatial distribution where the variability of the neural activity between the two condition is greater. Panel B shows the time course for selected ROIs ordered following the time point of the significant divergence of the neural activity.*

3.3.4 Time-frequency analyses during sustained attention

In a next analysis step, we focused on neural correlates of the sustained attention which ensues after the stimulus has started, and last until a target was finally presented in either stream. Because neuronal events cannot be expected to be precisely time-locked, but rather correspond to internal cognitive processes with unknown on- and offsets, we did not analyze evoked responses (ERFs) during this period, but focused on ongoing brain oscillatory activity as a signature of sustained attention. For this purpose, the signal during these epochs was Fourier-transformed on a single-trial level and power was estimated for a frequency range from 1-15Hz using a 300ms sliding window with multitapers methods. The resulting power spectra were then also normalized in respect to the baseline period before stimulus-onset.

Figures 9 and 10 show the average time-frequency responses of all our main regions in the left and right hemisphere, respectively.

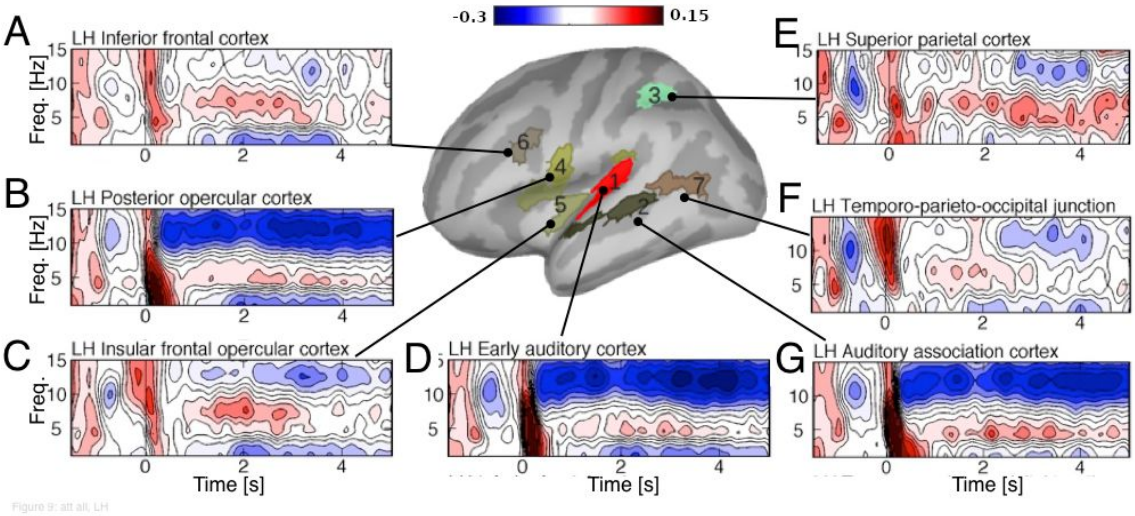

Figure 9. *Time-frequency spectrograms of the stimulus period with sustained attention, for the main regions of interest in the left hemisphere (LH). In general, the spectrograms were characterized by an initial broad-band response after the auditory stimulus was presented. After this initial evoked response had declined (at about 300ms), sustained attention was paid to either stimulus stream (attend environment or attend speech, here the data for both conditions was combined). As sustained attention is paid to the stimulus, most of the cortical areas show entrainment to a low-frequency activity at about 5Hz, which corresponds to the acoustic envelope of the auditory stimulus material. This increase in the low-frequency response was accompanied by a sustained inhibition of the alpha band around 8-12 Hz.*

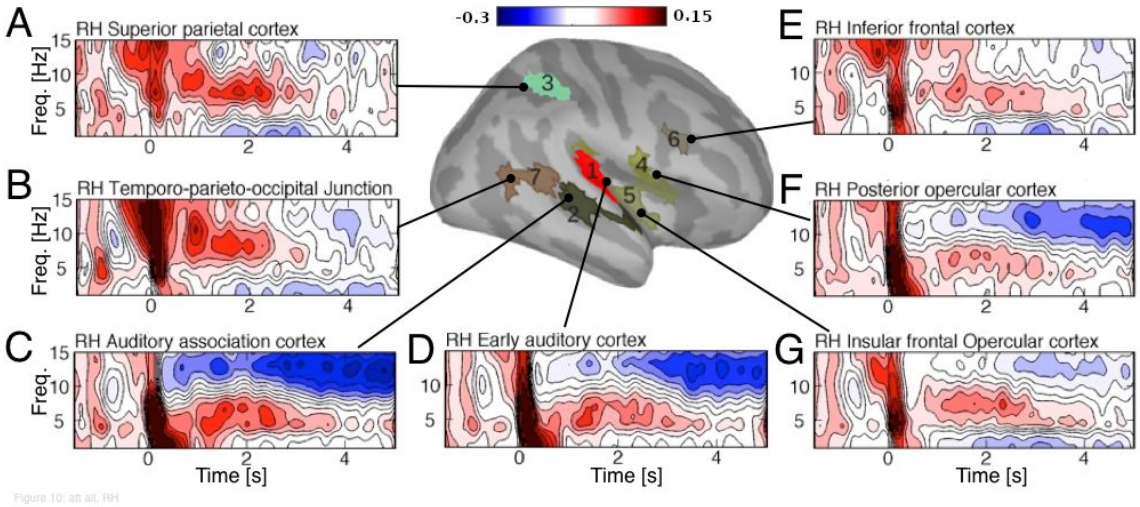

Figure 10. *Time-frequency spectrograms of the stimulus period with sustained attention, for the main regions of interest in the right hemisphere (RH, for a detailed description see Figure 9).*

When both sustained attention conditions are combined (see Fig.9 and 10, for left and right hemisphere), the spectrograms showed a strong initial broad-band response when the auditory stimulus was presented (0s). This broadband response was the strongest in early and higher auditory cortices – much stronger than in the parietal or frontal sites. This indicates that the broad-band response corresponds mostly to the evoked responses and event-related field changes presented earlier, which were also most prominent in these sensory auditory areas. After this initial evoked response had vanished (at about 300ms), sustained attention was paid to either stimulus stream. Here, the data for both conditions (i.e., 'attend environment' or 'attend speech' were combined). While sustained attention was paid to either stimulus, most of the cortical areas show entrainment to a low-frequency component at about 5Hz, which corresponds to the acoustic envelope of the auditory stimulus material. This entrainment is most likely due to early auditory cortices tracking closely the envelope of the auditory stimuli. The increase in the low-frequency response (entrainment) was accompanied by a substantial inhibition of the alpha band around 8-12 Hz.

The following Figures 11 and 12 show the time-frequency responses

specifically for the condition, in which sustained attention was deployed to the environmental stimulus stream. Again, the data is presented for the main regions of the left and right hemisphere, respectively.

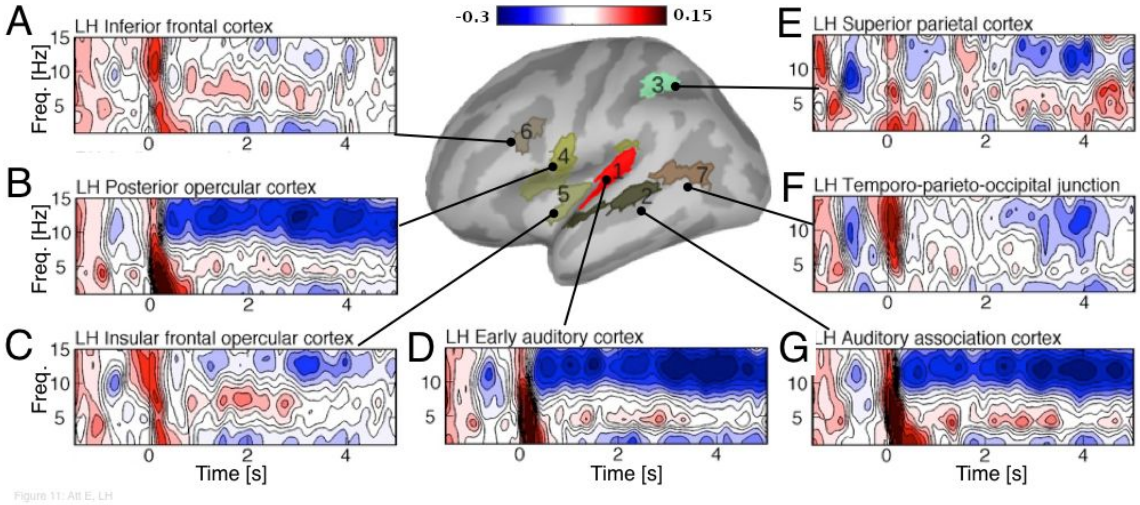

Figure 11. *Time-frequency spectrograms of the stimulus period with sustained attention directed to the stream with environmental noise, for the main regions of interest in the left hemisphere (LH, for a detailed description see Figure 9).*

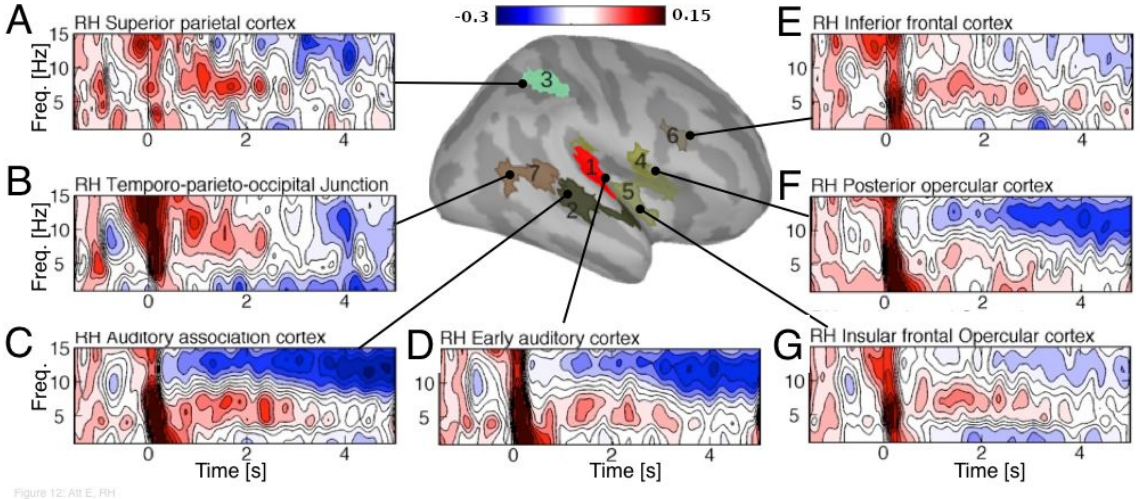

Figure 12. *Time-frequency spectrograms of the stimulus period with sustained attention directed to the stream with environmental noise, for the main regions of interest in the right hemisphere (RH, for a detailed description see Figure 9).*

As can be seen from the time-frequency spectrograms, the entrainment component at about 5Hz is less prominent. This could mirror the fact that subjects reported the 'attend environmental stream' condition to be more difficult. Also, the less prominent entrainment in this low frequency band could be a neural correlate of the weaker behavioral performance in this attentional condition. Further, it can be seen that the alpha suppression is nevertheless strong over the whole stimulus period. Even more, it becomes clear that the alpha band – a correlate of sustained sensory attention – becomes even stronger as the stimulus proceeds. This could indicate the sustained and increasing effort in this attentionally demanding condition.

Figures 13 and 14, finally, give the time-frequency responses specifically for the condition, in which the speech stream had to be attended (Fig. 13: left hemisphere, Fig. 14: right hemisphere).

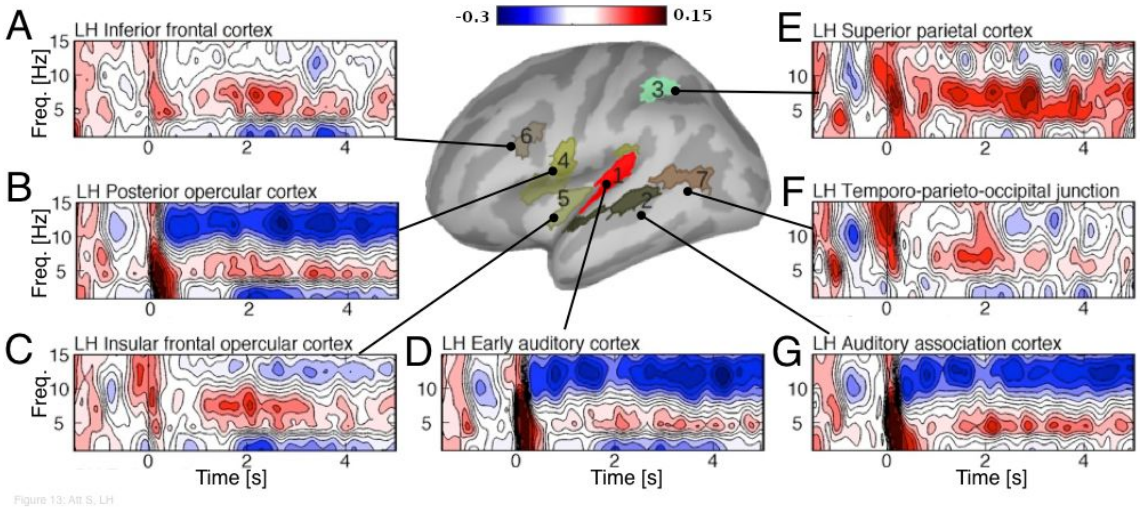

Figure 13. *Time-frequency spectrograms of the stimulus period with sustained attention directed to the stream with speech signal, for all the main regions of interest in the left hemisphere (LH, for a detailed description see Figure 9).*

