UCL

COGNITIVE ANALYSIS OF COMPLEX ACOUSTIC SCENES

Sundeep Teki

Wellcome Trust Centre for Neuroimaging

Institute of Neurology

A thesis submitted for the degree of Doctor of Philosophy

University College London

2013

DECLARATION

I, Sundeep Teki, hereby confirm that the work presented in this thesis is my own. Where information is derived from other sources, I confirm that this has been indicated in the thesis.

Sundeep Teki Date

ABSTRACT

Natural auditory scenes consist of a rich variety of temporally overlapping sounds that originate from multiple sources and locations and are characterized by distinct acoustic features. It is an important biological task to analyze such complex scenes and extract sounds of interest. The thesis addresses this question, also known as the "cocktail party problem" by developing an approach based on analysis of a novel stochastic signal contrary to deterministic narrowband signals used in previous work. This low-level signal, known as the *Stochastic Figure-Ground* (SFG) stimulus captures the spectrotemporal complexity of natural sound scenes and enables parametric control of stimulus features. In a series of experiments based on this stimulus, I have investigated specific behavioural and neural correlates of human auditory figure-ground segregation.

This thesis is presented in seven sections. Chapter 1 reviews key aspects of auditory processing and existing models of auditory segregation. Chapter 2 presents the principles of the techniques used including psychophysics, modeling, functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG). Experimental work is presented in the following chapters and covers figure-ground segregation behaviour (Chapter 3), modeling of the SFG stimulus based on a temporal coherence model of auditory perceptual organization (Chapter 4), analysis of brain activity related to detection of salient targets in the SFG stimulus using fMRI (Chapter 5), and MEG respectively (Chapter 6). Finally, Chapter 7

concludes with a general discussion of the results and future directions for research.

Overall, this body of work emphasizes the use of stochastic signals for auditory scene analysis and demonstrates an automatic, highly robust segregation mechanism in the auditory system that is sensitive to temporal correlations across frequency channels.

TABLE OF CONTENTS

ABBREVIATIONS

PAC	primary auditory cortex
ANOVA	analysis of variance
BOLD	blood oxygenation-level dependent
CRM	coordinate response measure
dB	decibels
ECD	equivalent current dipole
EEG	electroencephalography
EPI	echo planar image
ERP	event-related potential
fMRI	functional magnetic resonance imaging
FWE	family-wise error
FWHM	full-width-at-half-maximum
GLM	general linear model
HG	Heschl's gyrus
HRF	haemodynamic response function
Hz, kHz	Hertz, kilo Hertz
IC	inferior colliculus
ISI	inter-stimulus interval
IPS	intraparietal sulcus
LFP	local field potential
MEG	magnetoencephalography
MMN	mismatch negativity
MNI	Montreal Neurological Institute
MGB	medial geniculate body

MRI	magnetic resonance imaging
NMR	nuclear magnetic resonance
PT	planum temporale
RF	radio frequency
SEM	standard error of the mean
SFG	stochastic figure-ground (stimulus)
SPM	statistic parametric mapping
STG	superior temporal gyrus
STS	superior temporal sulcus
T	Tesla
T1	longitudinal relaxation time
T2	transverse relaxation time
TE	time-to-echo
TPJ	temporo-parietal junction
TR	time-to-repeat
VOI	volume of interest

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am most indebted to my supervisor Tim Griffiths for his expert guidance, support and encouragement throughout; for setting an example of how to think creatively with an emphasis on biological significance; for showing how to be organized, execute work with perfection, and maintain composure in every situation; and for inspiring me to become a scientist.

I am also grateful to my secondary supervisor Alex Leff for his help and support and guiding me into the higher-level world of speech and language.

Although not officially my supervisor, Maria Chait has been like one and her constant support and enthusiasm kept me going. I would like to thank her for showing how to critically analyse data and argue like a scientist, and for always being there as a friend.

I am also indebted to Sukhbinder Kumar for being the most reliable source of technical help and support and showing me how to program efficiently.

I would also like to thank Shihab Shamma, a wonderful collaborator and to-be post-doctoral advisor for introducing me to the world of modeling and neurophysiology and instilling a way of thinking based on single neurons.

Although we did not share an office, I always felt the support of the members of the Auditory Group at Newcastle University and would like to especially thank Manon Grube, Simon Baumann, Will Sedley, Tobias Overath, and Phil Gander for their companionship.

The Wellcome Trust Centre for Neuroimaging is an excellent department and I would like to especially thank Peter Aston, Marcia Bennett, Marina Anderson, David Bradbury, Janice Glensman, Sheila Burns, Letty Manyande, Isabel Stromboni, Eric Featherstone, Alphonso Reid, Ric Davis, Rachael Maddock, Chris Featherstone, Oliver Josephs, Nikolaus Weiskopf, and Gareth Barnes for making me feel at home and sorting out all my administrative, imaging, physics, MEG, computing and technical issues. I have also learnt a lot from the Principals of the department and am thankful to them for instructing how to design an imaging experiment and teaching the importance of hypothesis-driven scientific research.

I would also like to thank the fellows of the department and Queen Square at large, with particular thanks to Siawoosh Mohammadi, Rebecca Lyness, Sabine Joseph, Anna Jafarpour, Peter Smittenaar, Zoe Woodhead and Larissa Cuénoud for being wonderful friends and tolerating my company over many a meal! I am also grateful to Lara Li Hesse, Anahita Mehta, Nicolas Barascud, Lefkothea-Vasiliki Andreou and Lucile Belliveau from the Ear Institute for their support and cherished friendship.

David Blundred and Daniela Warr Schorri at the Education Unit have been most helpful in sorting out many administrative queries through the course of my PhD, for which I am most grateful.

The many volunteers who tolerated my experiments deserve a huge vote of thanks as well as the project students at Newcastle University and UCL

including Deborah Williams, Ayisha Siddiq, Madhurima Dey, Michael Savage, and Christopher Payne for collecting several behavioural datasets.

I also owe a big vote of thanks to my friends from the Goodenough College, who are too many to be named but know who they are: for their priceless companionship and support.

I am indebted to my parents and my brother for their constant support and motivation and encouraging me to pursue my academic dreams and making those few brief visits to home full of loving memories to help me go on in their absence.

Last but not least, this is for my fiancée Anu: without her steady commitment, support and encouragement this thesis would not have reached fruition.

Chapter 1. GENERAL INTRODUCTION

1.1 Sensory Coding in the Natural Environment

To obtain a coherent understanding of sensory coding and perception, it is vital to understand the structure of natural signals and how biological (neural) systems encode and process these complex stimuli. However, traditionally, neuroscientists and psychologists have used relatively simple, "controlled" stimuli in laboratory setups - sine-wave gratings, pure tones, spots, clicks, taps or periodic skin vibrations to probe the response properties of sensory neurons and characterize perceptual abilities. Although this approach represents a very successful method of understanding information processing at the early stages of sensory processing, it only offers a simplistic view of the complex sensory analysis that the brain performs in the real world. Furthermore, in the cerebral cortex, where information processing is highly nonlinear and under the influence of recurrent computation in the form of feedback signals from other cortical neurons, this approach offers limited utility. Cortical neurons encode specific spectrotemporal patterns from the input (auditory) stream, but it is challenging to discover these by simply probing one element at a time in a reduced stimulus space in the absence of appropriate context and behavioural relevance.

The auditory domain presents a rich mixture of signals that span a large bandwidth and are characterized by different spectrotemporal properties that vary significantly from one moment to another. Imagine a simple scenario of walking from home to office – we are immersed in an

environment buzzing with sounds of bird calls and rustling of leaves on trees (hopefully), the unavoidable din due to traffic and other people (unfortunately). We are frequently faced with such busy auditory environments and have to perform the complex task of making sense of it in real-time.

Considering the complexity of the rich (auditory) stimulus space, it is reasonable, therefore, to develop and use the sort of inputs that the sensory system is designed to process. This approach is applicable not only in the auditory domain but all aspects of sensory processing in general and has been adopted successfully by various interdisciplinary laboratories exploring related questions of sensory processing:

- What is the structure of natural signals in the environment and how can these be characterized using statistical principles?
- How are natural stimuli encoded by neurons?
- How are the different features combined to represent an object?
- How is sensory processing influenced by the interaction of the organism with the environment and behavioural goals?
- How robust is sensory encoding to noise and challenging environments?
- How can these principles be used to design synthetic stimuli, and build artificial devices to restore impaired sensory perception?
- How can research inform treatment and cure for clinical disorders of abnormal perception?

These questions lie at the heart of research in sensory systems neuroscience and the solutions require a multidisciplinary approach drawing

upon theory and methods from neurophysiology, psychophysics, neuroimaging, computational modeling, signal processing and complex data analysis.

This thesis addresses a fundamental question of information processing and perceptual organization in the auditory domain – what are the neural bases and mechanisms underlying our ability to group together elementary (spectrotemporal) features into discrete (auditory) objects and to segregate these objects from each other and from the background?

1.2 Natural scene analysis

The living world is a dynamic exhibition of several objects and stimuli and in order to survive, any organism must be able to perceive these signals accurately, make appropriate responses, assess the outcomes of the response and learn to make better and more flexible responses in the future. Embedded in an environment teeming with all kinds of sensory stimulation, an organism must have sophisticated sensory systems in place to make sense of the world. Al Bregman, a pioneer in the field of auditory perception, summarized the role of perception as below:

"The job of perception is to take the sensory input and to derive a useful representation of reality from it."

(Bregman, 1990)

1.2.1 Auditory scene analysis

The term, ‘auditory scene analysis’, owes its origin to Al Bregman who characterized auditory perceptual behaviour and examined our ability to separate objects and selectively attend to them in a stream of stimuli using a variety of perceptual paradigms (Bregman, 1990). Auditory scene analysis refers to the problem of separating the incoming mixture of sounds that reaches our ear and into individual perceptual objects.

Figure 1.1: A typical cocktail party.

The listener must follow the speech of one person in the presence of several other sounds. (Image from Breakfast at Tiffany's: Paramount Pictures).

Consider a typical listening environment that suitably describes the "cocktail party effect" (Cherry, 1953) as shown in figure 1.1. Audrey Hepburn is faced with the challenging task of listening to the person wearing the eye patch in the presence of several people who act as sources of background noise. In order to make sense of his speech, Audrey's brain has to decompose the mixture into discrete signals of interest and noise; maintain a stable representation of his voice and selectively attend to it over time.

Thus, there are two aspects to the cocktail party problem – firstly, a problem of sound segregation, and secondly, a problem of directing attention to the (segregated) sound of interest (McDermott, 2009). These two problems can be assumed to operate at two distinct levels of processing that may interact with each other – a bottom-up (primitive) low-level sensory process that is concerned with efficient coding of the stimulus features and deriving the properties of individual sounds; and a top-down (schema-based) cognitive process that operates at a higher level, presumably directly at the level of the grouped patterns of sounds and is concerned with allocation of attention to the target sound, switching attention between targets and maintaining a stable perceptual representation over time (Bregman, 1990, 2008). Bregman defined schemas as a set of brain processes for dealing with acoustic patterns. These schemas could be innate or could also be developed through learning and interaction with the environment. Schemas are essentially mechanisms that help in sound

recognition and categorization and thus assist in segregating these patterns from the background.

This problem is not unique to humans and is of importance to several species that must identify their mates, offspring, prey or predators in crowded environments. This suggests that over the course of evolution, the brain may have developed specialized mechanisms to perform auditory scene analysis that are robust in the presence of background noise.

1.2.1.1 What is an auditory object?

How do our senses treat environmental stimuli and form 'object-like' representations that are different for different objects yet remain stable for the same objects in time and space? Objects in general, can be considered as perceptual entities that are represented in the brain based on generic mechanisms that analyze and represent sensory information. The concept of such a perceptual object makes intuitive sense in vision but is a difficult notion in audition, touch and other senses.

Auditory objects can be defined as complex two-dimensional patterns in frequency-time space that are governed by grouping mechanisms in the frequency and time domains (Griffiths and Warren, 2004; Griffiths et al., 2012; Bizley and Cohen, 2013). Unlike visual objects, there are no clear edges or perceptual boundaries that distinguish one auditory object from another and the separation of information related to the object and to the rest of the background becomes a challenging task. Objects also need to satisfy properties of invariance or constancy in any one sensory domain – for

instance, a face is recognized as the same object when viewed from different angles or a voice is recognized as belonging to the same speaker even when it varies in loudness or pitch.

Generic principles of object analysis have been proposed for auditory objects (Griffiths and Warren, 2004; Bizley and Cohen, 2013) that are based on analysis of auditory patterns in frequency-time space. Auditory patterns can be grouped on the basis of several grouping principles that aid perceptual classification and are discussed in the following section.

1.2.1.2 Auditory grouping cues

Although it may appear that there is no structure in the signals in the world around us, they are often characterized by statistical regularities which the human brain may have learned over the course of evolution. For instance, most natural signals are characterized by a 1/f power spectrum and are represented by sparse perceptual codes in the primary visual (Olshausen and Field, 1996) and auditory cortices (Hromadka et al., 2008). A sparse code is a neural code in which each object is encoded by the strong activation of a relatively small population of neurons (Barlow, 1972). The ability to resolve complex acoustic mixtures is a challenging problem that may be rendered easier by the use of certain grouping principles or heuristics that exploit statistical regularities in the world (Bregman, 1990).

Auditory grouping can be considered to have two aspects – simultaneous grouping and sequential grouping. Simultaneous grouping refers to the task of determining which parts of the complex acoustic input

presented at the same time belong to which particular source. Natural sounds overlap in time and disentangling this mixture into separate sources presents a challenge to the sensory systems. Sequential grouping, on the other hand, is required for relating spectral components to their respective sources over time. Although there is an interaction between the two types of grouping processes, these are often investigated separately.

The Gestalt principles of grouping were postulated by a group of German psychologists in the early twentieth century to explain how units of visual experience are connected to one another (Koffka, 1935; Köhler, 1947). Gestalt refers to a 'pattern' and the psychologists developed an influential theory of how the brain generates mental patterns by forming connections between elements of sensory input based on the principles of similarity, continuity, proximity, and common motion. Visual objects can be grouped together on the basis of similarity and proximity which also apply for grouping of sounds based on similar features such as pitch and grouping on the basis of the location of sources. Sounds can also be classified as belonging to the same source based on principles of common onset and offset, harmonicity, and localization (Bregman, 1990; Darwin and Carlyon, 1995). A difference in these features between sounds can be used as a cue to distinguish between them.

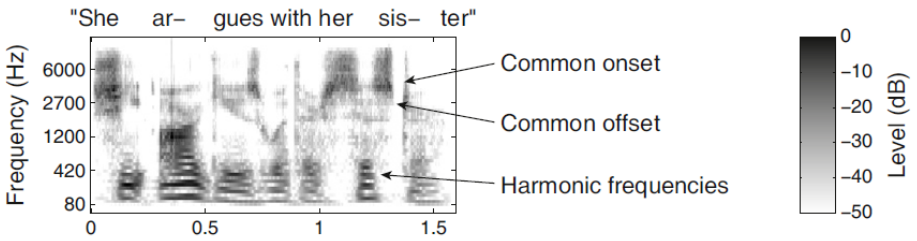

Figure 1.2: Grouping cues in audition.

The spectrogram of a speech sample is used to illustrate cues for grouping sounds together such as common onset, common offset and harmonic structure. Figure reproduced from McDermott, 2009.

In the case of complex tones (as opposed to pure tones), segregation can be achieved based on differences in a number of features including fundamental frequency, spectral shape or envelope (that determines timbre), spatial location, intensity, and amplitude envelope (Bregman, 2008).

1.2.2 Visual scene analysis

The problem of object separation, binding and perceptual representation is not unique to the auditory domain. The segmentation of visual scenes is a fundamental process of early vision and has received much more attention, especially by the Gestalt psychologists. Several principles of auditory grouping are indeed inspired by research into visual segmentation and principles of visual information processing (e.g. Julesz, 1962; Sporns et al., 1991), and continue to inspire models of auditory processing (King and Nelken, 2009).

The coherent dot motion paradigm has inspired several models of perceptual grouping in vision (Shadlen and Newsome, 1996). It consists of a number of dots whose direction of motion is parametrically controlled as shown in figure 1.3. The percentage of dots moving in a certain direction defines the coherence of those particular set of dots that comprise a "figure" which moves in a different direction from the remaining dots that move in random directions and comprise the "ground". Such stimuli have been instrumental in understanding the properties of direction- and orientation-selective cells in the primary visual cortex and inspired analogous versions of synthetic stimuli for examining auditory object formation (e.g. Overath et al., 2010).

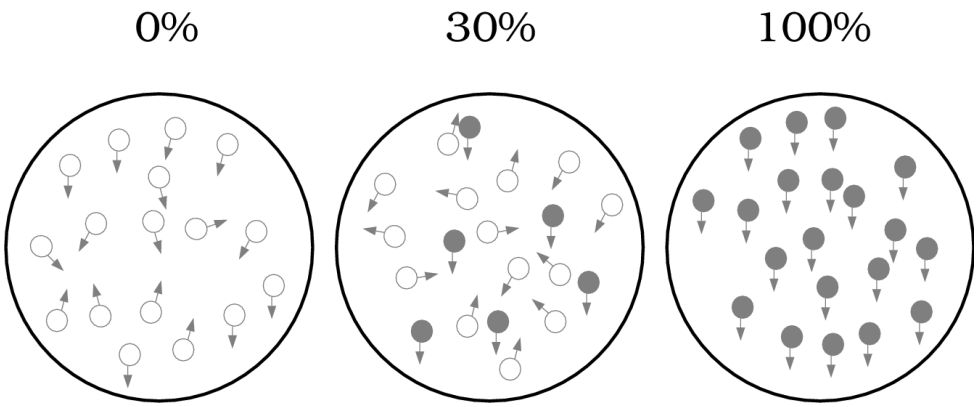

Figure 1.3: Visual coherent dot motion paradigm.

The paradigm involves manipulation of the number of dots moving in a certain direction whilst the remaining dots move in random directions. The three examples here indicate three different levels of coherence: 0%, 30%, and 100%.

Although visual scenes also contain multiple objects at different locations, the problem of segmentation is more pronounced for auditory signals. A prominent difference arises at the earliest level of processing – visual objects tend to occupy local regions on the retina whilst sounds are spread across the frequency map of the cochlea. Thus, there is considerable overlap in the representation of auditory objects at the initial stage of processing. Secondly, sound sources combine linearly to form a single input waveform at the ears, whilst visual objects occlude each other. Thirdly, the visual world is relatively static compared to the much more dynamic acoustic scenes. Another source of difference occurs at a higher level – auditory segregation is much more difficult to perform compared to visual segmentation and requires a significant "cognitive effort", that becomes worse with aging and hearing loss. These differences highlight the challenging aspect of auditory scene analysis and point towards different mechanistic bases of segregation in audition compared to vision.

The following section considers the anatomical and functional properties of the auditory system and how the underlying organizational principles of information processing in the auditory system inform our understanding of the mechanisms involved in auditory scene analysis.

1.3 The auditory system

The range of frequencies to which the auditory system best responds to varies from one species to another. In rats, it ranges from 0.25 – 70 kHz, from 0.125 – 60 kHz in cats, and from 0.02 – 20 kHz in humans. The processing of frequency can be considered to be the primary function of the

auditory system and serves as a major organizing principle. This operation is achieved by coordinated activity from the cochlea in the periphery to higher-order areas in the association auditory cortex. The next section briefly describes the anatomical pathways and flow of information between the various auditory processing stations.

1.3.1 Information flow from the cochlea to the cortex

The acoustic input that reaches our ears is processed by a network of structures that comprise the primary (lemniscal) ascending auditory pathway. At the level of the periphery, sound waves are mechanically transmitted through the outer and middle ear to the hair cells of the organ of Corti that is part of the cochlea of the inner ear. Hair cells span the entire length of the basilar membrane whose mechanical properties gradually vary along its length. This results in differential tuning of the hair cells such that they are tuned to progressively lower frequencies from the base to the apex of the cochlea. The cochlea thus acts as a frequency analyser (von Békésy, 1970) and this information is transmitted to the brainstem by auditory nerve fibres that synapse on the inner hair cells. In the cochlear nucleus complex, the input from the auditory nerve is shunted into a number of parallel ascending pathways that are characterized with separate trajectories and destinations (Fuchs, 2010). These tracts converge on the auditory midbrain, i.e. the inferior colliculus (IC) that serves as an obligatory relay station en route to the auditory cortex. The IC is organized into different nuclei that include the central (ICc), dorsal cortex (IC_{DC}) and lateral (IC_{L}) nuclei. The central IC nucleus is tonotopically organized and forms part of the “core

projection" whilst the other divisions of the IC constitute the non-tonotopic or diffuse ascending pathway, that comprise the "belt projection". The ICc projects to the ventral division of the auditory thalamus, the medial geniculate body (MGB) that mainly targets the tonotopically organized core areas of the auditory cortex. The dorsal divisions of the MGB receive afferents from the IC_{DC} and IC_{L} nuclei of the IC while a third magnocellular division of the MGB receives input from all three nuclei of the IC. The dorsal divisions of the MGB target the non-tonotopic belt areas that surround the core auditory cortex. Apart from the MGB, the auditory cortex also receives input from adjoining nuclei in the posterior thalamus that have auditory and multisensory properties (Hackett, 2011).

Thus, each cortical field receives inputs from the individual thalamic nuclei that have specific neurochemical properties. Each thalamic station thus sends distinct information to its cortical targets and can be assumed to form parallel information streams (Rodrigues-Dagaeff et al., 1989; Rouiller et al., 1989; Jones, 2003; Lee and Winer, 2008a).

1.3.2 Structural organization of the auditory cortex

In this section, the organizational structure of the auditory cortex is described with a specific focus on the anatomy of the human auditory cortex. Although with the advent of high-resolution functional and structural MR imaging, parcellation of human auditory cortical fields can be investigated in better detail, the core knowledge of the organizational principles of human auditory cortex is derived from cytoarchitectonic as well as physiological studies in animal models, especially the primates.

However, in spite of decades of research on this topic, the accurate definition of primary and secondary human auditory cortex still eludes us. There is considerable debate about the nomenclature used to describe human cortical fields and there is no established homology between core, belt and parabelt regions in the macaque auditory cortex and primary, secondary and association cortex in humans although a correspondence between these areas is assumed in the literature (see Baumann et al., 2013).

Early architectonic studies identified core auditory cortex in the temporal plane on the basis of a well-developed granular layer 4 (koniocortex), dense myelination, and thalamic connectivity (Fleschig, 1876, 1908; Campbell, 1905; Brodmann, 1909; von Economo and Koskinas, 1925; von Economo and Horn, 1930). Recent definitions of auditory cortex suggest that it comprises those areas of the cerebral cortex that receive significant thalamic input from one or more divisions of the MGB (Hackett, 2011). This definition constrains the auditory cortex to a group of adjoining regions in the superior temporal plane. In humans and higher primates, a significant portion of the auditory cortex is hidden beneath the Sylvian fissure (or lateral sulcus, as commonly denoted in other higher primates), separating the parietal and temporal lobes.

In all mammals that have been studied, the auditory cortex comprises more than one area as shown in figure 1.4. In cats and primates, more than ten areas have been identified which are classified into a central "core" region, whilst the secondary areas are grouped as "belt" and "parabelt" regions surrounding the core (Hackett, 2011). The core area

consists of a primary area (A1), and more anterior rostral (R) and rostrotemporal (RT) areas whilst the belt and parabelt subfields are named according to their respective anatomical locations (e.g. anterolateral, AL, caudomedial, CM; rostromedial, RM, and so on; see Figure 1.4).

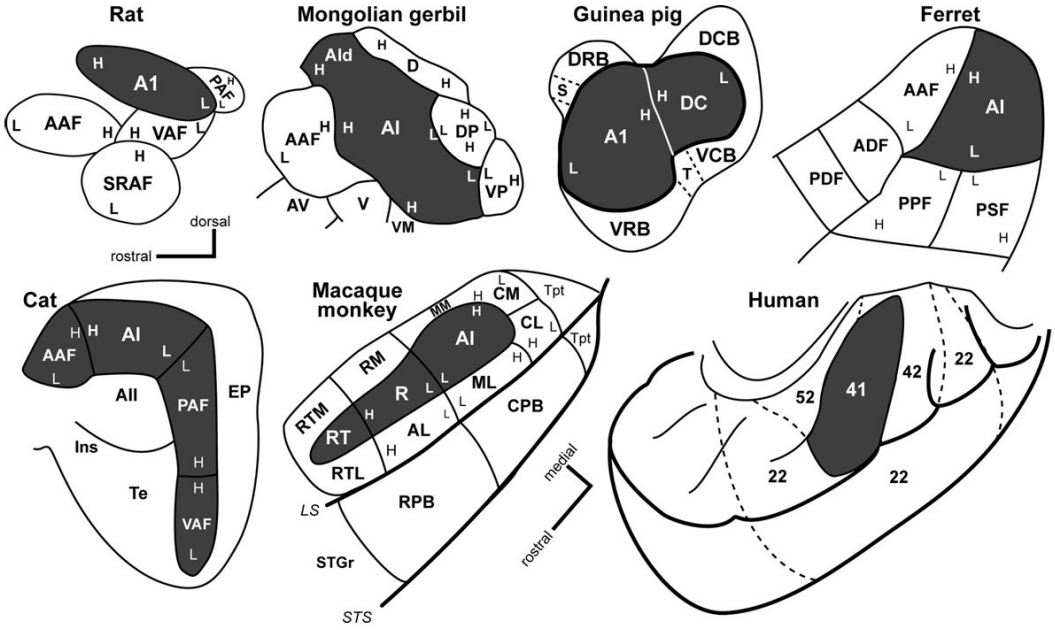

Figure 1.4: Organization of the auditory cortex in selected mammals.

Primary (core) areas are shaded while belt and parabelt areas are unshaded. Tonotopic gradients are depicted by H (high) and L (low) frequency. Figure reproduced from Hackett, 2011.

Abbreviations: AAF, anterior auditory field; A1, auditory area 1; AL, anterolateral area; CPB, LS, lateral sulcus; ML, middle lateral area; MM, middle medial area; PAF, posterior auditory field; R, rostral area; Ri, retroinsular area; RM, rostromedial area; RPB, rostral parabelt area; RT, rostrotemporal area; RTL, rostrotemporal lateral area; RTM, rostrotemporal medial area; STG, supeior temporal gyrus V, ventral division (medial geniculate); VAF, ventral posterior auditory field.

In humans, the homologous core, belt, and parabelt regions comprise some 30 functionally distinct subfields (Hackett, 2011; Clarke and Morosan, 2012). The human auditory cortex is considered to include the posterior portion of the superior temporal cortex, including the Heschl's gyrus (HG), the planum temporale (PT), and some areas in the posterior superior temporal gyrus (STG). These areas correspond to Brodmann areas (BA) 41, 42, 52, and 22 (Brodmann, 1909). Areas in the superior temporal sulcus (STS) and, more rostrally towards the planum polare (PP), at the temporal pole, are considered auditory-related areas (Hackett, 2011). In humans, the homologue of the core in non-human primates can also be classified into three distinct areas: a primary area in central HG, and two secondary areas in medial and anterolateral HG (Morosan et al., 2001; Rademacher et al., 2001); these are also referred to as areas Te1.0, Te1.1, and Te1.2, respectively (Morosan et al., 2001, 2005). Although there is consensus on the location of the centre of the human primary auditory cortex in the medial two-thirds of HG, its exact areal borders and number of subdivisions are still debatable (Clarke and Moroson, 2012; Baumann et al., 2013). This is partly complicated by the high inter-subject and inter-hemispheric variability with features such as forked or duplicated HG (estimated occurrence 41%, Rademacher et al., 1993). In the case of a duplicate HG, the primary auditory cortex usually covers parts of both gyri and the intermediate transverse sulcus.

The problem of the exact location of the human primary auditory cortex is also aggravated due to the lack of techniques to delineate PAC in vivo. Anatomical labels derived from post-mortem cytoarchitectonic maps

may be inaccurate in relation to different samples of in-vivo brains (Morosan et al., 2001). A recent approach to this problem involves the use of high resolution structural magnetic resonance imaging and analysis of the MR-tissue characteristics to precisely define the location of PAC in vivo. These techniques include high-resolution (800 μm) quantitative T1-mapping (Dick et al., 2012), mapping of longitudinal relaxation rate (R1; Sigalovsky et al., 2006; Lutti et al., 2013), as well as the complementary use of a combination of MR contrasts (T1 and T2) at high-resolution (700 μm; Wasserthal et al., 2013). Quantitative T1 and R1 mapping provide estimates of myelination and it has been shown that areas of high myelination co-localize with auditory koniocortex along the posteromedial two-thirds of the Heschl's gyrus (Sigalovsky et al., 2006; Dick et al., 2012).

1.3.3 Cytoarchitecture of auditory cortex

The primate auditory cortex is characterized with a distinctive cytoarchitecture: the core area displays typical features of primary cortex with a dense layer IV, indicating rich thalamocortical connections (Galaburda and Pandya, 1983). The ventral MGB projects mainly to layers IIIb and IV of the core whilst the dorsal MGB divisions send information to layers IIb of the belt and parabelt, avoiding layer IV (Hackett, 2011). Additionally, the core region is highly granular and myelinated, displays high metabolic activation patterns and stains profusely for the calcium binding protein, parvalbumin (Morel et al., 1993; Jones et al., 1995; Pandya, 1995; Kaas and Hackett, 2000). These properties are most pronounced for A1 and least prominent for area RT (Hackett et al., 1998a).

The cytoarchitectonic characteristics of the human homologue subfields share similarities with the core of the primate auditory cortex: all constituent areas show strong cytochrome oxidase, parvalbumin and acetylcholinesterase staining in cortical layers IIIc and IV (Rivier and Clarke, 1997; Clarke and Rivier, 1998; Hackett et al., 1998a; Hackett et al., 2001; Wallace et al., 2002). The secondary subfields in medial and lateral HG exhibit weaker metabolic activity in layer IV than the primary subfield (Wallace et al., 2002). The cytoarchitectonic properties of the human auditory association regions are however less well defined, making comparisons to primate belt and parabelt regions problematic (Hackett, 2011). Similar processing schemes have been demonstrated by diffusion imaging techniques in humans (Behrens et al., 2003; Behrens and Johansen-Berg, 2005) which confirm the functional architecture within human auditory cortex (Upadhyay et al., 2007, 2008).

1.3.4 Information flow within auditory cortex

Each cortical area is characterized by a unique pattern of connections but appears to follow certain anatomical relationships that help understand the nature of the interconnections between the cortical areas: (i) a particular area has reciprocal connections with other areas; (ii) adjacent areas in the cortex are more densely interconnected than anatomically segregated areas; (iii) neurons within a single area share dense connections; (iv) there is a systematic pattern of connections in cortical laminae and sublaminae (Kaas and Hackett, 1998; Read et al., 2001; Winer and Lee, 2007; Lee and Winer, 2008b; Hackett, 2011).

There are three major classes of cortical connections: feedforward projections that connect the output of one area to layer IV of another; feedback projections that arise from the infragranular layers of an area but avoid layer IV of the target area; and lateral connections that involve all layers and typically connect adjacent areas. These connections form a hierarchical network where feedforward inputs carry information from a lower to a higher hierarchical level and feedback projections transmit information from higher to lower hierarchical centres.

A general feature of the connectivity between the different auditory fields is that areas with similar thalamic inputs and cortical (usually lateral) connections belong to the same hierarchical level whilst areas with distinct thalamic inputs and clear feedforward or feedback projections are allocated to different hierarchical levels. In primates, along the medial-lateral axis, the belt region has been found to be densely interconnected with the core and parabelt areas, whilst the core and the parabelt regions are only weakly connected (Hackett et al., 1998a). The core areas send driving inputs to the surrounding regions in the belt but not to the parabelt areas. On the other hand, neurons in the parabelt project back to the core areas, possibly suggesting a feedback circuit (de la Mothe et al., 2006). These connectivity patterns are indicative of a hierarchical flow of information from core to belt to parabelt along the medial-lateral axis (see Figure 1.5; Kaas and Hackett, 1998, 1999; Rauschecker, 1998). Along the rostral-caudal axis, there is evidence that from PAC, information flows rostrally towards auditory and

auditory-related areas and caudally towards the temporo-parietal region (Figure 1.5; de la Mothe et al., 2006; Hackett, 2011).

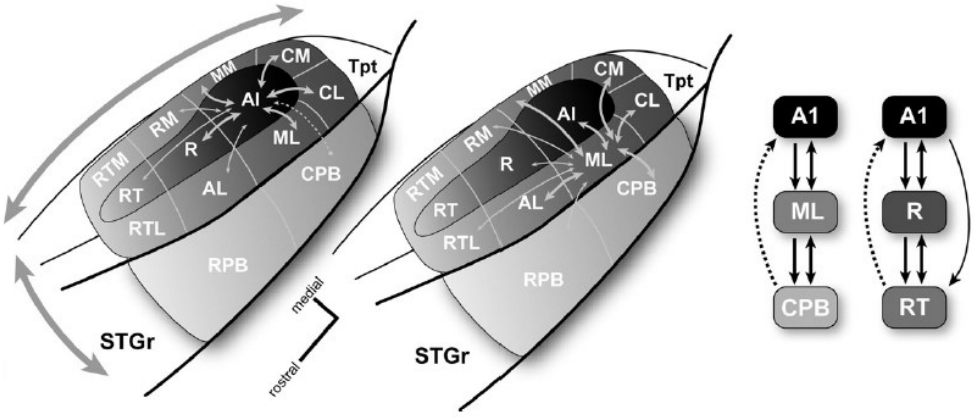

Figure 1.5: Local connections of core and belt areas in the primate.

Left: connections of the core area, A1 (left), and lateral belt area, ML (middle) in the primate.

Right: schematics of information flow along the medial-lateral axis (A1-ML-CPB) (core-belt-parabelt) and caudal-rostral axis in the core (A1-R-RT). Line thickness denotes the relative density of each projection. Dashed lines indicate feedback projections. Shading intensity (all panels) and large arrows (left panels) denote anatomical and physiological gradients along two major axes of information flow. Figure reproduced from Hackett, 2011.

1.3.5 Information flow beyond auditory cortex

Auditory processing is not only limited to the auditory cortex but also extends to auditory-related areas in the forebrain. The projections beyond the auditory cortex derive mainly from the belt and parabelt regions with only sparse projections from the core (Hackett, 2011). Information flows out of the auditory cortex in multiple directions but is influenced by the topographic flow of information within the auditory cortex as described previously. The principal pathways in the auditory network are known as *processing streams* (Kaas and Hackett, 1999, 2000; Rauschecker and Tian, 2000; Rauschecker and Scott, 2009), concordant with the term used to describe similar pathways in the somatosensory and visual domains (Mishkin, 1979; Ungerleider and Haxby, 1994). In the visual system in particular, processing streams have been extensively investigated: topographic connections between areas suggest the existence of two separate pathways known as the dorsal and ventral streams which are involved in the analysis of information related to 'where' and 'what', respectively. These pathways are also commonly known as the 'where' and 'what' pathways.

In the auditory system as well, there is evidence of two parallel processing streams that encode two different types of auditory information: the identity ('what') and the spatial location ('where') of the source (Kaas and Hackett, 1999; Romanski et al., 1999; Rauschecker and Tian, 2000; Tian et al., 2001). A rostrally directed stream with auditory-related targets in the temporal pole, ventral, rostral and medial prefrontal cortex, rostral

cingulate, parahippocampal cortex and the amygdala processes 'what' information. The 'where' information is processed by a caudally-directed stream that flows out from the caudal belt and parabelt regions into the temporoparietal junction, posterior parietal and secondary visual cortex, caudal and dorsal prefrontal areas, dorsal cingulate and parahippocampal areas (Hackett, 2011). There are two other less well defined streams that flow laterally from the belt and parabelt areas to the upper bank of the superior temporal sulcus and medially into insular areas in the lateral sulcus (Galaburda and Pandya, 1983; Hackett et al., 1998b; de la Mothe et al., 2006).

1.3.6 Tonotopy

Orderly topographic information processing pathways are a feature of several sensory cortical systems: neurons in the visual cortex show retinotopy, i.e., a one-to-one mapping of visual input from the retina to the cortex; and, there is an orderly representation of different body parts in the somatosensory cortex which gives rise to the cortical homunculus. The auditory system displays a similar mapping of frequency from the level of the cochlea to the auditory cortex, and this frequency-place code is referred to as tonotopy. The mechanical properties of the basilar membrane result in an orderly arrangement of a bank of bandpass filters that are tuned to progressively higher frequencies from the apex to the base of the membrane (von Békésy, 1960). This mapping of frequency to spatial position is preserved in the lemniscal ascending auditory pathway including the central nucleus of the IC, the ventral division of the MGB and the recipient cortical

fields. The non-lemniscal pathways that project from the dorsal or magnocellular divisions of the MGB, however, are not tonotopically mapped. Several cortical areas show tonotopic frequency gradients with the reversal of gradients often used to define the border between the distinct subfields.

Tonotopic maps in human auditory cortex have been demonstrated using fMRI (Formisano et al., 2003; Schönwiesner et al., 2002; Talavage et al., 2004; Langers et al., 2007; Humphries et al., 2010, Woods et al., 2010; Da Costa et al., 2011; Striem-Amit et al., 2011; Langers and van Dijk, 2012; Moerel et al., 2012; Herdener et al., 2013; Saenz and Langers, 2013). In macaques, fMRI has been useful to define tonotopic maps as well (Petkov et al., 2006; Baumann et al., 2010; Tanji et al., 2010; Baumann et al., 2013). Although it is possible to localize the boundary between fields A1 and R using tonotopy, the relation between frequency reversals and the location of the core koniocortex is still unclear. Probabilistic post-mortem cytoarchitectonic maps, also fail to clarify the localization (Morosan et al., 2001). Tonotopy can distinguish core areas A1, R and RT from each other and adjacent belt areas from each other but does not allow core and belt to be distinguished.

In primates, the core area A1 displays a tonotopic gradient from high to low along the rostral to caudal axis whilst subfield R shows a reverse gradient from low to high and the gradient in RT is similar to A1 (Figure 1.6). In spite of intense research on the topic, the exact configuration of the tonotopic gradients is currently under debate. Figure 1.6 demonstrates the

different proposed configurations in macaque and human auditory cortex. Accurate characterization of tonotopic maps in humans is limited due to the poor spatial resolution of fMRI relative to the size of the auditory subfields and the lack of a clear consensus regarding the precise location of the primary auditory cortex. A detailed discussion of the topic is beyond the scope of the thesis; these issues are aptly summarized by Baumann et al. (2013) and Saenz and Langers (2013).

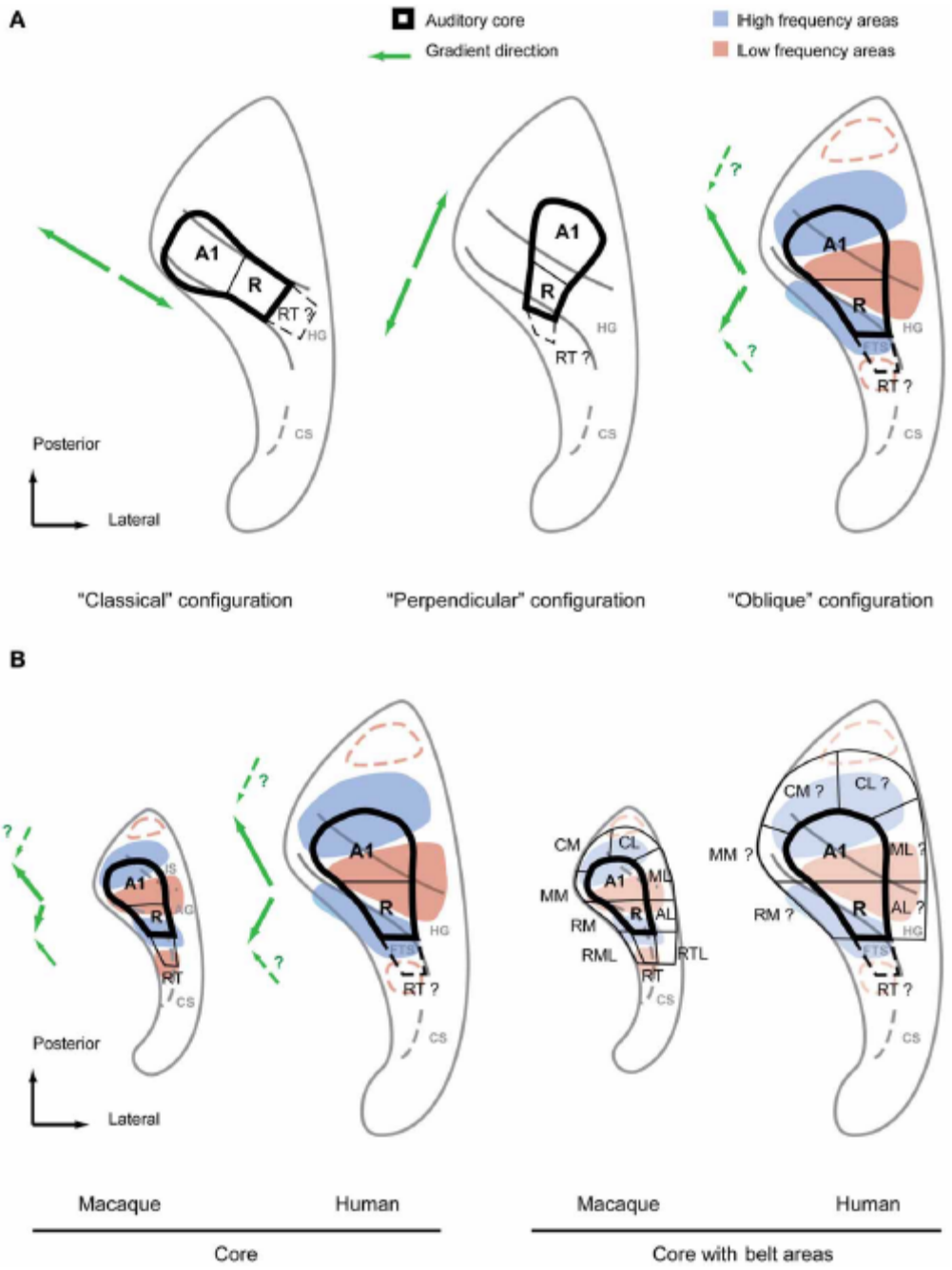
A
Auditory core
Gradient direction
High frequency areas
Low frequency areas
A1
R
RT ?
HG
CS
Posterior
Lateral
"Classical" configuration
"Perpendicular" configuration
"Oblique" configuration
B
CM
CL
ML
AL
MM
RM
RML
RTL
CM ?
CL ?
MM ?
ML ?
RM ?
AL ?
Macaque
Human
Core
Core with belt areas

Figure 1.6: Configurations of auditory cortical organization in humans and non-human primates.

(A) The two main configurations of auditory core fields under debate (left, middle) in comparison with the “oblique” configuration proposed by the authors (right). The main frequency response areas based on the summary of recent evidence (Formisano et al.,2003; Humphries et al., 2010; Woods et al.,2010; Da Costa et al.,2011; Striem-Amit et al.,2011; Langers and van Dijk,2012) are superimposed over this configuration. The suggested directions of the main gradient axes are indicated with green arrows next to each configuration. Additional anterior and posterior low frequency preference areas suggested by some studies are marked by red dashed lines.

(B) Core fields and frequency preference areas in the superior temporal plane of macaque and human according to oblique configuration (left). Location of auditory belt fields in macaques and presumed location of belt fields in humans (right). Main gradient directions from low to high of the frequency response areas are indicated with green arrows left of each scheme. IS, intercalated sulcus; AG, annectant gyrus; CS, circular sulcus; FTS, first transversal sulcus. Figure reproduced from Baumman et al., 2013.

1.3.7 Neurophysiological correlates of information flow

The topographical pattern of anatomical pathways also influences the physiological properties of the various auditory cortical subfields. One distinctive feature of auditory cortical processing is that the core areas appear to be specialized for encoding basic spectrotemporal features whilst the belt and parabelt areas process more complex acoustic attributes. This is exemplified in the case of frequency processing where simple sinusoidal stimuli are robustly encoded by core areas whilst the belt and parabelt areas respond more strongly to complex sounds and conspecific vocalizations (Rauschecker et al., 1995; Rauschecker, 1998; Rauschecker et al., 1997; Rauschecker and Tian, 2000; Tian et al., 2001).

Another prediction of the core-belt-parabelt serial processing model is that response latencies would increase, spectral integration (tuning bandwidth) would increase, and temporal precision would decrease (Rauschecker, 1998). Temporal precision relates to entrainment to periodic temporal events and there is evidence that it systematically decreases from A1 to ML to CPB (Hackett, 2011). Thus, there is a general rule that neurons become more broadly tuned and more temporally sluggish along the core-belt-parabelt axis, consistent with the information processing hierarchy from core to belt to parabelt. Within the core, however, latencies increase from A1 to R to RT (Recanzone et al., 2000; Bendor and Wang, 2008; Kusmierek and Rauschecker, 2009). This trend is also apparent in human auditory cortex where responses in medial HG are usually recorded ~20ms post stimulus onset, whilst central and lateral HG take longer time to respond,

with latencies of ~50ms and ~60-75ms, respectively (Liegeois-Chauvel et al., 1990). Responses in PT peak ~100ms (Brugge et al., 2008; Liégeois-Chauvel et al., 1991) and the latencies rise and become more variable in the STG and parietal operculum (Liégeois-Chauvel et al., 1991). There is also evidence for back-projections from the STG to HG (Brugge et al., 2003), which may serve a modulatory function.

1.3.8 Spectrotemporal receptive fields

Functional properties of auditory cortical neurons can be represented in terms of their spectrotemporal receptive fields (STRFs; Aertsen and Johannesma, 1981a, 1981b; Eggermont et al., 1981). The STRF is a summary of the cell's response properties and is represented by a kernel in the spectral and temporal domain and can be measured in many ways (Calhoun and Schreiner, 1995; deCharms et al., 1998). STRFs reflect both excitatory as well as inhibitory response characteristics and provide important clues to information processing in the cortical neurons (Elhilali et al., 2007).

A popular method for measuring STRFs is the "ripple analysis method" (Kowalski et al., 1996; Klein et al., 2000) where ripples refer to sinusoidally modulated spectrotemporal envelopes whose properties can be parameterized. Neurons in PAC respond strongly to ripple stimuli and exhibit selectivity to a narrow range of parameters that reflects their STRF characteristics. By varying the ripple velocity and density over a wide range of parameters, a complete description of the cell's spectrotemporal response properties can be obtained.

Interesting insights into auditory processing have been obtained using this method. Elhilali and colleagues (2007) demonstrated that neurons show stable STRFs to certain acoustic features when these are not behaviourally relevant but these change rapidly if the stimuli are made behaviourally relevant (Fritz et al., 2003, 2005, 2010; David et al., 2012). STRFs measured in awake monkeys have been shown to display on-excitation as well as off-excitation, providing an elegant neural code for spectrotemporal integration as in the case of natural sounds and conspecific vocalisations (Shamma & Symmes, 1985; Pelleg-Toiba & Wollberg, 1989; deCharms et al., 1998). STRFs have also been computed based on MEG responses to analyse encoding of speech in human auditory cortex (Ding and Simon, 2012).

The above section (1.3) provided a general framework of the functional anatomy of the auditory cortex, highlighting fundamental principles of information processing, such as the co-existence of serial and parallel processing streams and gradation of physiological properties along the ascending auditory pathways. These provide a foundation for a proper understanding of auditory processing in response to complex signals used in auditory scene analysis research considered in the following section.

1.4 Stimuli used in auditory scene analysis

Several stimuli and paradigms have been employed to study auditory scene analysis. These range from a simple sequence of two alternating tones (streaming), to a sequence of tones whose probability of occurrence is varied to elicit a mismatch response (oddball stimuli), to multi-tone stimuli

with embedded targets (informational masking stimuli), to complex naturalistic stimuli (acoustic textures) as well as natural vocalizations including speech. These stimulus paradigms have been most successful in uncovering the principles of auditory segregation. The acoustic details of each stimulus are presented in the next section followed by a review of the literature and findings based on the corresponding stimulus in section 1.5.

1.4.1 Streaming

Streaming refers to a stimulus as well as a phenomenon that is based on a sequence of two pure tones (A and B) of different frequencies alternating in time as shown in Figure 1.7. This pattern of tones forms triplets (ABA_ABA_ABA_ …) that are separated by short silent intervals and repeat over time. Although it appears to be a very simple acoustic pattern, this stimulus has distinct perceptual effects that have made it one of the most commonly used signals.