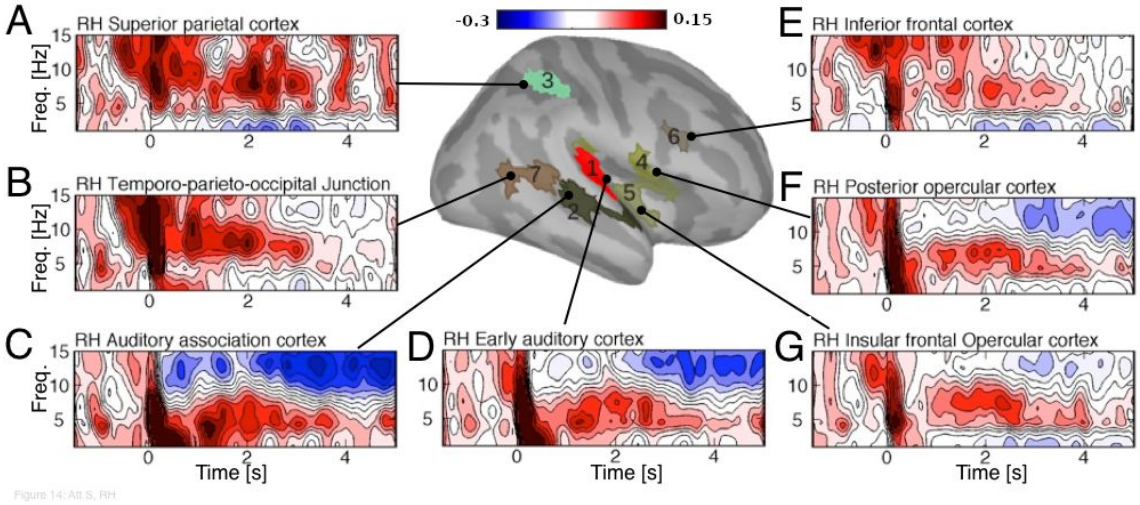

Figure 14. *Time-frequency spectrograms of the stimulus period with sustained attention directed to the stream with speech signals, for all the main regions of interest in the right hemisphere (RH, for a detailed description see Figure 9).*

In comparison to the attend-environment condition, the entrainment by the

stimulus envelope at about 5Hz was stronger for the attend-speech condition. This is especially true for the left hemisphere, where the signal is very dominated by this component. This is consistent with much of the left hemisphere's specialization to language-related processing. Interestingly, also the attentional suppression of the alpha-band in the sensory areas was more pronounced on the language-prone left hemisphere. In the right hemisphere (Fig. 14), we also see some entrainment to the attended speech stimuli, but to a lesser degree, and mostly in early auditory cortex and directly adjacent sites. Most interestingly, in terms of attentional control, the right hemisphere showed not only less stimulus entrainment and less alpha suppression (in comparison to the language-dominant left hemisphere) when attention was directed to speech-signals, but also a marked increase in alpha activity in the right parietal cortex, right inferior frontal cortex and the temporo-parietal junction. This is intriguing as these sites are known to be important sources of top-down control signals. The increase in alpha activity could therefore be a sign of suppression of these top-down sources in the hemisphere that less concerned with language, while the focus of attention is shifted to the language-circuits on the left side.

3.4 Discussion

In the present MEG-study, we set out to investigate the neural correlates of the object-based attention deployed in a complex naturalistic soundscape by inspecting the spatio-temporal dynamic of the cortical activity before and after the point in time when a behavioral target was presented. To do so we used a repetition detection task combined with a typical Posner-type cueing paradigm designed by us and first introduced in a behavioral study in the Chapter 2 of the thesis. With this new paradigm we tried to tackle the object-based level of

processing in auditory attention. The experimental manipulation of the attention resulted in a behavioral facilitation effects in the valid cueing trials and inhibition in the invalid cueing trials (both in respect to a neutral cueing condition). These behavioral effects were then represented at the level of cortical activity by a significantly stronger activation when the repetition target was embedded in the attended stream, while no evoked response was measurable when it was embedded in the unattended stream. Moreover, a similar result was shown when contrasting the time course of the evoked responses when the repetition segment was detected versus when it went undetected. Here it is noteworthy, that the biased competition between stimuli and strong top-down guidance can often lead to extreme perceptual biases, even to the point where not attended stimuli are not detected at all (a form of inattention blindness, see e.g., Drew et al., 2013).

Interestingly, the analysis of event-related neural responses revealed different significant temporal patterns depending on the selected region of interest, or, in other words, exhibited different temporal onsets based on the spatial distribution of the cortical activity. Specifically, in the attended versus not attended contrast the neural response timing manifested itself in a significant increasing activation when the repetition was in the attended stream starting at around 350ms in the auditory associative cortex, and only later showed similar effects in higher-level cortical areas, such as in insular frontal opercular cortex, temporo-parieto-occipital junction and superior temporal cortex. Contrasting the detected versus undetected neural response gave a similar spatial distribution of the cortical activity and its temporal evolution, but with an earlier time-point start of the stronger neural representation of the detected neural activity. Most importantly when the repetition target was present in the unattended stream or when it went undetected, the event related trace maintained an almost flat activity similar to the baseline window, indicating that the MEG response is not detectable as much as the participant's perception of the repetition. These can be interpreted as a very strong neural representation of the biased-competition model (Desimone & Duncan, 1995) in auditory domain within complex naturalistic

streams.

In both contrast, the evoked magnetic fields start to differentiate from each other later than the usual N1-response reported in previous studies using cocktail-party stimuli made of simple tones or patterns (Ahveninen et al., 2011; Elhilali et al., 2009b). We argue that this was most likely due to the inherent time necessary to process the task at the level of a complex auditory object, which required to bind together low-level features in higher-order objects, store those in working memory, process the incoming signals and recognize that some of the objects were actually repeating. We argue in fact that the timing of the evoked magnetic field is indeed a good marker candidate to disentangle a possible feature-based attention confound within our paradigm. Different auditory objects in fact carry inevitably different low level features that were grouped or segregated in the formation of the target object (i.e. the repetition) over time. However an important aspect is that the variability of the stimuli sound signals do not encourage the development of strategies based on specific low-level “local” features like harmonicity (Carlyon, 2004; Griffiths & Warren, 2004) or high level features like pitch, timber or even reconstruction of meaning. Instead a specific object-based strategy based on the identification of the repetition object - at each trial a formation over time of unique combinations of many features that makes an object - was the most likely solution to the task.

The most important aspect that has to be highlighted here is the fact that analyzing the ERF in response to repetition target onsets, in fact, shed also a light on the temporal dynamic of what we argue is the representation of the object-based attention processes - not only a mere representation of the attended object itself.

In paradigms, in which the task consist of attending one or the other stream (Ding & Simon, 2012; Lalor & Foxe, 2010; O’Sullivan et al., 2015; Power et al., 2012; Xiang et al., 2010) it has been demonstrated that the top-down attention mechanism is often based on the phase entrainment of the neural oscillations. Neuronal oscillators in auditory cortes follow closely the temporal dynamics of the envelope of the attended acoustic stream. The “selective entrainment

hypothesis" (Giraud & Poeppel, 2012; Zion-Golumbic & Schroeder, 2012) show that low-frequency and high-gamma entrainment are both involved in modulating attention, with low-frequency (delta-theta rhythm) having broader topographic distribution, involving not only low-level auditory areas, but also the anterior and inferior temporal cortex and language-processing cortices, as well as brain regions involved in the attentional control, such as the inferior parietal lobule and inferior frontal cortex, including the inferior-frontal junction (IFJ) (Baldauf & Desimone, 2014; Shinn-Cunningham et al., 2017; Zion Golumbic et al., 2013a). Our findings are novel in respect to the detailed source-localization we provided based on minimum-norm estimates of the original MEG signal, and the millisecond-precision temporal dynamic of the event-related responses. Our detailed description of events is, however, consistent with the interpretation of the selective entrainment hypothesis, since we found that the larger MEG response of the attended and detected condition extended beyond low-level and associative auditory cortex but also to regions involved in attentional processing like superior temporal cortex and temporo-parietal occipital junction. Especially the latter has been identified also in visual studies (Corbetta, Patel & Shulman, 2008) as responsible for reorienting attention both due to bottom-up as well as top-down processes.

Our time-frequency analyses during periods of sustained attention are in general compatible with previous studies. When sustained attention was paid to either stimulus stream most of the cortical areas showed strong entrainment by the acoustic envelopes of the auditory stimuli (for similar results of cortical entrainment see, e.g., Baldauf & Desimone, 2014; de Vries & Baldauf, 2019; Tabarelli et al., 2020). These entrainment effects are most likely due to early auditory cortices tracking closely the envelope of the auditory stimuli, and exhibiting phase-correlated rhythmic activations (Lakatos et al., 2013; Mesgarani & Chang, 2012; Morillon & Schroeder, 2015). This entrainment component was present consistently, and independently of the attentional condition, i.e. whether sustained attention was directed to the speech or environmental stimulus. However, the entrainment was clearer and stronger when attention was directed

to the speech stream, especially in early auditory cortex, associate auditory cortices and frontal auditory sites of the left hemisphere. This is expected, given the known left-lateralization of the language-processing system in humans. In trials, in which the environmental stream had to be attended, the low-frequency entrainment component to the stimuli was more balanced across both hemispheres.

When sustained attention was instead focused on the environmental stimulus stream, the entrainment component to the quasi-rhythmic environmental noise at about 5Hz was less strong. This is most likely related to the fact that subjects reported the ‘attend environmental stream’ condition to be more difficult. The less prominent entrainment to the stimulus’ envelope modulation could therefore also be a neural correlate of the weaker behavioral performance in this attentional condition.

The strong entrainment by the quasi-rhythmic acoustic stimuli in auditory cortex was accompanied by a strong inhibition of the alpha band, in all attentional conditions, potentially reflecting the sustained attentional effort (Bagherzadeh et al., 2017; Bagherzadeh et al., 2020; Baldauf & Desimone, 2016; Baldauf et al., 2016; Keefe & Störmer, 2020). This alpha-suppression was strongest when sustained attention was focused on the environmental stimulus stream, which was also behavioral slightly more difficult. In many regions of the auditory cortex, the alpha-suppression became even stronger as the stimulus proceeded (see similar in Baldauf & Desimone, 2016). This might reflect the behavioral results that subjects needed some time to track the attended stream, and that they tended to track the attended stream better the longer the stimuli went on.

Initially the decision to concentrate the analysis preferably on theta and delta frequency bands in relation to the modulation of alpha was justified by the set of stimuli used in our paradigm considering some evidence from previous works. In line with the selective entrainment hypothesis, and the amplitude modulation attentional effect, the phase of low frequencies has been known to track the temporal dynamics of the envelope of a quasi rhythmic stimulus, especially speech (Ding & Simon, 2012; Giraud & Poeppel, 2012; Kerlin, Shahin

& Miller, 2010; Lakatos et al., 2008; Zion Golumbic et al., 2013a). High gamma power instead has been linked to represent specific features of the speech at the segmental and diphonic level (Giraud & Poeppel, 2012; Luo & Poeppel, 2012; Zion Golumbic et al., 2013a). Our set of stimuli share the common property of a temporal quasi rhythmic characteristic of the amplitude envelope, in fact the speech stimuli consist of exotic languages with very different structures and with no access to any kind of meaning, and the environmental sound signals were amplitude modulated on the speech envelope, but do not share the same fine grained temporal structure of the speech signal. This complementary choice led to the identification of the low-frequency modulation as the candidate marker to a more equal comparison of the effects of the object-based attention with both sets of stimuli. Beyond these considerations is nonetheless interesting to note that when time-frequency decomposition was computed also for beta and gamma bands (see appendix “A”), it showed an increase of gamma activity in the auditory association cortex, especially in the attend speech condition. This can be referred to previous results in the literature (Zion Golumbic et al., 2013a) which point at gamma as a possible neural oscillation, crucial for segmental and diphonic level speech processing. In our results the fact that the increase is localized mainly in the auditory association cortex, and does not display a more widespread pattern, might suggest that is less directly linked to the attention mechanisms and more to the specific language processing.

In the case of crafted auditory scenes made of simple stimuli, a coherence analysis between the spectral features of the stimuli and the neural perceptual representation (Shamma et al., 2011) may help detect the involvement of feature-based attention processes that contribute to the foreground-background segregation through mapping the stimulus features in the neural response spectrogram. However with rich and complex naturalistic scenes made of multiple concurrent sounds the encoding of sounds features and neural response can be less direct. A better implementation would be to design an entirely new paradigm that is able to scale from simple stimuli to more complex scene in controlled steps (e.g. adding more sounds objects) in order to highlight the

processes for which a feature-based attention approach is preferred over a full object-based attention approach in order to solve the task.

In general, it was this combination of neural entrainment to the envelope of the attended stimulus stream and the alpha-suppression that persistently characterized the results of our time-frequency analyses, across all experimental conditions. Both of these components were systematically more left-lateralized in the condition, in which the speech stream was in the focus of attention. And both components seem to be congruent with the different task difficulties of the various conditions and the behavioral results.