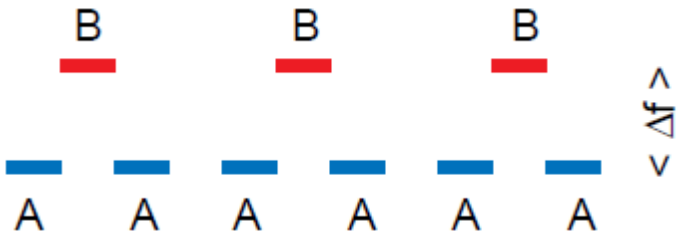

Figure 1.7: A schematic of the streaming stimulus.

The streaming stimulus consists of two tones, A and B, that alternate in time and are separated by a frequency separation (Δf). Figure reproduced from Carlyon, 2004.

At slow rates of presentation, the ABA triplets are perceived as repeating units with a galloping rhythm; but as the sequence becomes faster, the high frequency tones separate from the low frequency ones into two distinct isochronous sequences – one consisting of a slow sequence of high tones and the other, a faster sequence of low tones. This phenomenon is referred to as 'streaming' and the two sequences are called 'streams' (van Noorden, 1975; Bregman, 1990). Although the two streams occur at the same time, they are perceived independently. However, one can focus their attention to one stream only which forms the foreground whilst the other stream is relegated to the background.

Another feature of streaming is observed by varying the frequency difference between the two tones. As the spectral separation is increased, the percept tends to change to that of two separate streams. Thus, by varying the frequency separation and the rate of presentation, the perceptual effects can be modulated from that of a single 'integrated' percept to two divergent 'segregated' percepts.

Apart from spectral cues, non-spectral factors also influence stream segregation (Moore and Gockel, 2012). These cues include rate of fluctuation of temporal envelope (Grimault et al., 2002), timbre (Iverson, 1995), phase spectrum (Roberts et al., 2002), fundamental frequency (F0; Vliegen and Oxenham, 1999), lateralization cues such as interaural time differences (ITD; Darwin and Hukin, 1999; Stainsby et al., 2011), onset and offset asynchrony (Darwin and Carlyon, 1995), harmonicity (Moore et al., 1986), and ear of entry (Darwin and Carlyon, 1995).

The formation of streams is a type of auditory scene analysis – when the streams are segregated, the auditory system perceives two sound sources in the environment instead of one. Although this is a simplistic representation of the complex type of segregation performed in natural environments, the paradigm has been used extensively to study sequential grouping and illuminated several key principles of auditory perceptual organization.

1.4.2 Oddball stimulus

Another popular experimental tool to study auditory scene analysis is based on a sequence of two tones where the probability of occurrence of the two tones is manipulated: a *standard* tone is presented repeatedly amidst a few *deviant* tones that occur more rarely. This sequence of tones is known as an oddball stimulus where the oddball refers to the deviant tones.

This stimulus is classically used in electrophysiological studies in both humans and animals, usually in passive listening conditions. The tones elicit a clear evoked response which is averaged separately to obtain event related potentials for the standard and deviant tones respectively. The hallmark of perceptual responses is the significantly larger response for the deviant compared to the standard stimuli. The difference between the event related waveforms for the deviants and the standards is measured as the *mismatch negativity* (MMN) response. The MMN is defined as a negative waveform in the deviant ERP response that occurs 150-250ms after sound onset. The magnitude of MMN response varies as a function of the dissimilarity between the standard and the deviants as shown in figure 1.8.

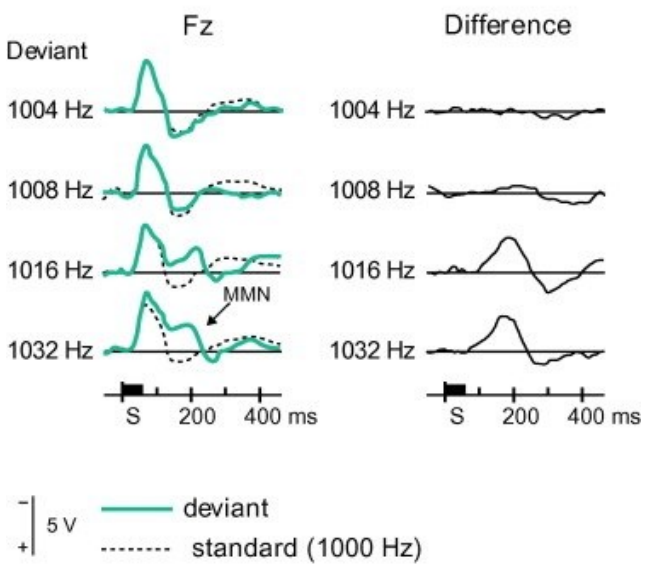

Figure 1.8: MMN response as a function of frequency change.

Left: The responses to deviant tones with increasing frequencies are indicated in blue while the response to the standard (1000Hz) is shown in dotted black lines.

Right: The difference between the deviant and standard response is plotted for each condition on the right. The magnitude of the MMN response increases as a function of the (spectral) dissimilarity between the standard and deviant tones.

This paradigm is especially pertinent for studying auditory scene analysis as the greater response to the deviant tone represents the encoding of a novel object in the acoustic environment (Näätänen et al., 1978, 2007). It has been widely adopted especially in clinical settings as well as special populations including newborn infants (Winkler et al., 2003b) and patients in permanent vegetative states (Boly et al., 2011) as MMN can be elicited even in the absence of task-directed attention. The MMN has been interpreted to reflect different kinds of mental representations (Winkler, 2007) but a prominent explanation is that it represents an error in predicting the incoming acoustic stimuli. This is discussed in greater detail in section 1.5.2 alongside a description of empirical results and theoretical models based on this stimulus paradigm.

An associated positive response often observed in MMN experiments is the P3 which comprises two distinct components: P3a and P3b. In contrast to MMN which reflects a pre-attentive automatic process, P3 requires active attentional processes. P3a is said to originate from stimulus-driven frontal attentional mechanisms during task processing whilst the P3b is related to subsequent memory processing with sources in the temporo-parietal cortex (Polich, 2007).

1.4.3 Informational Masking

In contrast to streaming signals that only comprise two frequencies as shown in Figure 1.9A, a more spectrally complex signal known as 'informational masking' (IM) stimulus has been developed to model natural acoustic scenes as shown in Figure 1.9B. The stimulus has been adopted by

several laboratories to explore aspects of auditory segregation that go beyond the primitive mechanisms required for streaming. IM refers to a type of non-energetic or central masking that is associated with an increase in detection thresholds due to stimulus uncertainty and target-masker similarity that is distinct from peripheral energetic masking (Pollack, 1975; Durlach et al., 2003). These multi-tone masking experiments require listeners to detect tonal target signals in the presence of simultaneous multi-tone maskers, often separated by a 'spectral protection region' (a certain frequency region around the target with little masker energy) that promoted the perceptual segregation of the target from the masker tones.

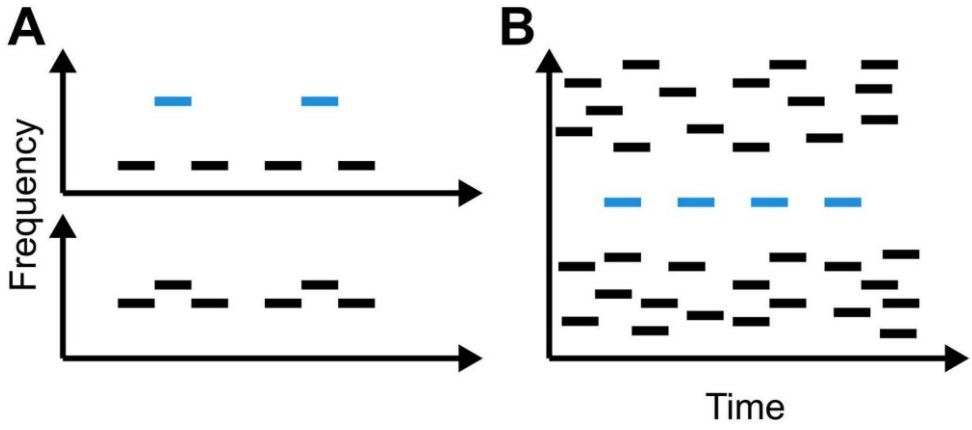

Figure 1.9: Schematic of the informational masking paradigm.

(A) Illustration of the streaming stimulus where the blue tones are the target.

(B) The IM stimulus consists of a target tone (blue) that is repeated regularly in the presence of other masking tones (black) in the background. The target tones are separated from the masking tones by a protective spectral region centred on the target frequency. Figure reproduced from Dykstra and Gutschalk, 2013.

1.4.4 Complex naturalistic stimuli

In recent years, there has been an increasing interest in the synthesis of signals that capture the properties of natural acoustic scenes more faithfully than streaming or multi-tone burst sequences as discussed previously. This line of work is motivated by a multidisciplinary interest in auditory scene analysis with increasing crosstalk between the fields of neuroscience, machine hearing, signal processing and audio engineering.

1.4.4.1 Acoustic Textures

From first principles, an auditory object can be designed based on a number of acoustic features that are constant within a given spectrotemporal space that defines the object. Based on this approach, Overath and colleagues (2008) designed an 'acoustic texture' stimulus based on randomly distributed linear frequency modulated ramps with varying trajectories as depicted in figure 1.10. The percentage of coherent spectrotemporal modulation, i.e., the proportion of ramps with identical direction and trajectory were systematically controlled, producing acoustic textures with different levels of spectrotemporal coherence. Boundaries between textures were created and their magnitude varied by juxtaposing acoustic textures of different coherence levels. In such a stimulus, it is possible to parametrically control and study the emergence of a novel object characterized by a different signature in frequency-time space.

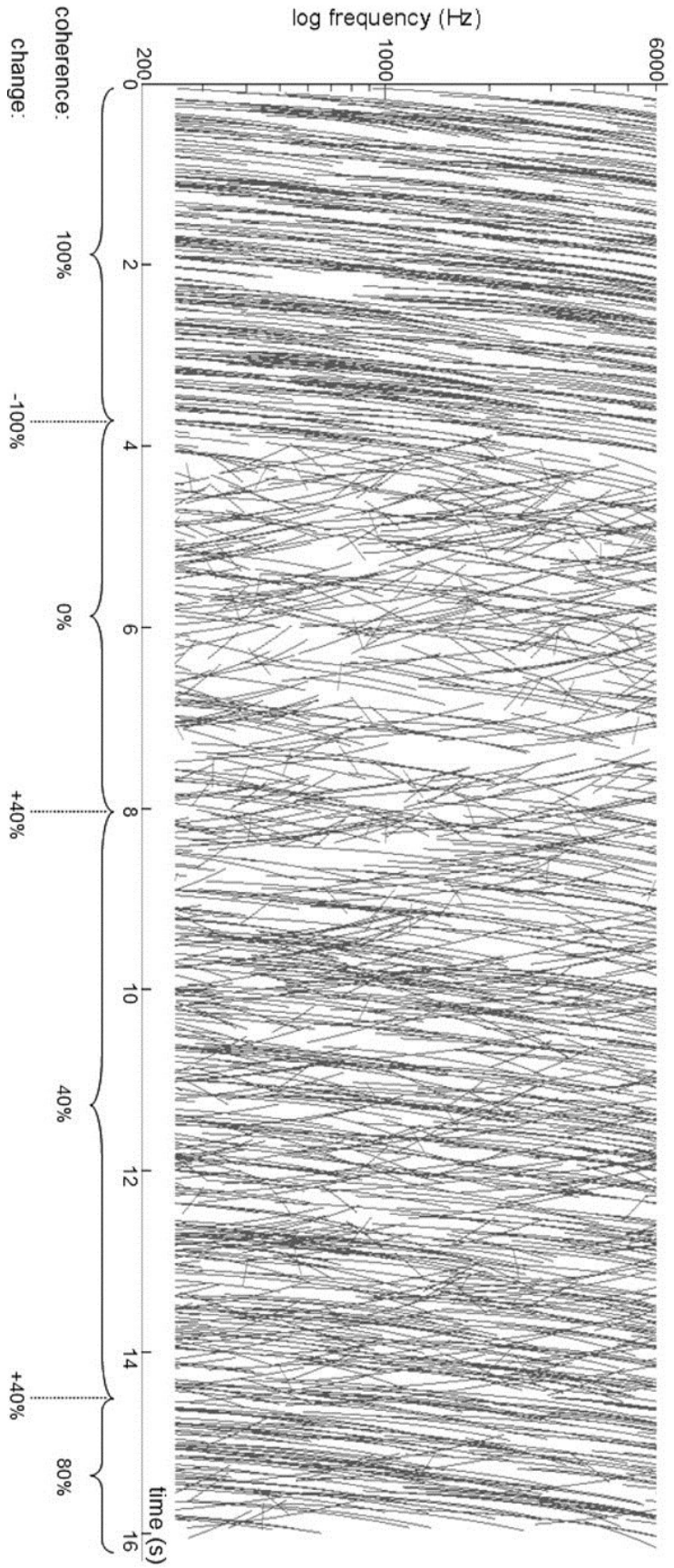
log frequency (Hz)
6000
1000
200
0
2
4
6
8
10
12
14
16
time (s)
coherence:
100%
0%
40%
80%
change:
-100%
+40%
+40%

Figure 1.10: Spectrogram of the acoustic texture stimulus.

Example of a block of sound with four spectrotemporal coherence segments showing absolute coherence values for each segment and the corresponding change in coherence between the segments. The absolute value of coherence as well as the change in coherence from one segment to another is indicated on the bottom. Figure reproduced from Overath et al., 2010.

1.4.4.2 Synthetic sound textures

McDermott and Simoncelli (2011) recently developed a synthetic 'sound texture' stimulus based on statistical measurements of natural stationary signals. Stationary signals are constant in their statistical parameters over time. They are based on sound textures in the real world such as rainfall, stream of water, swarm of insects, or the rustling of leaves that are characterized by temporal homogeneity, i.e. any two samples of such textures recorded at different times sound alike. Conceptually similar to visual textures that have been studied for decades (Julesz, 1962), these sound textures are formed from the superposition of many similar acoustic events, which are characterized by aggregate statistical properties. They processed such real-world textures with an auditory model containing filters tuned for sound frequencies and their modulations, and measured the statistics of the resulting decomposition that summarize the qualities of a sound. Textures provide a compact representation format for encoding sounds and were synthesized to match the statistics of natural sounds as shown in figure 1.11.

These synthetic stimuli provide another controlled approach for studying auditory scene analysis. The utility of this particular stimulus is that any sample of natural sound such as human speech can be taken as the input and a model of textures that captures the statistics of the chosen input can be produced (McDermott et al., 2013). This approach not only informs an investigation into auditory segregation capabilities but also allows a

better understanding of the encoding and analysis of such pseudo-natural sounds in the auditory system.

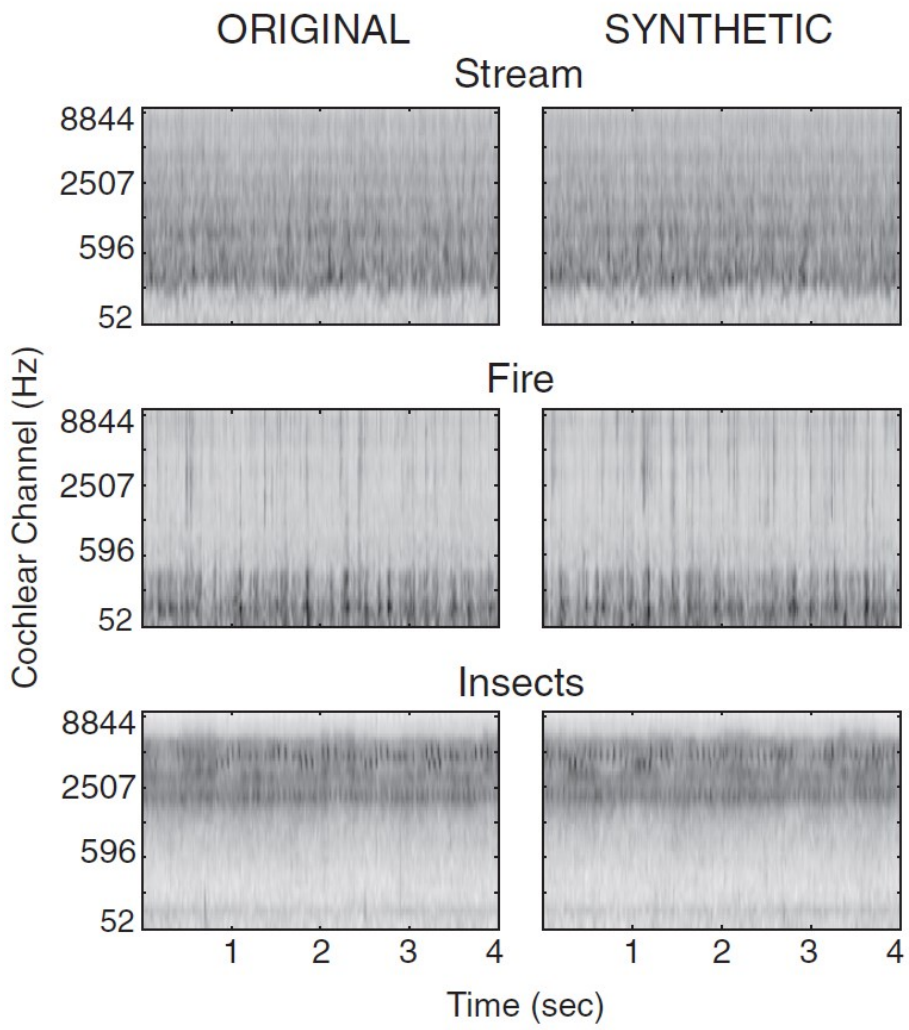

Figure 1.11: Synthetic sound textures.

Spectrograms of three sound textures (stream of water, fire, and insects) are shown with the original spectrogram on the left and the spectrogram of the synthetic texture on the right. Figure reproduced from McDermott and Simoncelli, 2011.

1.4.5 Speech and Animal vocalizations

Traditional speech recognition tasks involve two distinct speech samples spoken by different talkers and have been used in a variety of behavioural (Cherry, 1953), imaging (Ding and Simon, 2012) as well as multi-electrode surface recordings from human auditory cortex (Mesgarani and Chang, 2012) amongst other paradigms. In animal studies, spectrally rich conspecific vocalizations are commonly used, e.g. in tree frogs (Velez and Bee, 2011), zebra finches (Schneider and Woolley, 2013), marmosets (Miller et al., 2010) amongst others.

Speech recognition and speech intelligibility in noisy backgrounds represent a practical problem that affects a significant percentage of the population. It is also of clinical interest as speech recognition in busy and crowded settings becomes worse with aging, hearing loss as well as a number of neurological diseases such as dementia, dyslexia, schizophrenia amongst others. A number of standardized speech intelligibility tests have been developed such as the Modified Rhyme Test (MRT) which consists of a set of fifty six-word lists of rhyming or similar-sound monosyllabic English words where each word is constructed from a consonant-vowel-consonant sound sequence (e.g. *went, sent, bent, tent, rent, dent; hold, cold, gold, fold, told, sold)*. The six words in each list differ only in the initial or final consonant sound and the task of the listener is to identify which of the six words was actually spoken by the talker. A carrier sentence is usually used. The MRT measures errors in discrimination of both the initial as well as final consonant sounds.

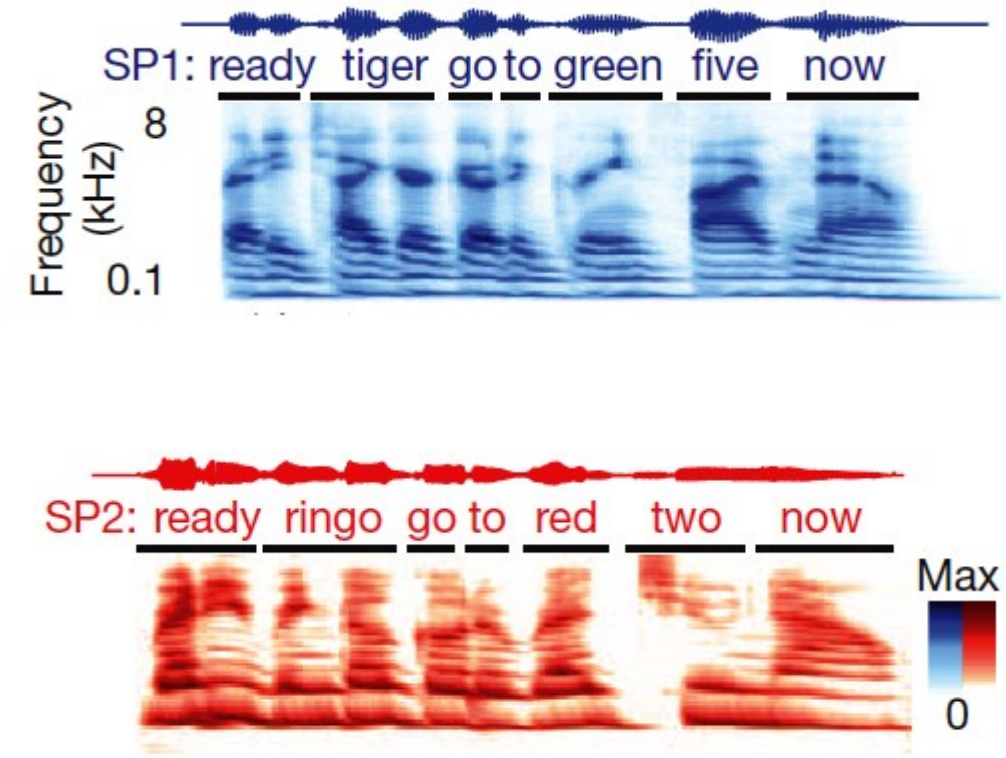

Figure 1.12: Examples of speech stimuli used to study segregation.

Spectrograms of two different CRM speech stimuli spoken by speakers 1 (above) and 2 (below) are shown. The acoustic waveform for each speech stimulus is indicated on the top of each spectrogram. Figure reproduced from Mesgarani and Chang, 2012.

Figure 1.12 illustrates the spectrograms of two speech samples spoken by different speakers. These stimuli are taken from a commonly used speech corpus for multi-talker communication, known as the Coordinate Response Measure (CRM; Moore, 1981; Bolia et al., 2000). The CRM task developed by Bolia and colleagues is an extension of standardized tests such MRT with particular relevance for military environments. It consists of a call sign (e.g. "tiger" in Figure 1.12A) and a colour-number combination (e.g. "red-two" in Figure 1.12B) embedded in a carrier phrase. The listener is assigned a call sign and is required to indicate the colour-number combination spoken by the talker whose speech contained his or her call sign. In the presence of multiple talkers speaking simultaneously with each speaking a different call sign and a different colour-number combination, this task represents a scene analysis problem as the listener must be able to discriminate between his or her call sign from a set of simultaneous call signs. This provides a measure of the listener's ability to selectively attend to a single channel whilst rejecting other irrelevant channels.

1.5 Literature review

This section presents a review of the experimental findings obtained from each stimulus paradigm as described in section 1.4 and is further organized in terms of evidence using different experimental techniques. Each stimulus section covers the results from psychophysics (human and animal behaviour), electrophysiological recordings in animals and direct intra-cortical or surface recordings in humans, non-invasive human functional imaging based on fMRI, MEG and EEG as well as computational

modeling. These approaches are complementary and the aim is to provide a holistic description of the current state of research in auditory scene analysis.

1.5.1 Streaming

The streaming paradigm is the most commonly used tool to study auditory scene analysis. First investigated in the 1970s by Bregman and colleagues (e.g. Bregman and Campbell, 1971), the stimulus has driven a lot of research on the mechanisms of auditory grouping and informed several models of segregation. The following sections briefly describe the results based on this stimulus from a variety of experimental techniques.

1.5.1.1 Psychophysics

Fission and fusion boundaries

Leon van Noorden examined the behaviour of human listeners in response to the repeating triplets (ABA_) of the streaming signal and characterized perceptual boundaries that govern integration, segregation and bistability (van Noorden, 1975). These boundaries are demarcated as shown in figure 1.13. The temporal coherence boundary is a limit beyond which listeners can never hear one stream whilst the fission boundary is limit beyond which listeners can never hear two streams. The area between the two curves represents an ambiguous region where the percept is bistable and switches between that of one or two streams.

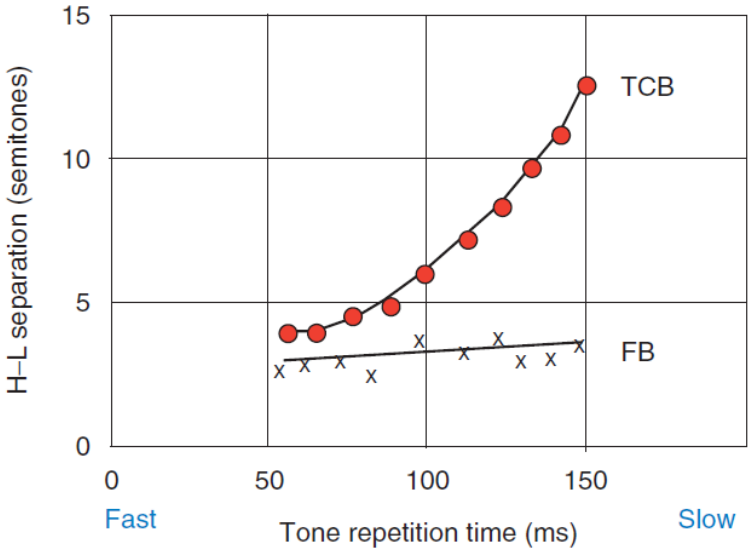

Figure 1.13: Perceptual boundaries in streaming.

The red dots form the temporal coherence boundary whilst the crosses indicate the fission boundary. Figure reproduced from Bregman, 2008.

These data highlight the important roles of frequency separation and tone presentation rate in determining the perceptual state of the listeners. A segregated stream percept is commonly obtained under conditions of high frequency separations and faster speeds of presentation. The effect of presentation rate, however, was negligible when trying to segregate the streams. The temporal coherence boundary however increases markedly with the presentation period and it was shown that the most important temporal factor governing this boundary is the time interval between successive tones of the same frequency rather than the interval between tones of different frequency or the actual duration of the tones (Bregman et al., 2000).

Furthermore, these two perceptual boundaries also reflect different mechanisms of segregation. A single stream percept is susceptible to interference by primitive, bottom-up grouping mechanisms whilst a top-down process that employs selective attention is used when trying to hear separate high and low frequency streams.

Buildup of streaming

Another commonly observed effect in streaming paradigms is the gradual increase in the tendency of listeners to report a segregated percept with repeated presentation of the streaming signal (Bregman, 1978; Anstis and Saida, 1985). Usually, the dominant percept at the beginning of the stimulation is that of a single stream and it takes some time for the percept to break and for listeners to report segregation. The time taken for a

streaming percept to emerge is known as the *buildup time* and is usually on the order of a few seconds. However, it is important to note that the segregation tendency can be partially or completely reset by sudden changes in the properties of the sequence or by switches in attention. The reset due to sudden changes is suggested to reflect the activation of a new sound source and causes the perceptual system to return to its original default state of a single integrated percept (Moore and Gockel, 2012).

Anstis and Saida (1985) attributed the buildup effect of streaming to frequency-shift detectors which integrate successive tones into a single stream. With repetition of the stimulus, these detectors are suggested to habituate and the breakdown of their integrative function results in the formation of two separate streams. Another suggestion put forward by Bregman (1990) is that the default perceptual state is that of one stream and stimulus repetition increases the evidence in favour of two distinct sources of sound that leads to the formation of two different streams.

Bregman's model of auditory streaming

Based on a number of experiments on streaming, Bregman (1990) formulated a model of auditory scene analysis. He postulated the existence of two types of brain mechanisms involved in grouping. The first is a primitive, bottom-up mechanism that is involved in encoding the sensory attributes of the incoming stimuli and grouping them on the basis of Gestalt principles such as continuity, common fate and good continuation amongst others. The second mechanism consists of a set of higher-level processes or

schemas that are usually learned through exposure to the acoustic environment. These schemas allow recognition of patterns in the environment and allow recognition of familiar words, languages, speakers, melodies amongst others. They can operate in conjunction with attention, for instance, when following a specific person's voice in a crowded room.

These ideas have guided auditory scene analysis research over the past few decades but fall short of providing detailed description of the actual physiological mechanisms and the neural substrates involved in each kind of grouping process. Bregman's work has been taken forward in a number of physiological experiments in both humans and animals that have now shed light on the mechanistic bases of auditory streaming. These are discussed in greater detail in the following sections.

Bistability

Another significant facet of the streaming paradigm is the ambiguity in perceptual reports of the listeners for intermediate values of frequency separation and presentation rate. This defines an 'ambiguity region' where the percept often flips between one or two streams (van Noorden, 1975). In figure 1.13, this corresponds to the area between the two perceptual boundary curves and results in alternation of the percept between one or two streams. This is analogous to many visual 'multistable' phenomena where the same stimulus results in ambiguous and mutually incompatible perceptual reports like the Necker cube (Necker, 1832) or Rubin's face/vase illusion (Leopold and Logothetis, 1999; Pressnitzer et al., 2011, 2012).

Bistability in streaming is almost always observed over a wide range of stimulus parameters (Denham and Winkler, 2006; Kashino et al., 2007). The flips, however, do not occur in a regular manner as it has been found that sometimes listeners do not report hearing both percepts at the same time (Pressnitzer and Hupe, 2006) whilst other studies have indicated that listeners are conscious of both streams at the same time (Bendixen et al., 2010). Pressnitzer and Hupe (2006) reported that the distribution of switches in the perceptual states during streaming is similar to that for visual multistability and that the dynamics of visual and auditory switching are almost identical when measured in the same group of listeners. Although the switching occurs on a random basis, it can be influenced by behavioural goals or task instructions.

Attention and streaming

The predominant view of the role of attention in streaming is that it is involved in selection of (and switching between) streams rather than in the process of stream formation itself which is designated as a primitive bottom-up process. In this object-based view of attention in auditory scene analysis, attention operates at the level of objects (or streams) that are already grouped by downstream sensory processing mechanisms. However, this view is contradicted by behavioural findings that paying attention to the high frequency tone for instance results in a much smaller frequency separation for segregation (van Noorden, 1975).

Psychophysically, the major focus of research is on the role of attention in the buildup of streaming. Carlyon and colleagues (2001) examined this by using a dual-task paradigm where the streaming signal is presented to the left ear of the listeners for 21 seconds whose task is to indicate whether they heard one or two streams. In a baseline condition, no sounds were presented to the right ear whilst in the dual-task condition, listeners were required to detect changes in the intensity of white noise presented to the right ear for the first 10 seconds of the sequence, and required to switch their attention to the left ear and the streaming task after these 10 seconds. The control condition presented the same stimuli but the task was based only on the streaming signals in the left ear. The results from the main condition of interest (task-switching) demonstrate that the probability of hearing two streams was significantly reduced after diverting attention from the right to the left ear compared to the control condition. The same effects were found even if attention was distracted using a visual or a numerical task (Carlyon et al., 2003). In a similar vein, Cusack et al. (2004) found that buildup of streaming is reset if attention is briefly diverted away from the streaming signals presented in the left ear. These findings show that the buildup of segregation depends on attention such that the tendency to report streaming is reduced by an absence of attention or switch in attention.

Effect of temporal regularity

A recent line of investigation has focused on the role of temporal regularity as a grouping cue in auditory scene analysis. The classical

streaming stimulus consists of tones presented at regular rates and it is not certain to what extent the regularity of the sequence affects perception. Bendixen and colleagues (2010) assessed the influence of pattern regularity in a streaming paradigm where listeners were required to indicate whether they heard one or two streams. The frequency and intensity of the tones was jittered by a small amount and regular patterns were imposed on these two features in either the A tones or B tones or both. Bistable percepts were reported as usual but it was observed that regular patterns in either the A tones or B tones, or both, increased the mean duration of the two-stream percepts relative to the condition when the patterns were irregular. The duration of single-stream percepts was not affected by this manipulation. The authors concluded that temporal regularities likely recruit central mechanisms that tend to stabilize auditory streams once they have been formed on the basis of primitive low-level mechanisms.

Recent work has further corroborated the role of temporal regularity in stream segregation (Andreou et al., 2011; Rajendran et al., 2013). Andreou and colleagues (2011) found that temporal regularity serves as an effective cue for segregation but its effect is limited to fast presentation rates (4-10 Hz) and low frequency separation (2 semitones) between the two tones of the streaming sequence. Rajendran et al. (2013) also employed a streaming paradigm where a limited amount of temporal jitter was added to the B tones and found that the percentage of trials associated with a two stream percept significantly increased with increase in temporal jitter.

1.5.1.2 Human functional imaging

Early behavioural work on auditory streaming since the 1970s provided a solid theoretical foundation for the assessment of neural correlates of streaming with the advent of modern imaging methods such as EEG, MEG and fMRI. This section provides a brief review of research on stream segregation based on these imaging techniques (Melcher, 2009; Gutschalk and Dykstra, 2013).

ERP evidence

Scalp-recorded ERPs were used to examine Bregman's model of two mechanisms involved in streaming: a bottom-up, pre-attentive grouping mechanism and a higher-order attention-dependent buildup mechanism. Winkler et al. (2005) used streaming signals that contained frequent omissions of the tones at early and late phases of the sequence to analyze whether the topography of the resulting ERPs support Bregman's model. Stimulus parameters were chosen to evoke either an integrated or a bistable percept in two separate conditions and listeners were required to indicate whether they heard one stream or not. A low tone was omitted to elicit deviant responses in both the early and the late phase of the sequence and results indicated early as well as late frontocentral negativity for the deviant responses in the ambiguous condition. The early difference waveform (N1) was elicited only when the listeners heard a single percept whilst the late difference waveform (P2b) was observed only when two streams were reported. Also, the P3a component that is related to attentional switching was evoked, presumably when the listeners heard the ambiguous percept.

Similarly, Snyder and colleagues (2006) observed that auditory evoked potentials, specifically the P2 and N1c in response to the streaming sequence increased in amplitude with increasing frequency separation and correlated with behavioural measures of streaming. Furthermore, a slowly rising positivity was also found through the course of the sequence whose time course varied similarly to the buildup of streaming.

Additionally, Sussman and colleagues combined the MMN and streaming paradigms to demonstrate that deviant stimuli embedded within a high-tone stream resulted in a mismatch response during the perception of two streams (Sussman et al., 1997; 2007). Also, the deviants occurred more at the end of sequence rather than the beginning, in line with the time course of the buildup. Significantly, this pattern of results was noticed whether the listeners attended to the stream or not, suggesting that attention is not required for buildup unlike the previously discussed behavioural studies. This result may be explained by the fact that attention only modulates buildup in the absence of robust segregation cues such as large frequency separations as used in the study (Sussman et al., 2007). Sussman, Winkler and colleagues validated the ERP correlates of streaming using the mismatch paradigm in adults (Winkler et al., 2003a), school-age children (Sussman et al., 2001) as well as newborn infants (Winkler et al., 2003b), thus demonstrating the utility of this approach.

MEG evidence

Gutschalk and coworkers (2005) measured auditory evoked neuromagnetic fields in response to the streaming signal in two separate experiments where the stimulus parameters were chosen to promote either an integrated/segregated percept or a bistable percept respectively. The first experiment revealed that changes in frequency separation and inter-stimulus interval (ISI) affected the magnitude of the auditory evoked fields in a manner that correlated with the degree of perceived stream segregation, i.e., the magnitude of P1m and N1m evoked by the B tones in the repeating triplet increased with larger frequency separations. This trend was also observed in the behavioural data where high correlations were found between the magnitudes of the P1m and N1m evoked fields and the reported ease of streaming. The second experiment, where an ambiguous percept was induced showed similar results to experiment 1: the magnitude of P1m and N1m covaried with the perceptual state and was larger for two vs. one stream percepts. Dipoles were fitted to the two evoked fields and were found to be localized in the non-primary auditory cortex in a majority of the subjects, without any significant lateralization of the sources.

fMRI evidence

Functional MRI has also been adopted to investigate the neural bases of auditory stream segregation. Although the poor temporal resolution of fMRI is too poor to track the fast dynamics of perceptual states during streaming, its high spatial resolution allows the examination of the brain regions that mediate streaming.

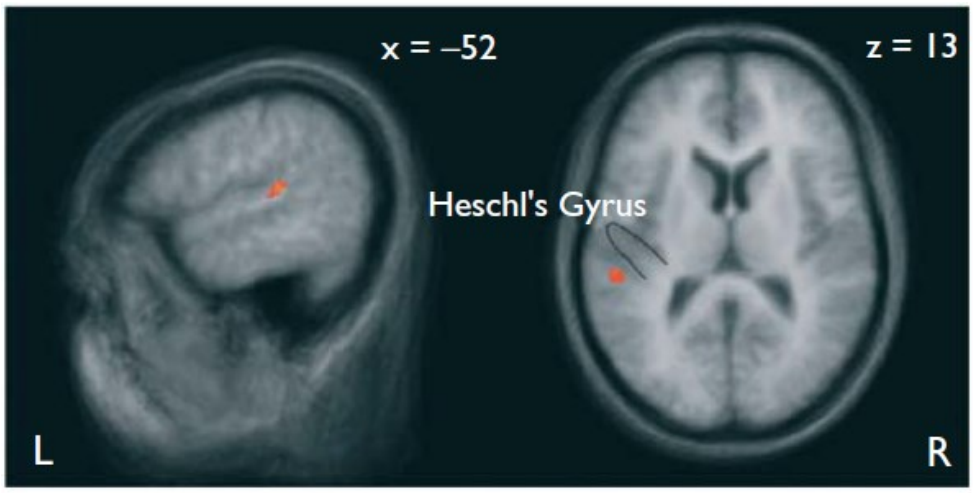

Figure 1.14: BOLD activation in PT for a contrast between 2 vs. 1 streams.

Group average data is displayed on a sagittal (x = -52) and axial (z = 13) section. Location of Heschl's gyrus is indicated in the axial section on the right. Figure reproduced from Deike, 2004.

Deike and colleagues (2004) used a variant of the streaming paradigm that consisted of harmonic tones with alternating spectral envelopes, or timbres (organ-like and trumpet-like). Listeners were instructed to detect low-probability targets that were distributed in either stream. Analysis of BOLD activity for a two stream vs. one stream contrast revealed significant clusters in the left superior temporal gyrus, posterior to the Heschl's gyrus as shown in figure 1.14. Analysis of individual auditory fields showed increased activity in the left posterior fields T2 and T3 for the same contrast. These data suggest that the left auditory cortex is involved in segregation of sounds based on spectral cues.

In another fMRI study, however, Cusack (2005) did not find any differences in activation in the auditory cortex when comparing BOLD activity during the percept of two vs. one stream in the ambiguous condition. Listeners were required to indicate whether they heard one or two streams where the stimulus parameters were modulated to result in a bistable percept. Significant difference between the two conditions was found instead in the parietal cortex, in the intraparietal sulcus (IPS) as shown in figure 1.15. This was the first evidence that areas outside the conventional auditory system may have a role in auditory streaming. Cusack interpreted the IPS activity to reflect attentional switching between the two streams in the bistable state. However, it is not certain whether the IPS activity is a cause or consequence of the perceptual shift from one to two streams (Shamma and Micheyl, 2010). These results, however, make sense in light of the role of IPS in visual binding (Xu and Chun, 2009) and were

further corroborated by Hill and colleagues (2011) who showed an effect of perceptual state during switching in the IPS.

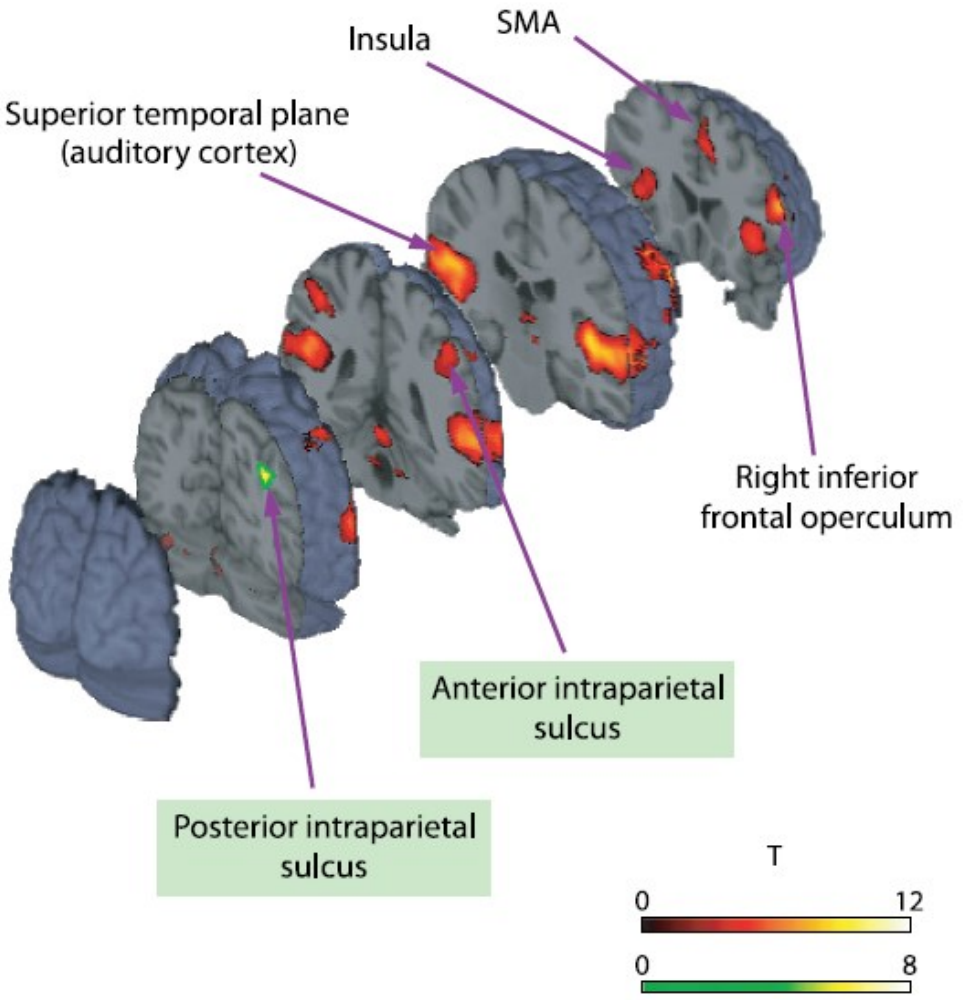

Figure 1.15: Intraparietal sulcus activation for a contrast of two vs. one streams.

Regions activated when sound was presented and task performed, relative to silence (red–yellow–white colours). When the percept was of 2 streams rather than 1, a right posterior IPS region was activated in the whole-brain analysis (green–yellow–white) and an ROI analysis found activity in the anterior IPS (as indicated by the light green shading). Figure reproduced from Cusack, 2005.

In another fMRI experiment on streaming, Wilson and coworkers (2007) scanned listeners while they reported their perception of sequences of alternating-frequency tone bursts separated by 0, 1/8, 1, or 20 semitones. They observed that at the null and small frequency separations, the sequences were heard as one stream with a perceived rate equal to the physical tone presentation rate. The corresponding BOLD activity was measured in the auditory cortex and found to be phasic in nature, with significant peaks at the onset and offset of the sequences. However, at larger frequency separations, BOLD activity related to the two segregated streams was more sustained and larger in magnitude. These results are consistent with an interpretation that the modulation of fMRI activity as a function of frequency separation mediates the encoding of simultaneous changes in perceived rate and the perceptual organization of the sequences into auditory streams.

Kondo and Kashino (2009) used an event-related fMRI design to probe the temporal dynamics of brain activity as a function of the direction of perceptual reversals, i.e. from one to two-stream percept and two to one-stream percept. They used different frequency separations and found that irrespective of the magnitude of the spectral separation, activations in the MGB and PAC were correlated with individual differences in perceptual predominance in streaming. This was determined by computing the correlation between the proportion of single-stream predominant durations and temporal precedence of BOLD activity in a region-of-interest analysis based on the MGB and PAC. The direction of the switches affected the BOLD activity in the MGB and the PAC asymmetrically: MGB was

activated earlier during switching from a non-predominant to predominant percept whilst PAC was activated earlier during switching in the opposite direction from a predominant to non-predominant percept. These data provide crucial evidence supporting the role of feedfoward and feedback pathways between the MGB and PAC for perceptual formation during streaming. In a subsequent study based on a similar event-related paradigm, Kondo and Kashino (2012) confirmed the role of the MGB and PAC for perceptual switching during streaming as well as verbal transformations during a repeated word presentation task. On the contrary, Schadwinkel and Gutschalk (2011) found that PAC as well as the early auditory processing centres in the inferior colliculus (IC) are also involved during perceptual reversals in streaming, albeit, here segregation was achieved on the basis of differences in interaural time differences (ITD).

1.5.1.3 Human neurophysiology

Bidet-Caulet et al. (2007) performed direct recordings from human auditory cortex during the streaming task. Depth electrodes were inserted into the temporal cortex of epileptic patients who were presented with stimuli whose onset asynchrony was manipulated to induce either streaming or grouping. They recorded electrophysiological responses to acoustically identical stimuli that corresponded to different percepts of one or two streams and found that transient and steady-state evoked responses as well as induced gamma band oscillations are larger for onset synchrony of the two concurrent sounds than in the case of onset asynchrony. Transient evoked responses were first elicited 60ms following sound onset in the

posterior lateral STG and spread over PT and lateral STG until 200ms. Next, induced gamma oscillations were modulated in nearby regions until 300ms and finally, steady-state responses were evoked for several hundreds of milliseconds in PAC as well as the anterolateral part of the HG. These results offer a direct insight into the neurophysiological mechanisms in play during auditory perceptual organization (Bidet-Caulet and Bertrand, 2009).

A more recent study investigated the neural correlates of auditory streaming by using intracranial EEG and recording from electrodes placed over the temporal, frontal and parietal cortex (Dykstra et al., 2011). Dykstra and colleagues found a number of areas spread across the superior temporal and peri-rolandic cortex, middle temporal gyrus as well as the inferior and middle frontal gyrus to be involved in auditory streaming, thus adding to the accumulating evidence in favor of a role for higher order non-auditory areas in auditory scene analysis.

1.5.1.4 Animal electrophysiology

Direct recording of single units and neuronal ensembles from animal models has provided the basis for several models of auditory stream segregation. Scene analysis has been studied in a number of species including macaques, ferrets, zebra finches, rats, bats, frogs as well as fish (Fishman and Steinschneider, 2010a). This section presents a brief review of the most significant findings from electrophysiological recordings in animal models.

Macaques

Fishman and colleagues (2001) performed seminal experiments in awake macaques using neuronal ensemble techniques (multiunit activity and current source density) to elucidate the cortical basis of the streaming phenomena. They investigated the nature of responses elicited by an alternating ABAB sequence as a function of the presentation rate. The A tones corresponded to the best frequency (BF) of the cortical region in A1 while the B tones were separated from this site by a certain frequency separation. They observed that at slow presentation rates, A and B tone evoked responses were generated at the stimulus presentation rate, thus suggesting that a single stream was perceived at these rates. However, at fast presentation rates, the B tone responses were found to be differentially suppressed and the A tone responses occurred predominantly at half the presentation rate, consistent with responses to a segregated stream percept. Furthermore, the magnitude of the suppression of the B tones increased with greater frequency separation. The authors suggested that the differential suppression of the BF and non-BF tones may be due to forward masking (Calford and Semple, 1995; Brosch and Schreiner, 1997) of B tones and that this suppression increases with frequency separation. It was also found that BF tones cause greater suppression of subsequent tones than non-BF tones.

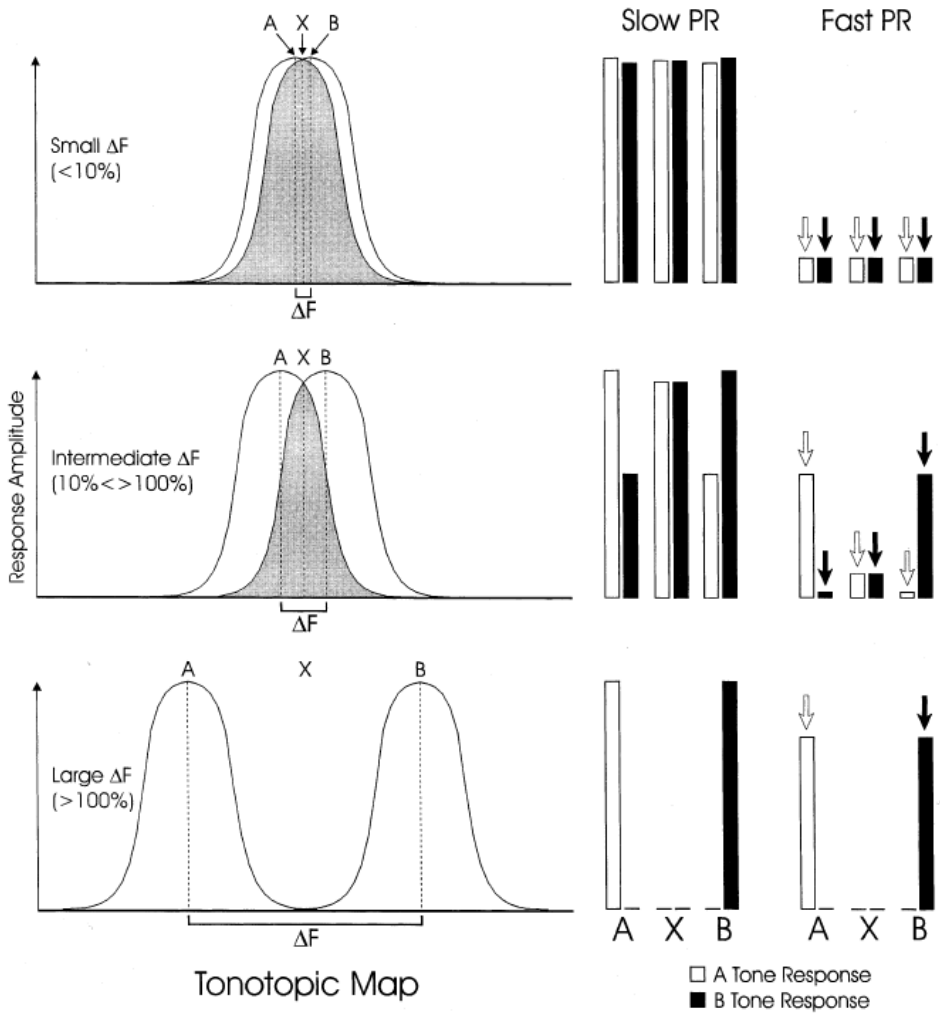
A X B
Slow PR
Fast PR
Small ΔF
(<10%)
ΔF
A X B
Response Amplitude
Intermediate ΔF
(10%<>100%)
ΔF
A
X
B
Large ΔF
(>100%)
ΔF
Tonotopic Map
A X B
A X B
A Tone Response
B Tone Response

Figure 1.16: Model of neural stream segregation in PAC (Fishman et al., 2001).

Bell-shaped curves labeled `A' and `B' represent spatial activity patterns evoked by `A' tones and `B' tones, respectively, along the tonotopic map. The region in between the `A' tone and `B' tone tonotopic locations is labeled `X'. Shaded regions represent locations where activity patterns generated by the tones overlap. Spatial distributions of activity under three different ΔF conditions are depicted (small, intermediate, and large). Hypothetical `A' tone and `B' tone response amplitudes at tonotopic locations `A', `B', and `X', marked by the dashed vertical lines, are represented by white and black bars shown in the right half of the figure under slow and fast PR conditions. Bar height is proportional to response amplitude. Under the intermediate ΔF condition at slow PRs, the overall evoked activity is relatively evenly distributed across tonotopic space. At fast PRs, non-BF tones are differentially suppressed in locations `A' and `B', while in regions equally responsive to both tones (`X'), amplitudes of responses to both tones are equally diminished. This results in the formation of spatially discrete foci of activity along the tonotopic map to which attention can be subsequently directed. Figure reproduced from Fishman et al., 2001.

Based on these results, the authors proposed a model of neural stream segregation in A1 which is presented in figure 1.16. This model is based on the fact that responses of the A and B tones are localized in small circumscribed areas along the tonotopic map at a location determined by their respective frequencies. They proposed that with increasing frequency separation, the responses of the A and B tones become spatially segregated along the tonotopic map and this corresponds to the percept of two separate streams as observed in human psychophysical experiments. The responses are also modulated as a function of the presentation rate where there is greater suppression of the responses at faster rates due to adaptation and forward masking.

Thus, the model emphasizes the role of frequency selectivity (responses to A and B tones are distinct and peak at corresponding tonotopic locations), forward masking (suppression of tones due to the preceding tone which is stronger for preceding BF rather than non BF tones) and adaptation (decrease in responses due to repeated stimulation; Fishman and Steinschneider, 2010a; Micheyl et al., 2007a). However, the link between these responses and perceptual state is indirect as no behavioural measurements were carried out. Another limitation of the model is that it is based only on A1 and does not inform if the same holds true for non-primary auditory cortical areas. Furthermore, the model falls short of explaining streaming of complex sounds with overlapping spectra. Izumi (2002) also showed that Japanese monkeys can discriminate between tone sequences based on frequency separation. Further experiments revealed that

increasing tone duration enhances the differential suppression of non-BF B tones (Fishman et al., 2004) in a way that resembles the findings from behavioural experiments (Beauvois, 1998; Bregman et al., 2000).

Another seminal study was carried out by Micheyl and colleagues (2005) who recorded single unit responses in the primary auditory cortex of awake rhesus monkeys in response to the streaming sequence. Using spike-count measures, they showed that the spiking data correspond well with human behavioural findings and mirrors the buildup of segregation with time as well as the effects of frequency separation and presentation rate. The major aspect of this work is the proposal of a model based on statistical variability of the neural responses to predict the probability of perceptual judgments. The central idea of the model is that evoked responses in PAC are "read out" by other neurons that act as binary classifiers and assume one of two possible states that correspond to the percept of one or two streams depending on the inputs received from the neurons in PAC. This classification is predicted to be based on measures of spike counts evoked by the A and B tones in a streaming triplet. If the number of spikes evoked by both tones exceeds a fixed threshold, a single stream response is generated and if the number of spikes evoked by only one of the tones exceeds the threshold, a two stream response is produced. The value of the threshold was determined on the basis of maximizing the fit between the data and the model predictions and did not depend on the frequency separation. The model predictions vs. the psychophysical findings are shown in figure 1.17.

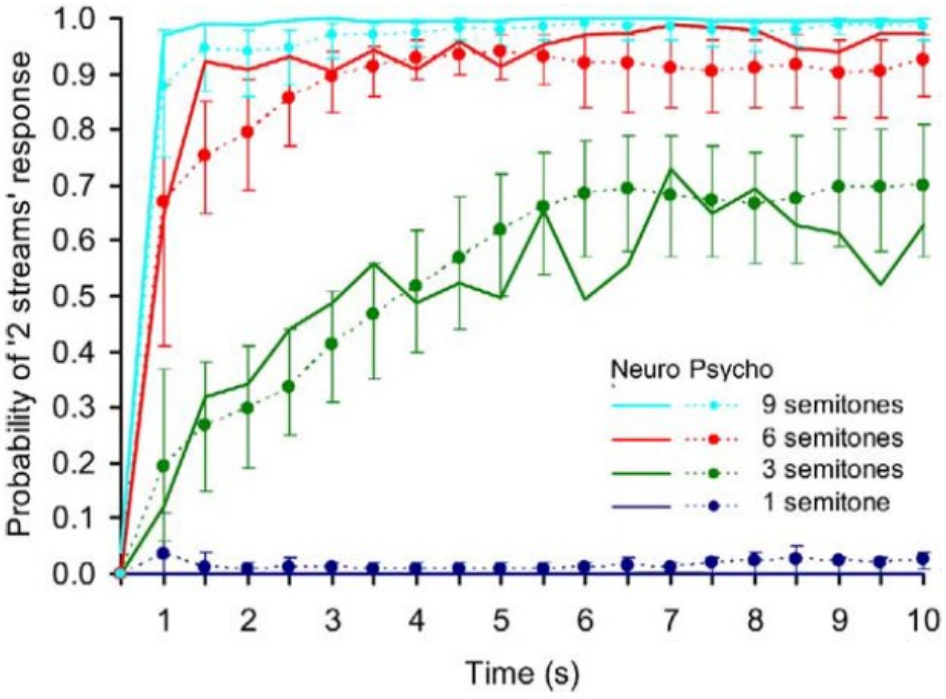

Figure 1.17: Comparison between psychometric and neurometric functions.

The psychometric functions are plotted here as dashed lines, to facilitate comparison with the neurometric functions, which are shown as solid lines. The error bars indicate 95% confidence intervals around the mean proportions estimated using statistical bootstrap. Figure reproduced from Micheyl et al., 2005.

Songbirds

The European starling, a species of songbird has also been shown to exhibit auditory stream segregation for synthetic pure tone sequences as well as discriminate between excerpts of its own song and songs from other avian species (MacDougall-Shackleton et al., 1998). More convincing evidence was presented by Bee and Klump (2004, 2005) who performed careful experiments similar to the studies in monkeys (Fishman et al., 2001; 2004) and evaluated neuronal responses in awake songbirds in response to the streaming sequence as a function of frequency separation and tone presentation time. Their data replicated the findings of Fishman and colleagues (2001, 2004) and consolidated the role of frequency selectivity and forward masking in sequential stream segregation.

Other species

Auditory scene analysis has been investigated in other species as well, including goldfish (Fay, 1998), bats (Kanwal et al., 2003), ferrets (Elhilali et al., 2009a), guinea pigs (Pressnitzer et al., 2008), tree frogs (Velez and Bee, 2011) amongst others. These studies generally reported findings that are congruent with the previously discussed literature. However, Pressnitzer et al. (2008) showed that single unit responses in the cochlear nucleus of the guinea pig exhibit frequency selectivity and forward suppression thus demonstrating that these features of streaming may already be active at the level of the peripheral auditory system (Hartmann and Johnson, 1991; Beauvois and Meddis, 1991, 1996; Denham and McCabe,

1997). Another significant finding was demonstrated by Elhilali and colleagues (2009a) who manipulated the streaming sequence and presented the two tones A and B synchronously while varying the frequency difference between the two. Psychophysical results from humans indicated that the resultant percept is generally of a single stream irrespective of the magnitude of frequency separation. This result is in contrast with the standard results that suggest that large frequency separations promote a segregated percept. On the basis of this finding, Elhilali et al. (2009a) presented this synchronous sequence to ferrets while recording from their primary auditory cortex and found that the neural responses also follow a similar pattern. These results led to the proposal of a model of scene analysis based on "temporal coherence" which is discussed in greater detail in chapter 4.

1.5.1.5 Computational Models

Early theoretical models of auditory stream segregation focused on peripheral processing (Hartmann and Johnson, 1991) to explain the findings from behavioural experiments. Beauvois and Meddis (1991) developed a computer model of streaming that was based on a few principles: (i) a peripheral spectral analysis feeding channels that are characterized by a bandpass frequency response; (ii) inherent "noise" in the system; (iii) a "leaky integration" principle that allows excitatory activity to slowly build up in the channels and decay slowly with time; and (iv) an attentional mechanism that selectively responds to the channel with the maximum activity.

In the model, each channel is modeled by two separate pathways. The excitation-level path computes the overall, smoothed excitation level in the channel whilst the filtered-signal path carries the unsmoothed filtered signal for later calculation at the output. These two pathways reflect the segregation of the auditory pathway at the level of the anteroventral and posteroventral cochlear nucleus respectively. They conducted a number of behavioural experiments and compared these results to the predictions of the model.

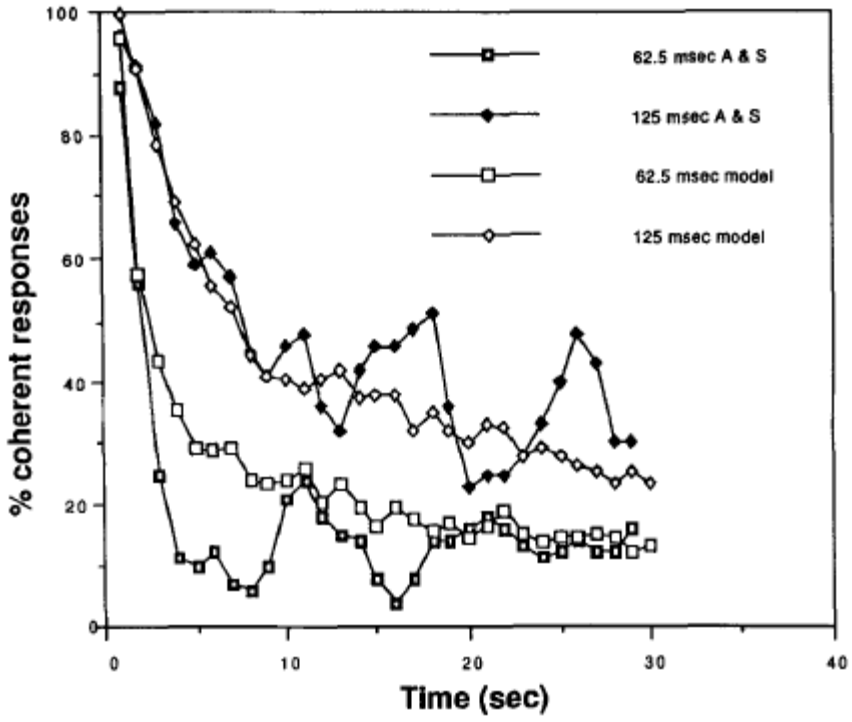

Figure 1.18: Comparison of the output of the Beauvois and Meddis (1991) model with the results of Anstis and Saida (1985).

The percentage of coherent responses in a streaming paradigm is indicated as a function of time. These results show the concordance between the simulations based on the model of Beauvois and Meddis (1991) with the experimental results obtained by Anstis and Saids (1985). Figure reproduced from Beauvois and Meddis (1991).

Figure 1.18 shows the predictions of the model for a particular simulation where the buildup of streaming was examined vis-à-vis the results of Anstis and Saida (1985). The input comprised of 30 seconds long alternating ABAB sequences where the two frequencies were 800 Hz (A) and 1200 Hz (B) and the tone repetition time was either 62.5 or 125ms. The model was able to replicate the buildup of streaming as reported by Anstis and Saida (1985). Beauvois and Meddis (1996) refined their model to include adaptation of the auditory nerve responses and demonstrated that the model can capture buildup of streaming as well as the temporal coherence and fission boundaries.

McCabe and Denham (1997) developed a model of streaming with an aim to simulate the functional properties of auditory processing during stream formation. It represents an advance over the model of Beauvois and Meddis (1991, 1996) in that inhibitory feedback is incorporated here to achieve a graded inhibition rather than arbitrarily suppressing the output of non-dominant channels by half. Furthermore, a background stream is also incorporated to capture the output for both the dominant channel and residual activity as shown in figure 1.20. The model consists of two sets of interacting neurons that comprise the foreground (F) and the background (B) with symmetrical connectivity structure. The foreground array however receives inhibitory input reflecting the background activity that suppresses responses in F where B is currently active, and the inverse of the foreground activity which suppresses responses in the channels where F was previously least active. Similarly, B receives inputs from F and the inverse of B and

competitive interactions between these channels results in graded inhibition. The behaviour of the model was shown to be consistent with a number of established psychophysical findings such as the effect of frequency separation, presentation rate and buildup of streaming. The physiological basis of the proposed circuitry is unclear but appears to be consistent with temporal processing of signals in the cortical rather than peripheral areas contrary to the models proposed by Beauvois and Meddis (1991, 1996).

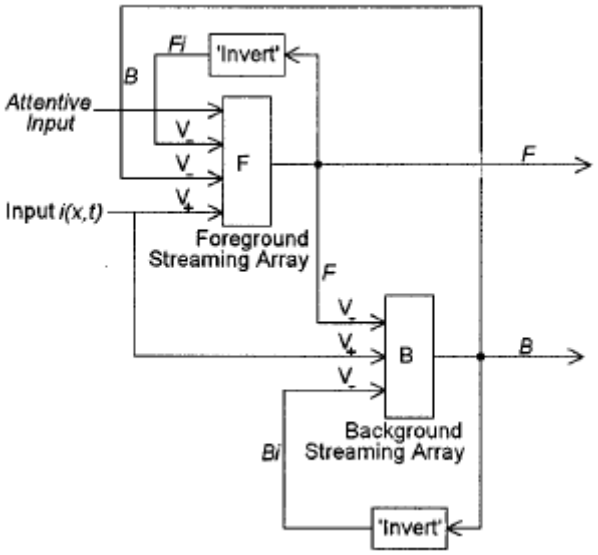

Figure 1.19: Model of auditory streaming by McCabe and Denham, 1997.

Model diagram showing the connectivity patterns between the foreground and background streaming arrays in a physiological model of streaming developed by McCabe and Denham (1997). Figure reproduced from McCabe and Denham, 1997.

Peripheral models of auditory stream segregation, however, are not able to explain aspects of streaming such as bistability. Results of studies based on animal and human neurophysiology further indicate that the primary auditory cortex is a key substrate for streaming. Denham and Winkler (2006) proposed a new model of streaming based on generative models (Friston, 2005). In such a framework, information processing is considered to operate at different levels in the cortical hierarchy with higher level areas passing predictions or expectations to lower level sensory areas involved in generating prediction errors based on the top-down predictions and the actual sensory input. Denham and Winkler (2006) suggested that the auditory system generates predictive models of the acoustic environment that compete which each other and form the bases of auditory perception. The model was based on four key processes: i) initial segregation based on primitive bottom-up cues as discussed in section 1.2.1; ii) predictive modeling that includes creation of alternate models of the acoustic input at different hierarchical processing levels; iii) competition between different (mutually exclusive, e.g. in case of bistable streaming signals) models of the input with a role for attention in the selection or biasing of such perceptual rivalry; and iv) neural adaptation that serves to reduce the inhibition of alternative models, eventually leading to the emergence of alternative perceptual states.

The role of predictive coding models in explaining various aspects of sensory analysis has since received wider attention (Friston and Keibel, 2009; Friston, 2010; Winkler et al., 2009, 2012; Bastos et al., 2012) and has

been successfully employed to explain the generation of the MMN response (Garrido et al., 2009) and pitch perception (Kumar et al., 2011), for instance. Recently, Mill and colleagues (2013) refined the predictive coding model of Denham and Winkler (2006) and provided a computational account of auditory perceptual organization that is based on competition between predictable representations of the sensory world. This predictive model (Mill et al., 2013) successfully replicated a number of phenomena related to streaming such as the emergence of, and switching between, one or two stream percepts; the influence of stimulus manipulations on perceptual dominance (Kondo and Kashino, 2009), rate of switching and phase durations of perceptual states; as well as the buildup of auditory streaming.

Finally, there are a number of other computational models of auditory scene analysis as well that are based on other principles such as neural networks (ARTSTREAM; Grossberg et al., 2004); synchrony between neural oscillations (Wang and Chang, 2008) and temporal coherence (Elhilali et al., 2009a; Shamma et al., 2011). The temporal coherence model is presented in greater detail in chapter 4.

1.5.2 Mismatch negativity

The mismatch negativity (MMN) is a differential ERP that is elicited when an oddball (deviant) stimulus is presented in a train of frequently repeating standard stimuli. It can be elicited by introducing a violation in a variety of acoustic features such as pitch, intensity, presentation rate, spatial location as well as by deviance from complex spectrotemporal rules as well as in other patterns of complex sequences such as speech (see Pulvermuller

and Shtyrov, 2006) and music (e.g. Tervaniemi et al., 2001). It has been widely used in basic and clinical (e.g. Leff et al., 2009; Schofield et al., 2009; Teki et al., 2013) research and has successfully revealed several facets of auditory processing, attention and memory.

The interpretation of MMN is still under debate but is generally considered to reflect an automatic, pre-attentive response that helps in detection of novel sound sources and segregation of the acoustic scene, even for task-irrelevant streams (Winkler et al., 2003c; Sussman, 2005; Sussman et al., 2005). One interpretation suggests that the MMN represents a sensory memory-mismatch trace (Näätänen et al., 1978; Näätänen, 1992). Although this view is still accepted, an emerging view links MMN with predictive coding (Friston, 2005): it reflects a process that updates the representations of detected regularities whose prediction is violated by the acoustic input (Winkler, 2007). Predictive coding models of MMN explain the generation of the MMN response as a generative process based on interactions between the different levels of a hierarchical network based in primary (generates bottom-up prediction errors) and secondary (generates top-down predictions) auditory cortices respectively (Garrido et al., 2009)

The MMN is believed to be a pre-attentive process as it can be evoked even in sleep, anesthesia or even minimal states of consciousness (e.g. Boly et al., 2011), and under certain conditions it can also be modulated by attention (Alain and Woods, 1997; Arnott and Alain, 2002; Sussman et al., 2003). In active paradigms, the MMN can also occur in conjunction with later ERP components linked with focused attention like

the N2b (~200-300ms after stimulus onset) and the P3b (~300-350ms after stimulus onset) which may be generated by sources in the anterior cingulate and prefrontal cortices (Crottaz-Herbette and Menon, 2006). The late ERP components can be used to index whether listeners actually attended to the sounds or not.

The neural architecture of the MMN response includes the primary auditory cortex, cortical areas in the PT and neighboring posterior STG and ventrolateral prefrontal cortex (Opitz et al., 2002; Schonweisner et al., 2007). These areas are argued to comprise a hierarchical network where PAC is involved in detection of acoustic changes, the secondary auditory areas mediate higher-order feature analysis, and the prefrontal cortex mediates attentional gating for salient changes (Schönweisner et al., 2007). More recently, this hypothesis has been incorporated in interacting predictive coding models of brain function where lower-level sensory areas are hypothesized to encode prediction errors whilst higher-level areas convey prediction signals to the lower-level areas (Friston, 2005; Garrido et al., 2009).

Investigations of the MMN response have also been performed in animal models and emphasize stimulus-specific adaptation (SSA) in PAC as a possible neuronal mechanism underlying acoustic change detection (Ulanovsky et al., 2003, 2004). SSA has been observed at the level of the cortex (Taaseh et al., 2011), thalamus (Anderson et al., 2009; Antunes et al., 2010) and the inferior colliculus but not in the cochlear nucleus (Ayala et al., 2012).

1.5.3 Informational masking

Informational masking refers to a type of masking that is distinct from energetic masking (EM). It is often referred to as non-energetic masking, where energetic masking is defined as masking that results from competition between target and masker at the level of the auditory periphery, i.e., overlapping excitation patterns in the cochlea or auditory nerve. On the other hand, IM represents a form of "perceptual" masking that is associated with an increase in detection thresholds due to stimulus uncertainty and target-masker similarity (Pollack, 1975; Leek et al., 1991; Durlach et al., 2003). Thus, EM is often associated with peripheral and IM with central masking.

Several paradigms have successfully exploited IM to study auditory perceptual behaviour. For instance, a common task requires the listener to detect a tonal target in the presence of simultaneous multi-tone maskers. The target is commonly a fixed-frequency tone and the masker is a complex of many tones selected randomly on each presentation that are constrained to lie outside a "spectral protection" region around the target (e.g. Kidd et al., 1994, Gutschalk et al., 2008; Elhilali et al., 2009a). The protection region is employed to minimize the effects of EM. The listeners are generally distracted by the masker and find it difficult to detect the target even though there is little masker energy around the target. Their performance can be improved, however, by a variety of procedures such as careful instructions, target cueing, practice, or by reducing the similarity between the target and the masker (e.g. Neff and Green, 1987; Kidd et al., 1994). More precisely,

the similarity between the target and the combination of target and masker needs to be manipulated so that it becomes easy to hear out the target as perceptually distinct from the masker (Durlach et al., 2003). This dimension of similarity-dissimilarity is closely related to the distinction between grouping and segregation in auditory scene analysis.

A number of psychophysical experiments have been carried out based on the IM stimulus and the results demonstrate that the detection of the target depends on the width of the spectral protection region and the density of the maskers (Kidd et al., 1994, 1995, 2003, 2011; Micheyl et al., 2007b; Gutschalk et al., 2008; Elhilali et al., 2009b). The bases of target detection in IM stimuli is predicted to rely on the same adaptation-based mechanisms as proposed for streaming sequences (Micheyl et al., 2007b). Some of the IM experiments carried out in humans and animals are discussed in detail below.

Gutschalk and colleagues (2008) devised a task where listeners were required to detect a stream of regularly repeating tones amidst a background of masking tones that were randomly organized in frequency and time. This stimulus is similar to those used by Neff and Green (1987) and Kidd et al. (2003). They measured brain activity using MEG and analyzed evoked fields in response to the perceptually detected and undetected target tones in the auditory cortex. They uncovered a response in PAC that was only present for the detected targets at a latency of 50-250ms as shown in figure 1.20. This response was termed as ‘awareness related negativity’ (ARN) that reflects conscious sound perception in the auditory cortex.

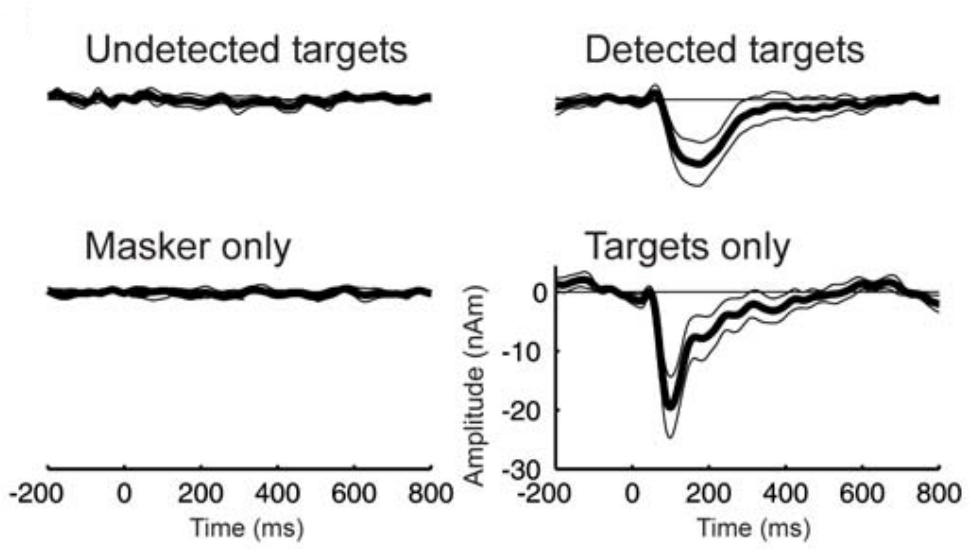

Figure 1.20: MEG source waveforms in response to the targets and maskers in an IM paradigm.

MEG source waveforms are shown here that are averaged over the different SOA conditions, hemispheres and listeners. There is essentially no response to maskers or undetected targets as shown on the left. There is a significant negativity in response to detected targets that is termed as the awareness related negativity. Figure reproduced from Gutschalk et al., 2008.

However, the source analysis based on dipole fitting cannot conclusively rule out the involvement of secondary auditory areas. In a subsequent experiment, Wiegand and Gutschalk (2012) used fMRI along with MEG to probe the neural substrates of the ARN in a similar paradigm. They found significantly stronger BOLD activity for detected vs. undetected targets in the core auditory cortex, with prominent activations in the medial part of the Heschl's gyrus.

In another IM experiment using MEG, Elhilali and colleagues (2009b) used a variant of the IM stimulus to examine the influence of listener's attentional state on the neural responses. Listeners had to perform two tasks: one based on the target tones where they were required to detect a frequency deviant in the repeating target sequence; and another complementary task based on the masker tones where they needed to detect an increase in the length of the maskers. Thus, with the same physical stimulation, the authors manipulated the attentional state of the listeners which was focused on different components in the acoustic scene. The MEG data revealed that attention strongly modulates the steady-state neural representation of the target stream and boosted the perception of the foreground signal. This effect was found to be mediated by the auditory cortex exclusively at the rate of presentation of the target stream (4 Hz) with a resolution of a fraction of a Hertz. The attentional enhancement was accompanied by an increase in coherence over distant channels reflecting an increase in neural synchronization.

Although this paradigm has several advantages over the streaming signal and revealed correlates of auditory perceptual awareness, it has a few limitations. The spectral protection region offers a cue to the listeners who can potentially solve the task by attending only to the energy in the limited frequency band surrounding the target. Although the stimuli are spectrally rich and span a broad frequency range, they are unlike natural sounds which consist of overlapping frequencies without any protection region around the signal of interest.

1.5.4 Speech

Speech and conspecific vocalizations represent natural communication signals with rich spectrotemporal content. Speech is an ideal stimulus to use in human studies on the cocktail party problem (Cherry, 1953; Billig et al., 2013). However, it is also contaminated with strong semantic content that may specifically invoke top-down processes and is not amenable to careful control of its spectrotemporal properties unlike the previously discussed signals. Sequences of tones as used in streaming, MMN and IM experiments are useful in probing low-level aspects of primitive stream segregation with parameterized control over stimulus features whilst speech offers an ecologically valid route to access attentional mechanisms involved in streaming. A number of psychophysical, imaging as well as electrophysiological experiments using speech and vocalizations in animal models have been conducted. It is beyond the scope of the present thesis to cover all bases and only a few studies that highlight brain bases of speech segregation are discussed below.

A major theory holds that speech perception relies on entrainment of cortical activity to multiple time scales in the speech signal that enable parsing of the input into different units of speech representation characterized by different frequencies, such as phonemes and syllables (Giraud et al., 2007; Lakatos et al., 2008; Giraud and Poeppel, 2012). This results in nesting of neural activity across multiple frequency bands that represents a general neural code for sensory perception. In the framework of a cocktail party scenario, selective entrainment to the attended speaker is important to track his or her speech over time. Several studies have explored the question of attentional control in the context of a multi-talker environment using techniques with high temporal resolution such as EEG, MEG and intracranial EEG to track the precise temporal dynamics of speech (Lee et al., 2013).

Luo and Poeppel (2007) measured MEG responses while listeners were listening to spoken sentences and analyzed the phase tracking dynamics. They found that the phase pattern of theta band (4-8 Hz) responses in auditory cortex reliably discriminates spoken sentences. This theta bandwidth is strongly represented in the speech envelope and is critical for accurate speech comprehension. Furthermore, the tracking ability was found to be correlated with speech intelligibility, i.e. theta phase tracking became less robust with decreased speech intelligibility. They suggested that a temporal window corresponding to the theta range (~200ms) segments the input speech signal and may be involved in processing syllables (mean duration of ~ 200ms). In a similar vein, Kerlin and colleagues (2010) also

found that selective attention enhances the discrimination of attended speech in auditory cortex in a frequency range from 4-8 Hz. Additionally, they demonstrated that a difference in alpha power (8-12 Hz) at parietal sites across hemispheres could predict the direction of auditory attention to speech. This is consistent with a role of the posterior parietal cortex in auditory spatial attention (Fritz et al., 2007).

Ding and Simon (2012) asked listeners in an MEG experiment to attend to one of two speakers while they manipulated the relative intensity between the attended and the background speakers. They analyzed phase-locked neural activity for any evidence of selective synchronization to the speech of the attended talker. Using a linear decoder, they found that the decoded envelope significantly correlated with the envelope of the attended speech. This correlation was insensitive to the intensity of the target speech as well as the relative intensity between the target and the masker. They further constructed a spectrotemporal receptive field (STRF) for each MEG sensor and examined the auditory evoked responses, M50 and M100. The M100 was localized in the secondary auditory cortex and was found to be stronger for the attended vs. the background speech unlike for the M50 response. Furthermore, both the evoked responses were insensitive to the intensity of the attended or the background streams suggesting that a robust object-based representation of the attended speaker was formed.

Mesgarani and Chang (2012) performed a similar experiment using multi-electrode surface recordings from the human auditory cortex. They showed that it is possible to reconstruct both the attended and the ignored

speech signal from the time course of high-gamma power and that the attended signal was more reliably reconstructed than the ignored one. The spectrogram obtained from the reconstruction of a single speaker was found to be remarkably similar to the spectrogram derived from the mixture of two speech signals when the same speaker was attended to. They also demonstrated that it is possible to decode both the attended words and speaker identity as well.

A similar experiment based on direct cortical recordings was conducted by Zion-Golumbic et al. (2013a) who examined both low frequency and high gamma neural representations of attended speech signals. They found that both low frequency phase and high gamma power concurrently track the envelope of attended speech and suggest that tracking in these two bands may represent separate neuronal mechanisms for speech perception. Attention modulated the perceptual representation in the auditory cortex by enhancing the tracking of the attended speech stream, although the ignored speech stream remained represented. In higher-order cortical areas, more selective representation of the attended speaker was observed but without any faithful representation of the ignored speech. Significantly, this selectivity evolved and became stronger with time. In a related experiment, the same authors demonstrated that vision can enhance the selective auditory cortical tracking of the attended speaker (Zion-Golumbic et al., 2013b). Visual cues represent a potent cue as they arrive before the corresponding acoustic signal and may serve to direct attentional

resources at precise moments in time when the speech signal is predicted to arrive.

1.5.5 Complex non-speech stimuli

More recently, complex acoustic signals have been used to examine the cocktail party problem (Nelken, 2004). These low-level stochastic signals are designed to simulate complex acoustic scenes that we are exposed to in our everyday lives. Such signals allow flexible parametric control over the acoustic properties of the stimulus that are based on stochastic variations in spectrotemporal space (see Figure 1.10; Overath et al., 2010) or are based on models of auditory perception that capture the statistics of stationary sounds in the environment (see Figure 1.11; McDermott and Simoncelli, 2011).

Overath et al. (2010) addressed a fundamental question of the formation and representation of an auditory object that is an essential prerequisite for subsequent segregation. They argued that from first principles, analysis of objects requires two fundamental perceptual processes (Griffiths and Warren, 2004; Griffiths et al., 2012; Bizley and Cohen, 2013). The first mechanism is required to detect boundaries between objects and is based on identifying variations in the statistical properties of individual objects at the edges in spectrotemporal space (Kubovy and Van Valkenburg, 2001; Chait et al., 2007, 2008). The second mechanism is required for invariant representation and maintenance of the segregated object (Griffiths and Warren, 2004). Although previous studies have investigated cortical bases of auditory edge detection (Chait et al., 2007,

2008) these did not address mechanisms pertaining to perceptual representation of the object. Here, the authors developed a novel stimulus based on spectrotemporal coherence to create objects and boundaries between them. This "acoustic texture" stimulus is conceptually similar to the visual coherent dot motion paradigm (Shadlen and Newsome, 1996) and comprised randomly distributed linear frequency-modulated ramps with different trajectories. The coherence between these ramps was manipulated to create different auditory objects and the transitions between ramps with different coherence represented boundaries between these objects.

Using a parametric fMRI design, they found that activity in the Heschl's gyrus, PT, temporo-parietal junction (TPJ), and superior temporal sulcus (STS) increased as a function of increasing change in spectrotemporal coherence at texture boundaries. For the representation of texture coherence, on the other hand, only activation in the secondary areas including PT and TPJ was observed. Another interesting result was that boundaries between textures associated with an increase rather than a decrease in coherence were found to be perceptually more salient, and resulted in greater neuronal activity. This phenomenon has also been observed in more recent work suggesting that appearance of an auditory object is more salient than its disappearance (Constantino et al., 2012).

A similar stochastic stimulus was developed by McDermott and Simoncelli (2011) that is based on capturing the statistics of real-world stationary sound textures such as a stream of water, the sound of fire or that produced by a swarm of insects (see section 1.4.4.2 for more details). In a

recent experiment based on such textures, McDermott et al. (2013) developed a 'cocktail party' texture that was based on the superposition of multiple recordings of different speakers. Four different versions of the textures with varying density or number of speakers (1, 7, 29, or 115) were created. Listeners were presented with three excerpts of textures (of which two were identical) and were required to indicate which excerpt was different from the other two (as in an AXB paradigm). Two different durations of the textures were used – 50ms and 2500ms. Results revealed that the shorter exemplars were highly discriminable for all conditions but varied for the longer exemplars, producing an interaction between duration and the density of the textures. These results are in line with other experiments in the same study where discrimination of different exemplars of the same texture declined with the duration of the textures. This is contrary to discrimination performance for samples of different textures where performance increased with duration. Overall, the results suggest that summary statistics for mixtures such as speech may have a role in encoding time invariant properties of speech like voice quality or speaker identity and thus may aid segregation based on these features.

1.6 Key problems addressed in this thesis

Auditory scene analysis has been a topic of intense investigation over the last several decades and with the advent of modern imaging and recording techniques as well as development of sophisticated acoustic stimuli, there has been considerable progress. However, much of the work is and continues to be inspired by simple deterministic stimuli such as streaming and oddball stimuli, and multi-tone complexes that constrain the interpretation of the experimental findings. It is difficult to ascertain if the principles and mechanisms of streaming derived from such simple paradigms apply for real world sounds as well.

More recently, there has been a trend towards the use of complex stimuli that are based on stochastic variation of certain acoustic features that define an object (Overath et al., 2010) as well as synthetic stimuli that capture the time-invariant statistical properties of natural sound textures (McDermott and Simoncelli, 2011). This doctoral thesis adds to the growing field of perceptual analysis of complex acoustic scenes based on realistic stimulus patterns. A novel stochastic stimulus to study low-level figure-ground segregation in a controlled way is reported here. This signal, referred to as *stochastic figure-ground* (SFG) stimulus is an approach to segregation in real-world acoustic scenes.

The SFG stimulus forms the central theme for all the studies reported here. A variety of parametric designs using complementary behavioural, modeling and functional imaging techniques were used to elucidate the brain bases and mechanisms of segregation in complex acoustic scenes. The

following sections provide a brief description of the motivation for the studies that comprise this thesis.

1.6.1 Chapter 3 - Study 1

What are the behavioural capabilities of segregation in the novel SFG stimulus and how robust is performance to spectrotemporal manipulations?

Segregation can be easily performed in stimulus paradigms based on streaming, oddball as well as informational masking stimuli which represent a relatively simple simulation of segregation in real world acoustic environments. These signals comprise of deterministic narrowband patterns that either do not overlap in time (streaming and oddball signals) or have a spectral protective region surrounding the target tone (IM signals). This study introduces a novel stimulus, known as the stochastic figure-ground (SFG) stimulus that improves upon the limitations presented by these signals. The stimulus consists of a sequence of chords with randomly varying pure tone components that change from one chord to another. The target is defined by a set of repeating frequency channels that can only be detected by binding across both frequency and time domains. This study investigated target detection performance in the SFG stimulus under a variety of different stimulus conditions that manipulated the spectrotemporal structure of the stimulus.

1.6.2 Chapter 4 – Study 2

What are the mechanistic principles underlying segregation in the SFG stimulus and does a computational model based on temporal coherence explain segregation in the SFG stimulus?

Study 1 characterized target detection behaviour in the SFG stimulus and performance was found to be robust to several spectrotemporal manipulations. Segregation in the SFG stimulus cannot be easily explained based on standard models of auditory stream segregation (Fishman and Steinschneider, 2010a). This study investigated the ability of a new model of auditory segregation based on temporal coherence between frequency channels (Shamma et al., 2011) to explain segregation in the SFG stimulus. Temporal coherence refers to the average cross-correlation between channels over a specific time window and emphasizes the role of time in auditory scene analysis. In this study, each of the SFG stimuli examined in study 1 was simulated according to the model and its predictions were compared relative to the behavioural results.

1.6.3 Chapter 5 – Study 3

Which brain areas are involved in detecting the emergence of a target in the SFG stimulus?

Studies 1 and 2 established the behavioural and mechanistic bases of segregation in the SFG stimulus. This naturally leads to the next question, i.e., which brain areas are involved in detecting the emergence of the "figure" in this complex stimulus? Are the same brain areas in the auditory cortex involved in segregation as found in studies based on streaming or are other brain areas recruited in the case of the more complex SFG signal? Study 3 explored the brain bases of segregation in the SFG stimulus using functional magnetic resonance imaging.

1.6.4 Chapter 6 – Studies 4 and 5

What are the temporal dynamics of segregation in the SFG stimulus?

Moving from functional magnetic resonance imaging to magnetoencephalography, the aim of these studies was to elucidate the temporal dynamics of segregation in the basic SFG stimulus as well as a stimulus with white noise alternating between successive SFG chords as characterized in study 1. MEG tracks brain activity with a temporal resolution on the order of milliseconds and was used to investigate segregation in the SFG stimulus in a passive paradigm based on a simple transition from "background" to "figure". This study investigated the profile of the evoked transition responses and examined the underlying sources, with a specific aim to understanding the role of auditory cortex in figure-ground analysis in the SFG stimulus.

Chapter 2. METHODS

This chapter outlines the experimental methods used to analyse the behavioural and neuroimaging data presented in this thesis. The first section (section 2.1) deals with psychophysical procedures and measures used to index behavioural performance that forms a core component of the thesis. The next section (section 2.2) deals with the technique of magnetic resonance imaging (MRI) – from the physics of the MRI signal to data acquisition and statistical analysis. The final section (section 2.3) presents another popular tool in cognitive neuroscience – Magnetoencephalography (MEG) that provides high temporal resolution to precisely track the dynamics of cortical activity during auditory perception.

2.1 Psychophysics

Psychophysics has a long history, going back to the late 19th century, when Gustav Fechner first formulated it as a field of research to relate physical stimuli (e.g. light or sound) to the corresponding sensations they produce. It generally refers to the application of behavioural techniques to the study of sensory processing in human or animal species. In the auditory domain, psychoacoustics is the preferred term for analysis of auditory behaviour.

Psychophysics is an important field of study with widespread applications – from the development of animal models of auditory processing to design of acoustic devices such as hearing aids or cochlear implants (Fastl and Zwicker, 2006; Shofner and Niemiec, 2010). There exist

several psychophysical procedures, some of which are briefly described in the following section.

2.1.1 Psychophysical procedures

Psychophysical procedures can be classified in different ways, but the most common classification depends on whether stimuli are presented at fixed levels or at levels that vary adaptively according to the listeners' behaviour.

2.1.1.1 Method of Constant Stimuli

This method allows full sampling of the psychometric function where several stimulus levels that bracket the threshold are pre-selected and presented multiple times. The listener's absolute threshold can be calculated from the psychometric function which is often defined as the stimulus level that results in 50% correct detection. One disadvantage of the method is that it is relatively inefficient and requires many trials to estimate a single point on the psychometric function.

2.1.1.2 Alternative forced-choice procedures

In these methods, the listener usually has the task of deciding whether a signal was present or not on one or more of the presented intervals. In a one interval-two alternative forced-choice procedure, the listener has to judge whether a signal was present or not by responding "yes" or "no". In a two interval-two alternative forced choice procedure, two

different stimuli may be presented but only of these contains the signal and the listener has to decide which interval contained the signal.

A variation of the above procedure, known as the AXB paradigm is marked by three levels (Goldinger, 1998). This task involves stimulus comparison rather than detection of a particular stimulus feature. Two identical signals are presented at different intervals (either at A or X, or, at X or B) and the listener is required to indicate which interval contained a different or "odd" signal (A or B).

2.1.1.3 Adaptive tracking

Here, the stimulus levels depend on the listener's performance on the previous trials unlike the fixed algorithms in classic forced-choice procedures. Also known as "up-down" procedure, the stimulus level is reduced following a set number of correct detections and increased following a number of misses, asymptoting at a particular accuracy level on the psychometric function (Levitt, 1971). A one-down/one-up (two-down/one-up) tracking rule is associated with a reduction in stimulus level after one (two) correct detections and an increase in stimulus level after a single miss, tracking the 50% (70.7%) correct detection point on the psychometric function.

Reversing the direction of stimulus level continues until a set number of reversals are obtained; and the step size may be decreased after a set number of reversals to obtain a finer estimate of the threshold. The chief advantage of the tracking procedure is that it is more efficient than the

method of constant stimuli as more samples are obtained closer to the listener's threshold. This is determined by the step size and the number of reversals used to define threshold.

2.1.2 Signal detection theory

Signal detection theory is a theoretical framework that allows one to quantify decision making under uncertainty. It is particularly useful in behavioural experiments looking at the detection of sensory signals in the presence of noise.

A simple example is considered here for illustration purposes. Imagine an acoustic stimulus that contains speech in the presence of loud masking white noise. The speech is the signal of interest that the listener has to detect over and above the noise. On certain trials, the listener may respond "yes" when the speech signal is present ("hit") or "no" when the signal was absent ("miss"). Alternatively, the listener may also respond "yes" when the signal was absent ("false alarm") or "no" when the signal was actually absent ("correct rejection"). These constitute four possible responses and are considered together for quantifying discrimination performance and bias.

A measure of discriminability based on these responses known as the d-prime (d') can be formulated which represents a true measure of the internal response and does not depend on any criterion adopted by the listener.

D-prime takes both hits and false alarms into account and is defined as:

$$d' = Z(\text{hit rate}) - Z(\text{false alarm rate}) \qquad \text{(Eq. 2-1)}$$

where, Z is defined as the inverse of the cumulative Gaussian distribution (MacMillan and Creelman, 2005).

2.2 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is based on the principles of nuclear magnetic resonance (NMR; Cohen, 1996; 1999; Bandettini and Wong, 1998) which is a technique used to measure microscopic chemical and physical data from individual atoms. The technique came to be known as MRI instead of NMR because of the negative connotations associated with the word 'nuclear' in the 1970s.

The beginnings of NMR can be traced back to the 1940s when Felix Bloch and Edward Purcell independently discovered the magnetic resonance phenomenon, for which they received the Nobel Prize in 1952. Since then, it was used primarily for physical and chemical molecular analysis. In the 1970s, Raymond Damadian demonstrated that tissues and tumors have different magnetic relaxation times, thus motivating the use of NMR for clinical purposes. He later developed field-focusing MRI technique whilst Peter Mansfield at the same time developed the echo planar imaging (EPI) technique (Damadian et al., 1977; Mansfield, 1977).

Atomic nuclei that contain an odd number of nucleons are unstable entities and behave like magnetic dipoles with a magnetic moment and a

spin. Such nuclei are capable of producing NMR signals as they align in an external magnetic field and 'precess' at a frequency proportional to the field strength. The transitions between energy states (parallel and anti-parallel to the external magnetic field) emit energy in the radio frequency range when the nuclei return to equilibrium. Such NMR signals are not produced by nuclei with even numbers of nucleons. The human body contains approximately 63% hydrogen atoms which are present predominantly in the form of water in tissue.

In order to obtain high resolution MR images, there are a few essential requirements. Firstly, a powerful external magnetic field is required that aligns hydrogen atoms parallel to the field. Magnetic field is measured in Tesla (T), where 1 Tesla = 10000 Gauss. This represents extremely high field strength in comparison to the Earth's magnetic field of 0.5 Gauss. Modern MR scanners used in human neuroscientific research produce fields that vary from 3 T to 11 T and the imaging experiment reported in this thesis was carried out in a 3T Siemens Allegra scanner. Another requirement is a high energy (radio frequency, RF) pulse of a specific frequency and duration to perturb the equilibrium state of the nuclei and induce net magnetization that results in the emission of energy as discussed below.

From producing NMR images of single slices through the human body, MRI was further developed to incorporate spatial information of the tissue by spatially varying the magnetic field. Modern MRI techniques produce high resolution images based on spatial variations in the phase and

frequency of the radio frequency energy being absorbed and emitted by protons in human tissue.

In a magnetic field of strength B, a proton that has a net spin can absorb a photon of frequency ν and are related by the following equation:

$$\nu = \gamma B \qquad \text{(Eq. 2-2)}$$

where, ν is the Larmor frequency in MHz, γ is the gyromagnetic ratio in MHz/Tesla and B is the strength of the external magnetic field in Tesla. For hydrogen atoms, the Larmor frequency is 42.58 MHz/Tesla.

The spin can be considered as a magnetic moment vector causing the proton to behave as a tiny magnet which can align with the external field in either a low energy state (where poles are aligned N-S-N-S) or a high energy state (where poles are aligned N-N-S-S). The proton can transition between these two energy configurations by absorbing a photon whose energy is equal to the difference in energy between the two states. This relationship is given by the following equation:

$$E = h \nu \qquad \text{(Eq. 2-3)}$$

where, E is the energy of the photon and h is the Planck's constant (h = 6.63 x 10^{-34} Joules/second).

Thus, there are two possible magnetization alignments in a three dimensional reference – a longitudinal magnetization (M_z), where the magnetic moment (along z-axis) is in alignment with the external magnetic

field, B, and a transverse magnetization (M_{xy}) in the x-y plane due to the precession of the nuclei along the z-axis. At equilibrium, the net magnetization is equal to the longitudinal magnetization and there is no transverse magnetization.