3.4.1 Final considerations and future research

The event-related analysis based on the selected ROIs provided new insights on the spatio-temporal dynamic of the neural correlates of object-based attention and its modulation within naturalistic soundscapes. These findings integrate well with results of a diverse body of previous studies comprising other MEG studies that made use of simple tones stimuli as well as overlapping, competing speech streams. With a relatively simple analysis we were able to find the neural correlates of the attentional modulation effects observed in the biased competition model of the neural representation of the perceptual objects in vision (Baldauf & Desimone, 2014; Desimone & Duncan, 1995; Kastner & Ungerleider, 2001). The model argues that the attention focus is determined by the interplay between exogenous salience of stimuli and endogenous observer goals. Evidence of such mechanisms in the auditory domain is starting to emerge from an increasing number of studies (Kaya & Elhilali, 2014; Lee et al., 2013; Mesgarani & Chang, 2012). This interplay between exogenous and endogenous factors, or in other words the interactions between bottom-up and top-down components of the attentional modulation, are a central argument in the ongoing debate about the exact role attention is playing in the auditory domain at

processing stages of object formation and object selection. Object-based auditory attention is a difficult concept to start with; it is tricky to operationalize (Shinn-Cunningham, 2008; Shinn-Cunningham et al., 2017), and to test empirically. Not only is it hard to define what an auditory object really is, but it also turned out challenging to decide which object a participant is currently paying attention to. The aim of the current study was to improve the characterization of the top-down object-based attention, at the level of cortical activity of reconstructed sources of MEG signal, through a novel paradigm that stresses the object level of processing of a complex naturalistic auditory scene. However, even with our analyses of evoked responses it is still difficult to sort out the previously mentioned interplay between object formation and object selection, or describe in detail the cascade of information flow that the ERF seems to suggest. To improve these results and fully exploit the advantages of the paradigm design, it could be helpful to further understand the role of coupled and synchronized brain oscillations, both in relation to the pre-repetition target and post-repetition target as well throughout the length of the entire stimulus period. Finally, for greater completeness a better description of the auditory attentional network and the low-level perceptual network could be obtained by extending our time-frequency analyses from mere power estimates to connectivity measures based instantaneous phase, potentially allowing signal transmissions between the involved brain structures based on cycle-by-cycle coupled signal phases.

4 Summary and General discussion

The main goal of this thesis is to provide insights into possible behavioral and neural signatures of object-based auditory attention involved in complex, naturalistic sound scenes comprised of a mixture of different sources. To this end we conducted one behavioral study and one MEG study reported in the Chapter 2 and chapter 3 respectively of this thesis. In the following section I will briefly summarize the results of both studies and organically discuss them in the context of the literature presented throughout the chapters.

4.1 Summary

The first study was conducted to characterize the auditory selective attention system with a theoretical and empirical focus on high-level attentional modulations on the processing level of *auditory objects*. In order to illustrate the attentional mechanisms that help achieve the amplification of the processing of auditory objects within a complex and natural multi-source situation, we introduced a new auditory repetition detection task. In this new auditory repetition detection task, participants were asked to detect brief repetitions of auditory objects within the acoustic stream of a complicated, mixed soundscape. The logic behind this new task is that such a repetition detection task requires the participants to fully process the acoustic stream - as a superordinate entity of individual objects – all the way up to a cognitive level that allows them to recognize a certain, temporally extended set of low-level features as an object and to understand that this set of features was repeated. Importantly, this attention task cannot be solved by attending to distinct low-level features itself, nor by spatial attention.

In line with studies of object-based attention in the visual domain, we found behavioral effects of attentional facilitation through measures of accuracy, reaction

times, sensitivity and response bias / criterion, when the repetition segment was embedded within valid cueing trials, while at the same time we found an effect of attentional inhibition through measures of accuracy, reaction times, sensitivity and criterion, when the repetition segment was presented in invalid cueing trials. Moreover, the careful design of the stimuli, in addition to a control experiment with stimuli of the same category (two speech streams) confirmed that our paradigm - used in the second study also in MEG – is a valid naturalistic soundscape differentiating from experiments with only speech streams in stationary synthetic noise and from experiments using simply two competing speech signals.

The second study presented in this thesis was based on the same behavioral paradigm and validated set of stimuli. This time, we focused on the neural activity during the repetition detection period by investigating the temporal dynamic of the cortical activity at the source level (cortical activations). These analyses were accomplished both in time (in form of evoked responses, ERFs) and frequency domain (in form of time-frequency spectrograms). All analyses have been conducted also on source-space, providing not only insights into the time courses of events but also their exact spatial localization in cortex. Combining the paradigm with such ERF- and time-frequency analyses of the source reconstructed MEG recordings allowed us to identify the neural correlate of the attentional effects of facilitation and inhibition of top-down auditory object selection. Moreover, via the identification of seven ROIs - spanning from low-level processing areas of the early auditory cortex to higher-level processing areas of inferior frontal cortex - we were able to characterize the temporal dynamics of a possible hierarchical information flow. Further we conducted analyses illustrating the exact cortical localization where the neural activity differentiated the most between attentional conditions, that is, when the target attended versus not, and when the target was detected versus not.

Overall both studies show an attentional modulation effect operating indeed on the object-level in a naturalistic auditory scene that closely resembles the cocktail-party problem in a more ecologically valid setting. Findings of the first study provided evidence of the facilitation and inhibition effects, respectively in valid and invalid cueing

trials, that lead to a more accurate and faster identification of the repetition target in the valid trials. Findings of the second study complemented the results of the first one by elucidating the exact spatio-temporal cortical dynamics of the identification of the repetition segment in one of the two naturalistic streams.

Taken together the results support the biased competition theory (Desimone & Duncan, 1995), for which selective attention is the central mechanism that biases processing of perceptual stimuli by facilitating the processing of important information and - at the same time - filtering out irrelevant information, in auditory domain at the object-based level.

4.2 Object-based attention in complex naturalistic auditory scenes

Where and how in the neural pathway an object-based representation of an attended sound emerges, is yet to be understood. A growing body of studies are shedding light on numerous aspects of the auditory attention system, and the auditory object-based attention system is certainly one of the topics that is receiving great consideration, since Shinn-Cunningham and colleagues brought back the matter at the center of the discussion (e.g., Shinn-Cunningham, 2008, p.). A relative agreement on the rules that govern the definition of what is an auditory object has been reached by seminal and recent works (Bregman, 1990; Carlyon, 2004; Griffiths & Warren, 2004): low-level features such as the exact timing of co-occurrence and frequency contiguity work at a “local” temporal scale of what has been called “syllable level”; “streaming” means instead the organization of the sounds by grouping auditory objects by higher order perceptual features (pitch, location, and also previous acquired experience) across discontinuities of the ongoing sounds. However the relationship between object formation and auditory selective attention remains a subject of ongoing debate

(Shinn-Cunningham et al., 2015). Some studies suggest that an auditory object or a stream – as a superordinate ensemble of objects, extending throughout time - can be formed just when it is attended (Alain & Woods, 1997), others suggest that the object formation is pre-attentive and that attention is deployed only subsequently (Bregman, 1990). Recent work has instead pointed out the role of top-down auditory attention systems involved both in the object selection but also in object formation (Best et al., 2008; Hill & Miller, 2010; Maddox & Shinn-Cunningham, 2012). In the last decade a growing body of work concentrated their efforts to study attention in a more natural, complex listening scenario resembling closely many everyday situations, with paradigms that required attending one stream among two competing acoustic signals (see e.g., Ding & Simon, 2012).

In the present thesis we designed a novel paradigm to tackle the auditory selective attention at the level of object processing: a repetition target has been embedded in one of the two naturalistic streams in such a way that, after being cued, the participant necessarily had to build up representations of the auditory objects composing the scene across time in order to accomplish the task by responding as fast as possible when they detected that some of the objects were repeated. In study one, we reported a behavioral attentional facilitation effect in the validly cued trials, and an inhibition effect in the invalidly cued trials. In the second study, we reported a significant stronger MEG response, mapped at the source cortical level, when the repetition segment was in the attended stream. The same has been found when the target repetition was detected, suggesting that the time course of the neural activity and its spatial distribution represent one aspect of the neural correlates of the object-based attention. Object-based attention can therefore be concluded to also operate at the object level, to recognize that a segment of one of the streams, formed by many objects, had been repeated.

Our provide new strong evidence for the “biased competition model” hypothesized and studied in visual domain, suggesting that a similar mechanism is also involved in the auditory naturalistic complex scenes.

4.3 Future directions

Our ERF analyses and the time-frequency analyses at the source-space level have revealed neural activity patterns that suggest further investigations with more detailed analysis. It is therefore worth investing in the analyses of the phase-spectrum of the oscillatory components of the auditory representation of the entire scene, both at the pre-repetition and post-repetition state of the system. Beyond our first approaches into frequency-domain analyses of this matter, in form of time-frequency representations of oscillatory power, this could be extended by computing estimates of the instantaneous phase state of various neural oscillators. Such a decomposition of phase states could allow investigating further cycle-by-cycle phase consistencies, in the form of coherence and / or phase-locking, both between pairs of neural sites, and between each neuronal population and the external, physical stimulus. This could provide a meaningful measure of functional connectivity patterns and would be complementing the current results presented in this thesis in a more robust way. Observed patterns of phase-based functional connectivity (e.g., coherence analyses based on the cross-spectra of all ROIs in source space) could even better characterize the information flow between low-level auditory areas and higher-level regions belonging to attentional networks. The combination of these additional analysis with the ERF results and the novel designed paradigm then has the potential to add a more precise description of the role of the auditory attention at the interplay between object formation and object selection.

5 References

Ahveninen, J., Hämäläinen, M., Jääskeläinen, I. P., Ahlfors, S. P., Huang, S., Lin, F.-H., … Belliveau, J. W. (2011). Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proceedings of the National Academy of Sciences*, *108*(10), 4182–4187.

Ahveninen, J., Kopčo, N. & Jääskeläinen, I. P. (2014). Psychophysics and neuronal bases of sound localization in humans. *Hearing Research*, *307*, 86–97.

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A. & Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, *124*, 906–917.

Alain, C. & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience: A Journal and Virtual Library*, *5*, D202–D212.

Alain, C. & Winkler, I. (2012). Recording Event-Related Brain Potentials: Application to Study Auditory Perception. In D. Poeppel, T. Overath, A. N. Popper, & R. R. Fay (Eds.), *The Human Auditory Cortex* (pp. 69–96). New York, NY: Springer New York.

Alain, C. & Woods, D. L. (1997). Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology*, *34*(5), 534–546.

Alho, K., Rinne, T., Herron, T. J. & Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: A meta-analysis of fMRI studies. *Hearing Research*, *307*, 29–41.

Alho, K., Salonen, J., Rinne, T., Medvedev, S. V., Hugdahl, K. & Hämäläinen, H. (2012). Attention-related modulation of auditory-cortex responses to speech sounds during dichotic listening. *Brain Research*, *1442*, 47–54.

Andersen, S. K., Fuchs, S. & Müller, M. M. (2011). Effects of feature-selective and spatial attention at different stages of visual processing. *Journal of Cognitive Neuroscience*, *23*(1), 238–246.

Assmus, A., Marshall, J. C., Noth, J., Zilles, K. & Fink, G. R. (2005). Difficulty of perceptual spatiotemporal integration modulates the neural activity of left inferior parietal cortex. *Neuroscience*, *132*(4), 923–927.

Bagherzadeh, Y., Baldauf, D., Lu, B., Pantazis, D. & Desimone, R. (2017). Alpha and gamma neurofeedback reinforce control of spatial attention. *Journal of Vision*, *17*(10), 385–385.

Bagherzadeh, Y., Baldauf, D., Pantazis, D. & Desimone, R. (2020). Alpha Synchrony and the Neurofeedback Control of Spatial Attention. *Neuron*, *105*(3), 577-587.e5.

Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature*

Neuroscience, *20*(3), 327–339.

Baldauf, D. (2015). Top-down biasing signals of non-spatial, object-based attention. *Journal of Vision*, *15*(12), 1395–1395.

Baldauf, D. & Desimone, R. (2014). Neural Mechanisms of Object-Based Attention. *Science*, *344*(6182), 424–427.

Baldauf, D. & Desimone, R. (2016). Mechanisms of spatial versus non-spatial, modality-based attention. *Annual Meeting of the Society for Neuroscience*.

Baldauf, D. & Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Research*, *50*(11), 999–1013.

Baldauf, D., Grossman, N., Hu, A.-M., Boyden, E. & Desimone, R. (2016). Transcranial alternating current stimulation (tACS) reveals causal role of brain oscillations in visual attention. *Journal of Vision*, *16*(12), 937–937.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P. & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312.

Bertrand, O., Tallon-Baudry, C. & Pernier, J. (2000). Time-frequency analysis of oscillatory gamma-band activity: Wavelet approach and phase-locking estimation. In *Biomag 96* (pp. 919–922). Springer.

Best, V., Gallun, F. J., Carlile, S. & Shinn-Cunningham, B. G. (2007). Binaural interference and auditory grouping. *The Journal of the Acoustical Society of America*, *121*(2), 1070–1076.

Best, V., Ozmeral, E. J., Kopco, N. & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*.

Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P.-E., Giard, M.-H. & Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *Journal of Neuroscience*, *27*(35), 9252–9261.

Biesmans, W., Das, N., Francart, T. & Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(5), 402–412.

Biesmans, W., Vanthornhout, J., Wouters, J., Moonen, M., Francart, T. & Bertrand, A. (2015). Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5155–5158.

Bizley, J. K. & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707.

Botvinick, M. M., Cohen, J. D. & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

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

Bregman, A. S. (1993). *Auditory scene analysis: Hearing in complex environments.*

Bressler, S., Masud, S., Bharadwaj, H. & Shinn-Cunningham, B. G. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, *78*(3), 349–360.

Broadbent, D. E. (1958). *Perception and communication*. Elsevier.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica*, *86*(1), 117–128.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception & Psychophysics*, *77*(5), 1465–1487.

Brungart, D. S., Chang, P. S., Simpson, B. D. & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, *120*(6), 4007. (world).

Buschman, T. J. & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science (New York, N.Y.)*, *315*(5820), 1860–1862.

Bush, G., Luu, P. & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, *4*(6), 215–222.

Buzsaki, G. (2006). *Rhythms of the brain*. Oxford University Press.

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., … Knight, R. T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, *313*(5793), 1626–1628.

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

Carlyon, R. P., Plack, C. J., Fantini, D. A. & Cusack, R. (2003). Cross-Modal and Non-Sensory Influences on Auditory Streaming. *Perception*, *32*(11), 1393–1402.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*. (world).