The magnetization at equilibrium (M_0) can be perturbed by the application of a radio frequency pulse whose energy is equal to the energy difference between the two spin states. In the situation where the spin system is saturated, longitudinal magnetization can be reduced to zero. The time taken for the longitudinal magnetization to return to its equilibrium value is known as the spin lattice relaxation time, often denoted as T1. This is given by the equation:

$$M_z = M_0 (1 - e^{-t/T1}) \qquad \text{(Eq. 2-4)}$$

Another effect of the RF pulse with Larmor frequency ν is the precession of the nuclei in phase, causing a net transverse magnetization in the x-y plane. However, this net magnetization begins to dephase because the constituent spin packets experience different magnetic fields and rotate at different Larmor frequencies. The time constant which defines the return to equilibrium of the transverse magnetization is known as the spin-spin relaxation time, T2 and is given by the equation:

$$M_{xy} = M_{xy0}\, e^{-t/T2} \qquad \text{(Eq. 2-5)}$$

However, the effective time for the transverse magnetization to reduce to its equilibrium value is governed by molecular interactions (which leads

to a pure T2 molecular effect) and variations in the external field, B (which leads to an inhomogeneous T2 effect) and the effective time constant is known as T2* which is given by the equation:

$$1/T2^* = 1/T2 + 1/T2_{\text{inhomo}} \qquad \text{(Eq. 2-6)}$$

The generation of NMR images makes use of several tissue properties: the NMR signal varies as a function of the proton density. Additionally, tissues have different magnetization characteristics that determine how rapidly the NMR signal decays. The signal decay is a function of both T1 and T2*.

The most common NMR imaging technique is the 'spin-echo' technique (Hahn, 1950). An initial RF pulse is applied to the tissue at equilibrium which results in tissue-specific T1 and T2 effects as discussed above. A second 'echo' RF pulse is used to cancel the spin phase differences of the nuclei rotated by the initial RF pulse, thereby reforming the transverse magnetization decay and neutralizing the effects of T2* dephasing due to extrinsic inhomogeneities in the external magnetic field. This results in better detection of the small inhomogeneities that actually reflect tissue magnetization differences. The time at which the decay signal is read out with an RF receiver coil is known as the 'time-to-echo' (TE). The spatiotemporal resolution of the MR images is limited by the biological properties of the tissue as well as the characteristics of the scanner and the imaging sequence (field strength and TE). An alternative technique in NMR is a gradient-echo technique that records the signal after the initial 90° RF

pulse without phase refocusing and is thus more susceptible to T2* effects; for this reason, it is commonly used in fMRI.

NMR images are obtained by periodically varying the field strength in a gradient along each dimension, so that resonant frequency is a function of spatial position. The NMR signal obtained at the RF receiver coil at time TE is a complex of different frequencies that is analysed using Fourier decomposition. Thus, spatial frequency encoding is determined by the amplitude and duration of the gradients. To obtain a complete three-dimensional image, all combinations of spatial coordinates are sampled along each axis. A planar image is constructed on a grid in the Fourier spatial frequency domain or 'k-space' using two orthogonal gradients: a 'read-out' gradient along the x-axis (G_x) that encodes the spatial frequency and a 'phase-encode' gradient along the y-axis (G_y) that advances the phase using a series of appropriate RF pulses. In this k-space, high spatial frequencies are represented in the periphery whilst low spatial frequencies are encoded in the centre. The path traversed through the k-space to acquire the data is known as the k-space trajectory. The time between successive phase-encoding pulses is referred to as the 'time-to-repeat' (TR). An orthogonal 'slice-selection' gradient in the z-axis (G_z) enables the sampling of successive tissue planes. This gradient is crucial for ensuring that only the protons in a single slice (of thickness determined by the bandwidth of the RF pulse) become resonant and thus undergo rotation and emit a signal. In the end, an inverse Fourier transform is applied to the signal in each plane to recover the spatial characteristics of the imaged tissue.

The strength of MRI for cognitive neuroscience applications is its high spatial resolution that enables the accurate localization of neural activity. The resolution is characterized by the size of a single image volume element (voxel) that is determined by the ratio of the volume of the image (field of view) and the number of sampling points during image acquisition. Voxel size is characterized by the product of the number of samples in the read-out and phase-encoded directions and the slice thickness. In the fMRI experiment conducted as part of this thesis, the in-plane resolution was 3.0 x 3.0 mm^2 and the slice thickness was 2 mm with 1 mm gap between slices (see section 5.2.4 for more details).

2.2.1 Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI) has heralded a revolution in systems and cognitive neuroscience by providing experimental access to neuronal ensembles involved in perception and cognition. Previous imaging techniques such as Positron Emission Tomography (PET) involved ingestion of radioactive tracers, whilst fMRI offers the benefits of non-invasive imaging of the whole brain with high spatial resolution. It also has the advantage of flexible data acquisition characteristics that can be adapted for the specific problem being addressed.

In this section, the principles of fMRI, its neurophysiological bases, scanning protocols for acquisition of auditory datasets, data pre-processing and statistical analysis steps carried out to obtain functional correlates of task-related brain activity are briefly reviewed.

2.2.1.1 Echo-planar imaging

To investigate physiological phenomena using fMRI, rapid image acquisition is required. This is achieved through echo-planar imaging (Mansfield, 1977) that enables ultrafast acquisition of the x-y plane using a single excitation pulse ('single shot') on the order of tens of milliseconds per volume. Rapid switching of frequency (G_x) and phase (G_y) gradients is performed to cover the entire plane.

In functional applications of EPI, gradient-echo rather than spin-echo acquisition sequences are used to refocus the NMR signal. As the spin-echo is omitted, the signal is more sensitive to local field inhomogeneities (T2*) including those produced by deoxyhaemoglobin and thus better suited for detection of metabolic dynamics. Gradient echoes are usually generated by an oscillating gradient (Logothetis, 2002) along the read-out direction that follows a zigzag trajectory in k-space. Here, TE is defined as the time from the excitation pulse to the centre of k-space that is approximately equal to T2* (Logothetis, 2002). EPI requires large gradient amplitudes and rapid gradient switching for rapid acquisition which necessitates the use of dedicated hardware for phase encoding, and high-speed analog-to-digital conversion.

2.2.1.2 Physiological basis of BOLD signal

The technique of fMRI based on measurement of the BOLD signal is aptly summarized by Ogawa (2012) in an article from a special volume of NeuroImage celebrating twenty years of fMRI as below:

"To perform a given or spontaneous task, the brain mobilizes localized specific sites which form a functional network specialized for the task. Synaptic activity in such localized sites is tightly coupled through astrocytes to vascular responses that can be detected by fMRI. The response time, on the order of seconds, is much slower than neural events, but one can plot the time course of the MRI signal and infer the task-related neural events in the brain that caused the response. This non-invasive way of measuring phenomena related to brain function has indeed widened the scope of brain research."

Seiji Ogawa, termed the image contrast "BOLD" (Blood Oxygenation Level Dependent) as it was dependent on the content of deoxyhaemoglobin in the blood (Ogawa et al., 1990a; Ogawa, 2012). He demonstrated in vivo that changes in blood oxygenation affected T2 and T2* weighted signals (Ogawa et al., 1990a, 1990b). However, the application of the BOLD signal in its present form can be traced back to early work by Linus Pauling who showed that the magnetic susceptibility of haemoglobin depends on the specific isotopes that are bound differently to oxygen-bound iron – oxyhaemoglobin is diamagnetic while deoxyhaemoglobin is paramagnetic (Pauling and Coryell, 1936). The NMR signal of paramagnetic deoxyhaemoglobin decays faster than oxygenated haemoglobin. This results in magnetic susceptibility differences between the haemoglobin-containing vasculature and the surrounding tissue. This leads to greater dephasing of the protons and a reduction in the corresponding T2*

signal. Neural activity is thus related to changes in T2* (BOLD) signal and the corresponding regional intensity changes in T2*-weighted images (Ogawa et al., 1992).

The characterization of BOLD signal and neuronal dynamics was carried out in a series of experiments by Nikos Logothetis and colleagues who combined acquisition of BOLD signal in anaesthetized monkeys with intracortical microelectrode recordings from the visual cortex (Logothetis et al., 2001; Logothetis, 2002, 2003). It was established that the BOLD haemodynamic response correlates most strongly with low-frequency components of the extracellular local field potentials (LFPs) rather than spiking activity of local neuronal ensembles (Logothetis, 2012). Extracellular field potentials primarily reflect local neuronal processing within a cortical ensemble rather than the output activity *per se.* LFPs represent several effects such as neuromodulation, interactions between interneurons and pyramidal cells which may be the underlying bases for the resulting haemodynamic signal (Logothetis, 2012). Currently, it is accepted that haemodynamic responses depend on the size of the activated populations and reflect enhanced regional neural activity (Logothetis, 2012).

2.2.1.3 Haemodynamic response function

The haemodynamic response function (HRF) characterizes the BOLD response and has distinct characteristic phases (Logothetis, 2002). It captures the varied and complex interactions between regional cerebral blood flow, blood volume and blood oxygenation. There is an initial 'dip' that may reflect increase in oxygen consumption which changes the ratio of

deoxyhaemoglobin to oxyhaemoglobin (Malonek and Grinvald, 1996). This is followed by an increase in regional blood flow to the active regions. Using PET, Fox and Raichle (1986) demonstrated that this represents a decrease in oxygen extraction fraction, i.e., an increase in blood oxygenation. This increased signal corresponds to the peak of the HRF which approximately takes 4-6 seconds from stimulation onset to reach the maxima and returns to baseline approximately 5-20 seconds after stimulus offset in primary sensory (including auditory) cortices (Belin et al., 1999; Hall et al., 1999). Increased blood flow results in vasodilation and an increase in local venous blood volume which causes a post-stimulus undershoot in the HRF (Buxton et al., 1998). These haemodynamic changes depend on the external field strength but the peak BOLD response is typically on the order of 1-1.5% in the auditory cortex (Talavage et al., 1999).

2.2.2 fMRI for auditory stimulation

2.2.2.1 Problems in auditory functional neuroimaging

Although fMRI has proved to be very important in revealing aspects of human auditory perception, it is not completely free of methodological issues. The main constraint of human fMRI for auditory research is the loud acoustic noise produced by the switching of the gradient coils. Continuous EPI sequences can result in sound pressure levels of 120 dB in the bore of the scanner (Ravicz et al., 2000; Price et al., 2001). The primary source of noise is due to the read-out gradients, with other low frequency ambient

noise produced by the helium cooling pump and air conditioning systems (Ravicz et al., 2000).

Furthermore, the spectrum of the gradient noise is broadband and ranges from 250 Hz to 4 kHz with a peak around 1-1.5 kHz (Hall et al., 1999; Ravicz et al., 2000; Chambers et al., 2001). This is a major problem as it overlaps with a critical frequency range for human auditory perception. At low frequencies (below 500 Hz), the ear canal is a major route of conduction of environmental noise, whilst at higher frequencies (greater than 500 Hz), direct conduction through the bones becomes the major route when ear protection is provided (Ravicz and Melcher, 2001). Ear defenders can provide 30-40 dB of passive attenuation of scanner noise but active noise cancellation systems can provide additional benefits (Chambers et al., 2001; Moelker and Pattynama, 2003), however, its benefits are limited by the bone conduction of noise. Scanner noise causes a BOLD response of a variable magnitude (Moelker and Pattynama, 2003) in the primary auditory cortex and to a lesser extent in non-primary auditory cortex whose magnitude increases nonlinearly with the duration of the acquisition sequence (Talavage et al., 1999).

The scanner noise poses several problems: most significantly it reduces the signal to noise ratio (SNR) as the BOLD response is a complex response to the auditory stimulus of interest and the undesirable scanner noise in the background. Furthermore, it also precludes the use of a 'pure' silent baseline which is necessary for cognitive subtraction analysis. The difference in haemodynamic response between an active condition and a

baseline condition with scanner noise is not equal to the difference between the active condition and silence (Gaab et al., 2007). The constant noise further constrains accurate modeling of physiological responses as it results in adaptation or habituation of the response to the stimulus whose relative magnitude varies across cortical fields (Di Salle et al., 2001). Another problem is that the BOLD response varies with the absolute level of the stimulus (Jäncke et al., 1998) and thus it becomes difficult to quantify the effects of the background noise (Belin et al., 1999; Edmister et al., 1999; Talavage and Edmister, 2004). Apart from the loud unpleasant experience for the listener, the scanner noise makes it difficult to hear the stimulus which changes the nature of the perceptual task to an auditory figure-ground discrimination task (Scheich et al., 1998). This may result in an interaction between task and background noise that is not modeled in the experimental design (Hall et al., 1999).

2.2.2.2 Auditory imaging protocols

In order to avoid the problems posed by the scanner noise, a number of alternative imaging approaches have been developed. These include the use of quiet acquisition sequences (Belin et al., 1999; Sander et al., 2003), enhanced passive and active noise attenuation techniques (Chambers et al., 2001; Ravicz and Melcher, 2001) and development of 'silent' or 'sparse' imaging protocols that circumvent the issue of the scanner noise (Belin et al., 1999; Hall et al., 1999; Talavage and Hall, 2012).

Belin and colleagues (1999) developed an event-related paradigm and introduced a silent period between successive volume acquisitions. The

silent period was 9 seconds long and the stimuli of interest were jittered within the silent phase so as to map different points of the haemodynamic response function. In this approach, all images are acquired at the end of the TR as opposed to continuous imaging protocols. Typical TR values are between 10-16 seconds which minimizes overlap between the HRFs of the stimuli and the scanner noise. Hall and coworkers (1999) implemented a similar approach but acquired a volume at the predicted peak of the haemodynamic response. Such 'sparse' imaging protocols significantly improve the SNR albeit at the cost of limited temporal resolution. Additional limitations include extended scanning times to obtain reasonable SNR, subject fatigue and movement and loss of attention.

Generally, the choice of the imaging protocol depends on the question being addressed. The biological significance of the task must also be considered: an auditory task in a continuous scanning paradigm effectively becomes a figure-ground discrimination task where the target sounds of interest must be discriminated from the ongoing scanner noise in the background. If it is required to precisely map the HRF at different time points, then a continuous acquisition sequence is preferable. However, if the question is geared towards elucidating the sensory representation of specific acoustic features, then it is best to use sparse imaging protocols. In the work presented in this thesis, the imaging experiment was based on continuous acquisition as many trials were required for each parameter of interest to obtain a suitable SNR.

2.2.3 Image analysis

The analysis of functional MRI data requires several sophisticated pre-processing algorithms to obtain a veridical measurement of the spatial extent of brain activity at the single-subject level as well as at the group level on a common spatial reference frame. Specifically, the imaged volumes need to be realigned to account for movement of the listeners, normalised to a standard spatial reference frame to allow between-subject comparisons, and smoothed to increase the SNR. Careful modeling of the experimental design and statistical analysis is essential to eliminate any false positives which can prove to be a major hindrance due to the multiple comparisons problem.

These pre-processing steps were carried out using Statistical Parametric Mapping (SPM8) software (http://www.fil.ion.ucl.ac.uk/spm) implemented in MATLAB 2010 (MathWorks Inc.). A brief description of the theoretical principles underlying these procedures is described below.

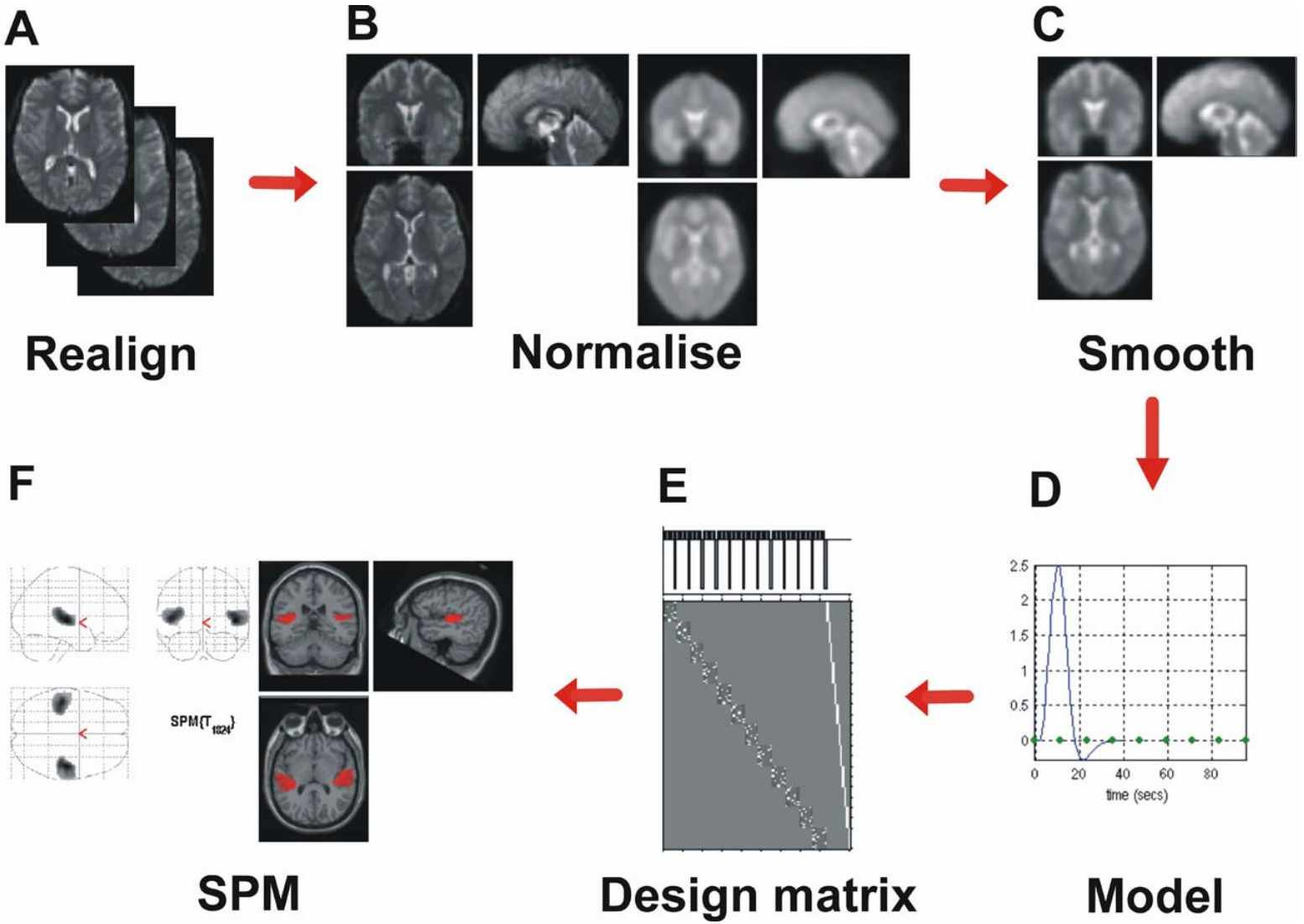
A
B
C
Realign
Normalise
Smooth
D
E
F
2.5
2
1.5
1
0.5
0
0
20
40
60
80
time (secs)
SPM{T1824}
Model
Design matrix
SPM

Figure 2.1: Steps for pre-processing and analysis of fMRI data.

(A) Raw brain images are first realigned to correct for subject movement and session effects, using an algorithm that minimises variance between images. (B) Realigned images are normalised to a brain template to transform them into a common stereotactic space and to correct for individual anatomical differences.

(C) Normalised images are smoothed using a Gaussian filter of specified full-width at half maximum. This step improves signal to noise ratio by increasing overlap between adjacent voxels, with corresponding reduction in spatial resolution.

(D) Data from smoothed images are analysed using a specified model: this includes convolution with a haemodynamic response function to account for the time course of cerebral blood flow in relation to neuronal activity.

(E) A design matrix is generated based on the general linear model, rows corresponding to scan number and columns to trials (effects or covariates of interest), with additional columns corresponding to effects or covariates of no interest (e.g., global cerebral blood flow for each subject). A software package (such as SPM) is used to estimate statistics on the design matrix. The parameter estimates in the column vectors are adjusted mean least squares estimates of the effects of interest (discounting effects of no interest); a contrast between experimental conditions is defined by a vector that represents a weighted sum of parameter estimates. Based on the null hypothesis that the effect of interest does not account for more signal variance than could be explained by chance (according to the assumptions of Gaussian random field theory), a *t* statistic can be derived at each voxel as the ratio of the contrast-weighted parameter estimates to the estimated standard error term for that voxel. The *t* statistics across brain voxels together constitute a statistical parametric map of brain activation for that contrast. Activations are thresholded at a specified significance level, typically $p < 0.05$ corrected for the effects of multiple comparisons across the brain volume or for the false discovery rate.

(F) A statistical parametric map (SPM) of the statistic can be plotted as 'glass brain' projections in axial, coronal and sagittal planes or rendered onto a structural

template (a canonical brain, group mean MRI, or the subject's own structural MRI) to indicate relationships of activation to brain anatomy.

2.2.3.1 Realignment and unwarping

fMRI time series are often contaminated due to the movement of the listeners. Head movement in particular is detrimental to accurate reconstruction of brain activity as it changes the location of a given voxel in a particular brain area. Thus, even tiny movements can result in misalignment across successive scans which can contaminate the data (Friston et al., 1995a) and contribute as much as 90% of the variance (Friston et al., 1996b). This can lead to misinterpretation of signal changes as brain 'activations' whose magnitude may be larger than the physiological response of interest. Smaller movements due to cardiac cycle variations are also a source of misalignment, especially in the brainstem structures. Thus, movement that may or may not be correlated to the experimental task poses a significant problem as it may be misattributed as activation and impair the detection of veridical brain responses. This makes motion correction of EPI particularly important to obtain true measures of brain activity.

Motion artifacts are reduced via procedures that realign successive images of a time series to a common spatial reference frame (usually the first image of the time series). This realignment is based on a least squares approach and a 6 parameter (three translations and three rotations) affine 'rigid-body' spatial transformation to calculate the movement associated with each scan (Friston et al., 1995a; Andersson et al., 2001). These parameters are used to 'reslice' the image to the new grid coordinates determined by the transformation (Grootonk et al., 2000). Additional motion artifacts due to magnetic inhomogeneities at air-tissue interfaces such as the

orbitofrontal cortex results in deformations in the sampling matrix (Andersson et al., 2001) and are further distorted by movement. This is accounted by an unwarping algorithm and the use of field maps which model these field inhomogeneities and associated geometric distortions (Hutton et al., 2002; Cusack et al., 2003).

2.2.3.2 Normalisation

Individual brains vary vastly in their anatomy and thus it is important to normalise imaged volumes from different individuals onto a common anatomical reference space. A nonlinear warping algorithm is used to coregister functional brain activity with structural scans. In SPM, the 'realign and unwarping' procedure creates a mean functional image that is used to estimate warping parameters to map it onto a standard stereotactic space. There are a number of standard neuroanatomical models that are based on either 'canonical' brains (Talairach and Tournoux, 1988; Toga et al., 1994) or average brains based on data from several individual brains (Evans et al., 1993; Roland and Zilles, 1994; Mazziotta et al., 1995). The normalisation is achieved via a 12-parameter affine transformation to obtain a spatial transformation matrix followed by a nonlinear estimation of spatial deformation patterns.

2.2.3.3 Smoothing

In the final stage of image pre-processing, the normalised data are smoothed by convolution with a Gaussian kernel of a specific width. This step is necessary to reduce noise (increase SNR) and effects due to residual differences in functional and gyral anatomy during inter-subject averaging

(Friston, 2003b). The Gaussian kernel typically has a full-width-at-half-maximum (FWHM) equal to 2-3 times the size of the voxel. The convolution improves the fit between the imaging data and the assumptions of Gaussian random field theory used for statistical analysis of brain activations as discussed below. The residual errors are rendered more normal, ensuring the application of parametric statistical tests. Generally, a kernel of 6mm FWHM is used at the individual subject level and a kernel of 8mm FWHM is appropriate at the group level.

2.2.4 Statistical analysis

The imaging experiment described in chapter 5 in this thesis was analysed using SPM8. The signal at every voxel is assumed to have a normal distribution under the null hypothesis of no regionally specific effects. This hypothesis is tested at each voxel using a mass-univariate approach based on General Linear Models (GLMs). It consists of a few steps which are described below in greater detail: i) specification of a GLM design matrix, ii) estimation of GLM parameters using classical or Bayesian approaches, and iii) assessment of results using contrast vectors to obtain Statistical Parametric Maps (SPMs) of regional brain activity.

2.2.4.1 General Linear Model

The GLM provides a theoretical framework for statistical analysis of functional imaging data using common parametric tests like Student's *t* test or analysis of variance (ANOVA). This method, the GLM, models the signal intensity in each voxel as the linear combination of effects of interest,

effects of no interest (or confounds) and error terms as given by the following matrix equation:

$$\mathbf{SX} = \mathbf{SG}\beta + \mathbf{SH}\gamma + \mathbf{Se} \qquad \text{(Eq. 2-7)}$$

where **X** is the data matrix comprising signal intensity values, **G** is a matrix reflecting the experimental variables as a linear combination of regressors, **β** is a matrix of parameter estimates for the effects of interest, **H** is a matrix including covariates of no interest or confounds such as motion parameters, **γ** is a matrix of effects of no interest, **e** is a matrix of normally distributed error terms and **S** is a convolution matrix that models the haemodynamic response function (Friston et al., 1995b). **G**, **H**, and **S** are specified in the design matrix which has one row for each scan and one column for each variable of interest. Effects of interest are modeled as box car functions. The parameter estimates in **β** are adjusted mean least-square estimates of the effects of interest and are contrasted against each other by appropriately weighted contrast vectors. A *t* statistic can then be generated for each voxel as the ratio of contrast-weighted parameter estimates to the estimated standard error term.

2.2.4.2 Random Field Theory

For testing the significance of the activations in each voxel, Gaussian Random Field Theory is invoked which assumes that under the null hypothesis, the statistical parametric maps of the parameter estimates for each voxel are distributed according to a certain probability distribution function, usually a *t* or *F* distribution. Any deviations of this distribution that

exceed a pre-specified statistical threshold can be attributed to the variables of interest with a probability of $1 - \alpha$, where α is the Type I error related to false rejection of the null hypothesis.

Normal correction methods for multiple comparisons are impractical in the case of fMRI data due to the vast number of observations, or voxels. Thus, an appropriate statistical framework is necessary to control the false positive rate. Conventional Bonferroni correction (where the false positive rate is simply divided by the number of independent observations) is impractical as it results in a very stringent statistical threshold. Furthermore, the signal intensity values in the voxels are not truly independent due to spatial correlations among neighbouring voxels and the spatial extent of the haemodynamic response function. The use of a conservative threshold decreases the likelihood of detecting true activation. Therefore, by convention, a significance threshold of $p < 0.001$ (uncorrected) is used for brain areas which are *a priori* predicted to be activated as a function of the experimental variables. Another solution for analysing activations in predicted brain areas is to restrict analysis to a discrete volume specified by that hypothesis, which is known as 'small volume correction'. For other brain regions, it is advisable to use a correction for multiple comparisons based on family wise error (FWE) rate (Logan and Rowe, 2007; Nandy and Cordes, 2007) or false discovery correction rate (Genovese et al., 2002).

2.2.4.3 Random-effects analysis

In analysis of fMRI data, the level of statistical inference is an important consideration (Friston et al., 1999). There are two principal types

of analyses: fixed-effects and random-effects analyses, which vary with respect to how data from multiple subjects is regarded at the level of the population at large.

In a fixed-effects analysis, the underlying assumption is that the variability in activation for a particular effect of interest is fixed and does not vary between subjects. Here, inter-subject variability is disregarded and time series of data from multiple subjects are treated as different sessions within a longer time series and only the error variance between scans is modelled. The number of degrees of freedom is high for fixed-effects analysis and is slightly less than the total number of brain volumes.

Random-effects analysis, on the other hand, treats the variability in activation between subjects as a random variable and allows inference about the average behaviour of a voxel across the population of subjects. Here, the degrees of freedom are equal to $n - 1$, where n is the total number of participants. Typically, 8-16 participants are required for obtaining reliable estimates of inter-subject variability. This is achieved through a two-step procedure, where contrasts between parameters of interest are estimated at the first level for each participant before evaluating these at the second level (for instance using a *t* test). The fMRI study reported in this thesis was based on random effects analysis.

2.3 Magnetoencephalography

"A recording of a component of the magnetic field vector as a function of time, at the head, is called a magnetoencephalogram", wrote David Cohen in the landmark paper in *Science* reporting the first measurements of the magnetic field around the human brain (Cohen, 1972). Since then, magnetoencephalography (MEG) has made significant advances and is regarded as an important non-invasive imaging method in cognitive neuroscience. The main attraction of MEG is its high temporal resolution, on the order of milliseconds, that enables precise tracking of brain dynamics during perception and cognition.

The electrical sources in the brain that produce the scalp potentials of the electroencephalogram (EEG) are also responsible for the magnetic field around the head. The principal advantage of using EEG or MEG over fMRI is that they are directly and instantaneously related to the actual neuronal generators, i.e., dendritic activity in pyramidal cells of the cortex (Cohen and Halgren, 2009). Spiking activity does not produce a magnetic field due to the random spatial arrangement of the underlying currents. This is in contrast to fMRI where the measured signal is only indirectly related to the underlying neuronal activity via neurovascular coupling. Furthermore, MEG is able to measure brain activity as it evolves every millisecond unlike the slow BOLD signal which takes up to 4-6 seconds to peak. These benefits have resulted in the worldwide adoption of MEG in research laboratories as well as hospitals as a powerful tool for basic and clinical neuroscience (Hari and Salmelin, 2012).

The sources of MEG as well as EEG signals are synchronous postsynaptic intracellular currents in the pyramidal cells of the cortex rather than spiking activity (Hari, 1990). This is because spikes produce magnetic quadropoles and the associated magnetic field decays at a faster rate with distance ($1/r^3$) as compared to the dipolar field produced by postsynaptic currents that decays as at a rate proportional to $1/r^2$. Also, action potentials are transient events that decay with a couple of milliseconds unlike postsynaptic potentials that evolve over tens of milliseconds during which several cells can contribute to the magnetic field strength.

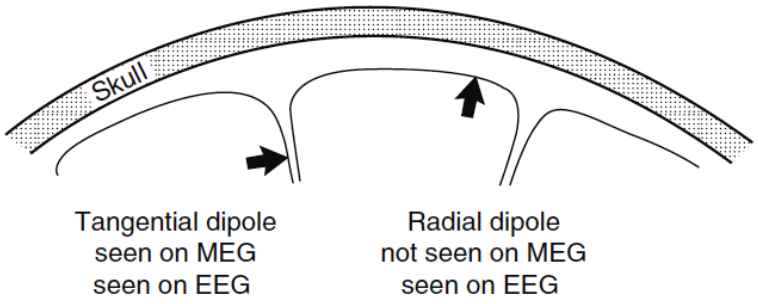

Figure 2.2. Sensitivity of MEG and EEG to tangential and radial dipoles.

MEG is sensitive to activity from dipoles oriented tangentially but not radially whilst EEG picks up signals from sources in both orientations relative to the skull. Figure reproduced from Cohen and Halgren, 2009.

The net current flow in pyramidal neurons is perpendicular to the cortical surface. The MEG signal, however, is most sensitive to dipoles that are tangential to the skull, in the sulci, whereas EEG can pick up signals from dipoles that are both tangential and radial to the skul as indicated in figure 2.2 (Hari, 1990; Hämäläinen et al, 1993). MEG and EEG thus complement each other because of their differential sensitivities to source orientations and locations.

2.3.1 Instrumentation

The magnetic field measured by the MEG is very weak with typical field strengths less than 10^{-12} T. This is much smaller than urban fluctuating magnetic background (10^{-7} T) or even the Earth's magnetic field of approximately 0.5 x 10^{-4} T. Thus, in order to pick up the tiny fields due to brain activity, a magnetic detector of high sensitivity and reduction of environmental magnetic interference is absolutely essential.

Early MEG recordings used an induction-coil magnetometer with a couple of million turns of copper wire around a ferrite core and the MEG signal (alpha rhythm) was obtained by averaging against an EEG reference signal (Cohen, 1968). The earliest MEG device based on SQUIDs, superconducting quantum interference devices (Silver and Zimmerman, 1965) was encased in a room with heavy magnetic shielding (Cohen, 1972).

Modern neuromagnetometers, however, contain an array of more than 300 SQUID sensors that operate at 4 K and are therefore immersed in a liquid helium dewar. Each SQUID is fed by a magnetic sensing coil which

is arranged in a spherical array over the head at grid points 2 or 3 cm apart. The spherical section is approximately 2 cm away from the scalp of the subject. The sensors allow simultaneous magnetic field measurements at several coil locations over the head, resulting in a continuous acquisition of magnetic field produced by brain activity.

2.3.2 Data analysis

MEG data is of high temporal resolution with typical sampling rates of 600 Hz or above. Analysis of such data requires computational resources including high memory and sophisticated processing software (Baillet et al., 2011) to process the raw data to obtain neuromagnetic measures of interest such as evoked field strengths, frequency-time response maps, source models of evoked activity or effective connectivity patterns based on hierarchical generative models using Dynamic Causal Modeling (Friston et al., 2003; Kiebel et al., 2009).

In terms of data processing, MEG analysis is not as standardized as is the case for fMRI for which automated analysis pipelines exist. There is a recent trend of trying to standardize the MEG analysis methods and develop good practical measures for conducting and reporting MEG research (Gross et al., 2013). Furthermore, there is also useful cross-talk between different methods communities to integrate algorithms that complement the strengths of different software such as FieldTrip (Oostenveld et al., 2011) or SPM (Litvak et al., 2011).

2.3.3 Data pre-processing and analysis

The MEG data reported in this thesis was collected using a CTF275 scanner at a sampling rate of 600 Hz and analyzed using SPM12 (Litvak et al., 2011). The first step in pre-processing involves converting the raw data into a format that is compatible with the particular software used to analyze the data. In order to save computational resources, it is advisable to downsample the data so that subsequent files are smaller in size and easier to work with. The cut-off frequency depends on the specific paradigm but it is common to downsample the converted data to 300 Hz to include potential high frequency components of interest. The next step involves defining data epochs of interest, i.e. defining time windows to divide the data into individual trials with a pre- and post-stimulus baseline period. A pre-stimulus baseline period (usually 500ms or longer) is specified to baseline correct the data. The baseline could be a silent period or irrelevant sounds as well (e.g. white noise in studies of pitch perception). The different stimulus conditions are also specified during epoching which allows comparison of brain activity across stimulus conditions.

Analysis of evoked data involves measuring components that are time-locked to the presentation of the stimulus. Auditory evoked potentials are usually measured in EEG and MEG experiments and include a variety of responses that reflect different cognitive processes. The M100 is an evoked response that is produced in response to the onset of a sound with an average latency of ~ 100ms. The M100 is typically mediated by sources in the auditory cortex (Lutkenhoner et al., 2003) and serves as a sanity check

in MEG studies of auditory perception. The MMN is another auditory evoked potential that has been studied in detail (Näätänen et al., 2007) as discussed in section 1.5.2. Other auditory evoked fields of interest include the M50, M200 and M300 (Nagarajan et al., 2010).

To obtain measures of evoked brain activity, the next pre-processing step usually involves applying a low-pass filter with a typical cut-off frequency of 30 Hz. The resultant MEG time-series data are averaged across all stimulus presentations to obtain mean evoked field strengths. The averaging procedure eliminates any induced components that are not time-locked to the stimulus. The averaged evoked fields are analyzed separately for each condition of interest and appropriate statistical tests are performed to obtain a summary of the evoked fields across stimulus conditions.

2.3.4 Source reconstruction

Source reconstruction of MEG time-series is an ill-posed problem: there exist an infinite number of solutions to the inverse problem of identifying brain sources that produce activity observed at the sensors. This problem can be resolved by making certain assumptions about the sources in order to constrain the solutions. A number of source modeling methods exist which make different assumptions about how the brain works: these include Variational Bayes Equivalent Current Dipole (VBECD; Kiebel et al., 2008), Multiple Sparse Priors (MSP; Friston et al., 2008), Minimum Norm Estimates (MNE, Hauk, 2004), and Beamforming (van Veen et al., 1997) amongst others.

An "imaging" (or distributed) approach implemented in SPM12 (Litvak et al., 2011) was used to reconstruct the sources of evoked power in the MEG experiment (see chapter 6). This approach involves projecting the sensor data into 3D brain space and considers sources to comprise of a large number of dipolar sources spread over the cortical sheet with specific locations and orientations. Source amplitude or power (evoked or induced) can be estimated for a specified time window and frequency range. This reconstructed activity is in 3D voxel space and can be analyzed using GLM-based statistical approach as implemented for making inference in fMRI data.

Distributed linear models have been used previously (Dale and Sereno, 1993) but the imaging approach in SPM incorporates two additional features which improve the accuracy of the localization procedure:

i) A Bayesian framework is incorporated in which several constraints (or priors) can be imposed and the best model can be determined through Bayesian Model Comparison (Friston et al., 2005)

ii) Spatial localization is improved by including the subject's own structural anatomy in the generative model of the data.

The next section briefly describes the steps involved in obtaining the inverse reconstruction.

2.3.4.1 Source space modeling

Data containing carefully defined epochs for each experimental condition is taken as the input for source space modeling. This involves generating individual head meshes describing the boundaries of different head compartments based on the subject's structural scan. A template head model can also be used in case a structural scan is not available which results in a precise head model. The resultant cortical mesh describes the locations of the sources of the MEG signal and can be specified to have different resolution. A "normal" mesh containing 8196 vertices is generally used for reasons of computational efficiency.

2.3.4.2 Coregistration

The coordinate space in which the MEG sensors are specified need to match the coordinate system of the corresponding structural MRI image (or MNI space) in order to make accurate interpretations about the sources of brain activity. Coregistration involves linking these two coordinate systems via a set of three anatomical landmarks (or fiducials) whose coordinates are known in both systems. These fiducial points include the left and right preauricular points and the nasion. Essentially, this step requires specifying the points in the structural image that correspond to the MEG fiducials.

2.3.4.3 Forward modeling

This step involves generating a forward model that captures the effect of the dipoles (on the cortical mesh) at the level of the sensors. The

result is specified as a matrix which has *N* sensors and *M* mesh vertices. Each column is called the "lead field" matrix corresponding to one mesh vertex. A number of forward models can be specified and for MEG, a single shell model is typically used. The lead field matrices are used for subsequent inversion of the data.

2.3.4.4 Inverse reconstruction

Here, an imaging approach based on the IID model based on classical minimum norm was used which assumes that out of all possible source configurations that can explain the measured data, the configuration with the minimum overall source power represents the most optimal solution, i.e., it assumes that the brain is an efficient machine and makes optimal use of its energy resources. Time window and frequency range of interest can be specified to localize evoked or induced power. Spatial priors can also be specified to simplify the model and the relative accuracy of each model can be determined through Bayesian model comparison.

Chapter 3. PSYCHOPHYSICS

Summary

In contrast to the complex acoustic environments we encounter in everyday life, research in auditory scene analysis is generally based on relatively simple signals such as the streaming paradigm. Study 1 presents a new synthetic stimulus designed to examine the detection of coherent patterns ("figures") from overlapping "background" signals. The stimulus incorporates stochastic variation of the figure and background that captures the rich spectrotemporal complexity of natural acoustic scenes. Figure and background signals overlap in spectrotemporal space, but vary in their statistics of fluctuation and the only way to extract the figure is by integrating the patterns over frequency and time. A series of behavioural experiments are reported which demonstrate that human listeners are remarkably sensitive to the emergence of such figures and can tolerate a variety of spectral and temporal perturbations. This robust behaviour is consistent with the existence of automatic auditory segregation mechanisms that are highly sensitive to correlations across frequency and time.

3.1 Introduction

This study considers the behavioural bases of segregation in a novel stochastic figure-ground (SFG) stimulus that is more representative of natural acoustic environments which consist of multiple sound sources, such as a busy street market or an orchestral performance. Although we do it effortlessly, the separation of such mixtures of sounds into perceptually distinct sound sources is a highly complex task. In spite of being a topic of intense investigation for several decades, the neural bases of auditory object formation and segregation still remain to be fully explained (Cherry 1953; McDermott, 2009; Griffiths et al., 2012).

The most commonly used signal for probing auditory perceptual organization is a sequence of two pure tones alternating in time that, under certain conditions, can "stream" or segregate into two sources (van Noorden, 1975; Bregman, 1990). Much research based on these streaming signals has been performed to elucidate the neural substrates and computations that underlie auditory segregation (Moore et al., 2012; Snyder et al., 2012; Denham and Winkler, 2013). A prominent model of auditory stream segregation was proposed by Fishman and colleagues who recorded multi-unit activity from the auditory cortex of macaques in response to a simple streaming sequence (Fishman et al., 2001, 2004). For large frequency differences and fast presentation rates, which promote two distinct perceptual streams, they observed spatially segregated responses to the two tones. This pattern of segregated cortical activation, proposed to underlie the streaming percept, has since been widely replicated (e.g. Bee and Klump,

2004, 2005; Gutschalk et al., 2005, 2007; Micheyl et al., 2005; Wilson et al., 2007; Bidet-Caulet et al., 2007; Dykstra et al., 2011) and attributed to basic physiological principles of frequency selectivity, forward masking and neural adaptation (Fishman et al., 2001; Micheyl et al. 2007a; Fishman and Steinschneider, 2010a). These properties are considered to contribute to streaming by promoting the activation of distinct neuronal populations in the primary auditory cortex (PAC) that are well separated along the tonotopic axis (Fishman et al., 2001; Carlyon, 2004; Micheyl et al., 2007a). Human imaging studies that directly correlated the perceptual representation of streaming sequences with brain responses also support the correspondence between the streaming percept and the underlying neural activity in PAC (Gutschalk et al., 2005; Snyder et al., 2006; Wilson et al., 2007, but see Cusack, 2005). However, similar effects have also been shown at the level of the cochlea (Pressnitzer et al., 2008), suggesting that segregation might occur earlier in the ascending auditory pathway rather than be achieved in the auditory cortex (Hartmann and Johnson, 1991; Beauvois and Meddis, 1991, 1996; Denham and McCabe, 1997).

A prominent shortcoming, however, of the streaming stimulus is that it uses relatively simple, temporally regular narrowband signals that do not model the rich spectrotemporal diversity of natural sound environments. To overcome these limitations, a more spectrally rich signal referred to as the "informational masking" (IM) stimulus (Neff and Green, 1987; Kidd et al., 1994, 1995; Durlach et al., 2003) has been examined by several groups. IM refers to a type of non-energetic non-peripheral masking that is associated with an increase in detection thresholds due to stimulus uncertainty and

target-masker similarity (Pollack, 1975; Durlach et al., 2003). These multi-tone masking experiments were based on the detection of tonal targets in the presence of simultaneous multi-tone maskers, often separated by a "spectral protection region" (a certain frequency region around the target with little masker energy) that promoted the perceptual segregation of the target from the masker tones (Neff and Green, 1987; Kidd et al., 1994, 2003, 2011). Results from such experiments suggest that performance significantly depends on the width of the spectral protection region (Micheyl et al, 2007b; Elhilali et al, 2009b), and has been hypothesized to rely on the same adaptation-based mechanisms as proposed in the context of simple streaming signals (Micheyl et al., 2007b; Fishman and Steinschneider, 2010a).

However, the sounds that we are generally required to segregate are distinct from the narrowband signals used in streaming and IM stimuli and are often broadband with multiple frequency components that temporally overlap with other signals. Indeed, the ability of models inspired by such simplistic paradigms to describe stream segregation is currently under debate. A new model of scene analysis was recently proposed based on the demonstration that when the two tones in a streaming signal are presented synchronously, listeners perceive the sequence as one stream irrespective of the frequency separation between the two tones (Elhilali et al., 2009a), a result that is inconsistent with the predictions based on adaptation-based models (Fishman et al., 2001; Micheyl et al., 2005). At the neural level, there was no difference in responses to the synchronous and alternating sequence of tones that still resulted in different perceptual states. The

authors suggested that in addition to separation in acoustic features (e.g. pitch, intensity, spatial location), "temporal coherence" between the elements that comprise a scene is essential for segregation such that temporally incoherent patterns lead to a segregated percept whilst temporal coherence promotes an integrated percept (Shamma et al., 2011, 2013; also see Fishman and Steinschneider, 2010b; Micheyl et al., 2013a, b).

In this study, a novel stimulus (Stochastic Figure-Ground; SFG) is introduced that consists of coherent ("figure") and randomly varying ("background") components that overlap in spectrotemporal space and vary in their statistics of fluctuation (see Figure 3.1). The frequency components that comprise the figure vary from one chord to another so that it can only be extracted by integrating across both frequency and time dimensions (see section 3.2.1 for more details). The insertion of a brief figure embedded in the random tonal background was used to simulate perception of a coherent auditory object in a noisy acoustic environment. A number of behavioural experiments were performed where two spectrotemporal dimensions of the figure were manipulated – the "coherence" or the number of repeating frequency components, and the "duration" or the number of chords present in the figure.

Psychophysics was used to investigate listeners' ability to detect the complex figures and test the segregation performance in the context of several spectral and temporal perturbations. The results demonstrate that listeners are remarkably sensitive to the emergence of such figures (see

Figure 3.2) and can withstand a number of spectrotemporal manipulations designed to potentially impair spectrotemporal integration (see Figure 3.3).

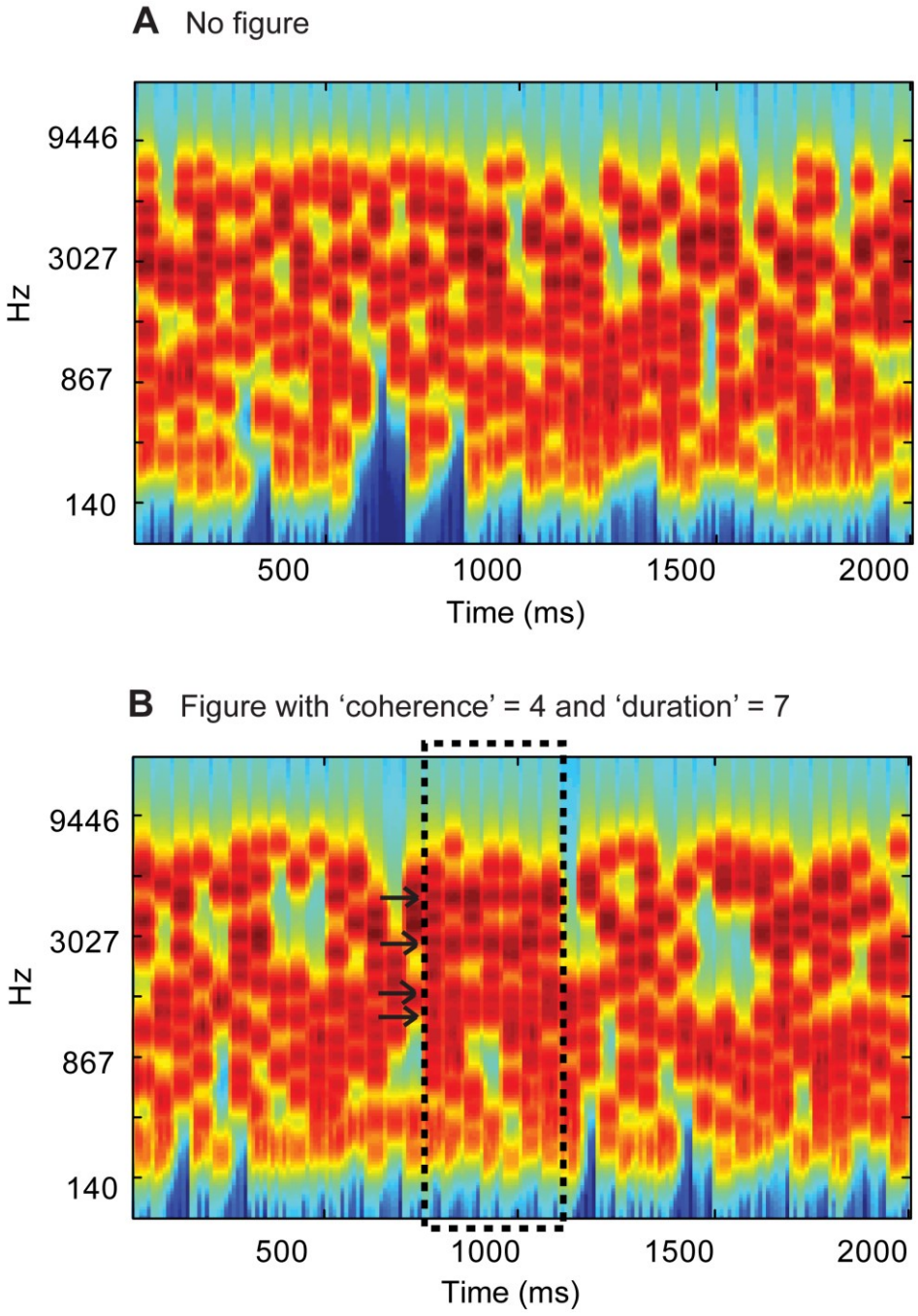
A No figure
9446
3027
867
140
Hz
500
1000
1500
2000
Time (ms)
B Figure with 'coherence' = 4 and 'duration' = 7
9446
3027
867
140
Hz
500
1000
1500
2000
Time (ms)

Figure 3.1: Stochastic figure-ground stimulus.

(A) Signals consisted of a sequence of 50-ms-long chords containing a random set of pure tone components.

(B) In 50% of the signals, a subset of tonal components repeated in frequency over several consecutive chords, resulting in the percept of a “figure” popping out of the random noise. The figure emerged between 15 and 20 chords (750 –1000 ms) after onset. The number of repeated components (the “coherence” of the figure) and the number of consecutive chords over which they were repeated (the “duration” of the figure) were varied as parameters. The plots represent auditory spectrograms, generated with a filter bank of 1/ERB (equivalent rectangular bandwidth) wide channels (Moore and Glasberg, 1983) equally spaced on a scale of ERB-rate. Channels are smoothed to obtain a temporal resolution similar to the equivalent rectangular duration (Plack and Moore, 1990).

3.2 Materials and Methods

3.2.1 Stochastic figure-ground stimulus

A novel synthetic stimulus was designed to model naturally complex situations characterized by a figure and background that overlap in feature space and are only distinguishable by their fluctuation statistics. Contrary to previously used signals, the spectrotemporal properties of the figure vary from one moment to another and the figure can only be extracted by binding the figure components across frequency and time.

Figure 3.1A shows the spectrogram of the SFG stimulus which consists of a sequence of random chords, each 50ms in duration with 0ms inter-chord-interval, presented for a total duration of 2000ms (40 consecutive chords). Each chord contains a random number (average: 10 and varying between 5 and 15) of pure tone components that are randomly selected from a frequency pool of 129 frequencies. These frequencies are equally spaced on a logarithmic scale between 179 and 7246 Hz such that the separation between successive components is equal to $1/24^{th}$ of an octave. The onset and offset of each chord are shaped by a 10ms raised-cosine ramp. In half of these stimuli, a random number of tones are repeated across a certain number of consecutive chords (e.g. in Figure 3.1B, four components marked by arrows repeat across seven chords) that results in the "pop-out" of the figure from the background. To eliminate correlation between the figure and background components, the figure was realized by first generating the random background and then adding additional, repeating components to the relevant chords. To avoid the confound that the

interval containing the figure might, on average, contain more frequency components, and to prevent listeners from relying on this cue, the remaining 50% of the stimuli (those containing no figure as in Figure 3.1A) also included additional tonal components, that were added over a number of consecutive chords (equal to the duration of the figure) at the same time as when a figure would have appeared. However, these extra components varied from one chord to the other and did not repeat to form a coherent pattern.

In the present study, the number of consecutive chords over which the tones were repeated ("duration") and the number of repeated frequency components ("coherence") was parametrically varied. The onset of the figure was jittered between 15-20 chords (750-1000ms) post stimulus onset.

3.2.2 Participants

All participants reported normal hearing and had no history of audiological or neurological disorders. Experimental procedures were approved by the research ethics committee of University College London (Project ID number: 1490/002), and written informed consent was obtained from each participant. For each experiment the number of listeners whose data are included in the final analysis is reported. In each experiment, a few listeners (2-3) were excluded because of their inability to reliably perform the task.

9 listeners (2 females; aged between 20 and 47 years; mean age: 26.9 years) took part in experiment 1. 9 listeners (6 females; aged between 22

and 28 years; mean age: 23.8 years) participated in experiment 2 based on the AXB design. 10 listeners (5 females; aged between 20 and 36 years; mean age: 25.7 years) took part in experiment 3. 10 listeners (5 females; aged between 23 to 31 years; mean age: 26.8 years) participated in experiment 6a. 27 listeners (Group 1: 9 listeners; 5 females, aged between 19 and 27 years; mean age: 21.1 years; Group 2: 10 listeners; 3 females; aged between 19 and 25 years; mean age: 21.3 years; Group 3: 8 listeners; 3 females; aged between 19 and 29 years; mean age: 22.4 years) participated in experiment 6b. 10 listeners (6 females; aged between 21-34 years, and mean age of 24.7 years) participated in experiment 4a with ramp step equal to 2 and another group of 10 listeners (3 females; aged between 20-30 years and mean age of 24.5 years) took part in experiment 4b with ramp step of 5. 10 listeners (5 females; aged between 22-31 years, mean age: 24.8 years) participated in experiment 5.

3.2.3 Stimuli

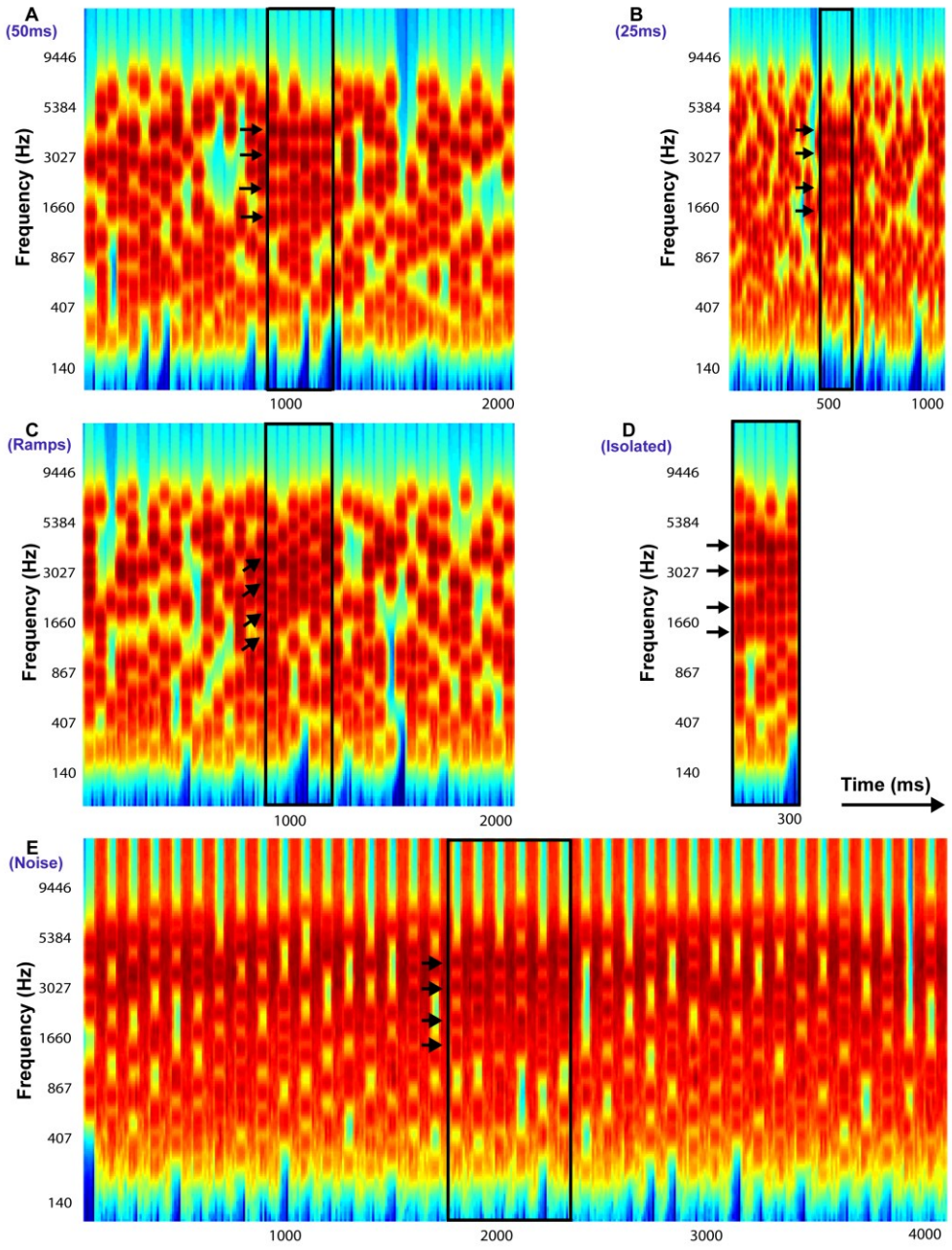
A
(50ms)
B
(25ms)
C
(Ramps)
D
(Isolated)
E
(Noise)
Frequency (Hz)
9446
5384
3027
1660
867
407
140
1000
2000
500
300
3000
4000
Time (ms)

Figure 3.2: Examples of Stochastic Figure-Ground stimuli.

All stimuli contain 4 identical frequency components (only for illustrative purposes: these were selected randomly in the experiments) with F_{coh} = 1016.7 Hz, 2033.4 Hz, 3046.7 Hz, and 4066.8 Hz repeated over 6 chords and indicated by the black arrows. The figure is bound by a black rectangle in each stimulus.

(A) **Chord duration of 50ms:** Stimulus comprises of 40 consecutive chords each of duration 50ms with a total duration of 2000ms.

(B) **Chord duration of 25ms:** Stimulus comprises of 40 consecutive chords each of duration 25ms with a total duration of 1000ms.

(C) **Ramped figures:** Stimulus comprises of 40 consecutive chords each of duration 50ms each (like A) but the frequency components comprising the figure increase in frequency in steps of 2*I or 5*I, where I = $1/24^{th}$ of an octave, represents the resolution of the frequency pool.

(D) **Isolated figures:** Stimulus comprises only of the "figure present" portion without any chords preceding or following the figure. The duration of the stimulus is given by the number of chords.

(E) **Chords interrupted by noise:** Stimulus comprises of 40 consecutive chords alternating with 40 chords comprising of loud, masking broadband white noise, each 50ms in duration. In experiment 6b, the duration of the noise was varied from 100ms to 500ms.

SFG stimuli in experiment 1 (figure 3.2A) consisted of a sequence of 50ms chords with 0ms inter-chord interval and 2 s duration (40 consecutive chords). The coherence of the figure varied between 1, 2, 4, 6 or 8 and the duration of the figure ranged from 2-7 chords. Stimuli for all combinations of coherence and duration (equal to 30) were presented in separate blocks where 50% of the trials (50 trials per block) contained a figure.

Stimuli for experiment 2 comprised 50ms chords with a coherence value of 6. Figure duration varied between 4, 8 and 12 (in separate blocks). Stimuli, all containing a figure, were presented in triplets as in an AXB design (e.g. Goldinger, 1998). The background patterns were different in all 3 signals but two of them (either A and X or B and X) contained identical figure components. Listeners were required to indicate the "odd" figure (A or B) by pressing a button. Three blocks of 60 trials each were presented for each duration condition.

Stimuli for experiment 3 were identical to those in experiment 1 except that the duration of each chord was reduced to 25ms (figure 3.2B). The coherence of the figure varied between 2, 4, 6 or 8 and the duration of the figure ranged from 2-7 chords resulting in a total of 24 blocks.

In experiments 4a and 4b, stimuli were similar to those in experiment 1 except that in this condition, the successive frequencies comprising the figure did not repeat from one chord to the next but rather increased in frequency across chords in steps of $2*I$ or $5*I$, where $I = 1/24^{th}$ of an octave is the resolution of the frequency pool used to create the SFG

stimulus (figure 3.2C). The coherence of the figure was 4, 6, or 8 and duration was 5, 7 or 9 chords resulting in a total of 9 blocks for each condition. In this experiment, however, the maximum duration of the figure (9 chords) is longer than the maximum duration of the figure in the remaining experiments (7 chords).

The stimuli for experiment 5 were identical to the stimuli used in experiment 1 except that they comprised of the figure chords only (3-7 chords or 150-350ms) without any chords that preceded or succeeded the figure as in previous experiments (figure 3.2D). The coherence of the figure was 2, 4, 6, or 8 chords and this resulted in a total of 20 blocks.

For experiment 6a, the stimuli were modified so that successive chords were separated by 50ms broadband noise bursts (figure 3.2E). The loudness of the noise was set to a level 12 dB above the level of the stimulus chords. The coherence of the figure was 2, 4, 6 or 8 and the duration of the figure ranged from 3-7 chords resulting in a total of 20 blocks.

In experiment 6b, the stimuli were identical to the previous experiment 6a except for the following differences: (a) coherence and duration were fixed at a value of 6; (b) the duration of the noise was varied in three different experiments in increasing order: group 1: 50, 100, and, 150ms; group 2: 100, 200, and, 250ms; group 3: 100, 300, and, 500ms respectively. The 100ms condition was chosen as an anchor and only those participants who performed above a threshold of d'=1.5 in this condition were selected for the whole experiment.

3.2.4 Procedure

Prior to the study, participants received training that consisted of listening to trials with no figures, easy-to-detect figures (high coherence and duration), difficult-to-detect figures (low coherence and duration) and a practice block of fifty mixed trials. In the actual experiments, the value of coherence and duration was displayed before the beginning of each block and participants were instructed to press a button as soon as they heard a figure (for the brief figures used here, these sounded like a 'warble' in the on-going random pattern). Feedback was provided. Blocks with different values of coherence and duration were presented in a pseudorandom order. The participants self-paced the experiment and each experiment lasted approximately an hour and a half. The procedure was identical for all experiments.

3.2.5 Analysis

Participants' responses were measured in terms of sensitivity (d prime, or d') and hit rates are also reported for certain conditions as mean ± one standard error.

3.2.6 Apparatus

All stimuli were created using MATLAB 7.5 software (The Mathworks Inc.) at a sampling rate of 44.1 kHz and 16 bit resolution. Sounds were delivered diotically through Sennheiser HD555 headphones (Sennheiser, Germany) and presented at a comfortable listening level of 60 to 70 dB SPL that was adjusted by each listener. Stimuli were presented

using Cogent (http://www.vislab.ucl.ac.uk/cogent.php). Listeners were tested individually in an acoustically shielded sound booth. The apparatus was identical for all experiments.

3.3 Results

3.3.1 Experiment 1: Chord duration of 50ms

In experiment 1, the basic SFG stimulus sequence was used to assess figure-detection (figure 3.1). Listeners' responses were analyzed to obtain d' for each combination of coherence and duration of the figure. The results (figure 3.3) show a clear effect of increasing performance with higher coherence and duration values. Hit rates (not shown) also mirrored d' with listeners achieving mean hit rates of 93 ± 2% for the most salient coherence/duration combination. It is important to note that the figure patterns were very brief (longest figure duration was 7 chords or 350ms), yet high levels of performance were observed (and with minimal practice). This is consistent with the idea that the SFG stimulus taps low-level, finely tuned segregation mechanisms.

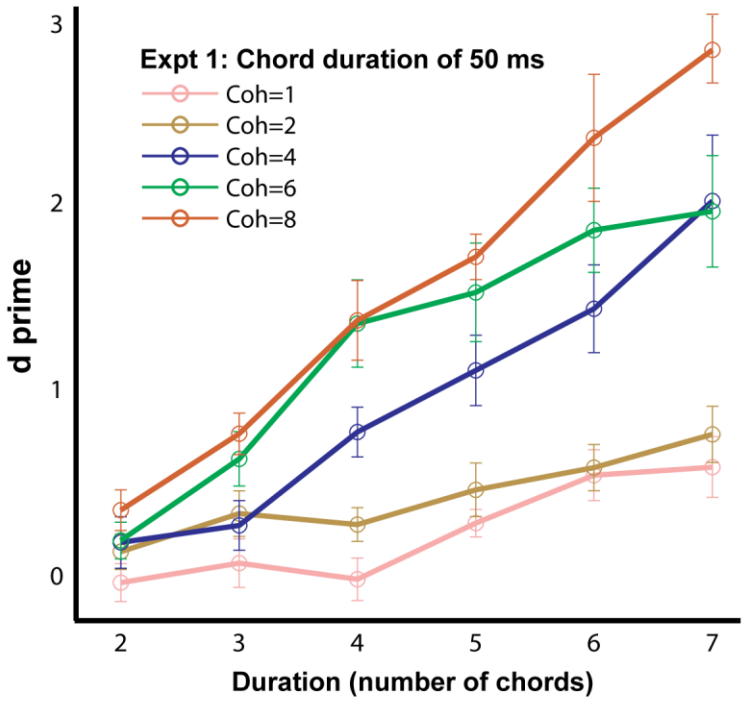

Figure 3.3: Behavioural performance in experiment 1.

The d' for experiment 1 (n = 9) are plotted on the ordinate and the duration of the figure (in terms of number of 50ms long chords) is shown along the abscissa. The coherence of the different stimuli in experiment 1 is colour coded according to the legend (inset). Error bars signify one standard error of the mean (SEM).

3.3.2 Experiment 2: Figure identification

What underlies such robust sensitivity to brief figure patterns? Since "figure-absent" and "figure-present" signals were controlled for overall number of components (see section 3.2.1), a global power increase associated with the emergence of the figure can be discounted as a potential cue. However, it is possible that the listeners' judgments are based on other features within the stimulus, such as the emergence of a figure might be associated with a change in the temporal modulation rate of a few frequency channels. The aim of experiment 2 was to investigate whether the detection of figures involves a specific figure-ground decomposition, namely whether the figure components are grouped together as a detectable "perceptual object" distinct from the background components, or whether listeners were rather detecting some low-level changes within the stimulus. To address this issue, stimulus triplets with different background patterns were created in which each stimulus contained a figure but the figure components were identical in two out of the three signals. Listeners were required to detect an "odd" signal that contained a figure that was different from the identical figure present in the other two signals in this AXB psychophysical paradigm. Results are shown in figure 3.4 and indicate that for the very short figure durations (4 chords, or 200ms) listeners had difficulty in discrimination ($d' = 0.31 \pm 0.18$; not significantly different from 0: $p = 0.12$, $t = 1.72$), but that performance increased significantly for a longer figure duration of 8 chords (400ms; $d' = 1.75 \pm 0.34$) and reached ceiling for a figure duration of 12 chords (600ms; $d' = 2.93 \pm 0.26$). These results indicate that figure detection in these stimuli may be associated with a

segregation mechanism that groups coherent components together as a distinct perceptual object.

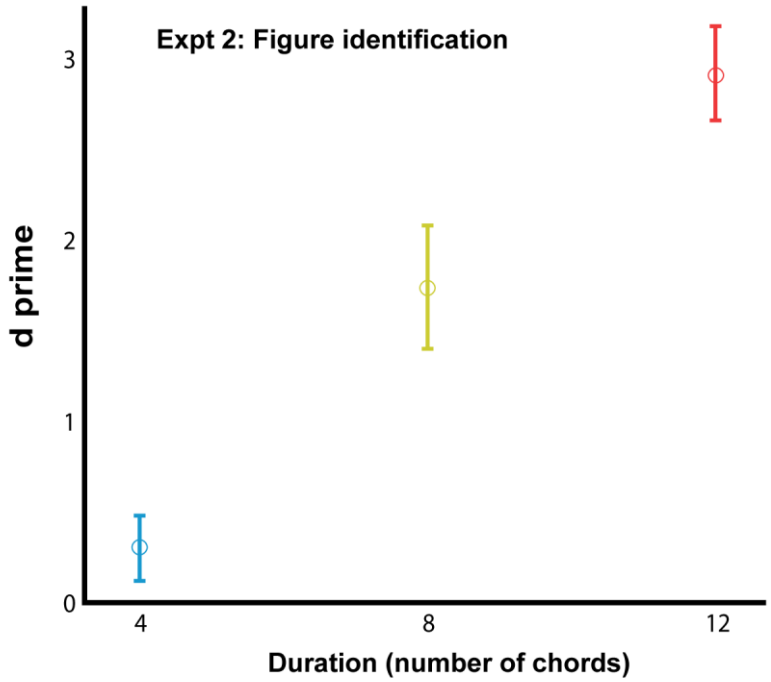

Figure 3.4: Behavioural performance in experiment 2.

The d' for experiment 2 (n = 9) are plotted on the ordinate and the duration of the figure (in terms of number of 50ms long chords) is shown along the abscissa. The coherence of the stimuli was fixed (equal to 6) and three different levels of duration were tested. The AXB figure identification task was different from the single interval alternative forced choice experiments: listeners were required to discriminate a stimulus with an "odd" figure from two other stimuli with identical figure components. Error bars signify one SEM.

3.3.3 Experiment 3: Chord duration of 25ms

In experiment 3, the duration of each chord was halved to 25ms, thereby reducing the corresponding durations of the figure and the stimuli (figure 3.2B). Here, the aim was to test whether figure-detection performance would be affected by such temporal scaling, i.e., whether figure-detection would vary as a function of the total duration of the figure (twice as long in experiment 1 vs. experiment 2) or the number of repeating chords that make up the figure (same in experiments 1 and 2).

Behavioural results (figure 3.5A) reveal good performance, as in experiment 1. Listeners achieved mean hit rates of 92 ± 3% for the highest coherence/duration combination used. An ANOVA with coherence and duration as within-subject factors and chord duration (50ms vs. 25ms) as a between-subject factor revealed no significant effect of condition ($F_{1,15} = 2$; $P = 0.174$), suggesting that performance relies on the number of repeating chords irrespective of their duration.

3.3.4 Experiment 4: Ramped figures

In the preceding experiments, figure components were identical across a certain number of chords. In experiment 4, figure components were manipulated such that instead of repeating across chords they were ramped, i.e., increasing in frequency from one chord to the next (figure 3.2C). The components in the frequency pool used to generate the SFG signals are separated equally by $1/24^{th}$ of an octave; and in the following two experiments, the frequency steps from one chord to the next were increased by 2 times (Experiment 4A; figure 3.5B – thick lines) or 5 times

(experiment 4B; figure 3.5B – thin lines) the frequency resolution (i.e., $2/24^{th}$ octave and $5/24^{th}$ octave respectively).

Robust performance was observed (maximum hit-rates of 0.97 and 0.83 were obtained for figures with coherence equal to 8 and duration equal to 7 for the two ramp levels of 2 and 5 respectively) and a comparison with experiment 1 using an ANOVA with coherence and duration as within-subject factors and stimulus type (repeating vs. ramp size 2 vs. ramp size 5) as a between-subject factor revealed a significant effect of condition: $F_{2,25} = 19$; $P < 0.001$.

Performance was found to be significantly worse for the ramp = 5 vs. ramp = 2 condition ($F_{1,18} = 21$, $P < 0.001$), but, remarkably, listeners exhibited above-chance performance even for the steeper slope condition. This suggests that the underlying segregation mechanisms are more susceptible to spectral than temporal perturbations (as in experiments 3, and 6 below) but can still integrate over dynamically changing, rather than fixed, figure components.

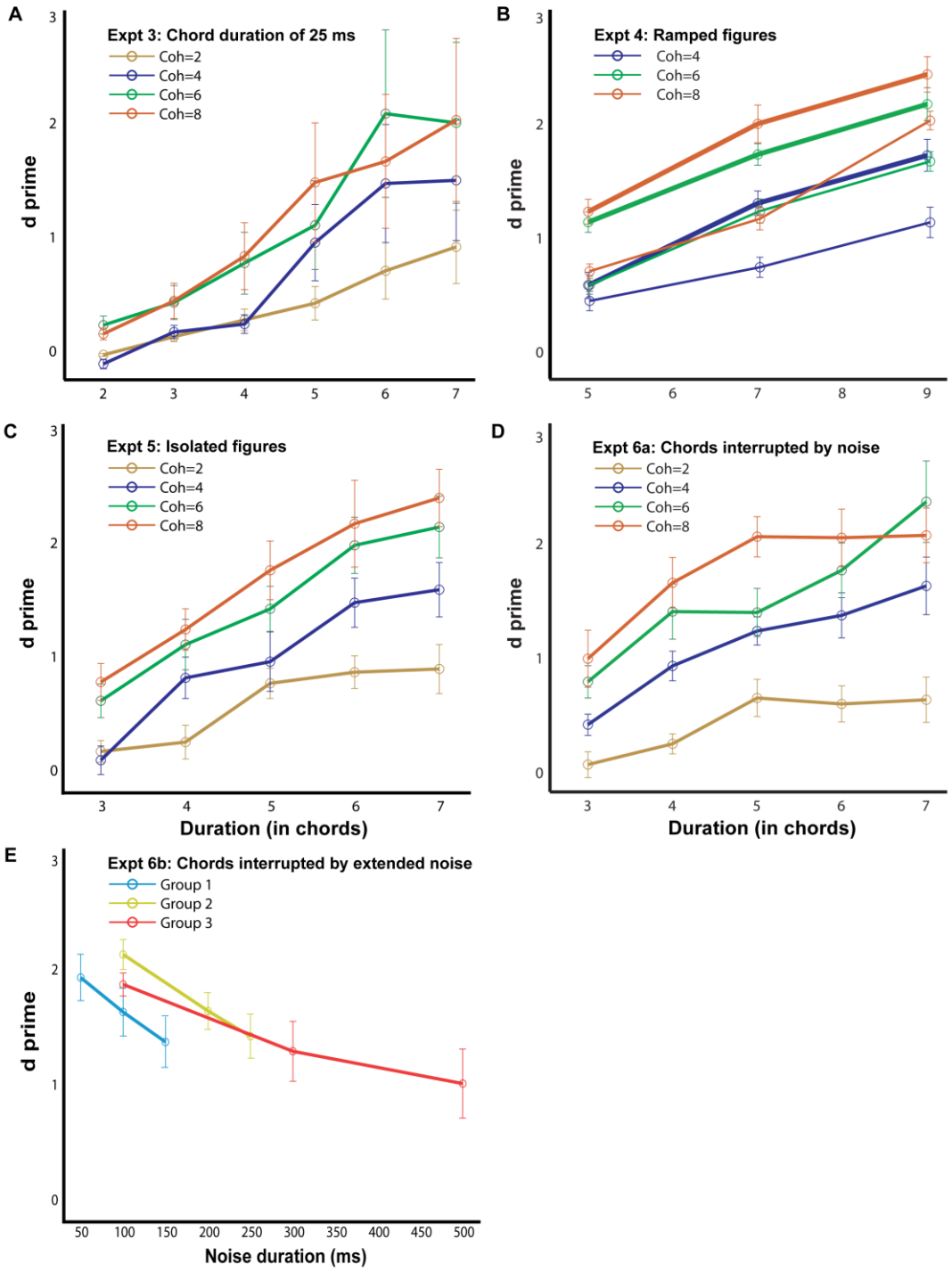
A
Expt 3: Chord duration of 25 ms
Coh=2
Coh=4
Coh=6
Coh=8
d prime
B
Expt 4: Ramped figures
Coh=4
Coh=6
Coh=8
d prime
C
Expt 5: Isolated figures
Coh=2
Coh=4
Coh=6
Coh=8
d prime
Duration (in chords)
D
Expt 6a: Chords interrupted by noise
Coh=2
Coh=4
Coh=6
Coh=8
d prime
Duration (in chords)
E
Expt 6b: Chords interrupted by extended noise
Group 1
Group 2
Group 3
d prime
Noise duration (ms)

Figure 3.5: Behavioural performance in experiments 3-6.

The d' for experiment 3, 4a (thick lines; ramp step = 2), 4b (thin lines, ramp step = 5), 5, 6a and 6b are shown here, as labelled in each figure (n = 10 for all conditions). The abscissa represents the duration of the figure (Figures 3.5A – 3.5D) and the duration of the masking noise in Figure 3.5E. Note that the maximum duration value in experiments 4a and 4b is larger (9 chords) than in the other experiments. Error bars signify one SEM.

3.3.5 Experiment 5: Isolated figures

Previous experiments consisted of stimuli that comprised a sequence of "background-only" chords, prior to the onset of the figure, and another sequence of "background-only" chords after figure offset. From first principles, segregation could be realized by adaptation to the ongoing background statistics and detection of the figure as a deviation from this established regular pattern. In order to test this hypothesis, in experiment 5, the "background" chords which preceded the occurrence of the figure were removed (figure 3.2D). The stimulus consisted simply of the chords that defined a brief figure of duration between 3-7 chords.

Similar to previous experiments, the results (figure 3.5C) show a strong effect of the coherence and duration, and performance improved with increasing salience of the figures with listeners achieving average hit rates of 89 ± 5% for the most salient condition. To compare behavioural performance with respect to experiment 1, an ANOVA with coherence and duration as within-subject factors and experimental condition (with background vs. no background) as a between-subject factor was used. This yielded no significant effect of condition: $F_{1,16} = 0.033$; $P = 0.859$, suggesting that the "background-only" chords which preceded the figure did not affect performance.

3.3.6 Experiment 6a: Chords interrupted by noise

Experiment 6 consisted of stimuli that contained 50ms of loud, broadband masking noise between successive 50ms long SFG chords (see Figure 3.2E), in an attempt to disrupt binding of temporally successive

components. If figure detection is accomplished by low level mechanisms that are sensitive to a power increase within certain frequency bands, the addition of the noise bursts would disrupt performance by introducing large power fluctuations across the entire spectrum, thus reducing the overall power differences between channels.

The results show decent behavioural performance (maximum hit rate of 0.93 was obtained for the most salient condition) which varied parametrically with the coherence and duration of the figure (figure 3.5D). An ANOVA with coherence and duration as within-subject factors and experimental condition (50ms repeating chords vs. 50ms chords alternating with white noise) as a between-subject factor revealed no significant effect of condition ($F_{1,17} = 0.004$; $P = 0.953$). Thus, interleaving the noise bursts between successive chords did not affect detection of the figures.

3.3.7 Experiment 6b: Chords interrupted by extended noise

A natural question that arises from the preceding experiment is – what are the temporal limits of segregation in SFG stimuli with interleaved white noise? To answer this question, the duration of the intervening noise bursts between stimulus chords was gradually varied in steps in three related experiments with different durations of noise for a particular combination of coherence (6) and duration (6). Results (figure 3.5E) indicate robust performance for all durations of noise up to 300ms and surprisingly, supra-threshold performance (d' = 1.00 ± 0.30; significantly different from 0: p = 0.01; t = 3.29) even for a noise duration of 500ms. This remarkable ability of listeners to integrate coherent patterns over 3s long (in the case of 500ms

noise bursts) suggests that higher-order mechanisms may be involved that are robust over such long time windows.

Temporal windows of integration, as long as 500ms, have rarely been reported in the context of auditory object formation. The results suggest the existence of a central mechanism that integrates the repeating pure tone components as belonging to a distinct object over multiple time scales. The long temporal windows of integration implicate cortical mechanisms at the level of the primary auditory cortex or beyond.

3.4 Discussion

In this study, a new stochastic figure-ground stimulus is introduced to examine segregation in complex acoustic scenes. Conceptually similar to the Julesz texture patterns (Julesz, 1962), the figure and background signals are indistinguishable at each instant in time and can be segregated only by integrating the patterns over frequency and time. An important perceptual characteristic of the SFG stimulus is the rapid buildup. For coherence levels of four components and above, as few as seven consecutive chords (a duration of 350ms) were sufficient to reach ceiling detection performance (Figure 3.3). This is in contrast to the much longer buildup times reported in streaming (~ 2000ms; Anstis and Saida, 1985; Micheyl et al., 2007b; Pressnitzer et al., 2008) and IM experiments (> 2s; Gutschalk et al., 2008) attributed to prolonged accumulation of sensory evidence, possibly requiring top-down mechanisms (Denham and Winkler, 2006). The shorter buildup times observed for SFG signals suggest that segregation may rely on partially different mechanisms from those that mediate streaming (Sheft

and Yost, 2008). All of these features make the SFG stimulus an interesting complement to streaming signals, with which to study pre-attentive auditory scene analysis.

The SFG stimulus represents a significant advantage over related IM stimuli that also consist of a number of components over a wide frequency range. However, the IM stimuli usually feature a band-stop ("protective") region surrounding the target tone to reduce (energetic) masking by the masker tones. Thus, the target channel will be excited but the neighboring channels will not be activated, and this could potentially provide a cue to the presence of a target in that frequency channel. Indeed, it has been shown that detection of targets in IM stimuli varies as a function of the width of the spectral protective region (Kidd et al., 1994; Micheyl et al., 2007b). Gutschalk and colleagues (2008) showed that the probability of correctly identifying the target within such signals increases with time, with d' ~2 achieved after several seconds of stimulation.

On the other hand, target detection in the SFG stimuli was found to be quick with minimal training (Experiment 1). These data also reveal the sensitivity of figure-detection to the underlying spectrotemporal characteristics, i.e., the coherence and duration of the figure. Performance increased monotonically with increasing number of components that comprised the figure as well as the duration of the figure. The SFG stimulus thus provides a convenient handle to assess the relative effects of spectral and temporal features by manipulating the coherence and the duration of the figure respectively.

Experiment 2 required more sensitive figure-ground discrimination abilities: out of three stimuli, listeners were required to identify the signal with an "odd" figure whilst the other two figures were identical. For a fixed coherence value (6 components), it was found that discrimination ability increased with the duration of the figure and listeners achieved significantly above-chance performance for a duration of 8 and 12 chords. These results suggest that the listeners were able to distinguish the figures as "perceptual objects" on the basis of high-level mechanisms that did not depend on differences in frequency or temporal modulation cues. Thus, target detection in the SFG stimulus truly represents a figure-ground discrimination task and not a simple feature discrimination task.

Experiments 3-6 examined discrimination abilities in the presence of a number of manipulations that modulated the temporal or spectral properties of the figure. It was found that the mechanism involved scales in time, in that detection depends on the number of components rather than their absolute duration (Experiment 3). Here, although the duration of each chord was reduced by half to 25ms (from 50ms chord length in experiment 1), there was no significant difference in performance. This chord presentation rate corresponds to 40Hz which is at the upper limits of temporal phase-locking values observed in the auditory cortex (Miller et al., 2002). It remains to be investigated, however, if figure-detection shows the same insensitivity to higher rates of presentation.

Experiment 4 involved a spectral perturbation where the slope of the figure components was manipulated. Instead of being linear and regularly

repeating across chords, the figures in this experiment comprised components that belonged to successively higher frequencies. Thus, the figure patterns formed linear (upward) ramps whose slope was varied in two separate experiments: the successive frequency steps were equal to $2/24^{th}$ and $5/24^{th}$ of an octave in experiments 4a and 4b respectively. Here, the figure-detection abilities were significantly impaired in comparison to detection of repeating figures in experiment 1 suggesting that the underlying mechanisms may be sensitive to the spectral shape of the patterns to be segregated. Although significantly worse performance was observed in comparison to experiment 1, the behaviour was still well above-chance suggesting that such patterns could still be reliably detected.

Results from experiment 5 with presentation of isolated figure components without any preceding or succeeding chords demonstrated rapid figure-detection abilities. Here, the duration of the stimulus was equal to the duration of the figure (ranging from 200 to 350ms) and similar performance to that in experiment 1 was observed. This suggests that the presence of the preceding stimulus chords which could possibly offer a predictive contextual cue is not crucial to segregation. These results point to the existence of a highly robust segregation mechanism that can operate over rather fast time scales.

Experiments 6a and 6b demonstrated the robustness of the mechanism to another type of spectrotemporal perturbation. Successive stimulus chords were interleaved with broadband white noise whose intensity was much higher than the SFG stimulus chords. In experiment 6a, the duration of the

noise was 50ms while it was gradually increased in a set of three following experiments from 100ms up to 500ms. The data obtained in experiment 6a revealed no significant difference in performance compared to experiment 1 suggesting that the insertion of noise had no major effect on discrimination. This suggests that the mechanism may be able to capture the temporal variation in the figure across noise segments and reject these as belonging to another (background) source. Subsequent experiments with longer duration of the intervening noise bursts were designed to explore the temporal limits of integration of such a high-level segregation mechanism. Even at the highest duration tested (500ms), performance was still found to be above-chance (d' ~1).

Overall, these set of psychophysical experiments point to the existence of a higher-order mechanism for sequential grouping that is clearly distinct from that proposed in the case of simple sequences of alternating tones (Fishman et al., 2001; Micheyl et al., 2007a). These results also advocate the use of complex signals such as the SFG stimulus has allowed a better understanding of the bases for segregation in realistic simulations of real-world sound scenes.

Chapter 4. TEMPORAL COHERENCE MODELING

Summary

In this study, the mechanistic bases of segregation in the stochastic figure-ground signals are examined. A number of features make the SFG stimulus different from the commonly used narrowband and temporally non-overlapping stimuli such as streaming sequences that have inspired a number of prominent models of auditory segregation. The broadband SFG stimulus, on the other hand contains overlapping frequency components that vary from one moment to another and figure-detection depends on integration over both frequency and time. The present models of stream segregation are not designed to explain segregation in such stochastic stimuli, thus necessitating the need for a novel conceptual framework for more complex signals. The temporal coherence model of auditory scene analysis (Shamma et al., 2011) which posits that temporally correlated elements bind together to form a single stream provides a potential solution. The SFG stimuli used in the behavioural experiments reported in chapter 3 were fed to the model and the resultant temporal coherence matrices were analyzed vis-à-vis the behavioural response patterns. A strong qualitative correspondence was found between the model simulations and behaviour thus supporting a role for temporal coherence as an organizational principle in auditory scene analysis.

4.1 Introduction

Natural sounds in the environment comprise of dynamic signals with temporally structured characteristics, i.e., they fluctuate at specific temporal rates. As a result, the salience of acoustic attributes of a source also varies similarly, be it pitch, intensity or location. Different sources are thus characterized with different temporal patterns of fluctuations that may serve as a cue to distinguish between them. This is related to the Gestalt principle of common fate, i.e., sounds that start and stop together belong to the same source. This is apparent in an orchestra, where a heterogeneous group of musical instruments is perceived as a single source due to the temporal correlations between the individual sources of music. Bregman uses the term 'sonic objects' to refer to groupings such as choirs or orchestras.

Standard models of sound segregation, however, propose that separation in "feature space" is essential for segregation (Fishman et al., 2001; Micheyl et al., 2005, 2007a; Fishman and Steinschneider, 2010a). Based on neurophysiological evidence, these models suggest that spatially segregated activation of brain areas that encode particular features such as pitch or intensity forms the basis of segregation. This has been suggested in a range of animal and human investigations of auditory streaming (Fishman and Steinschneider, 2010a; Snyder et al., 2012; Moore et al., 2012; Denham and Winkler, 2013).

A recent model of auditory scene analysis challenges these standard models and postulates that "temporal coherence" between sound tokens is essential for perceptual organization of the auditory environment (Shamma

et al., 2011). The model asserts that any sequence of temporally correlated acoustic features will bind together as a perceptual sound object and the lack of such temporal coherence provides the basis for segregation between two sound signals. This was demonstrated in the case of streaming signals which usually stream apart when the two constituent tones are well separated along the frequency dimension. However, when the two tones were made synchronous or temporally correlated, streaming was not observed as the two tones group together to form a complex that is perceived as a single stream even at large frequency separations (Elhilali et al., 2009a). Neural responses from ferret primary auditory cortex however did not show any difference between synchronous or alternating streaming signals, even though the two signals produced different perceptual reports (Elhilali et al., 2009a). These results present a convincing case against tonotopic separation as an essential correlate of stream segregation and highlight the importance of the temporal dimension. Recently, Micheyl and colleagues (2013a, b) further demonstrated that synchrony limits listeners' ability to perceive separate streams with reduced probability of segregated perceptual reports for synchronous compared to alternating tone sequences.

The potential use of temporal correlations as a cue for perceptual segmentation has been shown previously in a model of the auditory system (von der Marlsburg and Schreiner, 1986) where segregation was found to be dependent on the synchronous onset of the stimulus occurring independently in two input signals, leading to rapid and persistent decoupling of two coherent sets of neurons.

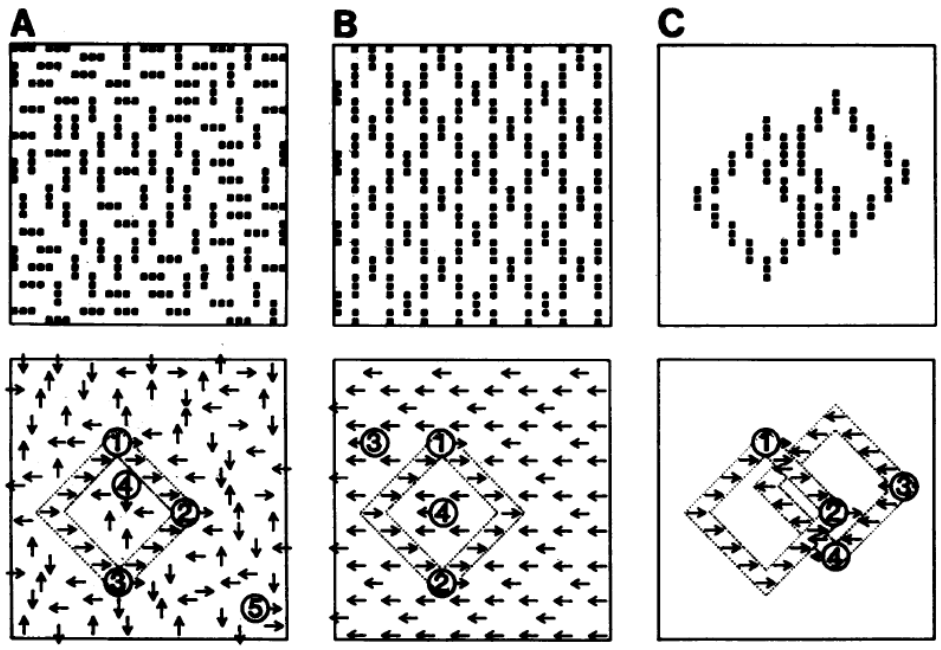

Figure 4.1: Visual figure-ground discrimination.

(Upper) Visual displays (pixel images) of oriented bars are shown.

(Lower) Direction of motion of the component bars is indicated by arrows; to help the reader, coherently moving ensembles of bars are enclosed by a stippled border. The stimuli consist of: a figure (diamond) coherently moving to the right and a background composed of bars moving in randomly selected directions (A); the same figure moving to the right and a coherent background moving to the left (B); two identical and overlapping, but differently moving figures (diamonds) (C). Figure reproduced from Sporns et al., 1991.

Temporal structure is not only important for auditory scene analysis; it is also a primary factor governing visual segmentation (Blake and Lee, 2005). Edelman and colleagues proposed that neurons responding to a particular object will be temporally correlated amongst themselves whilst being uncorrelated with neurons responding to other objects or to the background (Edelman, 1978; von der Malsburg, 1981). Sporns and coworkers (1991) developed a computational model to account for figure-ground segregation in visual scenes where a figure was defined on the basis of coherent motion of oriented bars as shown in figure 4.1. Based on Edelman's temporal correlation framework, Sporns et al. (1991) showed that the responses of the model were able to group elements corresponding to a coherent figure and segregate them from the background or another figure.

These considerations strongly suggest a role for temporal structure in the perceptual analysis of visual and acoustic environments. The SFG stimulus presented in chapter 1 represents one such complex signal that is conceptually similar to the visual coherent dot motion stimulus (see Figure 1.2; Shadlen and Newsome, 1996). The SFG signal consists of a series of chords with random pure tone components that vary from one chord to another. The perceptual target, i.e., the figure is defined on the basis of a certain number of frequency components that repeat synchronously over a certain number of chords, whilst the remaining channels contain random frequency components and are temporally uncorrelated. The principle on which the coherent figure (the constituent frequency components start and

stop at the same time) is defined suggests that the temporal coherence model (Shamma et al., 2011) may underlie segregation in the case of the complex SFG stimulus where previous models based on the streaming signals fall short (Fishman et al., 2001; Micheyl et al., 2005, 2007a).

In this study, the temporal coherence modeling framework is applied in the case of the SFG stimuli and it is examined whether temporal coherence between the frequency channels that comprise a figure corresponds with behavioural results from the experiments reported in chapter 1.

4.2 Temporal coherence model

The temporal coherence model is a spatiotemporal model which proposes that auditory stream segregation requires both separation in feature space and temporal incoherence between the responses of the corresponding channels. The model predicts that if the activity of auditory channels is positively correlated over time, then they define a single stream irrespective of the spatial distribution of the responses. On the other hand, channels that are uncorrelated or anti-correlated are assigned to different streams. This theory provides a general framework that can be applied to auditory dimensions other than frequency, such as intensity, spatial location and temporal modulations. The model, however, does not reject the importance of frequency selectivity (Fishman et al., 2001, 2004; Bee and Klump, 2004, 2005; Micheyl et al., 2005, 2007a). Sharp frequency selectivity is necessary for segregation: if the frequency tuning is broader than the frequency separation between two tones then this will always result in the perception

of a single stream. Thus, frequency separation is an important factor for segregation but not sufficient according to the coherence model.

The temporal coherence model consists of two distinct stages as shown in figure 4.2. The first stage analyzes the auditory spectrogram of the acoustic input and performs temporal integration through a bank of bandpass filters that are tuned to different physiologically plausible parameters that capture the rich variety of spectrotemporal receptive fields (STRFs) found in PAC (Chi et al., 2005; Elhilali and Shamma, 2008; Elhilali et al., 2009a, Shamma et al., 2011). STRFs summarize the response of a neuron to acoustic input and can be either broadly or sharply tuned. Auditory cortex contains a diverse range of STRFs that are tuned to a specific range of spectral resolutions (or "scales"; 0.125 to 8 cycles per octave) and a limited range of temporal modulations (or "rates"; 2-32 Hz).

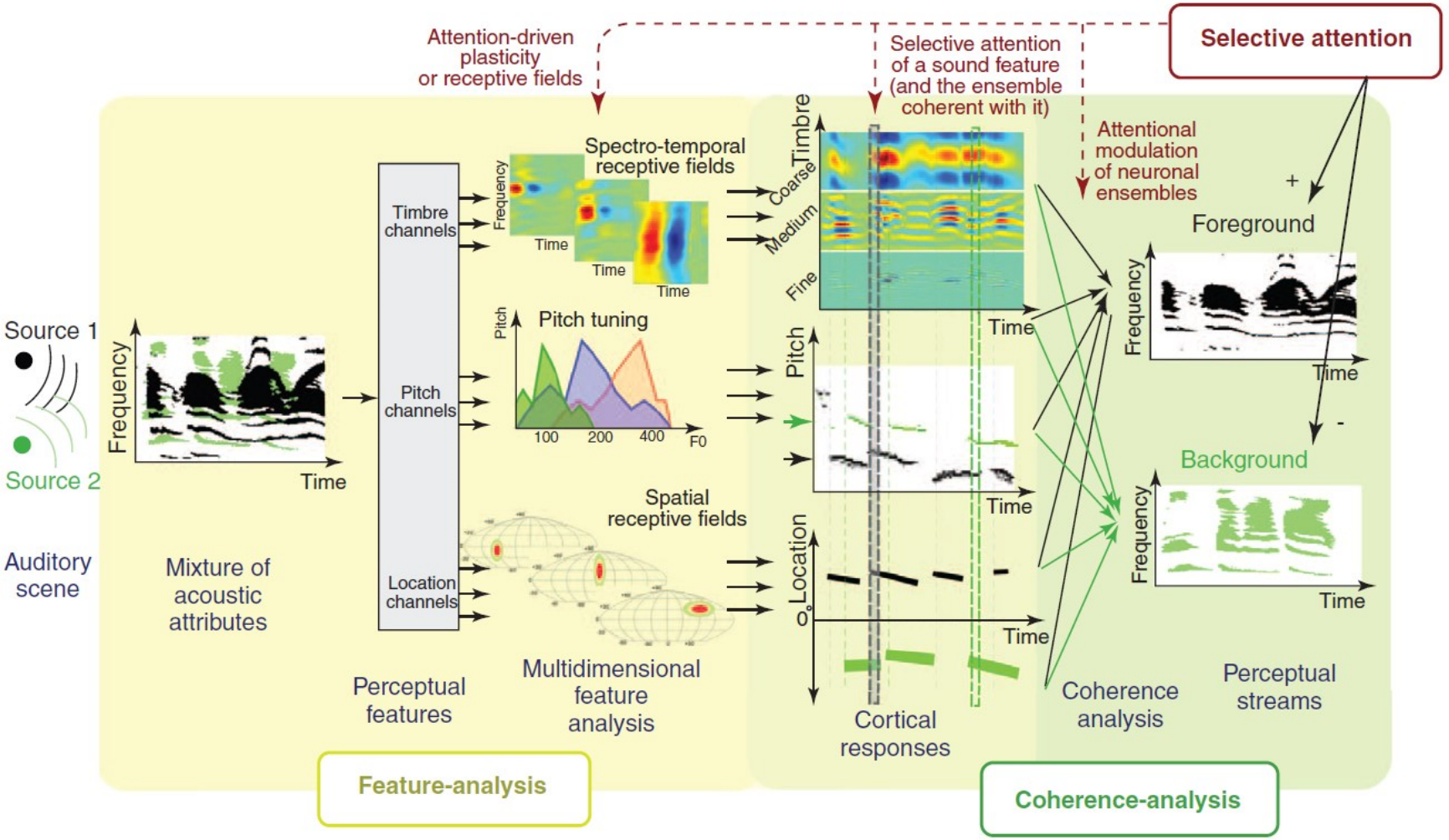
Selective attention
Attention-driven plasticity or receptive fields
Selective attention of a sound feature (and the ensemble coherent with it)
Attentional modulation of neuronal ensembles
Source 1
Source 2
Auditory scene
Frequency
Time
Mixture of acoustic attributes
Timbre channels
Pitch channels
Location channels
Perceptual features
Spectro-temporal receptive fields
Pitch tuning
Pitch
100
200
400
F0
Spatial receptive fields
Location
0°
Multidimensional feature analysis
Timbre
Coarse
Medium
Fine
Cortical responses
+
Foreground
-
Background
Coherence analysis
Perceptual streams
Feature-analysis
Coherence-analysis

Figure 4.2: A schematic of the temporal coherence model.