Ciaramitaro, V. M., Mitchell, J. F., Stoner, G. R., Reynolds, J. H. & Boynton, G. M. (2011). Object-based attention to one of two superimposed surfaces alters responses in human early visual cortex. *Journal of Neurophysiology*, *105*(3), 1258–1265.

Cohen, D. (1972). Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. *Science*, *175*(4022), 664–666.

Cohen, E. H. & Tong, F. (2015). Neural mechanisms of object-based attention. *Cerebral Cortex*, *25*(4), 1080–1092.

Cohen, M. X. (2011). It's about time. *Frontiers in Human Neuroscience*, *5*, 2.

Colflesh, G. J. H. & Conway, A. R. A. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic Bulletin & Review*, *14*(4), 699–703.

Conway, A. R. A., Cowan, N. & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*(2), 331–335.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., … Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. *Neuron*, *21*(4), 761–773.

Corbetta, M., Kincade, M. J., Lewis, C., Snyder, A. Z. & Sapir, A. (2005). Neural basis and recovery of spatial attention deficits in spatial neglect. *Nature Neuroscience*, *8*(11), 1603–1610.

Corbetta, M., Patel, G. & Shulman, G. L. (2008). The reorienting system of the human brain:

From environment to theory of mind. *Neuron*, *58*(3), 306–324.

Costa, S. D., Zwaag, W. van der, Miller, L. M., Clarke, S. & Saenz, M. (2013). Tuning in to sound: Frequency-selective attentional filter in human primary auditory cortex. *Journal of Neuroscience*, *33*(5), 1858–1863.

Critchley, H. D., Mathias, C. J., Josephs, O., O'Doherty, J., Zanini, S., Dewar, B.-K., … Dolan, R. J. (2003). Human cingulate cortex and autonomic control: Converging neuroimaging and clinical evidence. *Brain*, *126*(10), 2139–2152.

Cusack, R., Decks, J., Aikman, G. & Carlyon, R. P. (2004a). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 643.

Cusack, R., Decks, J., Aikman, G. & Carlyon, R. P. (2004b). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 643–656.

Dale, A. M., Fischl, B. & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179–194.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*(9), 327–333.

de Vries, E. & Baldauf, D. (2019). Attentional weighting in the face processing network: A magnetic response image-guided magnetoencephalography study using multiple cyclic entrainments. *Journal of Cognitive Neuroscience*, *31*(10), 1573–1588.

Degerman, A., Rinne, T., Särkkä, A.-K., Salmi, J. & Alho, K. (2008). Selective attention to sound location or pitch studied with event-related brain potentials and magnetic fields. *European Journal of Neuroscience*, *27*(12), 3329–3341.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual*

Review of Neuroscience, *18*, 193–222.

Deutsch, J. A. & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80.

Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.

Ding, N., Chatterjee, M. & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46.

Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164.

Ding, N. & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 12.

Ding, N. & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859.

Ding, N. & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735.

Duncan, J. (1984). Selective Attention and the Organization of Visual Information. *Journal of Experimental Psychology*, *113*(4), 501–517.

Egly, R., Driver, J. & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*(2), 161.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J. & Shamma, S. A. (2009a). Temporal coherence in the perceptual organization and cortical representation of auditory

scenes. *Neuron*, *61*(2), 317–329.

Elhilali, M. & Shamma, S. A. (2009). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, *124*(6), 3751. (world).

Elhilali, M., Xiang, J., Shamma, S. A. & Simon, J. Z. (2009b). Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. *PLoS Biology*, *7*(6), e1000129.

Engel, L. R., Frum, C., Puce, A., Walker, N. A. & Lewis, J. W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. *NeuroImage*, *47*(4), 1778–1791.

Eriksen, B. A. & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Falkenberg, L. E., Specht, K. & Westerhausen, R. (2011). Attention and cognitive control networks assessed in a dichotic listening fMRI study. *Brain and Cognition*, *76*(2), 276–285.

Fan, J., Flombaum, J. I., McCandliss, B. D., Thomas, K. M. & Posner, M. I. (2003). Cognitive and Brain Consequences of Conflict. *NeuroImage*, *18*(1), 42–57.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A. & Posner, M. I. (2002). Testing the Efficiency and Independence of Attentional Networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781.

Fischl, B., Sereno, M. I. & Dale, A. M. (1999). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, *9*(2), 195–207.

Fries, P. (2015). Rhythms for cognition: Communication through coherence. *Neuron*, *88*(1), 220–235.

Ghinst, M. V., Bourguignon, M., Beeck, M. O. de, Wens, V., Marty, B., Hassid, S., … Tiège, X. D. (2016). Left Superior Temporal Gyrus Is Coupled to Attended Speech in a Cocktail-Party Auditory Scene. *Journal of Neuroscience*, *36*(5), 1596–1606.

Ghitza, O. (2012). On the Role of Theta-Driven Syllabic Parsing in Decoding Speech: Intelligibility of Speech with a Manipulated Modulation Spectrum. *Frontiers in Psychology*, *3*.

Ghitza, O., Giraud, A.-L. & Poeppel, D. (2013). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, *6*.

Giraud, A.-L. & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … Essen, D. C. V. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*.

Gregoriou, G. G., Gotts, S. J., Zhou, H. & Desimone, R. (2009). Long-range neural coupling through synchronization with attention. In N. Srinivasan (Ed.), *Progress in Brain Research* (pp. 35–45). Elsevier.

Griffiths, T. D. & Warren, J. D. (2004). What is an auditory object? *Nature Reviews*

Neuroscience, *5*(11), 887–892.

Gutschalk, A. & Dykstra, A. R. (2014). Functional imaging of auditory scene analysis. *Hearing Research*, *307*, 98–110.

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M. & Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *Journal of Neuroscience*, *25*(22), 5382–5388.

Gutschalk, A., Micheyl, C. & Oxenham, A. J. (2008). Neural Correlates of Auditory Perceptual Awareness under Informational Masking. *PLOS Biol*, *6*(6), e138.

Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C. & Melcher, J. R. (2007). Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation. *Journal of Neuroscience*, *27*(48), 13074–13081.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J. & Lounasmaa, O. V. (1993). Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, *65*(2), 413–497.

Hämäläinen, M. S. & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42.

Hawley, M. L., Litovsky, R. Y. & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, *115*(2), 833–843.

Hill, K. T. & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, *20*(3), 583–590.

Hopf, J.-M., Boelmans, K., Schoenfeld, M. A., Luck, S. J. & Heinze, H.-J. (2004). Attention to features precedes attention to locations in visual search: Evidence from electromagnetic brain responses in humans. *Journal of Neuroscience*, *24*(8), 1822–1832.

Hugdahl, K., Westerhausen, R., Alho, K., Medvedev, S., Laine, M. & HäMäLäInen, H. (2009). Attention and cognitive control: Unfolding the dichotic listening story. *Scandinavian Journal of Psychology*, *50*(1), 11–22.

Ihlefeld, A. & Shinn-Cunningham, B. G. (2008). Spatial release from energetic and informational masking in a divided speech identification task. *The Journal of the Acoustical Society of America*, *123*(6), 4380–4392.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J. & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412.

Kanwisher, N., McDermott, J. & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Kastner, S., O'Connor, D. H., Fukui, M. M., Fehd, H. M., Herwig, U. & Pinsk, M. A. (2004). Functional Imaging of the Human Lateral Geniculate Nucleus and Pulvinar. *Journal of Neurophysiology*, *91*(1), 438–448.

Kastner, S. & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia*, *39*(12), 1263–1276.

Kaya, E. M. & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*.

Kaya, E. M. & Elhilali, M. (2017). Modelling auditory attention. *Phil. Trans. R. Soc. B*,

372(1714), 20160101.

Keefe, J. M. & Störmer, V. S. (2020). Alpha-band oscillations and slow potentials shifts over visual cortex track the time course of both endogenous and exogenous orienting of attention. *BioRxiv*, 2019.12.12.874818.

Kerlin, J. R., Shahin, A. J. & Miller, L. M. (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a “Cocktail Party.” *Journal of Neuroscience*, *30*(2), 620–628.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A. & Carter, C. S. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. *Science*, *303*(5660), 1023–1026.

Kidd, G., Arbogast, T. L., Mason, C. R. & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*(6), 3804–3815.

Kim, Y.-J., Tsai, J. J., Ojemann, J. & Verghese, P. (2017). Attention to multiple objects facilitates their integration in prefrontal and parietal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(19), 4942–4953.

Kimura, D. (1964). Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology*, *16*(4), 355–358.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358.

Kitterick, P. T., Clarke, E., O’Shea, C., Seymour, J. & Quentin Summerfield, A. (2013). Target identification using relative level in multi-talker listening. *The Journal of the Acoustical Society of America*, *133*(5), 2899–2909.

Kleiner, M., Brainard, D. & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

Krumbholz, K., Eickhoff, S. B. & Fink, G. R. (2007). Feature-and object-based attentional modulation in the human auditory "where" pathway. *Journal of Cognitive Neuroscience*, *19*(10), 1721–1733.

Lakatos, P., Barczak, A., Neymotin, S. A., McGinnis, T., Ross, D., Javitt, D. C. & O'Connell, M. N. (2016). Global dynamics of selective attention and its lapses in primary auditory cortex. *Nature Neuroscience*.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I. & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110–113.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C. & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, *77*(4), 750–761.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G. & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, *94*(3), 1904–1911.

Lalor, E. C. & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193.

Lee, A. K. C., Larson, E., Maddox, R. K. & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, *307*, 111–120.

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H., Larson, E., Hämäläinen, M. &

Shinn-Cunningham, B. G. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience*, *6*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnins.2012.00190/full

Lee, T.-W., Girolami, M. & Sejnowski, T. J. (1999). Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources. *Neural Computation*, *11*(2), 417–441.

Liu, T. (2016). Neural representation of object-specific attentional priority. *NeuroImage*, *129*, 15–24.

Liu, T., Stevens, S. T. & Carrasco, M. (2007a). Comparing the time course and efficacy of spatial and feature-based attention. *Vision Research*, *47*(1), 108–113.

Liu, T., Stevens, S. T. & Carrasco, M. (2007b). Comparing the time course and efficacy of spatial and feature-based attention. *Vision Research*, *47*(1), 108–113.

Lopes da Silva, F. (2013). Eeg and meg: Relevance to neuroscience. *Neuron*, *80*(5), 1112–1128.

Luo, H. & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010.

Luo, H. & Poeppel, D. (2012). Cortical Oscillations in Auditory Perception and Speech: Evidence for Two Temporal Windows in Human Auditory Cortex. *Frontiers in Psychology*, *3*.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P. & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*,

29(1), 43–51.

Macmillan, N. A. & Creelman, C. D. (1991). *Detection Theory: A User's Guide*. CUP Archive.

Maddox, R. K., Atilgan, H., Bizley, J. K. & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *ELife*, *4*.

Maddox, R. K. & Shinn-Cunningham, B. G. (2012). Influence of Task-Relevant and Task-Irrelevant Feature Continuity on Selective Auditory Attention. *Journal of the Association for Research in Otolaryngology*, *13*(1), 119–129.

Mangun, G. R. & Hillyard, S. A. (1991). Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(4), 1057.

Martínez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., … Hillyard, S. A. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neuroscience*, *2*(4), 364–369.

Maunsell, J. H. R. & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, *19*(22), R1024–R1027.

McDermott, J. H. & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940.

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

Mesgarani, N. & Chang, E. F. (2012). Selective cortical representation of attended speaker in

multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Moore, C. M. & Egeth, H. (1998). How does feature-based attention affect visual processing? *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1296.

Moore, T. & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(6921), 370–373.

Moore, T. & Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annual Review of Psychology*, *68*, 47–72.

Moran, J. & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60.

Morillon, B., Liegeois-Chauvel, C., Arnal, L. H., Bénar, C. G. & Giraud, A.-L. (2012). Asymmetric function of theta and gamma activity in syllable processing: An intra-cortical study. *Frontiers in Psychology*, *3*.

Morillon, B. & Schroeder, C. E. (2015). Neuronal oscillations as a mechanistic substrate of auditory temporal prediction: Neuronal oscillations and temporal predictions. *Annals of the New York Academy of Sciences*, *1337*(1), 26–31.

Müller, H. J. & Findlay, J. M. (1987). Sensitivity and criterion effects in the spatial cuing of visual attention. *Perception & Psychophysics*, *42*(4), 383–399.

Müller, M. M., Andersen, S., Trujillo, N. J., Valdés-Sosa, P., Malinowski, P. & Hillyard, S. A. (2006). Feature-selective attention enhances color signals in early visual areas of the human brain. *Proceedings of the National Academy of Sciences*, *103*(38),

14250–14254.

Nagarajan, S. S. & Sekihara, K. (2019). Magnetoencephalographic imaging. *Magnetoencephalography: From Signals to Dynamic Cortical Networks*, 239–258.

Nobre, A. C., Gitelman, D. R., Dias, E. C. & Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: Overlapping neural systems. *NeuroImage*, *11*(3), 210–216.

O'Craven, K. M., Downing, P. E. & Kanwisher, N. (1999). FMRI evidence for objects as the units of attentional selection. *Nature*, *401*(6753), 584–587.

Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9.

O'Sullivan, J. A., Shamma, S. A. & Lalor, E. C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *Journal of Neuroscience*, *35*(18), 7256–7263. Retrieved from Scopus.

Peelle, J. E. & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*.

Peelle, J. E., Gross, J. & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex*, *23*(6), 1378–1387.

Petkov, C. I., Kang, X., Alho, K., Bertrand, O., Yund, E. W. & Woods, D. L. (2004). Attentional modulation of human auditory cortex. *Nature Neuroscience*, *7*(6), 658–663.

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

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Posner, M. I. & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*(1), 25–42.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B. & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503.

Raz, A. & Buhle, J. (2006). Typologies of attentional networks. *Nature Reviews Neuroscience*, *7*(5), 367–379.

Rizzolatti, G., Riggio, L., Dascola, I. & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1), 31–40.

Roberts, T. P., Ferrari, P., Perry, D., Rowley, H. A. & Berger, M. S. (2000). Presurgical mapping with magnetic source imaging: Comparisons with intraoperative findings. *Brain Tumor Pathology*, *17*(2), 57–64.

Rossi, A. F. & Paradiso, M. A. (1995). Feature-specific effects of selective visual attention. *Vision Research*, *35*(5), 621.

Ruusuvirta, T., Huotilainen, M. & Näätänen, R. (2007). Preperceptual human number sense for sequential sounds, as revealed by mismatch negativity brain response? *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(12), 2777–2779.

Saenz, M., Buracas, G. T. & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*(7), 631–632.

Sàenz, M., Buraĉas, G. T. & Boynton, G. M. (2003). Global feature-based attention for motion and color. *Vision Research*, *43*(6), 629–637.

Schoenfeld, M. A., Hopf, J.-M., Merkel, C., Heinze, H.-J. & Hillyard, S. A. (2014). Object-based attention involves the sequential activation of feature-specific cortical modules. *Nature Neuroscience*, *17*(4), 619–624.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1), 1–46.

Schwartz, A., McDermott, J. H. & Shinn-Cunningham, B. G. (2012). Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America*, *132*(1), 357–368.

Schwedhelm, P., Baldauf, D. & Treue, S. (2017). Electrical stimulation of macaque lateral prefrontal cortex modulates oculomotor behavior indicative of a disruption of top-down attention. *Scientific Reports*, *7*(1), 1–10.

Schwedhelm, P., Baldauf, D. & Treue, S. (2020). The lateral prefrontal cortex of primates encodes stimulus colors and their behavioral relevance during a match-to-sample task. *Scientific Reports*, *10*(1), 1–12.

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. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186.

Shinn-Cunningham, B. G., Best, V., Kingstone, A., Fawcett, J. M. & Risko, E. F. (2015). Auditory selective attention. *The Handbook of Attention*, *99*.

Shinn-Cunningham, B. G., Best, V. & Lee, A. K. (2017). Auditory object formation and selection. In *The auditory system at the cocktail party* (pp. 7–40). Springer.

Shulman, G. L., Remington, R. W. & Mclean, J. P. (1979). Moving attention through visual space. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 522.

Siegel, M., Donner, T. H., Oostenveld, R., Fries, P. & Engel, A. K. (2008). Neuronal synchronization along the dorsal visual pathway reflects the focus of spatial attention. *Neuron*, *60*(4), 709–719.

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176.

Simon, J. Z. (2015a). The encoding of auditory objects in auditory cortex: Insights from magnetoencephalography. *International Journal of Psychophysiology*, *95*(2), 184–190.

Simon, J. Z. (2015b). The encoding of auditory objects in auditory cortex: Insights from magnetoencephalography. *International Journal of Psychophysiology*, *95*(2), 184–190.

Sprague, T. C. & Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature Neuroscience*, *16*(12), 1879–1887.

Störmer, V. S. & Alvarez, G. A. (2014). Feature-based attention elicits surround suppression in feature space. *Current Biology*, *24*(17), 1985–1988.

Störmer, V. S., Cohen, M. A. & Alvarez, G. A. (2019a). Tuning attention to object categories: Spatially global effects of attention to faces in visual processing. *Journal of Cognitive Neuroscience*, *31*(7), 937–947.

Störmer, V. S., McDonald, J. J. & Hillyard, S. A. (2019b). Involuntary orienting of attention to

sight or sound relies on similar neural biasing mechanisms in early visual processing. *Neuropsychologia*, *132*, 107122.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Sussman, E. S., Bregman, A. S., Wang, W. J. & Khan, F. J. (2005). Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(1), 93–110.

Sussman, E. S., Horváth, J., Winkler, I. & Mark, O. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics*, *69*(1), 136–152. Retrieved from Scopus.

Tabarelli, D., Keitel, C., Gross, J. & Baldauf, D. (2020). Spatial attention enhances cortical tracking of quasi-rhythmic visual stimuli. *NeuroImage*, *208*, 116444.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. (2011). Brainstorm: A user-friendly application for meg/eeg analysis [Research Article]. Retrieved February 5, 2020, from Computational Intelligence and Neuroscience website: https://www.hindawi.com/journals/cin/2011/879716/

Tadel, F., Bock, E., Niso, G., Mosher, J. C., Cousineau, M., Pantazis, D., … Baillet, S. (2019). MEG/EEG Group Analysis With Brainstorm. *Frontiers in Neuroscience*, *13*.

Taulu, S. & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine & Biology*, *51*(7), 1759.

Theeuwes, J. (2013). Feature-based attention: It is all bottom-up priming. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), 20130055.

Theeuwes, J., Kramer, A. F., Hahn, S. & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, *9*(5), 379–385.

Trappenberg, T. P., Dorris, M. C., Munoz, D. P. & Klein, R. M. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, *13*(2), 256–271.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*(4), 242–248.

Treisman, A. M. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, *3*(6), 449–459.

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

Treue, S. & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579.

Ungerleider, S. K. and L. G. (2000). Mechanisms of Visual Attention in the Human Cortex. *Annual Review of Neuroscience*, *23*(1), 315–341.

Varela, F., Lachaux, J.-P., Rodriguez, E. & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, *2*(4), 229–239.

Vouloumanos, A. & Werker, J. F. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, *7*(3), 270–276.

Voytek, B., Samaha, J., Rolle, C. E., Greenberg, Z., Gill, N., Porat, S., … Gazzaley, A. (2017). Preparatory encoding of the fine scale of human spatial attention. *Journal of Cognitive*

Neuroscience, *29*(7), 1302–1310.

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological Reviews*, *90*(3), 1195–1268.

Wegener, D., Ehn, F., Aurich, M. K., Galashan, F. O. & Kreiter, A. K. (2008). Feature-based attention and the suppression of non-relevant object features. *Vision Research*, *48*(27), 2696–2707.

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

Womelsdorf, T., Fries, P., Mitra, P. P. & Desimone, R. (2006). Gamma-band synchronization in visual cortex predicts speed of change detection. *Nature*, *439*(7077), 733–736.

Wood, N. & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 255.

Woods, D. L. & Alain, C. (1993). Feature processing during high-rate auditory selective attention. *Perception & Psychophysics*, *53*(4), 391–402.

Woods, K. J. P. & McDermott, J. H. (2015). Attentive Tracking of Sound Sources. *Current Biology*, *25*(17), 2238–2246.

Woods, K. J. P. & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, *115*(14), E3313–E3322.

Xiang, J., Simon, J. & Elhilali, M. (2010). Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *Journal of Neuroscience*, *30*(36), 12084–12093.

Xu, M., Baldauf, D., Chang, C. Q., Desimone, R., & Tan, L. H. (2017). Distinct distributed

patterns of neural activity are associated with two languages in the bilingual brain. *Science advances*, *3*(7), e1603309.

Zatorre, R. J., Belin, P. & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.

Zhang, W. & Luck, S. J. (2009). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience*, *12*(1), 24–25.

Zhang, X., Mlynaryk, N., Japee, S. & Ungerleider, L. G. (2017). Attentional selection of multiple objects in the human visual system. *NeuroImage*, *163*, 231–243.

Zimmerman, J. E., Thiene, P. & Harding, J. T. (1970). Design and operation of stable rf-biased superconducting point-contact quantum devices, and a note on the properties of perfectly clean metal contacts. *Journal of Applied Physics*, *41*(4), 1572–1580.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013a). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron*, *77*(5), 980–991.

Zion Golumbic, E. M. Z., Cogan, G. B., Schroeder, C. E. & Poeppel, D. (2013b). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party.” *Journal of Neuroscience*, *33*(4), 1417–1426.

Zion-Golumbic, E. & Schroeder, C. E. (2012). Attention modulates ‘speech-tracking’at a cocktail party. *Trends in Cognitive Sciences*, *16*(7), 363–364.

Zotkin, D. N., Shamma, S. A., Ru, P., Duraiswami, R. & Davis, L. S. (2003). Pitch and timbre manipulations using cortical representation of sound. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP ’03). 2003 IEEE International Conference On*, *5*, V–517. IEEE.

6 Appendix A

Time-Frequency of beta and gamma waves

Time-Frequency analysis of beta and gamma frequencies

We further investigated the ongoing brain oscillatory activity as a signature of sustained attention beyond the initial 1-15Hz frequency range to look for beta and gamma neural oscillatory activity, in the same regions of the previous analysis. For this purpose, the signal of the epochs time-locked to the stimulus onset was Fourier-transformed on a single-trial level and power was estimated for a frequency range from 15-100Hz using a 300ms sliding window with multitapers methods. The resulting power spectra were then also normalized in respect to the baseline period before stimulus-onset.

Figures 15 and 16 show the average time-frequency response for left and right hemispheres respectively when both conditions, attending to the speech and attending to the environment, are combined. A generalized beta suppression is visible in every region for both hemispheres with gamma activity neural oscillation more prominent in the auditory association cortex, coherent with the temporal integration hypothesis in speech processing (Poeppel, 2003). The only visible lateralization effect consists of a stronger gamma suppression in the right hemisphere for the temporo-parietal-occipital region and insular frontal opercular cortex.

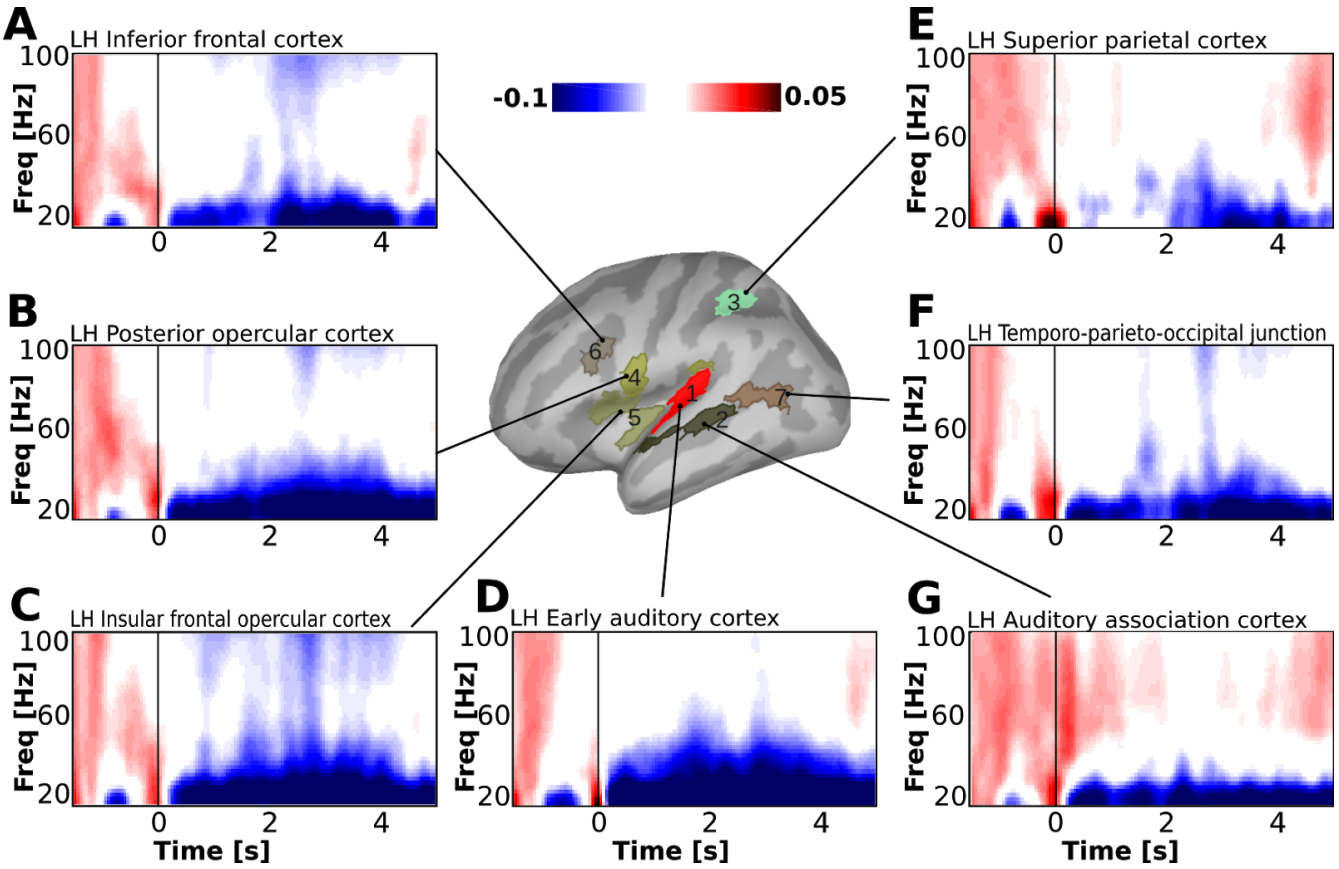

Figure 15. *Time-frequency spectrograms for both conditions combined (attend speech and attend Environment) of the stimulus period with sustained attention, for the main regions of interest in the left hemisphere (LH). In general, the spectrograms were characterized by a generalized beta inhibitory activity in every region, with gamma entrainment more prominent in auditory association cortex*.