From left to right: Multiple sound sources constitute an auditory scene, which is initially analysed through a feature analysis stage. This stage consists of a cochlear frequency analysis followed by arrays of feature-selective neurons that create a multidimensional representation along different feature axes. The figure depicts timbre, pitch and spatial location channels. Note that for computational convenience and illustration purposes, these feature maps are shown with ordered axes when in fact such orderly representations are neither known nor are essential for the model. The outcome of this analysis is a rich set of cortical responses that explicitly represent the different sound features, as well as their timing relationships. The second stage of the model performs coherence analysis by correlating the temporal outputs of the different feature-selective neurons and arranging them based on their degree of coherence, hence giving rise to distinct perceptual streams. Complementing this feed-forward bottom-up view are top-down processes of selective attention that operate by modulating the selectivity of cortical neurons. This feature-based selective attention translates onto object-based attentional mechanisms by virtue of the fact that selected features are coherent with other features that are part of the same stream. Figure reproduced from Shamma et al., 2011.

The next level of the model incorporates a temporal coherence analysis stage which computes a "windowed" correlation between each pair of channels by taking the product of the filter outputs corresponding to the different channels. The coincidence analysis is performed over a range of time scales of the order of tens to hundreds of milliseconds that is consistent with experimental findings from cortical recordings (Kowalski et al., 1996; Miller et al., 2002). A dynamic coherence matrix which consists of the cross-correlation values as a function of time is obtained that represents the output of the model. The diagonal entries represent the mean power in the input channels and do not predict perceptual representation. The off-diagonal elements of the matrix indicate the presence (or absence) of coherence across different channels, and are predictive of the perceptual representation of the input stimulus.

4.3 Temporal coherence analysis of SFG stimuli

The temporal coherence model was run for a range of temporal modulation rates: 2.5, 5, 10 and 20Hz for experiments 1, 4, 5, and 6a and 5 respectively, and 10, 20 and 40Hz for experiment 3 (see section 3.2.3). Additionally, a rate of 3.33Hz corresponding to the rate of presentation of 300ms white noise segments was used in experiment 6b. These rates cover the range of physiological temporal modulation rates observed in the auditory cortex (Miller et al., 2002). A single spectral resolution scale of 8 cycles per octave (corresponding to the bandwidth of streaming; 4 cycles per octave for experiment 4b where larger frequency steps are required to extract a ramped figure) was used.

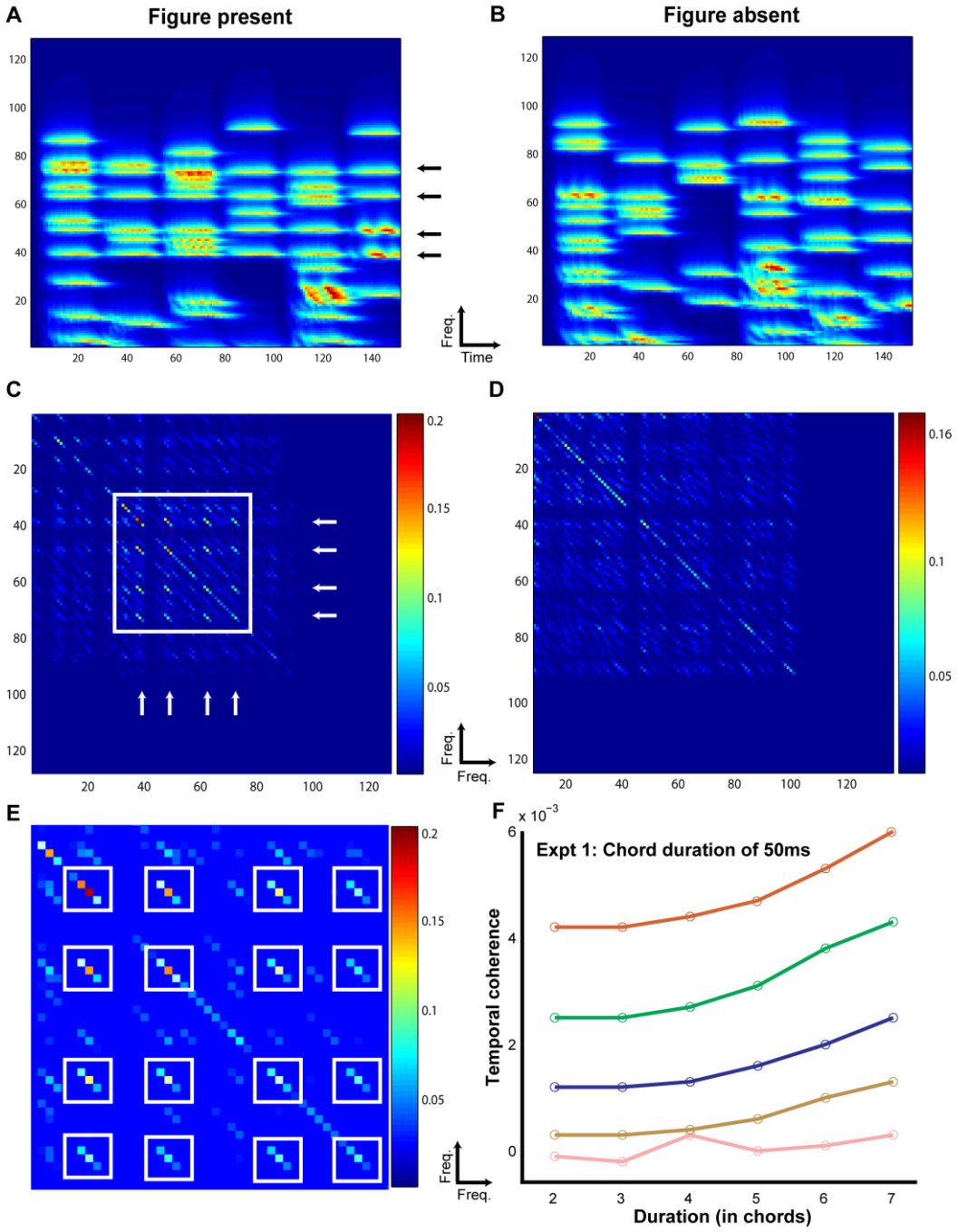
A
Figure present
B
Figure absent
Freq.
Time
C
D
Freq.
Freq.
E
Freq.
Freq.
F
x 10^-3
Expt 1: Chord duration of 50ms
Temporal coherence
Duration (in chords)

Figure 4.3: Temporal coherence modeling of the basic SFG stimulus.

The protocol for temporal coherence analysis is demonstrated here for experiment 5. The procedure was identical for modeling the other experiments. A stimulus containing a figure (here with coherence = 4) as indicated by the arrows (4.3A) and another, background only (figure absent) stimulus (4.3B) was applied as input to the temporal coherence model. The model performs multidimensional feature analysis at the level of the auditory cortex followed by temporal coherence analysis which generates a coherence matrix for each stimulus as shown in figures 4.3C and 4.3D respectively. The coherence matrix for the stimulus with figure present contains significantly higher cross-correlation values (off the diagonal; enclosed in white square) between the channels comprising repeating frequencies as indicated by the two orthogonal sets of white arrows in figure 4.3C. A magnified plot of the coherence matrix for the figure stimulus is shown in figure 4.3E where the cross-correlation peaks are highlighted in white boxes. The strength of the cross-correlation is indicated by the heat map next to each figure. The stimulus without a figure, i.e., which does not contain any repeating frequencies, does not contain significant cross-correlations. This process is repeated for 500 iterations (N_{iter}) for all combinations of coherence and duration. The differences between these two coherence matrices were quantified by computing the maximum cross-correlation for each set of coherence matrices for the figure and the ground stimuli respectively. Temporal coherence was calculated as the difference between the average maxima for the figure and the ground stimuli respectively. The resultant model response is shown for each combination of coherence and duration in figure 4.3F.

The analysis was conducted by entering the SFG stimulus for each experimental condition to the input stage of the model. For experiments 1 and 3, the entire stimulus was fed to the model input and for the remaining experiments a stimulus without the pre- and post-figure chords was entered. This was based on the prediction that the background chords before and after the figure onset contribute little to the cross-correlation matrix unlike the chords comprising the figure that contribute prominently to the net temporal coherence. The simulations were performed separately for the stimuli containing a figure and without a figure and repeated across five hundred iterations. To establish differences between the resultant coherence matrices, the maximum value of the cross-correlation across all time points was computed. This spectral decomposition helps to examine whether channels are correlated with each other (whereby the channels with repeating figure components could possibly be bound together as one object, or the "figure"), and not significantly correlated with each other (the channels with random correlation between channels may be perceived as belonging to the background). The difference in the average values of the maxima between the figure and the ground stimuli was calculated as the model response and plotted like the psychophysical curves (see figure 3.5) to obtain the model responses (see figures 4.3 and 4.4).

4.4 Results

It is difficult to account for listeners' performance in Experiments 1 and 2 based on the standard, adaptation-based models proposed in the context of the streaming paradigm (Hartmann and Johnson, 1991; Beauvois

and Meddis, 1991, 1996; Denham and McCabe, 1997; Fishman et al., 2001; Micheyl et al 2005, 2007a; Fishman and Steinschneider, 2010a). The figure and background in the SFG stimuli overlap in frequency space, thus challenging segregation based on activation of spatially distinct neuronal populations in PAC. Furthermore, the psychophysical data clearly indicate that performance strongly depends on the number of simultaneously repeating frequency components, suggesting a mechanism that is able to integrate across widely spaced frequency channels, an element missing in previous models based on streaming. Instead, the behavioural results are consistent with a temporal coherence model of segregation (Shamma et al., 2011).

The temporal coherence model is based on the idea that a perceptual "stream" emerges when a group of (frequency) channels are coherently activated against the backdrop of other uncorrelated channels (Shamma et al, 2011). In the SFG stimuli, the "figure" (defined by the correlated tones) perceptually stands out against a background of random uncorrelated tones. The temporal coherence model postulates that the figure becomes progressively more salient with more correlated tones in the different frequency channels. To measure this coherence, a correlation matrix across all channels of the spectrogram was computed. In principle, the correlation between the activations of any two channels at time *t* should be computed over a certain time window in the past, of a duration that is commensurate with the rates of tone presentations in the channels; this may range roughly between 2 and 40 Hz depending on the experimental session. Consequently,

to estimate the perceptual saliency of the figure segment in our stimuli, the correlation matrix was simultaneously computed for a range of temporal rates, and the largest correlation values are reported.

The computations incorporated a spectrotemporal analysis postulated to take place in the auditory cortex (Chi et al., 2005; Elhilali and Shamma, 2008). Specifically, temporal modulations in the spectrogram channels were first analyzed with a range of constant-Q modulation filters centred at rates ranging from 2 to 40 Hz (computing in effect a wavelet transform for each channel). The correlation matrix *at each rate* is then defined as the product of all channel pairs derived from the same rate filters. The maximum correlation values from each matrix were then averaged and was assumed to reflect the coherence of the activity in the spectrogram channels, and hence the saliency of the figure interval. Note that, as expected, the rate at which the maximum correlations occurred for the different experiments (reported in figures 3.5 and 4.4) approximately matched the rate of the tones presented during the figure interval.

Experiment 1

Figure 4.3 illustrates the modeling procedure and results for stimuli from Experiment 1 (see methods for details of the model). The model successfully accounted for the behavioural data in that, an average cross-correlation based measure was able to systematically distinguish "figure-present" from "figure-absent" (or background) stimuli in a manner that mirrored the behavioural responses. The model's measure of temporal

coherence showed a similar profile and increased with the coherence and the duration of the figure for the different experimental conditions (figure 4.4).

Experiment 3

Model predictions for experiment 3 that presented the SFG chords at a faster rate of 40 Hz (chord length was reduced to 25ms) were consistent with the experimental findings (figure 4.4A). Thus, correlations across the spectrogram channels remained significant, but now occurred at higher rates than in experiment 1 (40 Hz versus 20 Hz), reflecting the faster rate of tone presentations in the figure.

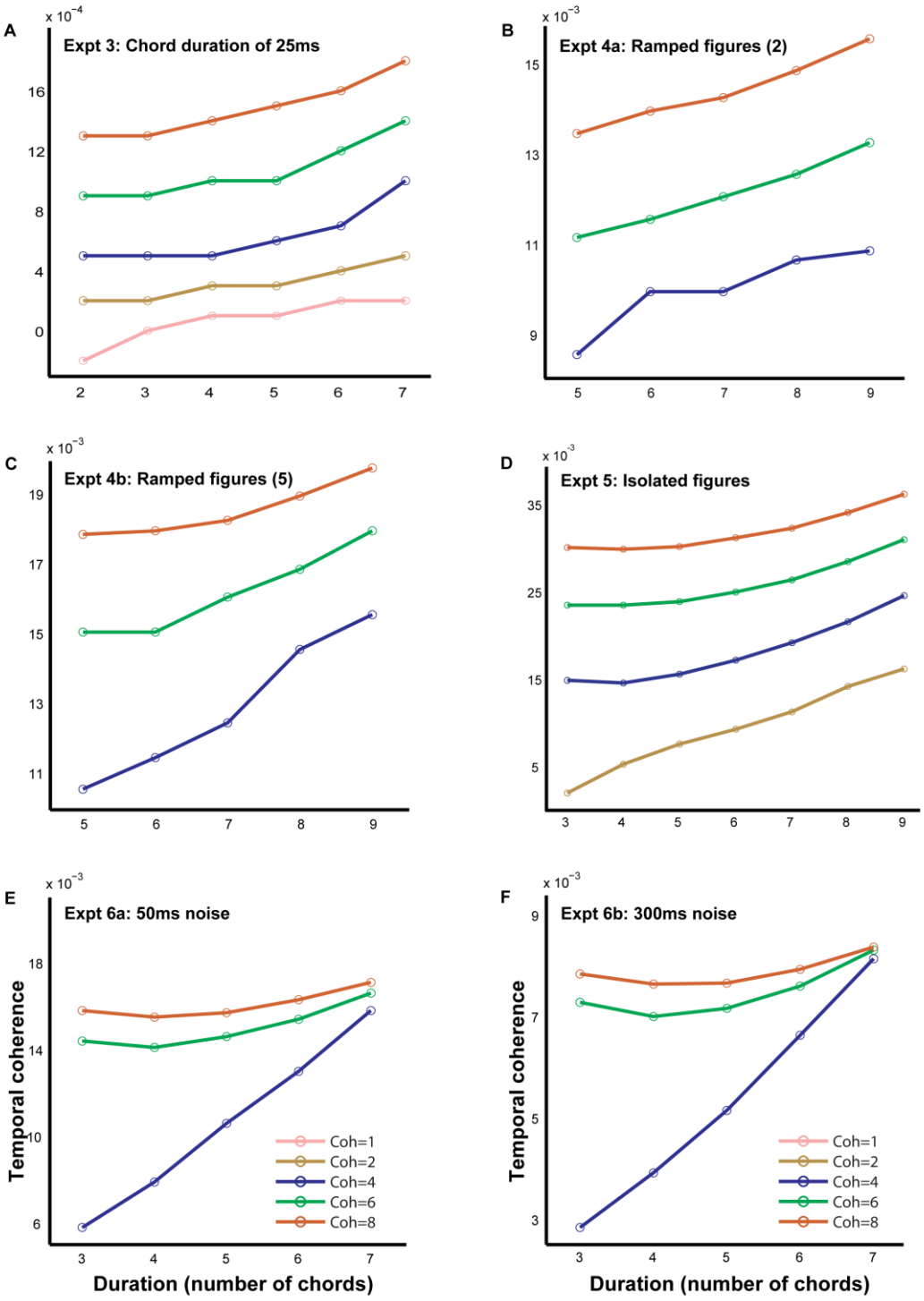
A
x 10^-4
Expt 3: Chord duration of 25ms
16
12
8
4
0
2
3
4
5
6
7
B
x 10^-3
Expt 4a: Ramped figures (2)
15
13
11
9
5
6
7
8
9
C
x 10^-3
Expt 4b: Ramped figures (5)
19
17
15
13
11
5
6
7
8
9
D
x 10^-3
Expt 5: Isolated figures
35
25
15
5
3
4
5
6
7
8
9
E
x 10^-3
Expt 6a: 50ms noise
Temporal coherence
18
14
10
6
3
4
5
6
7
Coh=1
Coh=2
Coh=4
Coh=6
Coh=8
Duration (number of chords)
F
x 10^-3
Expt 6b: 300ms noise
Temporal coherence
9
7
5
3
3
4
5
6
7
Coh=1
Coh=2
Coh=4
Coh=6
Coh=8
Duration (number of chords)

Figure 4.4: Temporal coherence modeling results for other SFG stimuli.

The output of the temporal coherence modeling procedure is shown for the remaining psychophysical experiments:

(A) Experiment 2 with 25ms chords modeled at a rate of 40 Hz;

(B, C) Experiments 4a and 4b with ramped figures with step size of 2 and 5 respectively modeled at a rate of 10 Hz;

(D) Experiment 5 with isolated 50ms chords modeled at a rate of 20Hz;

(E, F) Experiments 6a and 6b with chords interrupted by noise of duration 50ms and 300ms modeled at 20 and 3.33 Hz respectively.

Experiment 4

Experiment 4 consisted of SFG stimuli where the figure was defined on the basis of linear ramps, i.e., the successive channels that comprised the figure were not identical but rather increased in frequency in steps of 2/24th and 5/24th of an octave in two separate experiments. As with previous experiments, there were significant correlations among the channels predicting the saliency of the ramped figures. However, the optimal rate at which the correlations occurred here was slightly lower (at 10Hz; see figures 4.4B and 4.4C) than that observed in experiment 1 (20 Hz), perhaps because two 50ms chords are integrated as a single unit to define the ramp.

Experiment 5

Here, the basic SFG stimulus from experiment 1 was manipulated such that the chords that comprised the pre-figure and post-figure segments were removed and the figures were presented in isolation. Modeling for this experiment replicated the results of experiment 1 in that the correlations increased with the coherence and duration of the figure and showed maximum response at 20Hz (see figure 4.4D), corresponding to the rate of presentation of the chords comprising the figure.

Experiment 6

Experiment 6 measured figure-detection performance in a version of the SFG stimulus that comprised an alternating sequence of SFG chords and loud masking white noise. Model predictions in this experiment (see figures 4.4E) are broadly consistent with the findings in that detection became

easier with more coherent tones, and with longer figure intervals. The reason is simply because the noise weakens but does not eliminate the correlation amongst the tones, at least when computed at slower rates. Furthermore, the temporal correlations were found to be modulated even in experiment 6b where these are measured over longer windows (or slower rates - e.g. 3.33 Hz as in figure 4.4F).

4.5 Discussion

Models of auditory scene analysis have tried to explain segregation on the basis of peripheral channeling (Hartmann and Johnson, 1991; Beauvois and Meddis, 1991, 1996; Denham and McCabe, 1997); physiological principles of frequency selectivity, forward masking, and adaptation at the level of the auditory cortex that results in spatially segregated activation of neuronal ensembles (Fishman et al., 2001; Micheyl et al., 2005, 2007a; Fishman and Steinschneider, 2010a), and more recently on the basis of predictive coding mechanisms (Denham and Winkler, 2006; Winkler et al., 2009; Mill et al., 2013). Although these theories propound different principles to account for segregation, they have one common feature: all theories are based on a simple pattern of alternating tones that stream apart into separate perceptual streams following an initial percept of a single stream (van Noorden, 1975; Bregman, 1990; Moore et al., 2002, 2012). The proposed "central" models of segregation that invoke cortical mechanisms successfully account for several features of the streaming paradigm such as the buildup of streaming, and switching between perceptual states.

However, it is not known if these models can predict segregation in more complex signals other than streaming.

Frequency selectivity and tonotopic mapping form a strong element of these models. Shamma and colleagues (2011) proposed a new model of segregation that credits frequency selectivity as a significant factor but gives primary importance to "temporal coherence", i.e., the temporal relationship between sound tokens. The following sections describe the various aspects of the temporal coherence model.

4.5.1 Segregation based on temporal coherence

The temporal coherence model proposes that segregation is achieved not only on the basis of separation in feature-space but rather by the temporal relationship between different elements in the scene, such that temporally coherent elements are bound together, whilst temporally incoherent channels with independent fluctuation profiles are allocated to separate sources (Shamma et al., 2011). Specifically, the model involves two processing stages: firstly, a feature analysis stage that performs multidimensional feature analysis by distinct populations of neurons in the auditory cortex that are tuned to a range of physiologically relevant temporal modulation rates and spectral resolution scales (Chi et al., 2005; Elhilali and Shamma, 2008). Auditory features such as pitch, timbre and loudness are analyzed at this stage, and the output is fed to a second stage that computes temporal coherence.

Elhilali and colleagues (2009a) first demonstrated the relevance of temporal coherence for auditory streaming. In a psychophysical experiment, they showed that human listeners are more likely to report a percept of one stream when the tones of the streaming signal are made synchronous. This phenomenon was observed even if the frequency separation between the two tones was more than an octave. To investigate the neural bases of these findings, they performed recordings from ferret auditory cortex in response to the alternating and synchronous sequence of tones. They observed that the cortical responses to the two types of sequences were equally spatially segregated, irrespective of their temporal relationship and the differences in the perception of the two sequences. These results emphasize the fundamental importance of the temporal dimension in the perceptual organization of sound, and suggest that spatially segregated response patterns in the auditory cortex are not sufficient to explain streaming (also see Shamma and Micheyl, 2010; Shamma et al., 2011; Shamma et al., 2013).

The temporal coherence model has also been applied to model the perceptual organization of sound in signals more complex than streaming, such as music (Pressnitzer et al., 2011). For pieces of music characterized by a number of instruments playing at the same temporal rhythm, the predictions of the temporal coherence model were consistent with the perception of a single rich harmony. On the other hand, for a musical excerpt with several instruments playing independent melodies at distinct levels of rhythm, the coherence matrix showed off-diagonal activation

patterns that suggests the presence of incoherent sources (Pressnitzer et al., 2011).

In the present study, the stimulus consisted of coherent figures with a few temporally correlated channels in the presence of a number of channels with random fluctuation patterns that comprised the background. The results of the temporal coherence analysis for the SFG stimuli (Figures 4.3 and 4.4) demonstrate that temporal coherence varies significantly as a function of the coherence and the duration of the figure. Thus, temporal coherence is sensitive to the salience of the figure in a manner that is consistent with the behavioural results. These results, however, do not offer conclusive evidence in favour of the temporal coherence model but are strongly supportive of such a mechanism. The data indicate temporal coherence to be a correlate of stimulus salience by which the brain picks out the most salient sounds in complex scenes: a process that may not be computed by dedicated structures but could be achieved by binding across distributed feature channels. Similar accounts of binding in vision based on temporal structure also exist (Fahle, 1993; Alais et al., 1998; Treisman, 1999; Blake and Lee, 2005).

4.5.2 Neural bases of temporal coherence analysis

The brain mechanisms and substrates of temporal coherence analysis are yet to be determined. It is possible that temporal coherence may be computed by cells that show strong sensitivity to features across distant channels (e.g. that encode pitch, intensity, or spatial location). Alternatively, neurons that can multiplex information from distinct channels could encode

coherence (Elhilali et al., 2009a; Shamma et al., 2011, 2013). Elhilali and colleagues (2009a) sought such cells in the primary auditory cortex of the ferret but were unable to demonstrate any. In their study, they observed spatially segregated activation of neurons in the cortex for synchronous tone sequences with high temporal coherence, which was similar to the neuronal response patterns for the temporally uncorrelated alternating tone sequences. These data suggest that the locus of temporal coherence computations may be outside the auditory cortex. The behavioural and modeling data are indicative of a highly robust mechanism that is sensitive to correlations across frequency and time. Together, these lines of evidence suggest that a higher-order region in auditory-related areas or beyond that receives inputs from the cortex may be responsible for encoding coherence. Single-unit activity as examined by Elhilali and co-workers (2009a) may not be the ideal technique to answer this question and measurement of neural ensemble activity may shed some light on the problem. In humans, techniques like EEG and MEG with high temporal resolution could highlight coordinated brain activity across distinct channels that represent temporal coherence.

An important consideration relates to the involvement of oscillatory mechanisms such as gamma oscillations that have been implicated in object binding (Gray et al., 1989; Tallon-Baudry and Bertrand, 1999). In their original study, Elhilali and colleagues (2009a) suggest that coherence is a stimulus-driven phenomenon that may not be dependent on oscillatory mechanisms that help synchronize and bind activity across distant cortical sites. In other words, response to coherence might be expected to be time

locked (evoked) rather than induced at the point where the initial mechanism occurs. Another important feature of the temporal coherence model is the role of selective attention as illustrated in figure 4.2 which is predicted to operate at the output of the coherence analysis stage. The role of attention is described in greater detail in the following section.

4.5.3 Attention and temporal coherence

Attention is an important feature in auditory scene analysis and has been shown to be involved in selection of streams and the perceptual (schema-based) organization of the auditory scene (Bregman, 1990; Fritz et al., 2007; Snyder et al., 2012). The role of attention in stream formation, however, is considered to be modulatory. It can influence stream formation by sharpening the responses to different features, thus altering the neural representation. This has been demonstrated in several neurophysiological studies that showed task- and attention-dependent modulation of cortical STRFs (Fritz et al., 2003, 2005, 2010; David et al., 2012). Another way of influencing streaming is by modulating the temporal coherence of neuronal ensembles (Elhilali et al., 2009b; Shamma et al., 2011). In an MEG study, Elhilali and colleagues used an IM paradigm where the target consisted of a regularly repeating tone and found that responses were enhanced specifically at the attended rate. This resulted in enhanced phase coherence between neuronal populations that may help facilitate temporal coherence analysis (Shamma et al., 2011). Attention may also influence temporal coherence by acting on specific features which may serve as an anchor and bind other acoustic attributes of the same source that are temporally

correlated with that feature. For instance, attending to the spatial location of a speaker may provide a cue to bind other aspects such as pitch and timbre. Along the same lines, cueing a particular feature may also aid attentional selection of that feature which may enhance subsequent coherence analysis. In the case of the SFG stimuli, pre-cueing a frequency component that subsequently forms one of the coherent channels of the "figure" may improve temporal binding and consequently target-detection behaviour.

In a recent study, Shamma et al. (2013) measured cortical STRFs in ferret auditory cortex in the passive and behaving states. The stimuli consisted of a pair of tones that were either alternating or synchronous and transitioned to a random cloud of tones that enabled STRF measurements. In the passive state, the average STRFs were similar for both the sequences of tones. However, when the animal began to attend globally to the stimuli, a segregated pattern of responses was observed for the alternating and synchronous tone sequences. The STRFs for the alternating tones were significantly suppressed below the passive level, whilst the STRFs for the synchronous tones were enhanced. The latter was attributed to mutually positive interactions between neurons whilst inhibitory interactions decrease the responsiveness during the alternating tone presentation. Thus, temporal coherence and attention interact to enhance the perceptual representation of foreground streams and diminish the representation of background streams.

Chapter 5. FUNCTIONAL MAGNETIC RESONANCE IMAGING

Summary

In this study, functional MRI was used to investigate the neural substrates that are involved in the processing of salient figures in the stochastic figure-ground stimulus. The behavioural results and modeling simulations suggest the existence of a segregation mechanism that is highly sensitive to correlations in frequency and time. Such a mechanism may possibly be based at the level of the auditory cortex or beyond. Here, fMRI was used in a passive listening paradigm to examine the brain bases of stimulus-driven segregation in the SFG stimulus. Listeners were required to perform an irrelevant task while listening to a continuous stream of the SFG stimulus with brief figures embedded in the sequence. The coherence and duration of the figures was parameterized to investigate brain areas that are sensitive to the pop-out of the figures. Results demonstrate significant activations in the intraparietal sulcus (IPS) and the superior temporal sulcus related to bottom-up figure-ground decomposition. No significant activation was observed in the primary auditory cortex. These results are consistent with accumulating evidence suggesting a role for the IPS in structuring sensory input and perceptual organization of the auditory scene.

5.1 Introduction

Auditory figure-ground segregation refers to listeners' ability to extract a particular sound of interest from a background of other simultaneous sounds, for instance, the sound produced by the drums in an orchestra. Auditory segregation involves a set of processes that include grouping of simultaneous figure components from across the spectral array (Micheyl and Oxenham, 2010), grouping of sequential figure components over time (Moore and Gockel, 2002), and extraction of the grouped components from the background (de Cheveigné, 2001).

Several studies have examined the neuronal mechanisms underlying these processes based on signals such as streaming, informational masking stimuli, and oddball stimuli amongst others (see section 1.4). Based on such investigations in both human and animal experiments, a distributed network of areas along the auditory pathway has been implicated in segregation. Peripheral channeling models of segregation (Hartmann and Johnson, 1991; Beauvois and Meddis, 1991, 1996; Denham and McCabe, 1997) received physiological support from a streaming experiment in guinea pigs (Pressnitzer et al., 2008) where neuronal activity in the cochlear nucleus reflected sensitivity to frequency separation and presentation rate as demonstrated in human psychophysical experiments (van Noorden, 1975; Bregman, 1990). Further up the ascending pathway, the medial geniculate body (MGB) in the thalamus was also implicated in an fMRI experiment that examined the role of perceptual reversals during streaming (Kondo and Kashino, 2009, 2012). They found activation of the MGB specifically

during switching from a non-predominant to a predominant percept. The primary auditory cortex has been implicated in a number of studies in human (Deike et al., 2004, 2010; Bidet-Caulet et al., 2007; Wilson et al., 2007; Dykstra et al., 2011) and animal (Fishman et al., 2001, 2004; Bee and Klump, 2004, 2005; Micheyl et al., 2005) experiments. The auditory cortical responses demonstrate sensitivity to parameters such as frequency separation and tone presentation rate that determine the perceptual representation of streaming sequences. Non-primary auditory cortex including areas in the planum temporale (PT) have been shown to be involved in mediating both primitive and cognitive aspects of segregation including attentional modulation (Gutschalk et al., 2005; Alain, 2007; Schadwinkel and Gutschalk, 2010; Ding and Simon, 2012; Mesgarani and Chang, 2012; Zion-Golumbic et al., 2013a).

A more striking result was reported by Cusack (2005) who did not observe any activity in the primary auditory cortex related to the perceptual representation of two vs. one stream in an fMRI experiment. Instead, he found that activity in the intraparietal sulcus (IPS) most strongly corresponded to this contrast and suggested that it may be involved in attentional switching between streams in a bistable configuration. Hill et al., (2011) provided further evidence that the IPS is involved in mediating perceptual representation during streaming in another fMRI study of streaming.

However, a limiting factor in understanding the neural computations occurring at these different levels and relating existing experimental results

to listeners' experience in natural environments is that the stimuli used thus far have been rather basic, lacking the spectrotemporal richness of natural sounds. Most studies of segregation have used relatively simple stimuli consisting of sequentially presented, regularly alternating tones (Shamma and Micheyl, 2010) or static harmonic sounds (Alain, 2007).

In this study, the SFG stimulus was used in an fMRI experiment to examine the brain areas that underlie segregation in this complex signal (see section 3.2.1 for stimulus details). Unlike previous signals, this stimulus is not confounded by figure and background signals that differ in low-level acoustic attributes, or by the use of a spectral 'protective region' around the figure. Here, at each point in time, the figure and background are indistinguishable and the only way to extract the figure is by integrating over time (over consecutive chords) and frequency (identifying the components that change together). Behavioural results (see section 3.3) demonstrate that listeners are remarkably sensitive to the emergence of such figures. In this passive fMRI paradigm, the salience of the figure was systematically varied by independently manipulating the number of repeating components and the number of repeats in order to investigate the neural bases of the emergence of an auditory object from a stochastic background as occurs during the automatic parsing of natural acoustic scenes.

5.2 Materials and methods

5.2.1 Participants

Fourteen participants (9 female; mean age = 27.4 years) with normal hearing and no history of neurological disorders took part in the fMRI experiment. None of these subjects participated in the psychophysics experiments reported in chapter 3. Experimental procedures were approved by the Institute of Neurology Ethics Committee (London, UK), and written informed consent was obtained from each participant. The data for one subject were excluded from analysis due to a technical problem. All listeners completed the passive listening block. A subset of seven participants (3 female; mean age = 28.8 years) also subsequently completed an 'active detection' block to assess performance on the figure-detection task in the scanner.

5.2.2 Stimuli

5.2.2.1 Passive listening block

A key feature of the present experimental design is the brief duration of the figure. Whereas previous studies used relatively long, on-going figure-ground stimuli and, in many cases, instructed listeners to actively follow one of the components (Scheich et al., 1998; Cusack, 2005; Gutschalk et al., 2005, 2008; Wilson et al., 2007; Elhilali et al., 2009b), in this imaging experiment listeners were kept naïve to the nature of acoustic stimulation. They were presented with very short figure stimuli, embedded in an on-going random background. Figure duration (a maximum of 6 repeating chords – 300ms) was determined on the basis of a behavioural

experiment to find an optimal value that produced reliable detection. Such an experimental design was used in order to tap the bottom-up, segregation mechanisms rather than subsequent processes related to selective attention.

The stimuli were created in the same way as the psychophysical stimuli (see section 3.2.1) with the following differences: the results of the psychophysics experiments (see Figure 3.3) identified two parameters as the most informative to study the underlying brain mechanisms because performance on those conditions spanned the range from non-detectable to easily detectable: i) fixed coherence with four components and varied duration, and ii) fixed duration of four chords and varied coherence. The stimuli in the fMRI experiment thus consisted of signals with a fixed coherence level of four components with five duration levels (2-6) and signals with a fixed duration of four components with five coherence levels (1, 2, 4, 6, and, 8), resulting in nine stimulus conditions. Due to temporal resolution considerations related to the slow BOLD haemodynamics, and the need for a larger interval between events of interest, the duration of the signals was increased to 2750ms (as opposed to 2000ms in the psychophysical experiments), with the figure appearing between 1250-1500ms (25-30 chords) post onset. 66% of these signals contained a brief figure. Additionally, a small proportion (15%) of decoy stimuli consisting of 200ms white noise bursts (ramped on and off with 10ms cosine-squared ramps) were randomly interspersed between the SFG stimuli. Overall, listeners heard 40 repetitions of each of the nine different stimuli.

In order to avoid effects of transition between silence and sound, the stimuli were presented in succession without any gaps. The resulting continuous stream consisted of an on-going tonal background noise with occasional figures. This signal was intermittently interrupted by brief noise bursts which listeners were required to detect. Stimuli were presented via NordicNeuroLab electrostatic headphones at a sound pressure level of 85-90dB.

5.2.2.2 Active detection block

An active detection block was used to assess listeners' performance on the task in the presence of the scanner noise. Signals with a fixed coherence level of four components and five duration levels (2-6) and with a fixed duration level of four components with five coherence levels (1, 2, 4, 6, and 8) were used. Listeners heard eight repetitions of each stimulus condition. The order of presentation of the different stimuli was randomized with an ISI between 500-1250 ms. After every eighth stimulus, the ISI was increased to 12 s (to allow an analysis of a sound vs. silence constrast).

5.2.3 Procedure

The experiment lasted two hours and consisted of a 'passive listening' block followed by an 'active figure-detection' block. Each block consisted of three runs of 10 minutes each. Participants completed both blocks in a single session; they were allowed a short rest between runs. In the 'passive listening' block, the listeners were kept naïve to the stimulus structure and the aims of the experiment: they were instructed to look at a

fixation cross and detect noise bursts using a button box. In the 'active detection' block, listeners were instructed to detect the figures that popped out of the random tonal noise (50% of the signals). Crucially, this task was explained to the participants before the start of the 'active detection' block to ensure that they were unaware of the existence of figures during the 'passive listening' block. Pilot experiments did suggest that while the figures are readily detectable after a short practice, naïve listeners performing the decoy task remained unaware of their presence.

Before beginning the task, subjects completed a short practice session (about 10 minutes) in the MRI scanner and received feedback. To facilitate learning, feedback was also provided during the session proper.

5.2.4 Image acquisition

Gradient weighted echo planar images (EPI) were acquired on a 3 Tesla Siemens Allegra MRI scanner using a continuous imaging paradigm with the following parameters: 42 contiguous slices per volume; time to repeat (TR): 2520ms; time to echo (TE): 30ms; flip angle α: 90°; matrix size: 64 x 72; slice thickness: 2 mm with 1 mm gap between slices; echo spacing: 330μs; in-plane resolution: 3.0 x 3.0 mm^2. Subjects completed three scanning sessions and a total of 510 volumes were acquired. Field maps were acquired for each subject with a double-echo gradient echo field map sequence (short TE = 10.00ms and long TE = 12.46ms) to correct for geometric distortions in the EPI due to magnetic field variations (Hutton et al., 2002, Cusack et al., 2003). A structural T1-weighted scan was also acquired after the functional scan (Deichmann et al., 2004).

5.2.5 Image analysis

Imaging data were analyzed using Statistical Parametric Mapping software (SPM8; Wellcome Trust Centre for Neuroimaging, London, UK; see section 2.2.3). The first two volumes were rejected to control for saturation effects and the remaining volumes were realigned to the first volume and unwarped using the field maps. The realigned images were spatially normalized to stereotactic space (Friston et al., 1995a) and smoothed by an isotropic Gaussian kernel of 5 mm FWHM.

Statistical analysis was conducted using the general linear model (Friston et al., 1995b; see section 2.2.4). Onsets of trials with fixed coherence and fixed duration were orthogonalized and parametrically modulated by coherence and duration values respectively. These two conditions were modeled as conditions of interest and convolved with a hemodynamic boxcar response function. A high-pass filter with a cut-off frequency of 1/128 Hz was applied to remove low-frequency signal variations.

A whole-brain random-effects model was used to account for within-subject variance (Penny and Holmes, 2004). Each subject's first-level contrast images were entered into second-level t-tests for the primary contrasts of interest – "effect of duration" and "effect of coherence". Functional results are overlaid onto the group-average T1-weighted structural scan.

5.3 Results

The aim of the fMRI analysis was to identify the brain areas whose activity increases parametrically with an increase in the corresponding changes in coherence (while keeping duration fixed) and duration (while keeping coherence fixed) respectively.

The primary purpose of the active detection block (presented after the passive listening block) was to ensure that subjects were indeed able to detect the figures despite the loud, interfering MRI scanner noise and to compare their performance to that obtained outside the scanner. Because of the differences in stimulus presentation between the passive and active blocks, as well as other perceptual factors such as attentional load and focus of attention, a comparison of the activation patterns in the two blocks is not straightforward.

5.3.1 Psychophysics

Figure 5.1 shows the figure-detection performance obtained in the scanner ('active detection' block) alongside the results from the behavioural study. Listeners performed worse in the scanner than in quiet conditions (a difference of about 20%). This may be due the interfering scanner noise as well as lack of sufficient practice. It was important to keep listeners naïve for the passive half of the fMRI study and instructions for the figure-detection task were provided after the passive block, while listeners were already in the scanner. Moreover, as a consequence of the experimental design, listeners also encountered overall fewer 'easy' signals (those with a

fixed coherence of six and eight components and long duration) which could have contributed to some improvement with exposure.

Crucially, the data illustrate that the figures are easily detectable even in a noisy scanner environment and that the parametric modulation produced linear response patterns.

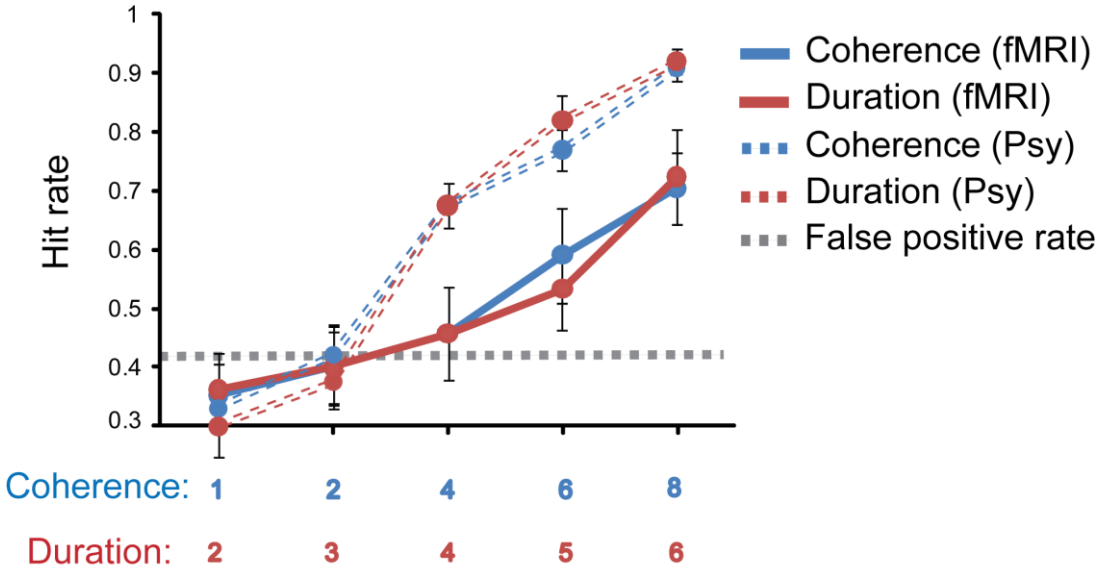

Figure 5.1: Comparison of behavioural performance in the psychophysics and fMRI experiments.

Behavioural performance on the figure detection task obtained in the scanner with continuous image acquisition (solid lines) presented along with data from the same stimuli obtained in quiet (dashed lines; see psychophysical study, Fig. 2). Hit rate is shown as a function of fixed coherence (4 components) and increasing duration (in red) and as a function of fixed duration (4 chords) and increasing coherence (in blue). The dashed line represents the mean false-positive rate. Error bars represent one SEM.

5.3.2 fMRI results

The purpose of the passive listening block was to identify the brain areas whose activity is modulated as a function of the figure salience. A decoy task was used to ensure that subjects remain vigilant and attend to the sounds whilst distracting them from the stimulus of interest. Performance on the decoy task was at ceiling for all listeners. As the primary aim was to examine predominantly bottom-up segregation mechanisms, it was essential that listeners were naïve to existence of the figures. Indeed, when interrogated at the end of the block, none of the subjects reported hearing salient sounds pop out of the background.

5.3.2.1 Effects of duration

The analysis of parametric changes in BOLD activity to figures associated with a fixed coherence and varying duration showed significant bilateral activations in the anterior IPS (figure 5.2A), the superior temporal sulcus (STS; figure 5.2B) as well as the right planum temporale (figure 5.2B). Additionally, the MGB was also found to respond to figures with increasing duration (figure 5.3).

5.3.2.2 Effect of coherence

The analysis of the effect of increasing the coherence of the figures while keeping the duration fixed showed significant bilateral activations in the posterior IPS (Figure 5.4A), and the STS (Figure 5.4B).

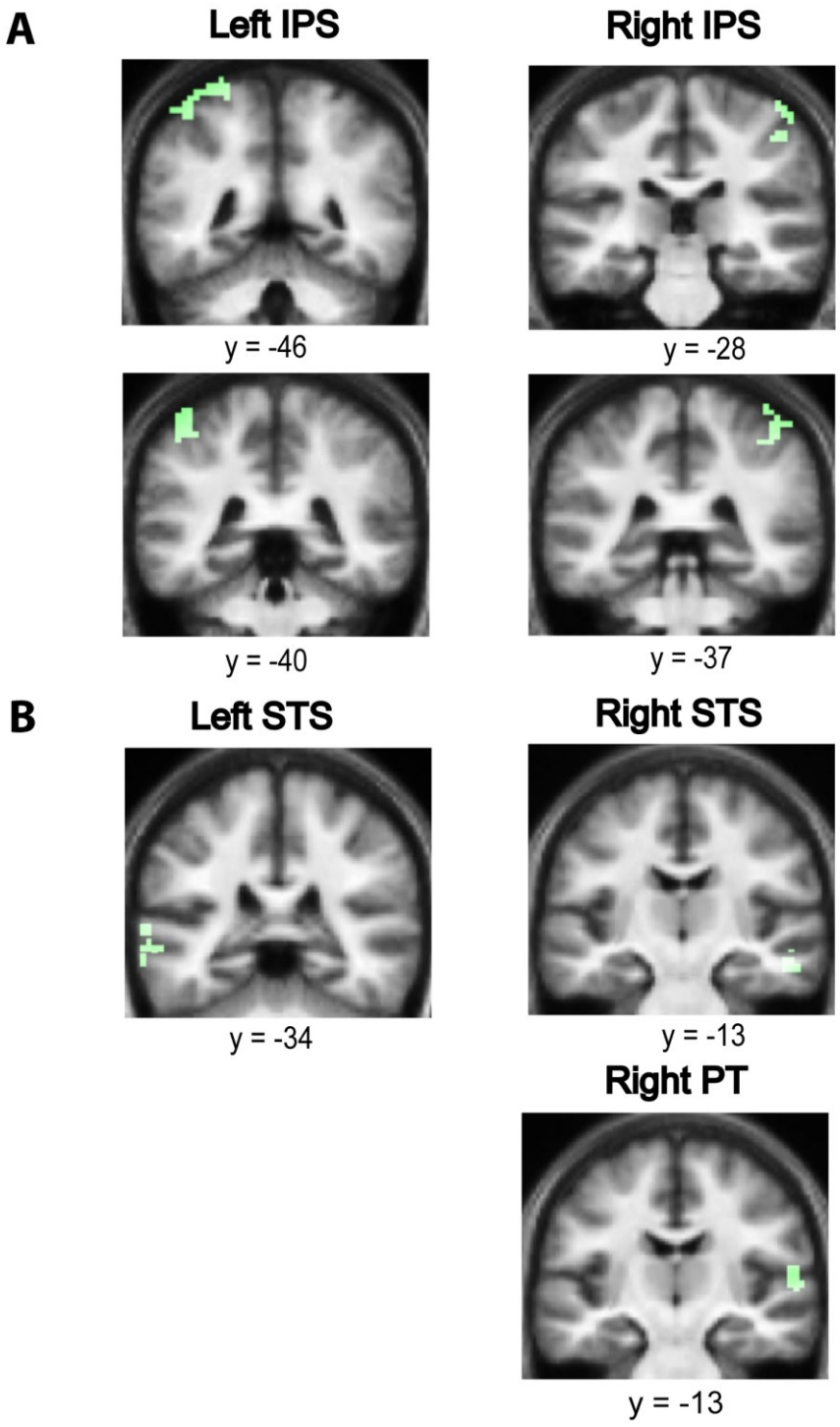
A
Left IPS
Right IPS
y = -46
y = -28
y = -40
y = -37
B
Left STS
Right STS
y = -34
y = -13
Right PT
y = -13

Figure 5.2 The effect of duration on segregation in the SFG stimulus.

(A) Areas in the anterior IPS showing an increased hemodynamic response as a function of increasing duration of the figures with fixed coherence (in green). Significant clusters for the effect of duration were found in the anterior IPS bilaterally. Results are rendered on the coronal section of the subjects' normalized average structural scan and results are shown at $p < 0.001$ uncorrected.

(B) Areas in the STS and PT showing an increased hemodynamic response as a function of increasing duration of the figures with fixed coherence (in green). Significant clusters for the effect of duration were found in the STS bilaterally and in the right PT. Results are rendered on the coronal section of the subjects' normalized average structural scan which is tilted (pitch = -0.5) to reveal significant clusters in the superior temporal plane, at $p < 0.001$ uncorrected.