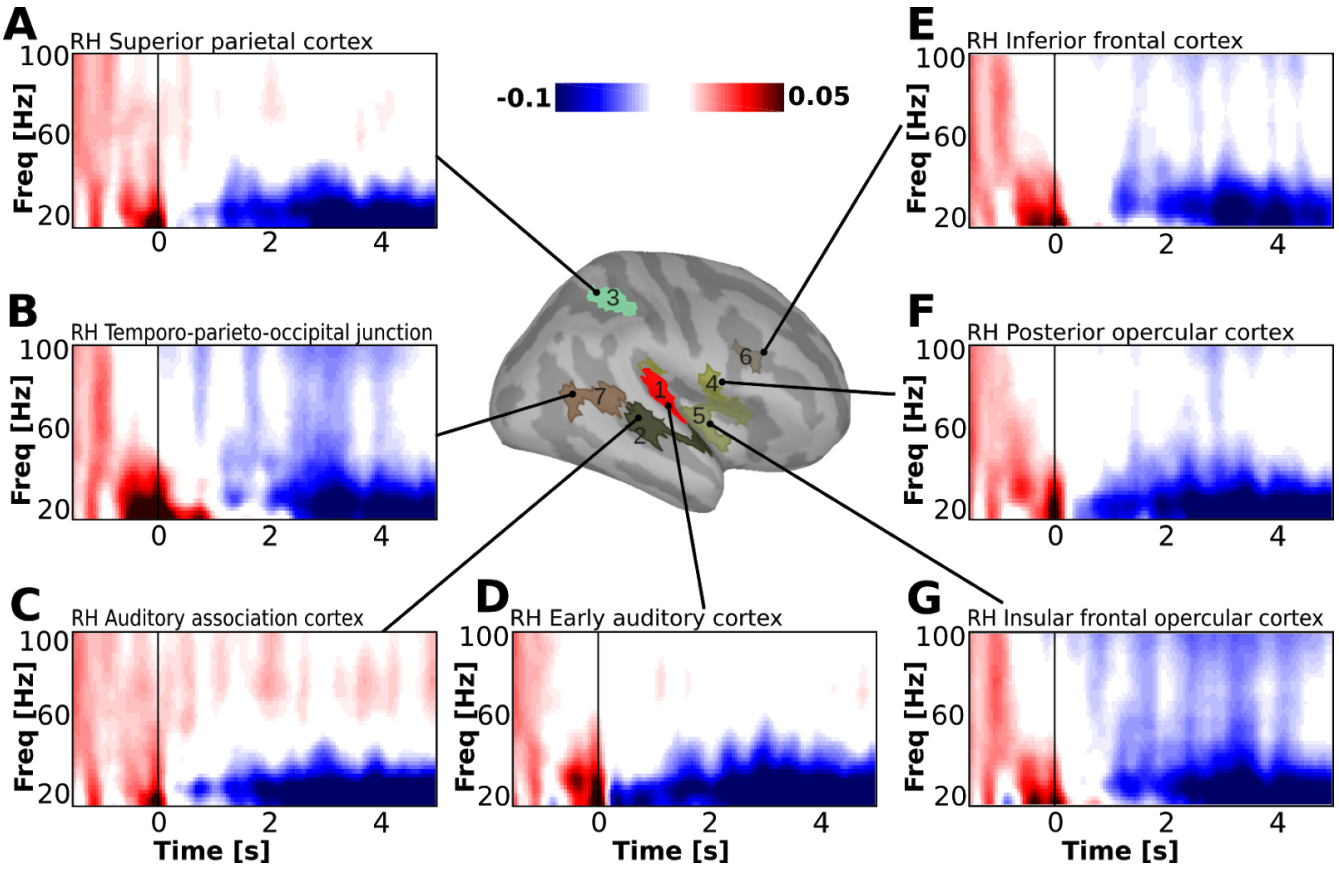

Figure 16. *Time-frequency spectrograms for both conditions combined (attend speech and attend Environment) of the stimulus period with sustained attention, for the main regions of interest in the right hemisphere (RH)*

Figures 17 and 18 display the average time-frequency spectrograms just for the condition in which attention was likely deployed to the environment signal. No particular difference between the two hemispheres can be observed within this condition that is characterized by a widespread inhibition effect of beta, but also gamma. An exception seems to be a sustained activity in the left auditory association cortex in the first second of the stimuli presentation within low-gamma band.

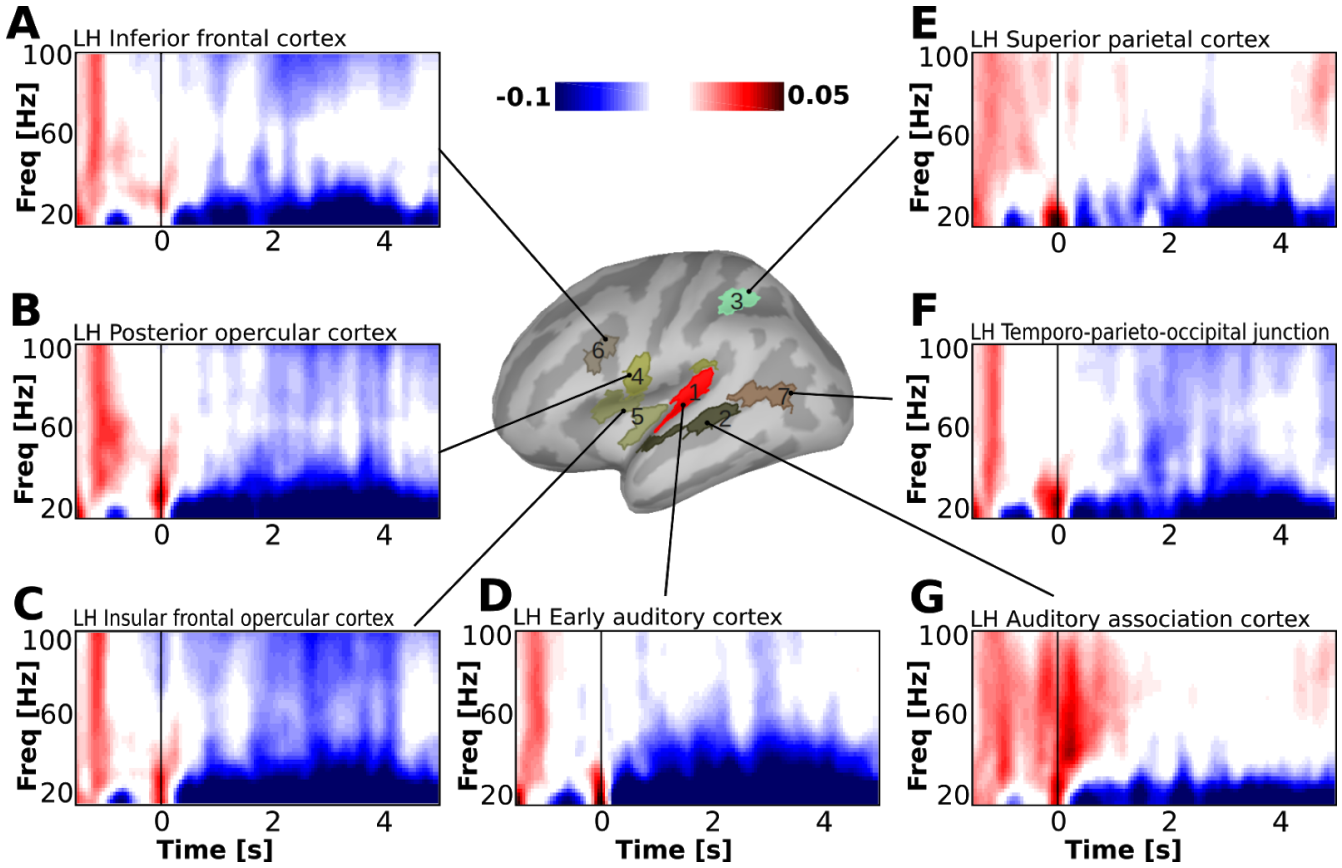

Figure 17. *Figure 18. Time-frequency spectrograms for attend environment condition of the stimulus period with sustained attention, for the main regions of interest in the left hemisphere (LH)*

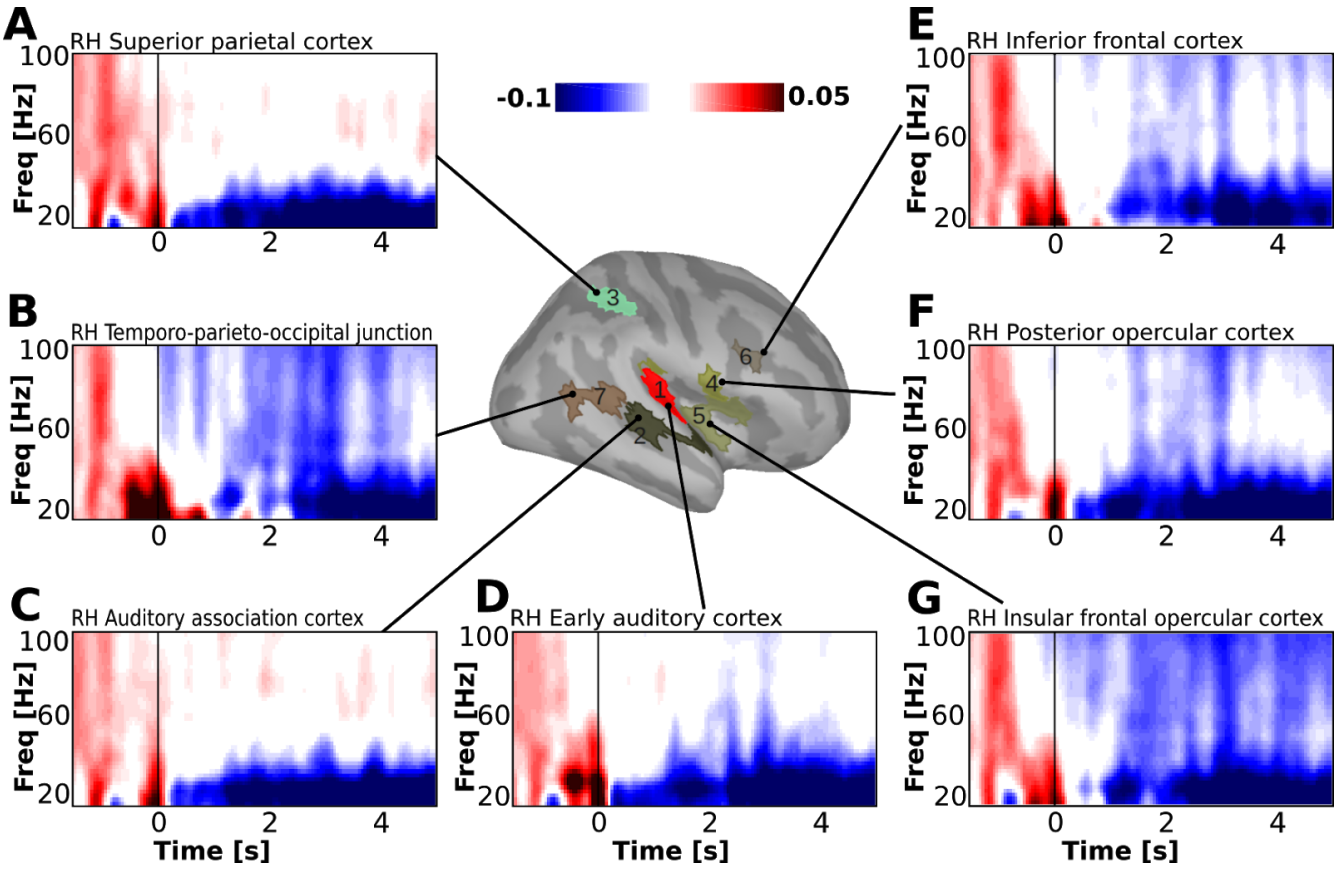

Figure 18. *Time-frequency spectrograms for attend environment condition of the stimulus period with sustained attention, for the main regions of interest in the right hemisphere (RH)*

The last two figures, 19 and 20, represent the time-frequency response for the condition in which attention was directed to the speech signal. Compared with the condition in which attention was deployed toward the environmental signal, there is a noticeable generalized reduction in the inhibitory activity in all the regions for the gamma band, and a strong reduction of the inhibitory activity of beta band as well. Strong sustained neural oscillations in gamma band is particularly evident in the auditory association cortex of both hemispheres, compared with the *attend environment* condition, with a more evident entrainment in the left hemisphere, arguably due to the lateralization effect of the speech stimulus. Beta suppression is more prominent in the left hemisphere (especially for superior parietal cortex) compared to the right hemisphere.

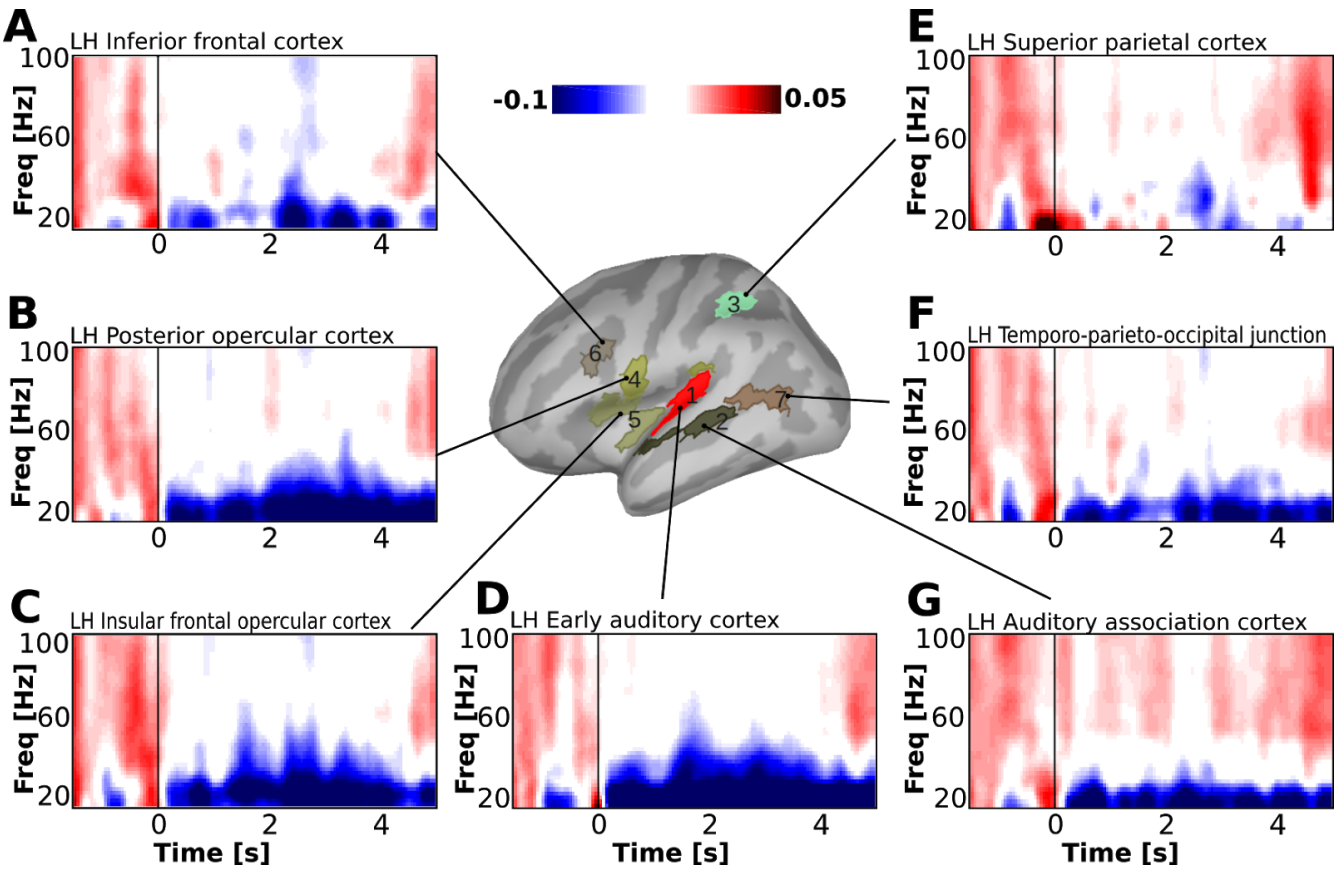

Figure 19. *Time-frequency spectrograms of beta and gamma bands for attend speech condition of the stimulus period with sustained attention, for the main regions of interest in the left hemisphere (LH).*

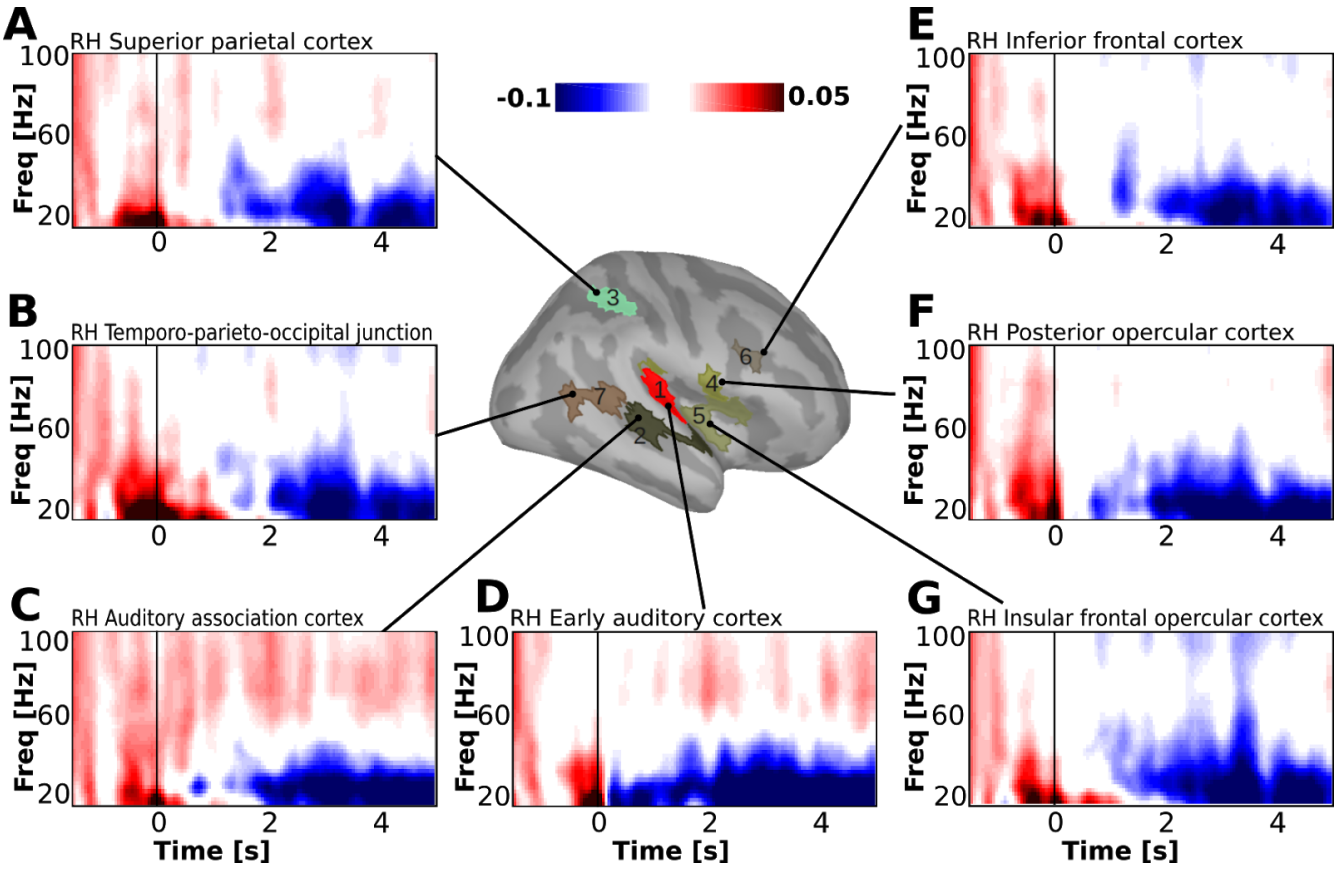

Figure 20. *Time-frequency spectrograms of beta and gamma bands for attend speech condition of the stimulus period with sustained attention, for the main regions of interest in the right hemisphere (RH)*

7 Appendix B

Visualization of neural time-courses for preliminary exploration of lateralization effects

In order to better convey the visualization of the time course of the neural activation for a possible lateralization effect in each condition, for which the ERF analysis was previously computed, we proceeded as follows. Two time windows of 400ms have been selected representing an early neural response that spans from 0 to 400ms, and a late response from 400ms to 800ms. The time course is related to the stimulus onset for the *attend speech* versus *attend environment* conditions and is related to the repetition onset for the *repetition detected* versus *repetition not detected* conditions as well for the *repetition attended* and *repetition not attended* conditions. The average of the signal within each time interval for each region of interest has been computed across all participants and plotted into bar graphs with the aim to highlight possible differences of neural activation between left and right hemisphere. The regions number from 1 to 7 corresponds to: 1) Posterior opercular cortex; 2) Early auditory cortex; 3) Auditory association Cortex; 4) Superior parietal cortex; 5) Insular frontal opercular cortex; 6) Inferior frontal cortex; 7) Temporo-parieto-occipital junction.

Between the attend speech and the attend environment conditions we can notice a very similar pattern of activation through the corresponding regions of interest within each time interval suggesting none or marginal difference between the two sets of stimuli in the first second of sound processing with a prominent activation of posterior opercular cortex, early auditory cortex *a*nd auditory association cortex. While the late time interval for both, speech and environment, shows a very balanced activation through the two hemispheres in all the regions, the early time window marks a lateralization effect toward the right hemisphere in all the regions.

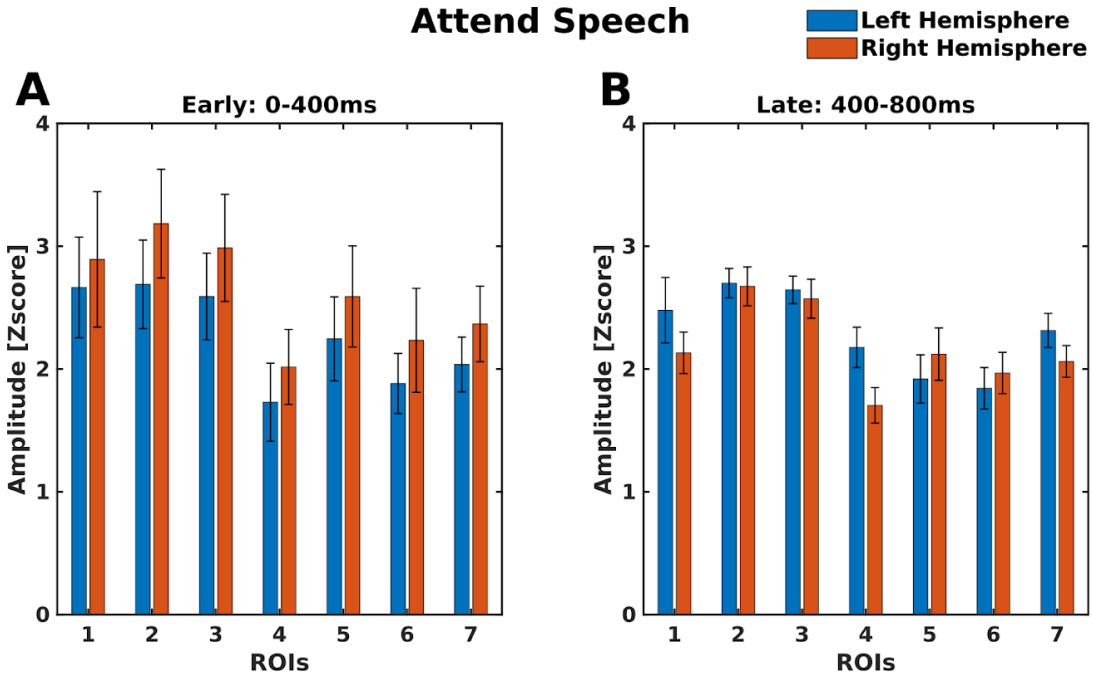

Figure 21. *Average neural activity in the attend speech condition for each region of interest in each hemisphere (blue bars represent the left hemisphere and the red bars represent the right hemisphere). Panel A represents the early time bin of 400ms and panel B represents a late interval still of 400ms. The region of interest from 1 to 7 are the follows: 1) Posterior opercular cortex 2) Early auditory cortex 3) Auditory association cortex 4) Superior parietal cortex 5) Insular frontal opercular cortex 6) Inferior frontal cortex 7) Temporo-parieto-occipital junction.*

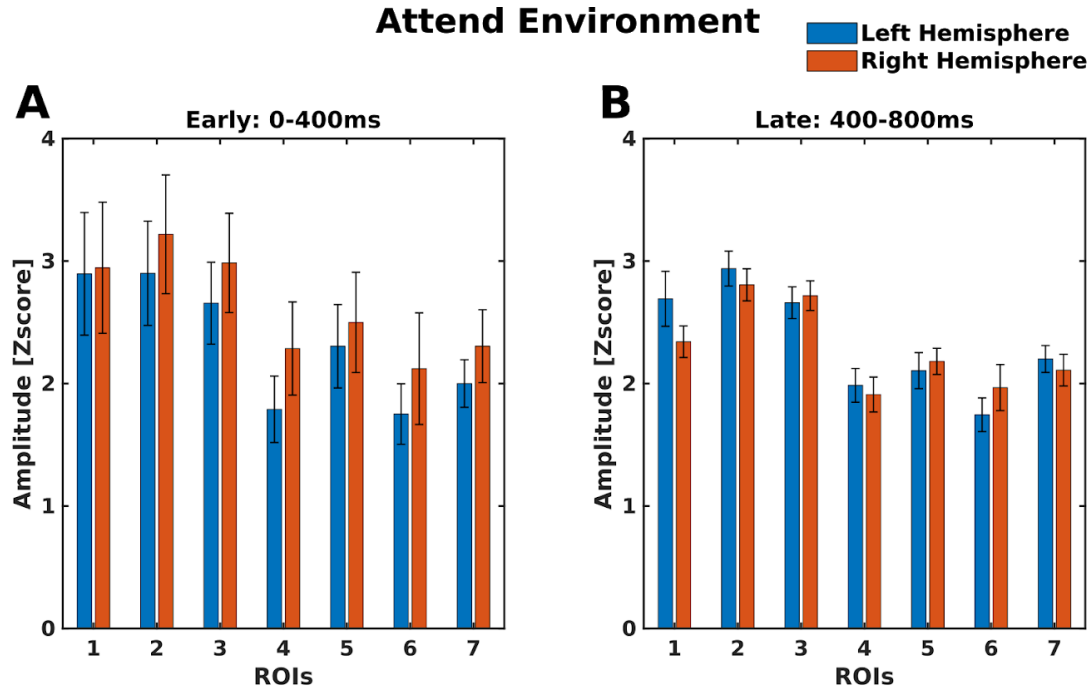

Figure 22. *Average neural activity in the attend environment condition for each region of interest in each hemisphere.*

The repetition attended (Figure 23) and repetition not attended (Figure 24) conditions highlight the contribution of the late time interval when the target was embedded in the attended stream. Figure 24 shows indeed a flat activity pattern for both time intervals and both hemispheres, instead in Figure 23 is noticeable an increase of the neural activity in the late bin in respect of the early bin with a lateralization effect toward the right hemisphere for all the regions, especially insular frontal opercular cortex, inferior frontal cortex, temporo-parieto-occipital junction, except for the superior parietal cortex that shows a slightly prominent activation in the left hemisphere.