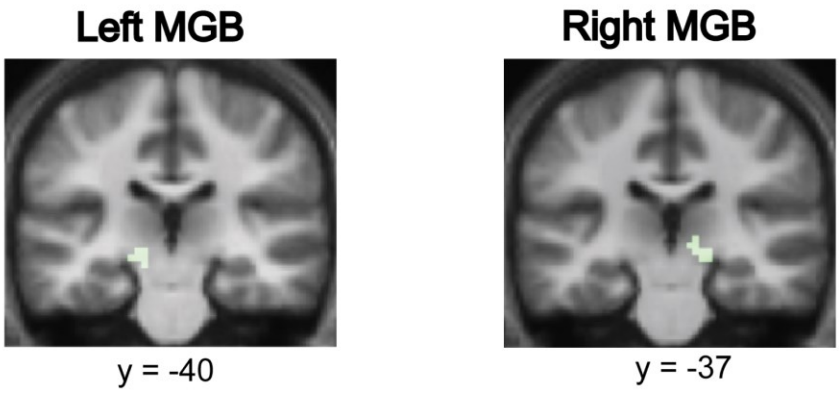

Figure 5.3: MGB activations for effects of duration.

Areas in the MGB showing an increased hemodynamic response as a function of increasing duration of the figures with fixed coherence (in green). Significant clusters for the effect of duration were found in the MGB bilaterally (also see Table 5.1). Results are rendered on the coronal section of the subjects' normalized average structural scan at a threshold of $p < 0.001$ (uncorrected).

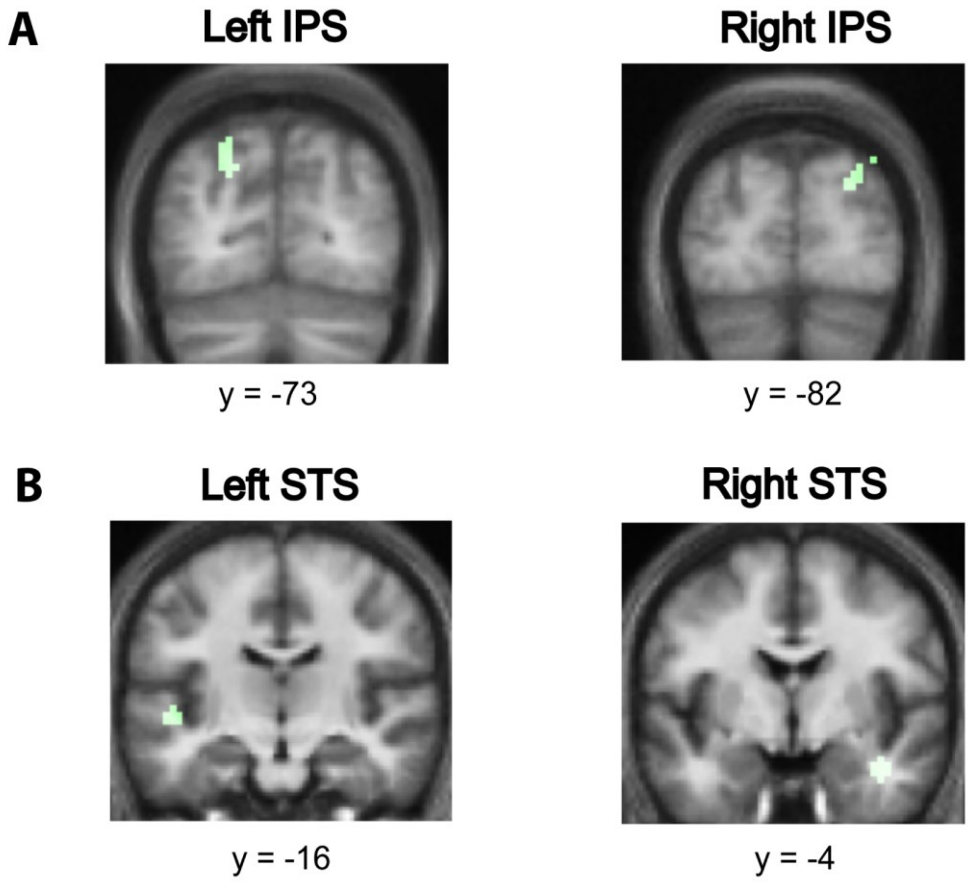

Figure 5.4: The effect of coherence on segregation in the SFG stimulus.

(A) Areas in the posterior IPS showing an increased hemodynamic response as a function of increasing coherence of the figures with fixed duration (in green). Significant clusters for the effect of duration were found in the posterior IPS bilaterally. Results are rendered on the coronal section of the subjects' normalized average structural scan at $p < 0.001$ uncorrected.

(B) Areas in the STS showing an increased hemodynamic response as a function of increasing coherence of the figures with fixed duration (in green). Significant clusters for the effect of duration were found in the STS bilaterally and in the right PT. Results are rendered on the coronal section of the subjects' normalized average structural scan to reveal significant clusters in the superior temporal plane at $p < 0.001$ uncorrected.

Contrast	Brain area	x	y	z	t-value	z-score
Effect of duration	Left IPS	-42	-46	64	5.14	3.67
		-48	-40	61	4.89	3.56
	Right IPS	51	-28	61	5.17	3.68
		45	-37	64	4.24	3.25
	Left STS	-57	-34	-2	4.42	3.34
	Right STS	60	-13	-11	4.06	3.16
	Right PT	60	-13	10	4.96	3.59
	Left MGB	-15	-25	-8	4.85	3.54
	Right MGB	18	-25	-8	4.92	3.57
Effect of coherence	Left IPS	-21	-73	46	4.99	3.60
		-24	-73	37	4.36	3.31
	Right IPS	27	-82	31	3.69	2.96
	Left STS	-48	-16	-5	3.43	2.81
	Right STS	39	-4	-26	3.77	3.00

Table 5.1: **MNI coordinates for effects of duration and coherence.**

Coordinates of local maxima for effects of duration and coherence are shown at a threshold of $p < 0.001$ (uncorrected).

Abbreviations: IPS, intraparietal sulcus; STS, superior temporal sulcus; PT, planum temporale; MGB, medial geniculate body.

5.3.2.3 Auditory cortex activations

As the two main conditions of interest revealed no significant clusters in the auditory cortex, a more stringent analysis was performed using the probabilistic cytoarchitectonic maps for primary auditory cortex - TE 1.0, TE 1.1 and TE 1.2 (Morosan et al., 2001), which are incorporated in the SPM Anatomy toolbox (http://www.fz-juelich.de/inm/inm-1/spm_anatomy_toolbox). A volume of interest analysis was performed that did not reveal any significant clusters ($p > 0.05$, FWE) when examined with the maps of the different cortical fields.

5.4 Discussion

The results of the psychophysics experiments reported in chapter 3 suggest the existence of a mechanism that is sensitive to correlations in frequency and time, and associated with a rapid buildup on the order of a few hundreds of milliseconds. The aim of the present study was to explore the neural substrates that mediate such robust segregation using functional MRI. A parametric design was employed to examine the sensitivity of the underlying mechanisms to spectral and temporal factors: the coherence and the duration of the figure were manipulated respectively whilst keeping the other dimension fixed. This approach was used to identify the brain areas that are sensitive to the coherence and the duration of the figure respectively. Another aspect of the design involved keeping the listeners naïve to the existence of the figures in the sound stream as the primary aim of this study was to elucidate the bottom-up, stimulus-driven bases of segregation as highlighted by the behavioural experiments. Listeners were

instructed to perform a decoy task based on detection of noise bursts that interspersed the SFG stimulus. As predicted by the behavioural experiments, the figures were expected to "pop-out" from the background and the purpose of the fMRI experiment was to identify brain areas that detect these salient figures in the absence of directed attention to the stimuli of interest. The results of the study demonstrated that the auditory cortex is not sensitive to the emergence of the salient figures; instead, parietal areas in the IPS exhibited significant sensitivity to the appearance of figures. These results are discussed in the next section in the light of previous work based on examination of neural responses to conventional stimulus paradigms such as streaming and IM stimuli.

5.4.1 Auditory cortex and segregation

Classically, auditory segregation has been investigated using two classes of stimuli. Simultaneous organization has been studied using signals consisting of multiple concurrent components where properties such as harmonic structure (tuned vs. mistuned: Alain, 2007; Lipp et al., 2010), spatial location (McDonald and Alain, 2005), or onset-asynchrony (Bidet-Caulet et al., 2007; Sanders et al., 2008; Lipp et al., 2010) were manipulated to induce the percept of a single source vs. several concomitant sources. Using such signals, human electroencephalography (EEG) and magnetoencephalography (MEG) experiments have identified responses in non-primary (Alain, 2007; Lipp et al., 2010) and primary auditory cortex (Bidet-Caulet et al., 2007), that co-vary with the percept of two sources.

The other class of stimuli used to study scene organization, is the streaming paradigm (van Noorden, 1975; Bregman, 1990; Shamma and Micheyl, 2010). Streaming refers to the process by which sequentially presented elements are perceptually bound into separate 'entities' or 'streams', which can be selectively attended to (Elhilali et al., 2009a). Human EEG and MEG experiments have demonstrated a modulation of the N1m (or M100) response, thought to originate from non-primary auditory cortex, depending on whether stream segregation takes place (Gutschalk et al., 2005; Snyder and Alain, 2007; Schadwinkel and Gutschalk, 2010; Snyder et al., 2012). fMRI studies have additionally identified activations in earlier areas along the ascending auditory pathway such as the MGB (Kondo and Kashino, 2009, 2012) and the primary auditory cortex (Wilson et al., 2007, Deike et al., 2004, 2010; Schadwinkel and Gutschalk, 2010) that are correlated with the streaming percept, in line with neurophysiological evidence from animal experiments (Fishman et al., 2001; 2004; Bee and Klump, 2004, 2005; Micheyl et al., 2005; Pressnitzer et al., 2008).

Stimulus-driven stream segregation has been suggested to be mediated by basic response properties of auditory neurons: frequency selectivity, forward suppression and adaptation, resulting in the activation of distinct neural populations pertaining to the figure and background (Fishman et al., 2001; Micheyl et al., 2007b; Snyder and Alain, 2007; Shamma and Micheyl, 2010; Fishman and Steinschneider, 2010a). Such mechanisms have been observed in primary auditory cortex (Fishman et al.,

2001; 2004; Bee and Klump, 2004, 2005; Micheyl et al., 2005) as well as in the cochlea (Pressnitzer et al., 2008). Segregation thus occurs in a distributed network over multiple stages in the ascending (and possibly descending) auditory pathways as well as areas outside the auditory cortex (Cusack, 2005; Dykstra et al., 2011; Hill et al., 2011).

Consistent with results from Cusack (2005), but contrary to a majority of the studies reviewed in section 1.5.1.2, no significant activation was found in PAC. This difference could be due to methodological issues (see also Cusack, 2005), as well as the more complex nature of the SFG stimulus. Furthermore, the fMRI results obtained by Cusack may reflect non-time locked activation due to the slow dynamics of the BOLD signal. It is possible that the IPS activations in the present study also represent induced activity patterns and not evoked activity time-locked to the appearance of the figure. In most studies that demonstrated activity in PAC to be correlated with the percept of one or two streams, stimulus parameters were modulated to produce streaming and any effect on primary cortex activity may be due to altered stimulus representations. On the other hand, Cusack (2005) used stimuli that produced a bistable percept, without any corresponding changes in the physical properties of the stimulus. The lack of activation differences in primary auditory cortex in his experiment is consistent with sensory rather than perceptual representation at that level. In the present study, the absence of activity in PAC could stem from the fact that adaptation-based mechanisms in primary auditory cortex, considered to underlie stream segregation, are not (or not sufficiently) activated by the

stochastic SFG stimuli. Alternatively, the activation of PAC in previous studies could be due to selective attention to the target stream (Bidet-Caulet et al, 2007; Elhilali et al, 2009b). The present experimental design, on the other hand, incorporated short figures and naïve subjects to specifically focus on automatic, bottom-up, stimulus-driven mechanisms instead of top-down attentional influences.

Furthermore, significant BOLD activity was also observed in the STS, which has previously been implicated in the perception of complex stimuli with a stochastic structure. It has also been shown to be involved in the analysis of spectral shape (Warren et al., 2005), changing spectrum over time (Overath et al., 2008), and detecting increasing changes in spectrotemporal coherence within acoustic 'textures' (Overath et al., 2010; see section 1.4.4.1). In sum, these studies suggest a role for STS in the 'abstraction' of features over spectrotemporal space that is relevant to the perception of distinct categories. STS is also involved in the analysis of stimuli with rich harmonic content such as voices that possess semantic information (Belin et al., 2000; Kriegstein and Giraud, 2004).

An important point to consider regarding the difference in experimental designs between previous studies and the current paradigm is that sequentially presented patterns were used in previous work where the target could be segregated by selective attention to the particular channel. Here, the target spanned a wide bandwidth and necessarily involved mechanisms that integrated the patterns over large frequency ranges.

5.4.2 IPS and auditory perceptual organization

The activation of IPS in the current study is a result that stands contrary to most previous work on streaming. Only a few studies have reported activation patterns related to segregation in the IPS (Cusack, 2005; Hill et al., 2011). The role of the IPS in auditory perceptual organization was first suggested by Cusack (2005) in a study where he measured BOLD activation patterns during the presentation of perceptually bistable streaming sequences and correlated changes in the BOLD response with listeners' perceptual reports. This allowed him to have a behavioural index on listeners' perceptual states and track on-going brain activity correlated with the switches between the two states. A simple contrast that looked at differences in activity for epochs associated with a two vs. one stream percept revealed significant activity only in the IPS. In another study, Hill and colleagues (2011) used bistable streaming stimuli and asked listeners to report switches to a grouped and split percept. Using fMRI, they found that maintenance of auditory streams is represented in the primary auditory cortex whilst the perceptual state is represented in higher-level cortical regions including the precuneus and the right IPS. These results are consistent with the results obtained by Cusack (2005) and suggest that the IPS may track the number of distinct objects after they have been segregated by auditory cortex or it may allow broad behavioural goals to influence streaming mechanisms.

Consistent with these findings, bilateral IPS activation was observed in the present study that is related to pre-attentive, stimulus-driven figure-

ground decomposition. The IPS activity was observed for both effects of interest, i.e., it increased as a function of coherence as well as duration. This pattern of activity was, however, spatially segregated within the IPS such that anterior IPS mediated the effects of duration whilst posterior IPS was involved in mediating the effects of coherence respectively. Such differential activation patterns are consistent with previous reports of functional dissociation within the IPS (e.g., Rushworth et al., 2001a; 2001b; Rice et al., 2006; Cusack et al., 2010). The two parameters, coherence and duration, together can be considered to represent the salience of the figure and the activity in IPS may be related to salience detection. This is in agreement with several studies of visual attention that consistently implicate the parietal cortex. Accumulating evidence also suggests that the IPS is crucial for encoding object representations (Xu and Chen, 2009), binding of sensory features within a modality (Friedman-Hill et al., 1995; Donner et al., 2002, Shafritz et al., 2002; Kitada et al., 2003; Yokoi and Komatsu, 2009), and across modalities (Bremmer et al., 2001; Calvert, 2001; Beauchamp et al., 2004; Miller and D'Esposito, 2005; Buelte et al., 2008; Werner and Noppeney, 2010).

The implication of IPS in auditory segregation adds a new dimension in the light of classic models of auditory scene analysis based on mechanisms within the core 'auditory system' (Fishman et al., 2001; Micheyl et al., 2007b; Snyder and Alain, 2007; Shamma and Micheyl, 2010). A critical issue that remains to be determined is whether IPS is causally responsible for segregation or whether it reflects the output of

perceptual organization occurring in primary or secondary auditory cortices (Shamma and Micheyl, 2010; Shamma et al., 2011). It is likely that the IPS activation observed by Cusack (2005) may result from the application of top-down attention during a subjective task or may be related to switching attention between the streams in the bistable state. Although IPS has been implicated in voluntary and involuntary control and shifts in auditory attention (Molholm et al., 2005; Watkins et al., 2007; Salmi et al., 2009; Hill et al., 2010), it is unlikely that the activation observed in the present experiment relates to top-down application of attention, or the active shifting of attention between objects. Listeners were naïve to the existence of the figure, and, when questioned, none reported noticing the figures. Additionally, the finding that different parametric modulations (duration vs. coherence) engage different fields in the IPS is inconsistent with a simple account in terms of subjective attention. These results are therefore in line with the suggestion that IPS plays an automatic, stimulus-driven role in segregation, and provide additional evidence implicating areas beyond the auditory cortex in auditory scene analysis.

5.4.3 IPS and Temporal coherence

The modeling results from chapter 4 suggest a role for temporal coherence in mediating segregation in the complex SFG stimulus. Across a variety of experiments (see chapter 3), temporal coherence was found to co-vary in a manner similar to the psychophysical response curves. Although these results do not offer conclusive causal evidence in favour of the temporal coherence theory of segregation, nevertheless, they offer

substantial bases to speculate a crucial role for temporal coherence. The neural substrates of temporal coherence analysis, however, remain unknown. In their recordings from ferret A1, Elhilali and colleagues (2009a) did not find any evidence of cells that show sensitivity to temporal coherence. This begs the question of the neural bases of temporal coherence analysis. In the light of the current fMRI results, it is tempting to speculate a role for the IPS that is considered next.

The parietal cortex receives bottom-up auditory input from the temporoparietal cortex (Pandya and Kuypers, 1969; Divac et al., 1977; Hyvärinen, 1982; Cohen, 2009) as well as top-down attentional input from the prefrontal cortex (Anderson et al., 1985; Barbas and Mesulam, 1981; Petrides and Pandya, 1984; Stanton et al., 1995) and is thus in an ideal position to integrate both stimulus-driven and top-down signals. The IPS has been implicated in both bottom-up and top-down attention and is a key structure implicated in saliency map models of visual search (Koch and Ullman, 1985; Itti and Koch, 2001; Walther and Koch, 2006, 2007) where low-level feature maps may combine with top-down cognitive biases to represent a global saliency map (Gottlieb et al., 1998; Geng and Mangun, 2009; Bisley and Goldberg, 2010). Furthermore, IPS (and its monkey homologue, lateral intraparietal; area LIP) has been implicated in mediating object representations, binding of sensory features within and across different modalities, as well as attentional selection as reviewed above. These lines of anatomical evidence present a sound basis to consider that the IPS may analyze input signals from the auditory cortices and compute

temporal coherence. An alternative possibility is that secondary auditory cortices involved in complex sound processing, such as the PT that was found to be active in the current study, may already process temporal structure and project the output to be represented in the IPS where selective attention may come into play as discussed in section 4.5.3. The temporal coherence model proposes that selective attention to a particular acoustic feature such as frequency may help bind together other temporally correlated features such as intensity or spatial location in order to encode the auditory object as a coherent whole.

It remains to be seen, however, whether the IPS actually represents a neural correlate of the figure percept. It is likely that such a perceptual representation depends on the computation of temporal coherence across multiple channels that are initially represented in the auditory cortex and biased by the IPS to attend to particular features within that object. Neurophysiological recordings from parietal neurons might in future determine whether such sensory analysis (before perceptual representation) involves parietal neurons or is established in auditory cortex first.

Chapter 6. MAGNETOENCEPHALOGRAPHY

Summary

In this study, MEG was used to examine the temporal dynamics of figure processing in the SFG stimulus. Behavioural results suggest a rapid buildup mechanism which may not be captured by the slow haemodynamics of fMRI which did not reveal any sources in the PAC; instead, IPS was found to be sensitive to the coherence and duration of the figure. Here, the high temporal resolution of MEG was used to study the evolution of figure processing with a specific focus on the auditory cortex. A passive design was used where listeners performed an irrelevant visual task whilst listening to the SFG stimuli that involved a simple transition from background to a figure with different levels of coherence. Two separate experiments with different SFG stimuli were conducted: these included the basic SFG stimulus and a variant with white noise between successive stimulus chords as reported in chapter 3. The results demonstrate robust evoked transition responses that consisted of an early peak and a later sustained component, the amplitude of which varied as a function of the coherence of the figure. Source reconstruction of evoked power revealed that PAC as well as IPS responded to the emergence of salient figure segments in both stimulus conditions. Analysis of the sustained phase of the response to the basic stimulus found activity in IPS that was not present during the early phase, suggesting a specific role for IPS in the perceptual representation of coherent figures after initial encoding in PAC.

6.1 Introduction

The detection of novel changes in the acoustic environment and separating out relevant sounds are fundamental auditory tasks. These processes usually occur over a short time scale, i.e. less than a couple of hundred milliseconds. It is indeed essential to respond quickly to new sources of sound and produce appropriate behavioural responses like approaching (e.g. attending to a crying baby) or avoiding (e.g. a lion in a jungle) the sound source. To process acoustic scenes on such a fast timescale requires mechanisms that are highly sensitive to salient changes in the acoustic environment. It has been shown that encoding sound onsets occurs quite rapidly with a latency of ~100ms (Lutkenhoner et al., 2003). However, the detection of a target signal in the presence of several simultaneous signals is a more complex task and the temporal dynamics of segregation in such complex sound scenes remains to be fully elaborated.

In humans, this question has been examined using functional imaging techniques with high temporal resolution such as EEG and MEG (Nagarajan et al., 2010) as well as direct intracortical recordings that can accurately track brain activity every millisecond. Studies of auditory segregation using EEG and MEG (Snyder and Alain, 2007) are based on simple stimulus paradigms such as streaming, oddball sequences, and informational masking paradigms. The earliest paradigm used in human neurophysiological experiments was based on auditory deviant responses classified as the MMN response (Naatanen et al., 2007). It is a pre-attentive differential evoked response that occurs 150-250ms following sound onset

when a violation in an acoustic pattern is detected, either passively or consciously (see section 1.4.2). Based on alternating patterns of high and low frequency tones, Sussman and colleagues (1999) demonstrated that stronger MMN is elicited for violations that are more readily perceived when the spectral separation between the tones is large, and suggested that streaming does not require directed attention. Beyond ERP experiments, Gutschalk et al. (2005) used MEG and demonstrated activity in the auditory cortex that varied as a function of the spectral separation in a streaming paradigm. They found that the amplitude of the P1 and N1 responses varied according to the perceptual state of the listeners, and were stronger in the case of segregated compared to single stream percepts. Bistable streaming paradigms have also been employed (Snyder et al., 2006; Hill et al., 2012; Szalardy et al., 2013) which revealed a positive-difference wave around 60-100ms post B-tone onset and localized in the auditory cortex, for segregated vs. integrated percepts. Dykstra et al. (2011) on the other hand, used surface recordings from the human cortex and found that distributed brain areas including the temporal, frontal and parietal cortices covaried with frequency separation in an active streaming task.

Further neurophysiological evidence supporting a role for the auditory cortex comes from the informational masking paradigm (see section 1.4.3). IM refers to a form of non-energetic masking and reflects computations at the level of the central auditory system rather than the periphery. Multi-tone complexes are used which consist of a regularly repeating target tone that tends to "pop out" from the random masking

tones. Interestingly, the buildup of this pop out effect mirrors the buildup observed in the streaming signals, suggesting a common underlying neural mechanism (Micheyl et al., 2007a). Using this paradigm, Gutschalk and colleagues (2008) demonstrated activity in the auditory cortex specifically for detection of target tones and no response for missed targets: this perceptual response was termed the awareness related negativity (ARN). Similarly, Elhilali and coworkers (2009b) demonstrated evoked activity in the auditory cortex that was left lateralized when the target was attended (cf. Deike et al., 2004, 2010), and right lateralized when the masker was attended.

The consensus from these and other neurophysiological studies in humans (reviewed in section 1.5) based on low-level sequences of tones as well as higher-level speech streams (Ding and Simon, 2012; Mesgarani and Chang, 2012; Zion-Golumbic et al., 2013) points to a role for the auditory cortex in sensory stimulus-specific processing at an early stage as well as perceptual representation of the target signal, which can be modulated by attention (Fritz et al., 2007).

The aim of the present study was to examine the nature of figure-ground analysis in the SFG signal that represents a more ecologically valid representation of natural sound scenes. Specifically, the aim was to examine the buildup of segregation in the SFG stimulus that involved a simple transition from background to figure. A passive design was used where listeners' attention was directed to an unrelated visual task to look at bottom-up correlates of segregation as previously found in the fMRI study.

The results of the fMRI experiment (see section 5.3.2) did not reveal any activation in the primary auditory cortex as a function of increasing coherence and duration of the figures, which may be attributed to the slow nature of the haemodynamic response function. Instead, activity in the IPS was modulated by the two spectrotemporal parameters of the figure. The role of IPS in auditory scene analysis has also been suggested by previous fMRI work (Cusack, 2005; Hill et al., 2011) but remains to be seen in EEG or MEG data. These results motivated the use of IPS as a spatial prior in source reconstruction of phase-locked power following the transition to a figure.

Here, MEG was employed to specifically examine the evolution of time-locked activity during figure processing in the SFG stimulus. With respect to the underlying mechanisms, a basis for the temporal coherence theory (Shamma et al., 2011) was predicted on the basis of modeling as described in chapter 4: it was hypothesized that both auditory cortex and IPS would be involved in coherence computations following the transition to a figure with an early role in encoding stimulus features in the auditory cortex and a later, possibly top-down role for the IPS in representation of temporal coherence based on inputs from the auditory cortex.

6.2 Materials and methods

6.2.1 Participants

6 listeners (3 females; mean age = 24.5 ± 3.8 years) and another 5 listeners (5 females; mean age = 24 ± 4.7 years) took part in two separate

psychophysics experiments based on the 'basic' and the 'noise' versions of the SFG stimulus respectively.

23 participants (12 female; mean age = 23.9 ± 6.2 years) with normal hearing and no history of neurological disorders took part in the MEG experiment. Experimental procedures were approved by the Institute of Neurology Ethics Committee (London, UK), and written informed consent was obtained from each participant. The data from three participants was excluded from analysis due to excessive movement during the scan.

6.2.2 Stimuli

6.2.2.1 SFG stimulus

The stimuli in the MEG experiments were slightly modified from the original SFG stimulus (see section 3.2.1) due to the sound delivery constraints imposed by the Etymotic tubes. These tubes act as low-pass filters and their frequency response tails off after ~ 2.5 kHz. Thus, the bandwidth of the SFG signals was reduced from ~ 7.2 kHz to 2.5 kHz for the MEG experiments. Two variations of the SFG stimuli were used: a faster version of the stimulus with 25ms chords (figure 3.2B) and a version of the stimulus with 25ms of white noise present between successive stimulus chords, each 25ms long (figure 3.2E; see section 3.2.3). These stimuli were used as it was previously demonstrated that there is no significant difference in performance between the original SFG stimulus based on 50ms long chords and these two versions of the stimuli respectively (see section 3.3). The use of the SFG stimuli with 25ms chords further helped in reducing the total duration of the experiment.

In order to test whether the reduced bandwidth did not affect figure detection, behavioural experiments were conducted for each version of the stimulus as described below. The stimuli for the psychophysics experiments were similar to those used in chapter 3: the figure was flanked by a background segment on either side.

In the MEG, however, a modified version of the stimuli was used: instead of a background-figure-background design, a simpler version with just one transition from background to figure was used as shown in figures 6.1 and 6.2 for the basic and the noise versions of the stimuli. This design was used to specifically examine evoked responses at the transition from background to a figure with different coherence levels (0, 2, 4, and 8 coherent components respectively) as well as activity related to maintenance of the figure percept. The way the stimulus was designed involved generating a background segment for the total duration of the stimulus, and incorporating additional components that were correlated in the figure segments (coherence = 2, 4, or 8) and uncorrelated in the background segment (coherence = 0) that served as a control. Thus, there was an increase in energy following the transition but the different stimuli were balanced with respect to their spectral energy profiles after the transition.

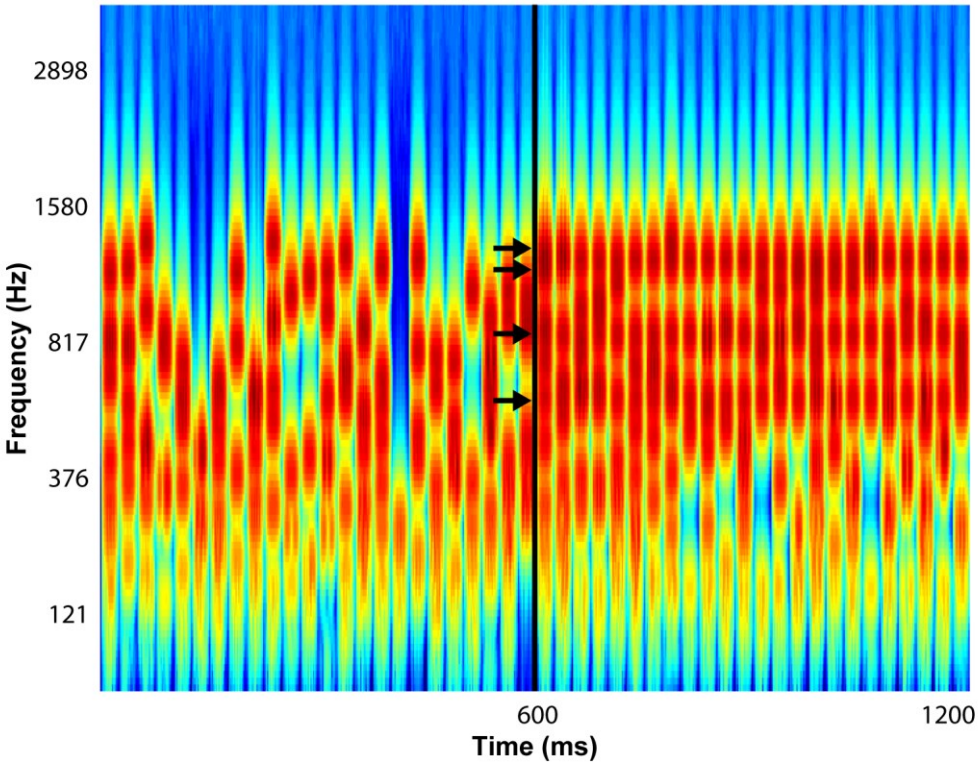

Figure 6.1: Spectrogram of the basic SFG stimulus used in MEG.

The stimulus consists of a series of 25ms long chords presented consecutively without any gap. The first 600ms of the stimulus consists of a background segment following which there is a transition to a coherent figure segment that is 600ms long. In the above example, the transition is indicated by the black vertical line. The coherence of the post-transition figure segment is equal to 4 and the repeating components are indicated by the black arrows.

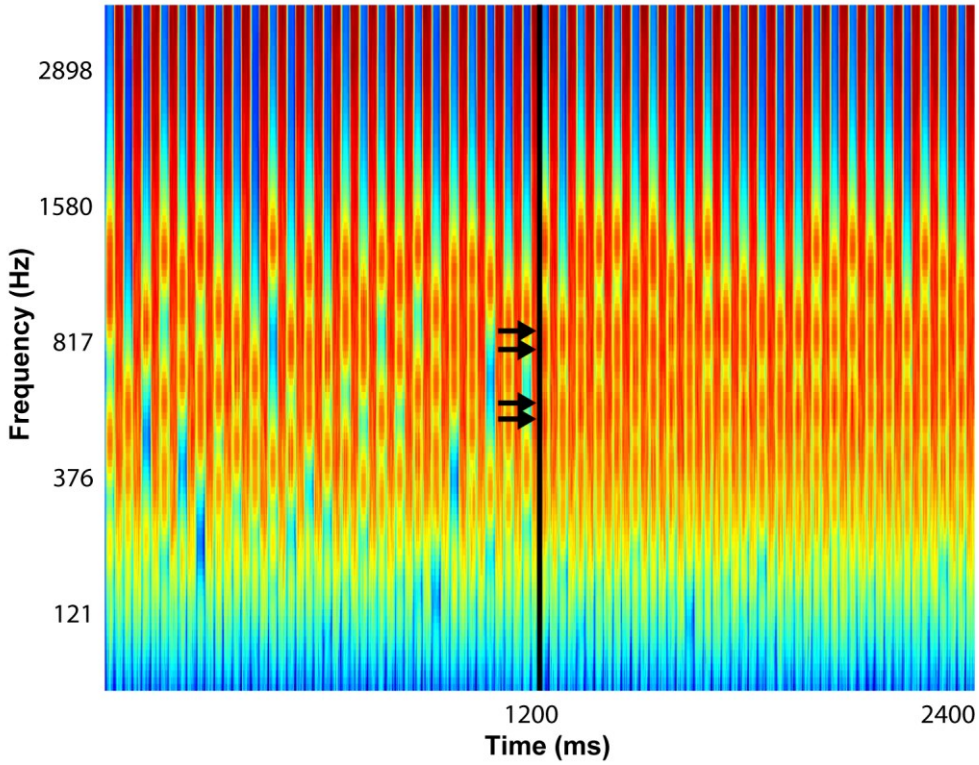

Figure 6.2: Spectrogram of the noise SFG stimulus used in MEG.

The noise stimulus consists of a series of 25ms long chords that alternate with 25ms of white noise. The first 1200ms of the stimulus consists of a background segment following which there is a transition to a coherent figure segment for another 1200ms. In the above example, the transition is indicated by the black vertical line and the coherent figure components (equal to 4) are indicated by the black arrows.

All acoustic stimuli were created using MATLAB 7.5 software (The Mathworks Inc.) at a sampling rate of 44.1 kHz and 16 bit resolution. Sounds were delivered diotically through Etymotic tubes and presented at a comfortable listening level of 60 to 70 dB SPL that was adjusted by each listener. Sounds were presented using Cogent (http://www.vislab.ucl.ac.uk/cogent.php).

6.2.2.2 Visual stimulus

An incidental visual task was used to engage participants' attention whilst passively listening to the SFG stimuli. Participants were instructed to pay attention to a series of images of landscapes and geographical landmarks as shown in figure 6.3. Each trial consisted of a set of three randomly chosen images from a set of 177 images. The task involved pressing a button if the third image was identical to the first image and to withhold responses if they were dissimilar. Each image was displayed for a random interval between 2 and 5s. The inter-trial interval varied randomly between 2 and 5s and the presentation of the images was not synchronized with the acoustic stimuli. All visual stimuli were presented from another machine using Cogent.

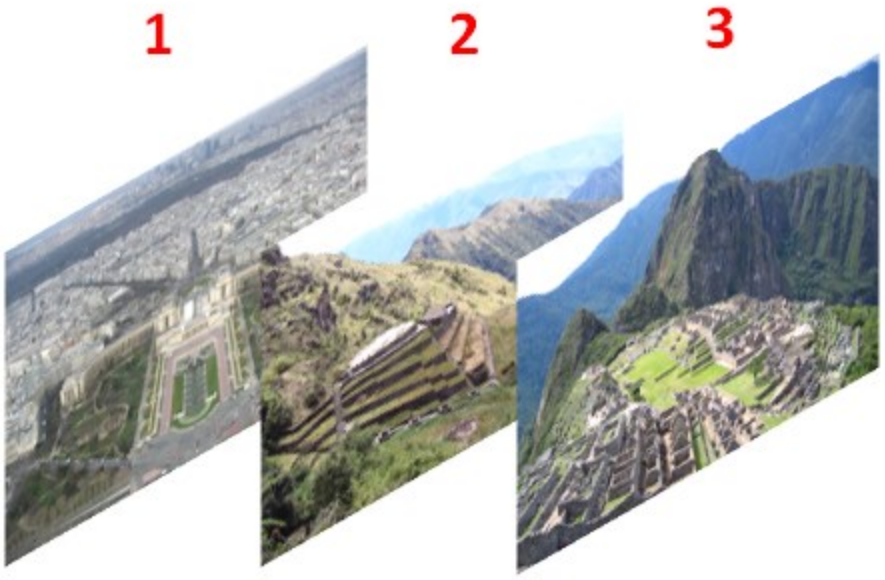

Figure 6.3: Visual task paradigm.

Listeners were required to pay attention to a series of 3 images of landscapes on each trial and press a button if the third image was identical to the first one. The presentation of the images was not synchronized to the acoustic stimulus.

6.2.3 Procedure

6.2.3.1 Psychophysics

Prior to the main experiment, participants received training that consisted of listening to trials with no figures, easy-to-detect figures (high coherence and duration), difficult-to-detect figures (low coherence and duration) and as well as a practice block. In the main sessions, the value of coherence and duration was indicated before each block and participants were instructed to press a button as soon as they detected a figure. Feedback was provided. Blocks with different values of coherence (2, 4, and 8) and duration (2-7) were presented in a pseudorandom order. The participants self-paced the experiment and each experiment lasted approximately an hour and a half. The procedure was identical for both experiments based on the 'basic' and the 'noise' versions of the SFG stimulus. The psychophysics was performed on a separate set of participants independently from the MEG experiment.

6.2.3.2 Magnetoencephalography

A functional source-localizer session was used at the start of the experiment that required participants to listen to a series of 100ms long pure tones (frequency equal to 1000 Hz) for approximately three minutes. The number of tones varied (between 180 and 200) with a random inter-stimulus interval that ranged between 700 and 1500ms. Listeners were required to attend to the sounds and report the total number of tones presented. This 'auditory localizer' session allowed an examination of MEG sensors that responded most robustly to sound onset which presumably reflect auditory

cortical activation.

The MEG experiment lasted between 1.5 and 2 hours and consisted of 8 blocks. Half of these blocks involved presentation of the basic SFG stimulus whilst the noise SFG stimulus was presented in the remaining four blocks. The order of the presentation of the basic and noise SFG stimuli was counterbalanced between subjects. For both stimulus conditions, the coherence of the post-transition segment was selected from one of 4 values (0, 2, 4, or 8). The number of control stimuli in each block (coherence of post-transition segment equal to 0) was equal to the number of stimuli with higher coherence levels (2, 4, or 8) combined to counterbalance the total number of transitions to a background and figure segment respectively. The duration of the basic SFG stimulus was 1200ms (600ms ground and 600ms figure segments) whilst the duration of the noise SFG stimuli was 2400ms (1200ms ground and 1200ms figure segments) respectively. Each block lasted between 8-10 minutes and subjects were allowed a short rest between blocks.

Importantly, the listeners were kept naïve to the stimulus structure and the aims of the experiment: they were instructed to perform a visual memory task based on a series of images of landscapes as depicted in figure 6.3. Feedback on the visual task was provided at the end of each block.

6.2.4 Data acquisition and analysis

MEG signals were recorded using a CTF-275 MEG system (axial gradiometers, 274 channels; VSM MedTech, Canada) at a sampling rate of 600 Hz.

A functional localizer session preceded the experimental blocks and the data from this session was divided into 700ms epochs, including 200ms pre-stimulus baseline period. A low pass filter with a cut-off frequency of 30Hz was applied to the baseline-corrected epochs. Visual artifacts were rejected using an in-built algorithm in Fieldtrip (Oostenveld et al., 2011). The M100 response was identified for each participant and the 40 strongest channels at the peak of the M100 (20 in each hemisphere) were selected as the channels that respond most robustly to sound. These channels were considered to reflect activity in the auditory cortex.

Data from the main experimental blocks consisted of a 500ms pre-stimulus baseline, and a 200ms post-stimulus period for the basic and noise SFG stimuli whose duration was equal to 1200 and 2400ms respectively. Data from approximately 100 epochs was averaged and low-pass filtered at 30Hz. In each hemisphere, the RMS of the field strength across the 20 channels, selected in the functional source localizer run, was calculated for each participant. These evoked responses were further processed using DSS (Denoising Source Separation; de Cheveigné and Parra, 2013) which is a procedure similar to ICA (Independent Component Analysis) that identifies the most reproducible components in electrophysiological time-series data. These data from each participant were averaged to obtain the group-RMS plots as shown in figures 6.5 and 6.14.

6.2.5 Source modeling

Source reconstruction of evoked power was performed using the 'Imaging' approach implemented in SPM12 (Litvak et al., 2011; Wellcome

Trust Centre for Neuroimaging). This method is based on an empirical Bayesian approach and the data can be inverted using a number of algorithms which have different assumptions about the data. Here, the IID approach based on a classical minimum norm algorithm was used to identify distributed sources of brain activity underlying the transition from a background to a coherent figure. The underlying principles of the method are described in detail in section 2.7.4.

For both stimulus conditions, data from the initial transition phase as well as a later sustained phase were localized separately. The assumption behind this design was to identify brain areas that may be differentially involved in initial figure-ground processing vs. perceptual representation of the figure.

The results were written out as 3D NIfTI images and analyzed using GLM-based statistical tests using Random Field theory. This approach is similar to the second level analysis used in fMRI to make inferences about region- and trial-specific effects (see section 2.6). For reconstruction of average evoked power, a time window of 300ms and a low frequency range from 0 to 48Hz was specified for both the early and late windows of interest. The data for all conditions was inverted together and separate NIfTI images were processed for each condition. The resultant 3D images were smoothed by using a Gaussian kernel with 5mm FWHM and taken to second-level analysis for statistical inference.

6.3 Results ('basic SFG')

The aim of the MEG analysis was to examine the temporal dynamics of figure processing in the SFG stimulus with a specific focus on the evoked responses at the transition from background to a figure with different coherence levels. An additional aim was to perform source reconstruction and identify brain areas involved in processing the transition to a coherent figure. The lack of parametric BOLD activity in the auditory cortex in the fMRI study provided another aim: to better understand the nature of processing in the auditory cortex and determine whether the lack of activation in the fMRI study was due to the slow nature of the underlying haemodynamics. The fMRI results also provided an *a priori* hypothesis for a specific role for the IPS in mediating the transition to figures and representation of the temporal coherence associated with the salient figures.

All analyses was conducted separately for the two versions of the stimuli used: the 'basic' SFG stimulus with 25ms long chords, and the 'noise' version of the SFG stimulus with 25ms long white noise segments between successive stimulus chords of the same duration (see section 6.2.2.1). Results from the psychophysical experiments, analysis of evoked MEG responses and source modeling are described in the following sections for the two studies respectively.

6.3.1 Psychophysics

Prior to the MEG experiments, psychophysics was used to examine figure-detection performance in a version of the SFG stimulus with reduced

bandwidth. This was done because of the limitations of the Etymotic tubes used for sound delivery in the MEG that have a low-pass filter characteristic. Thus, it was important to ensure that figure-detection performance is not affected by bandwidth and whether the results are qualitatively similar to those reported in Chapter 3. In these experiments, the frequency range of the stimulus varied from ~200Hz to ~2.5kHz and the duration of each chord was 50ms.

Behavioural results based on the 'basic' version of the SFG stimulus from 6 participants are shown in figure 6.4. The results indicate that listeners are sensitive to figures that span a narrower bandwidth. Performance for detection of figures with coherence equal to 2 was below chance but it increased monotonically for figures with coherence of 4 and 8 respectively.

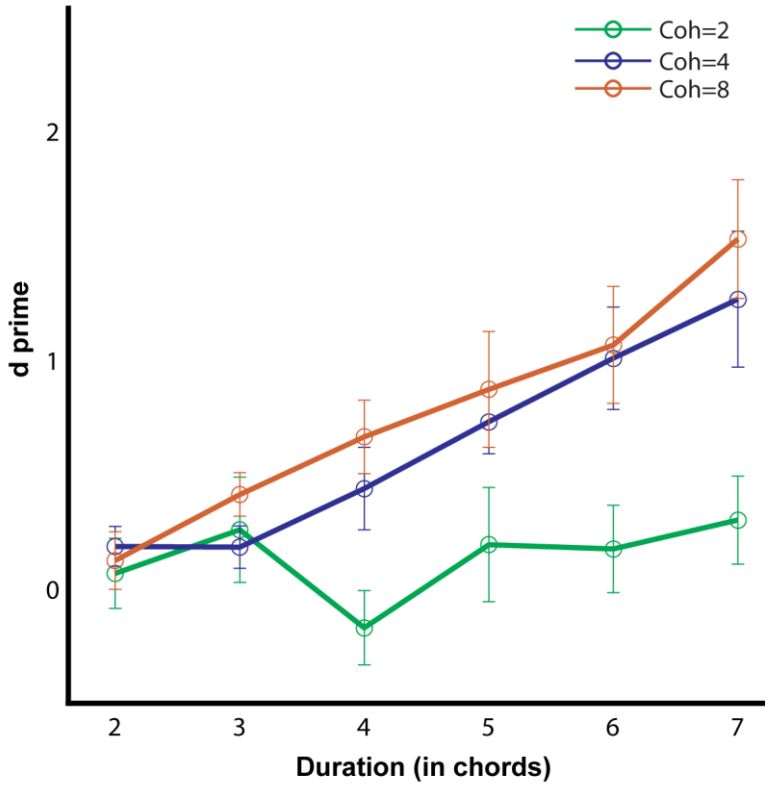

Figure 6.4: Figure-detection performance for the 'basic' SFG stimulus.

Behavioural results (d'; n=5) are plotted on the ordinate and the duration of the figure (in terms of number of 50ms long chords) is shown along the abscissa. The coherence of the stimuli was 2, 4, or 8 and six different levels of duration were tested. Listeners were required to press a button as soon as they heard a figure pop out from the background. Error bars signify one SEM.

6.3.2 Auditory evoked-fields

Figure 6.5 illustrates the group-RMS of auditory-evoked responses to the transition between the background and the figures separately for the left and right hemispheres respectively. The evoked field strengths are shown separately for the different coherence levels including the control condition that did not involve any change in coherence after transition.

These data demonstrate a strong response after transition that peaks 256.7 (251.7), 293.3 (288.3), and 298.8 (335) ms in the left (right) hemispheres after transition to a figure with coherence equal to 8, 4, and 2 respectively. The peak response is followed by a sustained phase of activity that persists throughout the duration of the figure that is followed by an offset response. The evoked responses are scaled according to the coherence of the figure with stronger responses for transition to a figure with higher coherence levels even though there is no difference in intensity.

Furthermore, the latencies at which the evoked field strengths for each of the three coherence levels were found to become significantly different from the field strength for the control condition (coherence = 0) were found to approximately parallel behavioural latencies for supra-threshold detection of the figures. For coherence of 8, the field strength became significantly different after 120ms which approximately corresponds to a figure with duration of 5 chords, for which d' of 0.87 ± 0.25 were achieved. For coherence of 4, the corresponding evoked field latency was 165ms which corresponds to a figure whose duration is equal to 6.6 chords. The d' for the detection of figures with coherence equal to 4 and

duration equal to 6 was 1.00 ± 0.22. These results suggest that the brain takes at the most as much time as the duration of the figure to detect the emergence of a salient figure even without the application of directed attention.

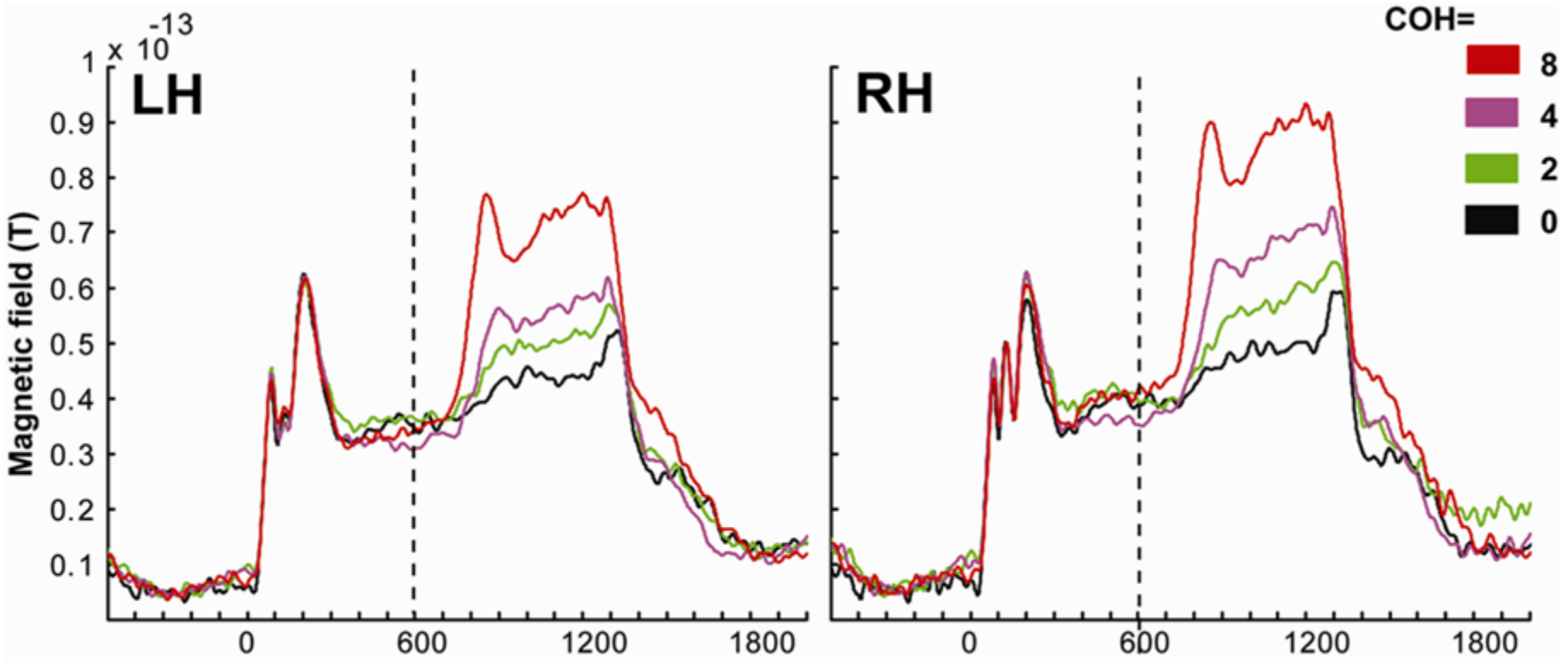
LH
RH
COH=
8
4
2
0
Magnetic field (T)
x 10^-13
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
600
1200
1800

Figure 6.5: Evoked field strengths in response to a transition from background to figure in the basic SFG stimulus.