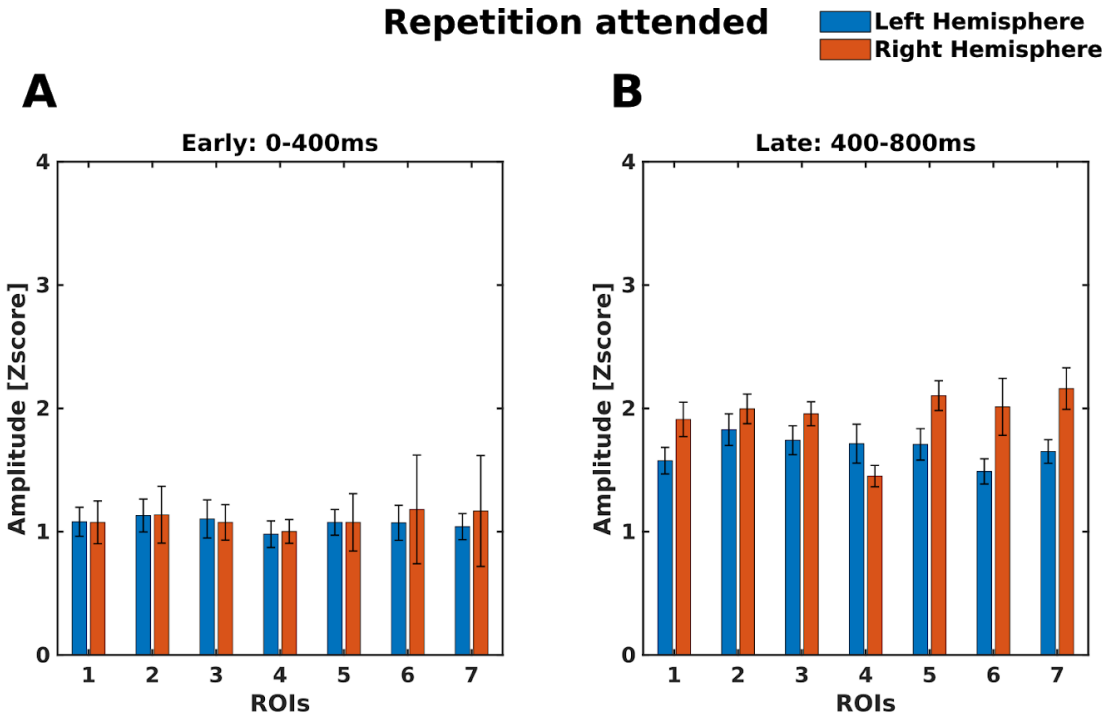

Figure 23. *Average neural activity in the repetition attended condition for each region of interest in each hemisphere (blue bars represent the left hemisphere and the red bars represent the right hemisphere. Panel A represents the early time bin of 400ms and panel B represents a late interval still of 400ms. See Figure 21 for region's labels.*

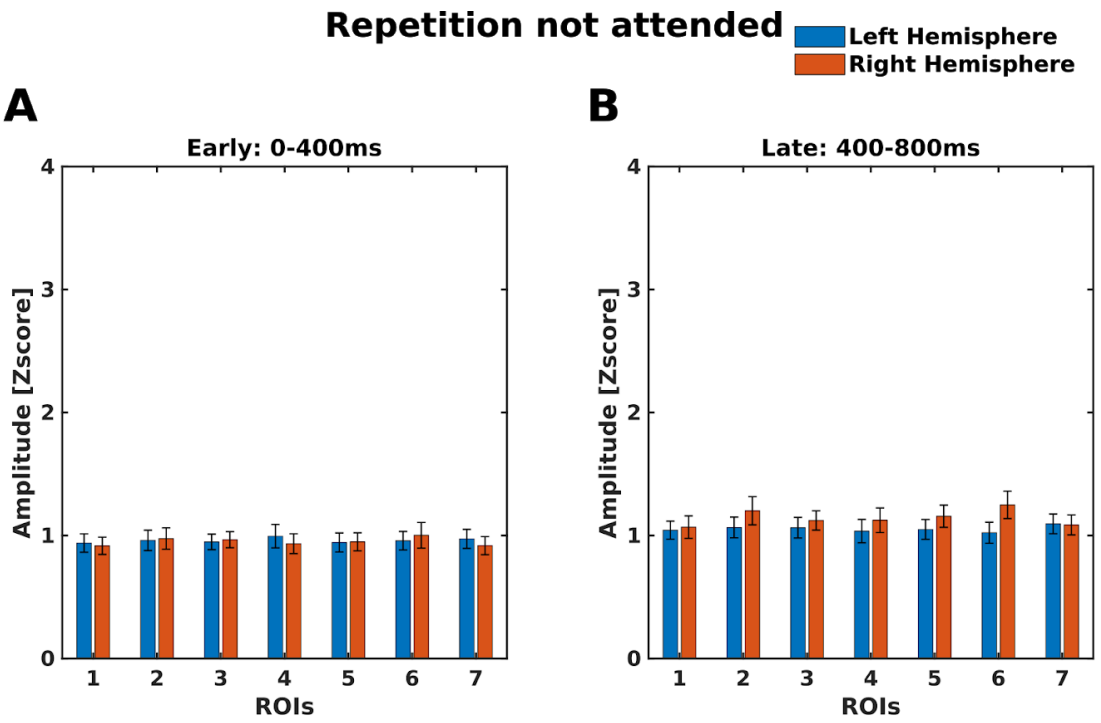

Figure 24. *Average neural activity in the repetition not attended condition for each region of interest in each hemisphere (blue bars represent the left hemisphere and the red bars represent the right hemisphere. Panel A represents the early time bin of 400ms and panel B represents a late interval still of 400ms. See Figure 21 for region's labels.*

Finally Figure 25 and 26 represents the conditions in which the repetition was detected versus not detected. The activation patterns through the different factors is similar to the conditions in which the target was in the attended stream, versus not attended stream, with figure 26 showing no variation of neural activity through regions and hemispheres in either of the two time intervals. Figure 25 highlights a slight increase of neural activity in auditory regions as well as insular frontal opercular cortex, inferior frontal cortex, temporo-parieto-occipital junction, that becomes more prominent in the late bin, lateralized to the right hemisphere. Also when the repetition is detected, the superior parietal cortex displays more activity in the left hemisphere.

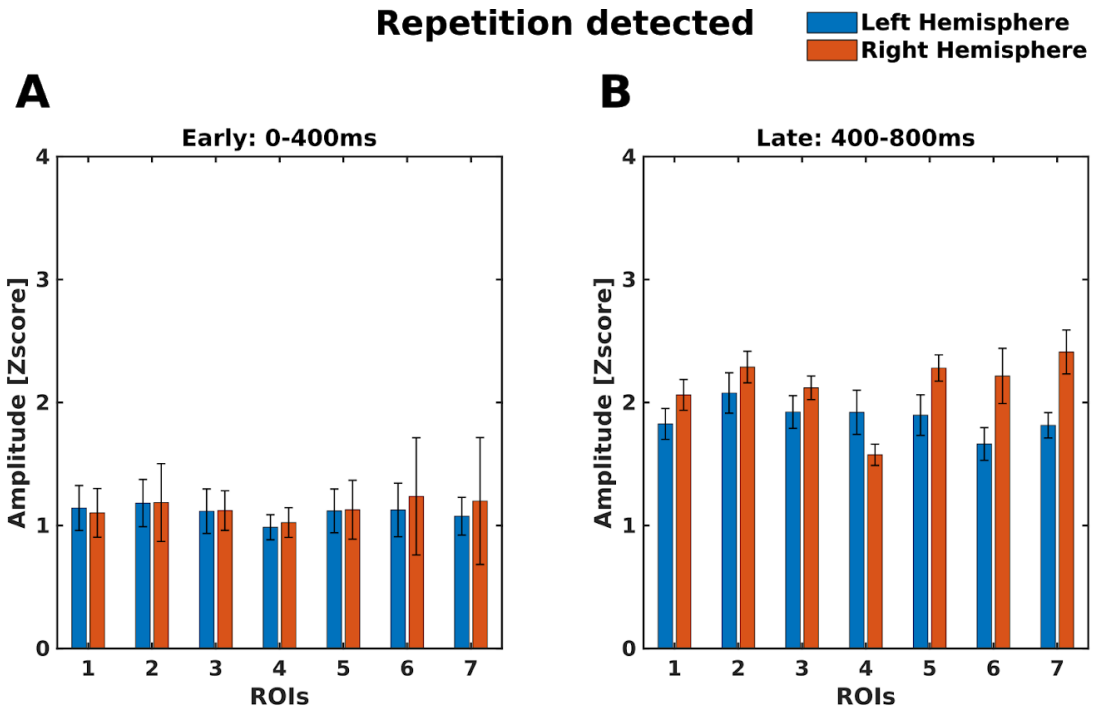

Figure 25. *Average neural activity in the repetition detection condition for each region of interest in each hemisphere (blue bars represent the left hemisphere and the red bars represent the right hemisphere. Panel A represents the early time bin of 400ms and panel B represents a late interval still of 400ms. See Figure 21 for region's labels.*

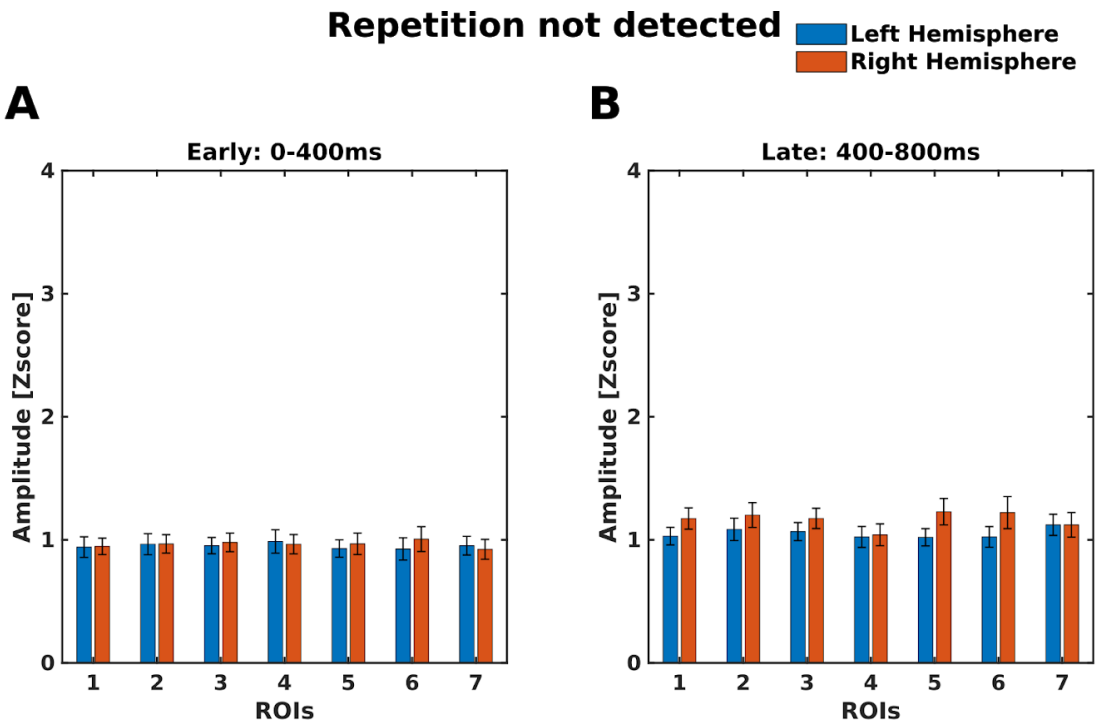

Figure 26. *Average neural activity in the repetition not detected condition for each region of interest in each hemisphere (blue bars represent the left hemisphere and the red bars represent the right hemisphere. Panel A represents the early time bin of 400ms and panel B represents a late interval still of 400ms. See Figure 21 for region's labels.*

Discussion

The new visual representation of the previously computed ERF results emphasize two important aspects that support the hypothesis that here object-based attention is effectively the unit to which attention has been deployed to solve the task. First, computing and visualizing the average neural activity of two time windows of the evoked responses, accentuate the role of the late time bin in the detection of the repetition that was observed in the classic ERF representation and statistical analysis before. This way to convey the results, therefore, suggest and are in line with the hypothesis that an high level processing of the sound scene is indeed necessary to solve the task like discussed previously in the chapter 3 and 4 of this thesis. Second, the speech and the environment condition with neural activity time-locked at the initial presentation of both stimuli, and averaged in the aforementioned time intervals, display

a very similar pattern of activity toward all the corresponding regions of interests in both hemispheres, suggesting a comparable neural response to both set of stimuli, at least in the first second in which participant are exposed to the sound scene, Importantly, both the ERF analysis and the average representation of the evoked activity in two different time bins presented here, strongly support the behavioral results obtained before that shows no significant difference between the set of stimuli. However it has to be noted that a generalized lateralization effect toward the right hemisphere is present, but needs further statistical analysis and could be potentially better conveyed through refactoring the dataset in more conditions (e.g. contrasting speech and environmental sounds within the repetition detection condition and within the repetition not detection condition). The lateralization effect that emerges here from the event related neural activity is also directed toward the opposite hemisphere (right) to the one emerged in the time-frequency analysis of theta and alpha (left). With the two types of neural signals analysis, intrinsically of different nature, being the first an evoked response and the second a time-frequency decomposition, becomes difficult to trace a possible comparison of the two emerging patterns. However in the time-frequency analysis, as a richer signal, is legit to affirm that the lateralization effects (Giraud & Poeppel, 2012; Morillon et al., 2012; Zatorre et al., 2002) can be due to the processing of speech stream specifically. In particular, left lateralization of theta band, is congruent with the hypothesis of theta representing an integrative function rather than a sampling function like gamma and high-gamma (Giraud & Poeppel, 2012; Poeppel, 2003) and with the neural entrainment to the envelope (Zion Golumbic et al., 2013a; Zion-Golumbic & Schroeder, 2012). Instead the lateralization effects of the event related neural activity is less likely to be determined by the speech specificity (Poeppel, 2003).