The magnetic field strength in Tesla is plotted on the ordinate and time in milliseconds is plotted on the abscissa. The dotted black line separates the background from the following figure segments whose coherence is colour coded as indicated in the legend on the top right. The left and right panels indicate the resultant evoked field strengths in the left and right hemispheres respectively.

6.3.3 Source modeling

Source reconstruction of evoked power was performed for two distinct phases of the post-transition response: an early phase from 0-300ms following the transition from ground to figure; and a late phase from 300-600ms after the transition. In the basic stimulus, the transition occurred midway through the stimulus, i.e., 600ms following sound onset.

IID source modeling algorithm was used to identify distributed sources for the early and the late (sustained) components separately. The modeled data was converted into NIFTI images that were taken to second-level and analyzed using conventional GLM-based statistical methods. Three different parametric tests were used:

i) ANOVA: to examine areas that are sensitive to increasing coherence;
ii) 2-samples t-test: to investigate brain areas that specifically mediate the perceptual effects of figure processing without any confound related to intensity differences between the two different levels of stimuli; and,
iii) 1-sample t-test: to identify regions that show sensitivity to the onset of the salient figures.

The results for the source reconstruction of evoked power during the early and the late components are summarized in tables 6.1 and 6.2 respectively.

Contrast	Brain areas	x	y	z	t-value	z-score
Effect of coherence	R HG	52	-14	16	4.91	4.53
COH8 vs. COH4	R HG	48	-22	6	3.88	3.51
COH8 vs. COH2	R HG	50	-16	16	3.91	3.53
COH8 vs. COH0	R HG	60	-18	8	4.03	3.62
COH8	R HG	48	-26	2	7.14*	4.79*
	L HG	-50	-26	26	5.88*	4.29*
	R IPS	54	-52	28	5.44	4.08
	L IPS	-50	-48	36	5.09	3.91
	R IFG	42	28	-14	7.73*	5.00*
COH4	R HG	50	-20	2	7.01*	4.74*
	L HG	-54	-20	14	6.89*	4.70*
	R IPS	44	-60	48	3.91	3.26
	L IPS	-50	-52	38	4.09	3.36
COH2	R HG	50	-24	2	7.57*	4.94*
	L HG	-60	-18	22	6.22	4.43
	R IPS	48	-56	36	4.71	3.72
	L IPS	-50	-48	36	4.15	3.40
COH0	R HG	42	-28	20	5.95*	4.31*
	L HG	-54	-22	22	6.03*	4.35*
	R IPS	50	-62	26	5.78	4.24
	L IPS	-48	-46	36	4.34	3.51

Table 6.1: MNI coordinates for reconstruction of evoked power in the early transition phase of the basic SFG stimulus.

Source coordinates of activity during the early phase of the transition (0-300ms following transition) to a figure specifically in the auditory cortex and the IPS are shown for the different contrasts as indicated. Asterisk indicates statistical significant at $p < 0.05$ (FWE) whilst other results hold at $p < 0.001$ (uncorrected).

For the early phase of the transition (first 300ms following the transition), the ANOVA analysis revealed a main effect of coherence in the right auditory cortex including HG and STS as shown in figure 6.6. However, due to the nature of the stimulus design, these areas cannot be said to purely mediate perceptual analysis of the figures as there were greater number of chords in the post-transition segments (see stimulus design in section 6.2.2). To analyze areas that are activated as a function of coherence irrespective of such energetic confounds, 2-samples t-tests were performed for the contrasts shown in figure 6.6. The data indicate greater activation in the right auditory cortex for coherence level of 8 relative to each of the other coherence levels.

Analysis for a main effect of each of the individual coherence levels revealed bilateral sources of activity in the auditory cortex as shown in figure 6.7. Also, significant clusters of activity were found in the IPS following a small-volume correction for representation of each of the four levels of coherence as shown in figure 6.8.

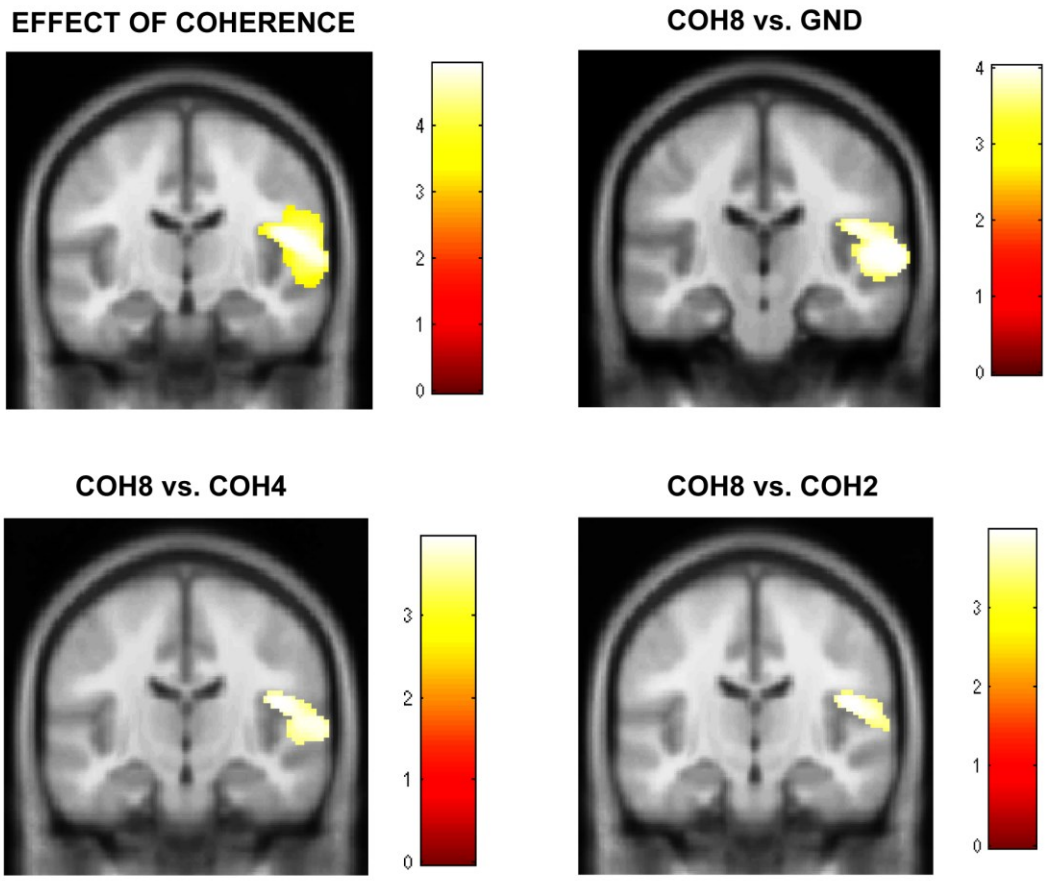

Figure 6.6: Activity in the auditory cortex as a main effect of coherence and difference in coherence levels during the early phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image. COH8, COH4, COH2, GND refers to the coherence value of a figure with 8, 4, 2, and 0 repeating components respectively. This nomenclature applies to figures 6.6 - 6.19.

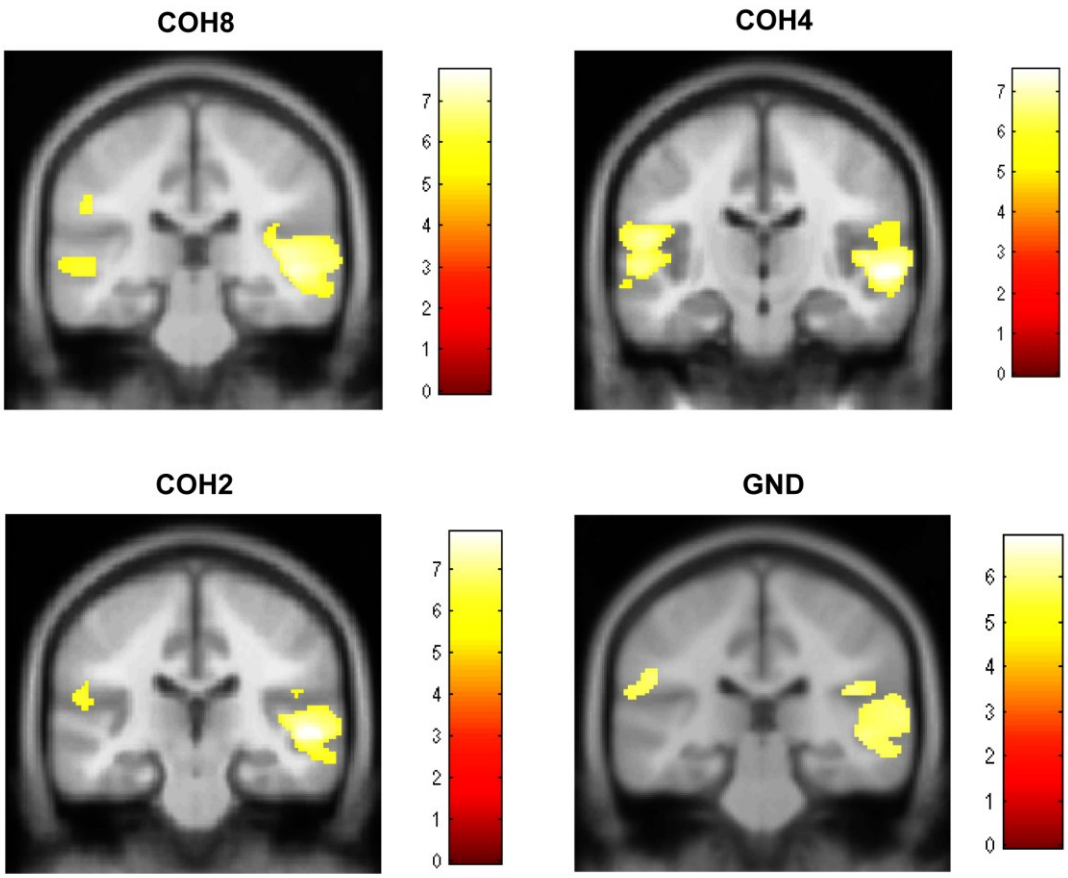

Figure 6.7: Activity in auditory cortex related to representation of figures with different coherence levels (1-sample t-test) during the early phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

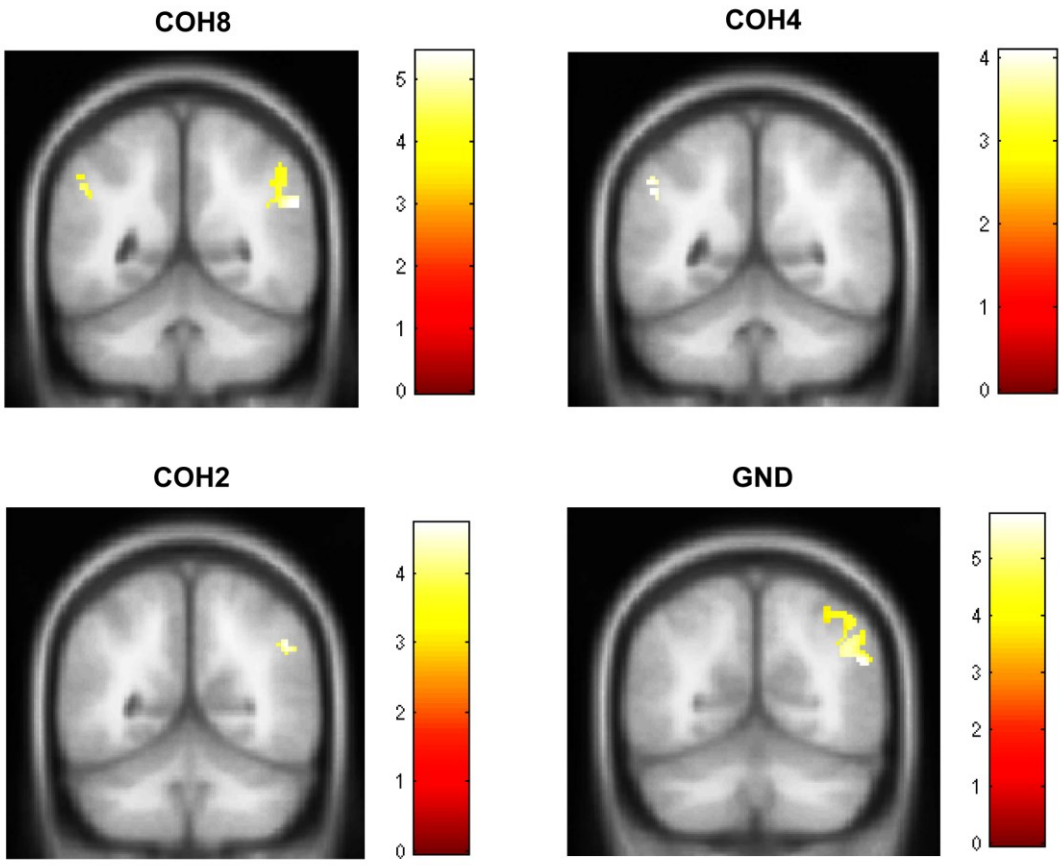

Figure 6.8: Activity in IPS related to representation of figures with different coherence levels during the early phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. A small-volume correction using a mask for IPS in the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to obtain these results. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

For the later sustained phase of the transition (from 300-600ms following the transition), the ANOVA analysis revealed a main effect of coherence bilaterally in the auditory cortex including the STS in the right hemisphere shown in figure 6.9. Analysis of brain regions that are purely activated as a function of coherence revealed clusters of activity in the auditory cortex as depicted in figure 6.9. Figure 6.10 shows activations in the IPS as a main effect of coherence. Activity in the IPS was also found to occur as a function of difference in coherence levels as shown in figure 6.10. Analysis for a main effect of each of the individual coherence levels revealed sources of activity in the auditory cortex as shown in figure 6.7as well as the IPS as shown in figure 6.8. The MNI coordinates of the sources for each of the above analyses are summarized in table 6.2.

Contrast	Brain areas	x	y	z	t-value	z-score
Effect of coherence	R HG	62	-16	6	7.35*	6.30*
	L HG	-52	-24	12	5.84*	5.25*
	R IPS	54	-52	30	4.98	4.59
	L IPS	-48	-62	34	3.98	3.77
COH8 vs. COH4	R HG	60	-16	4	3.54	3.24
COH8 vs. COH2	R HG	60	-16	10	4.63	4.05
	L HG	-52	-22	14	4.36	3.85
	R IPS	54	-54	30	3.99	3.59
COH8 vs. COH0	R HG	62	-16	6	6.07	4.96
	R IPS	54	-52	32	4.81	4.17
	L IPS	-50	-52	42	3.65	3.33
COH4 vs. COH0	R IPS	42	-58	28	3.38	3.12
COH8	R HG	50	-18	0	8.65*	5.29*
	L HG	-46	-26	16	9.71*	5.58*
	R IPS	50	-54	32	6.08*	4.37*
	L IPS	-50	-50	42	4.20	3.43
COH4	R HG	50	-22	2	8.90*	5.36*
	L HG	-52	-24	14	7.36*	4.87*
	R IPS	44	-60	34	5.17	3.95
	L IPS	-50	-48	36	4.96	3.85
COH2	R HG	50	-24	2	12.00*	6.11*
	L HG	-54	-20	16	6.49*	4.54*
	R IPS	56	-52	28	5.12	3.93
	L IPS	-50	-50	36	5.21	3.97
COH0	R HG	48	-28	2	6.80	4.66
	R IPS	54	-52	30	5.33	4.03
	L IPS	-50	-48	36	4.47	3.59

Table 6.2: MNI coordinates for reconstruction of evoked power in the late sustained phase of the basic SFG stimulus.

Source coordinates of activity during the late sustained phase of the transition (300-600ms following transition) to a figure specifically in the auditory cortex and the IPS are shown for the different contrasts as indicated. Asterisk indicates statistical significant at $p < 0.05$ (FWE) whilst other results hold at $p < 0.001$ (uncorrected).

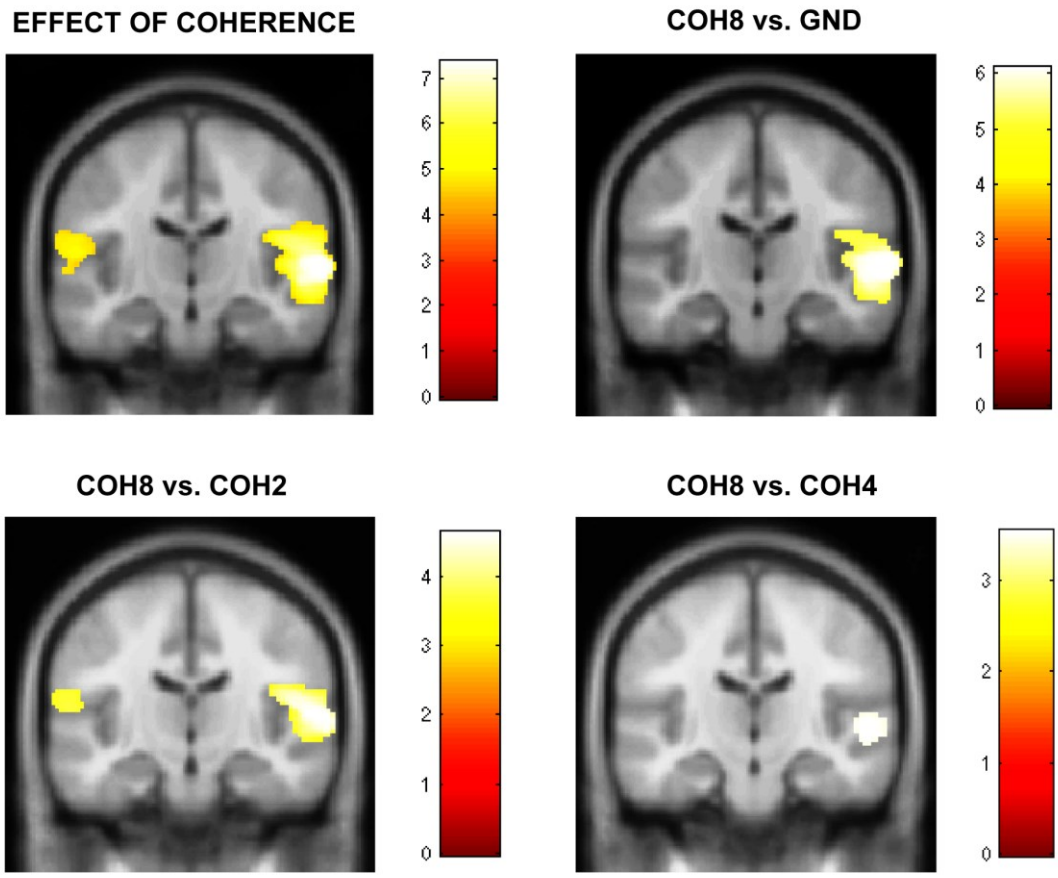

Figure 6.9: Activity in the auditory cortex as a main effect of coherence and difference in coherence levels during the late phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

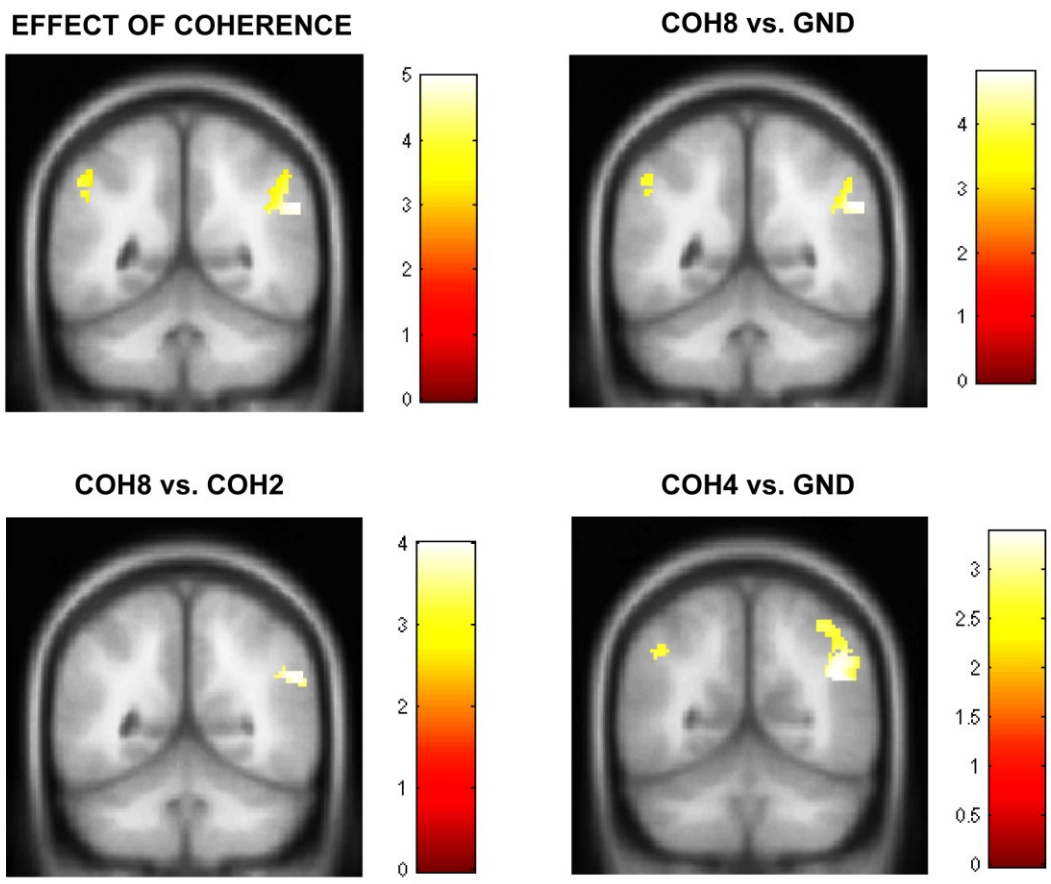

Figure 6.10: Activity in IPS as a main effect of coherence and difference in coherence levels during the late phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. A small-volume correction using a mask for IPS in the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to obtain these results. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

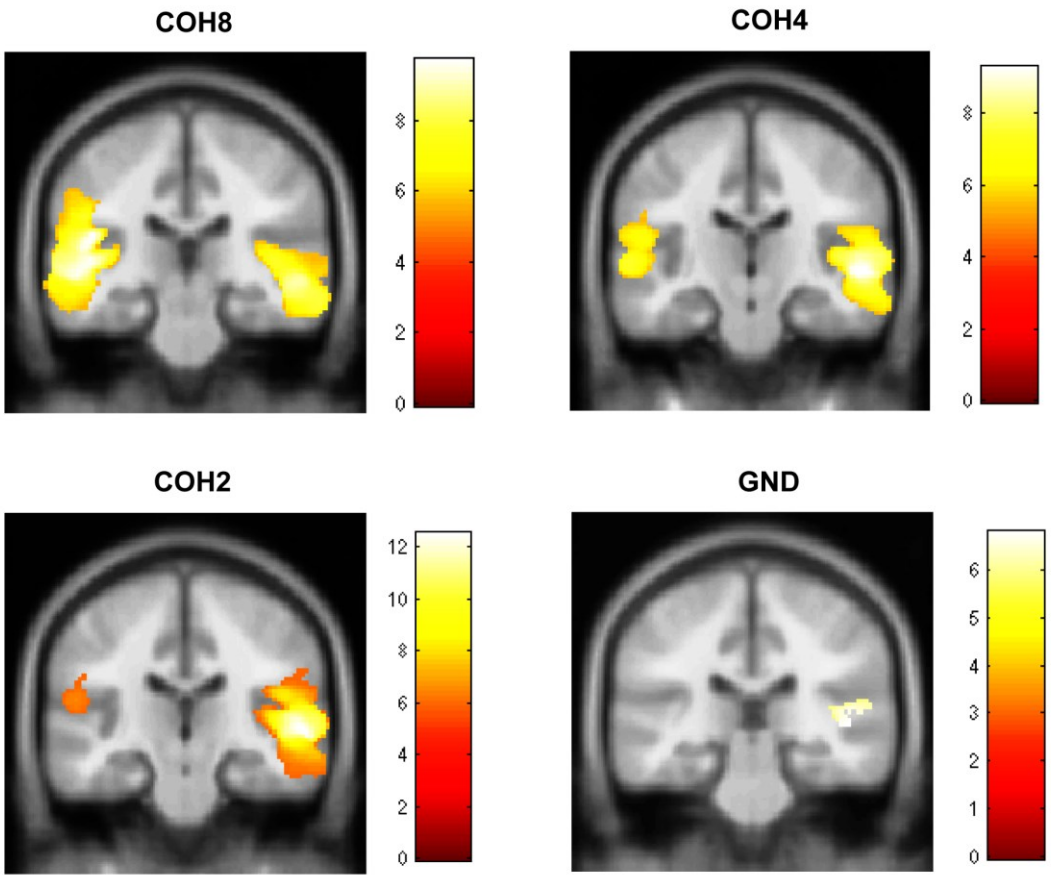

Figure 6.11: Activity in auditory cortex related to representation of figures with different coherence levels during the late phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

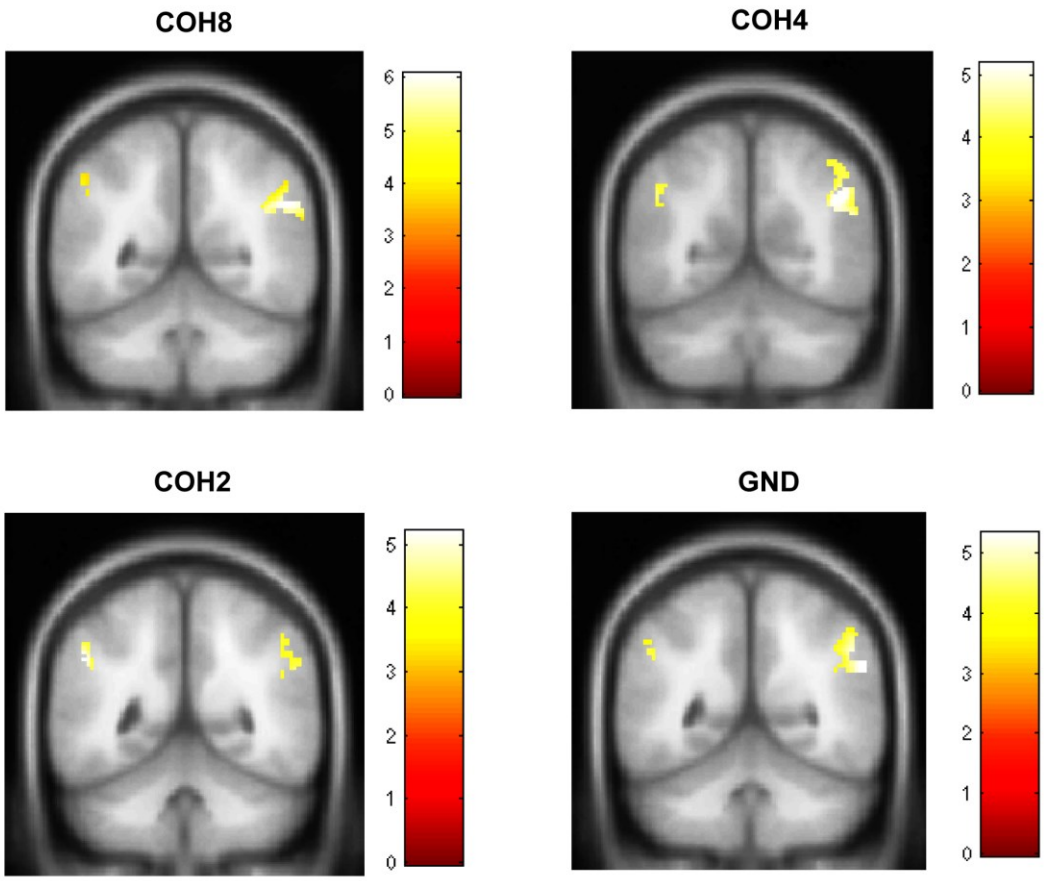

Figure 6.12: Activity in IPS related to representation of figures with different coherence levels during the late phase of the basic SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. A small-volume correction using a mask for IPS in the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to obtain these results. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

6.4 Results ('noise SFG')

6.4.1 Psychophysics

Behavioural results based on the 'noise' version of the SFG stimulus with a reduced bandwidth are shown in figure 6.13. The results from 5 participants indicate that listeners are still sensitive to figures that consist of alternating SFG and white noise chords. Performance for detection of figures increased monotonically with increasing coherence of the figures.

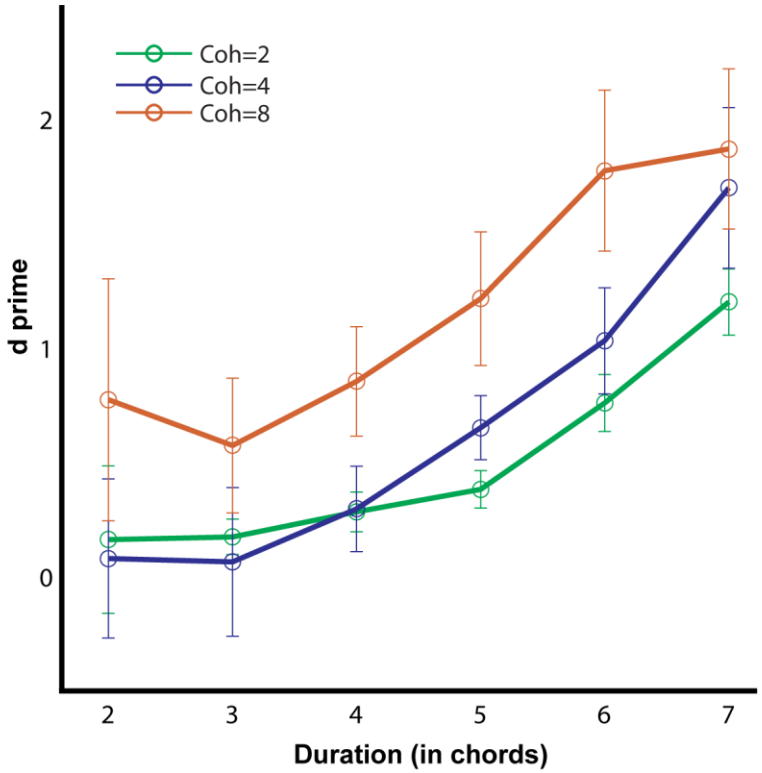

Figure 6.13: Figure-detection performance for the 'noise' SFG stimulus.

Behavioural results (d'; n=6) are plotted on the ordinate and the duration of the figure (in terms of number of 50ms long chords) is shown along the abscissa. The coherence of the stimuli was 2, 4, or 8 and six different levels of duration were tested. Listeners were required to press a button as soon as they heard a figure pop out from the background. Error bars signify one SEM.

6.4.2 Auditory-evoked fields

Figure 6.14 illustrates the group-RMS of auditory-evoked transition responses for the different coherence levels in the left and right hemispheres respectively. The data reveal an early transition response followed by a sustained component for transition to figures with coherence of 4 and 8. For transition to figures with coherence level of 2, the responses were not significantly different from the control condition whilst the responses for the higher coherence levels (4 and 8) increased as a function of coherence.

In this condition as well, the latencies at which the evoked field strengths (for transition to coherence = 4 and 8 only) became significantly different from the field strength for the control condition mirrored the behavioural latencies for supra-threshold detection of the figures. For coherence of 8, the field strength became significantly different after 173ms which corresponds to 100ms of SFG chords (remaining duration is white noise). This is equal to the duration of a figure with 4 repeating chords for which d' of 0.85 ± 0.20 were obtained. For coherence of 4, the corresponding evoked field latency was 347ms (comprising of 175ms of SFG chords) which corresponds to a figure whose duration is equal to 6.9 chords approximately. The d' for the detection of figures with coherence equal to 4 and duration equal to 7 was 1.70 ± 0.31.

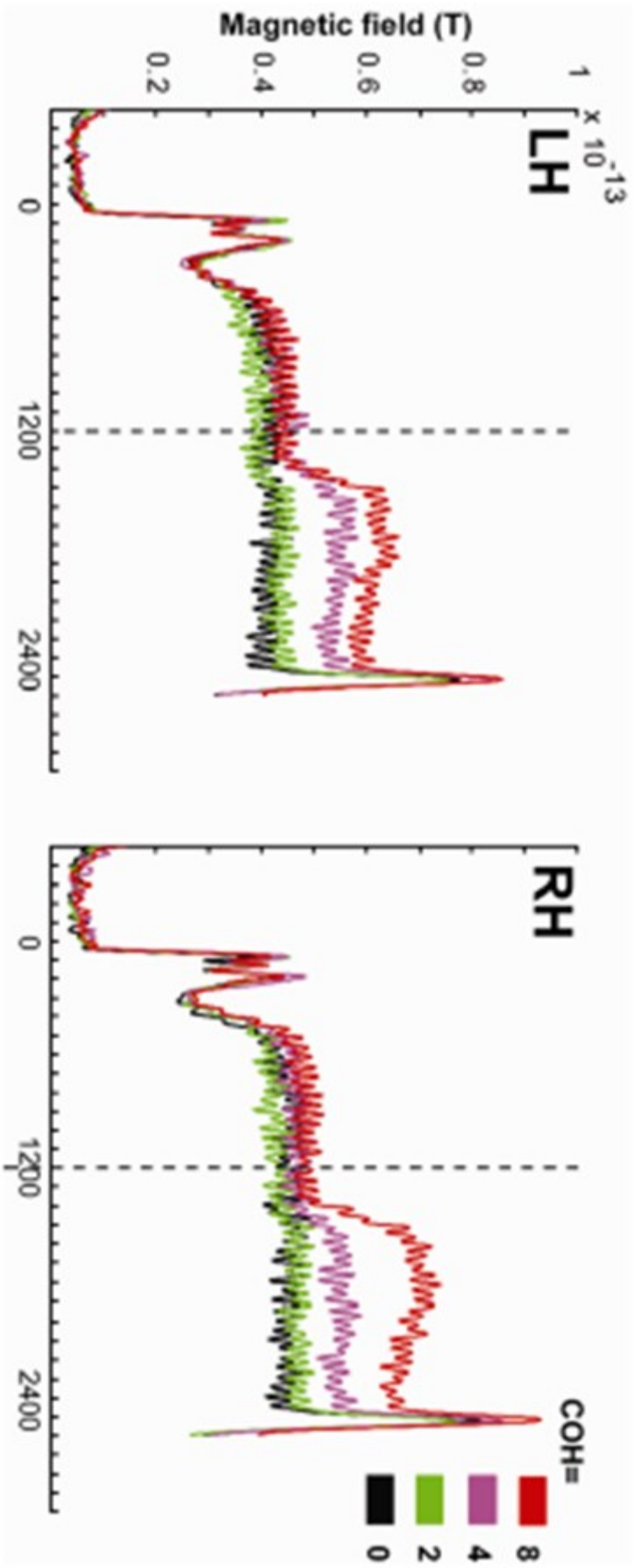
Magnetic field (T)
0.2
0.4
0.6
0.8
1
x 10^-13
LH
0
1200
2400
RH
0
1200
2400
COH=
0
2
4
8

Figure 6.14: Evoked field strengths in response to a transition from background to figure in the noise SFG stimulus.

The magnetic field strength in Tesla is plotted on the ordinate and time in milliseconds is plotted on the abscissa. The dotted black line separates the background from the following figure segments whose coherence is colour coded as indicated in the legend on the top right. The left and right panels indicate the resultant evoked field strengths in the left and right hemispheres respectively.

6.4.3 Source modeling

Similar source analyses as that performed for the basic SFG stimulus was done. The sources underlying the early and the late components were analyzed separately based on 300ms long windows: from 0-300ms for the early component and from 900-1200ms post-transition for the late component. The transition from background to figure occurred at 1200ms following sound onset.

IID source reconstruction method was used to identify the sources for the early and the late components separately. The modeled data was converted into NIFTI images that were taken to second-level and analyzed using three different parametric tests:

iv) ANOVA: to examine areas that are sensitive to increasing coherence;

v) 2-samples t-test: to investigate brain areas that specifically mediate the perceptual effects of figure processing without any confound related to intensity differences between the two different levels of stimuli; and,

vi) 1-sample t-test: to identify regions that are sensitive to the onset of the salient figures.

The results for the source reconstruction of evoked power during the early and the late components are summarized in tables 6.3 and 6.4 respectively.

Contrast	Brain areas	x	y	z	t-value	z-score
Effect of coherence	L IFG	-40	20	-14	3.89	3.71
COH8 vs. COH0	L IFG	-40	20	-14	3.86	3.52
COH4 vs. COH0	L IPS	-38	-58	-50	3.50	3.23
COH2 vs. COH0	L IPS	-28	-70	34	4.90	4.28
COH8	R HG	60	-26	10	5.55*	4.23*
	L HG	-56	-16	2	6.18*	4.52*
	R IPS	48	-56	36	4.98	3.93
	L IPS	-50	-48	36	4.23	3.51
COH4	R HG	48	-28	2	6.14*	4.50*
	L HG	-46	-30	18	6.10*	4.48*
	R IPS	58	-54	24	5.28	4.09
	L IPS	-46	-58	38	5.07	3.98
COH2	R HG	60	-26	10	8.03*	5.24*
	L HG	-60	-30	10	5.94*	4.41*
	R IPS	56	-52	28	4.36	3.59
	L IPS	-50	-48	36	5.03	3.96
COH0	R HG	48	-28	2	6.51	4.66
	L HG	-48	-16	16	6.26	4.56
	L IPS	-38	-58	50	3.50	3.23

Table 6.3: MNI coordinates for reconstruction of evoked power in the early transition phase of the noise SFG stimulus.

Source coordinates of activity during the early phase of the transition (0-300ms following transition) to a figure specifically in the auditory cortex and the IPS are shown for the different contrasts as indicated. Asterisk indicates statistical significant at $p < 0.05$ (FWE) whilst other results hold at $p < 0.001$ (uncorrected).

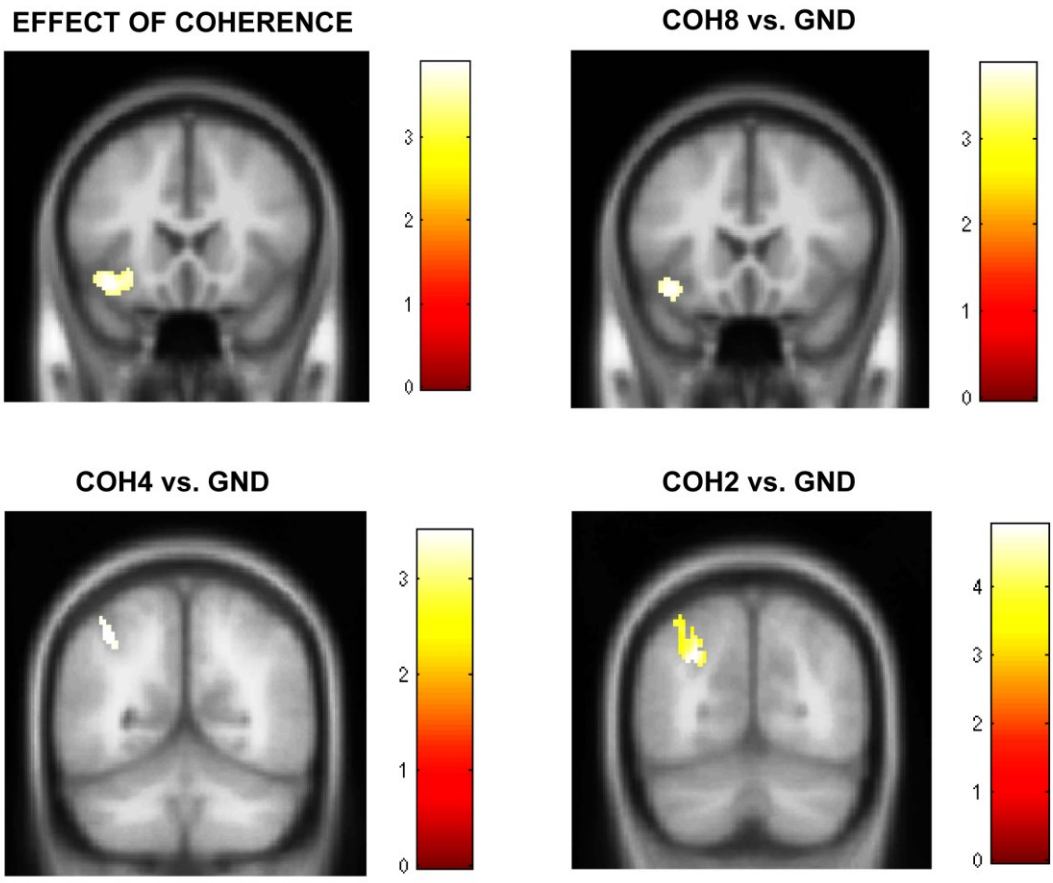

Figure 6.15: Activity in the inferior frontal and parietal cortex related to representation of figures with different coherence levels during the early phase of the noise SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. A small-volume correction using a mask for IPS in the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to obtain these results. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

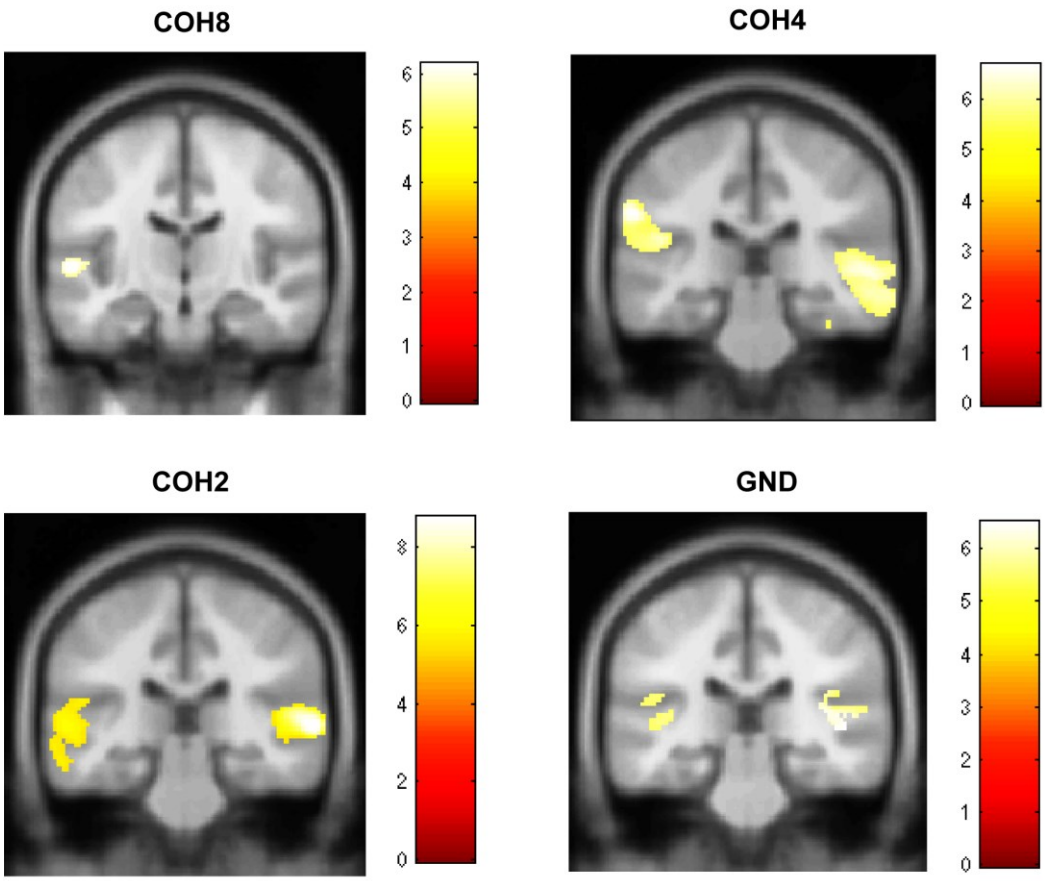

Figure 6.16: Activity in auditory cortex related to representation of figures with different coherence levels during the early phase of the noise SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

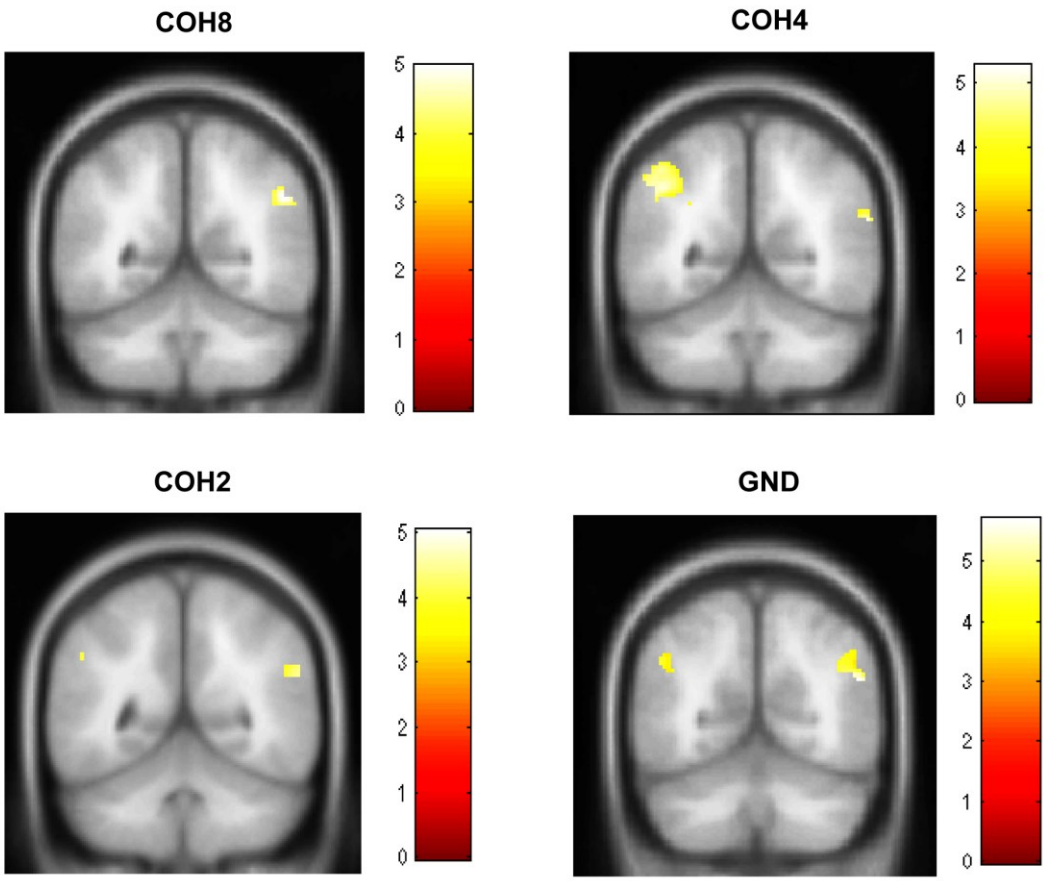

Figure 6.17: Activity in IPS related to representation of figures with different coherence levels during the early phase of the noise SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. A small-volume correction using a mask for IPS in the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to obtain these results. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

For the later sustained phase of the transition (900-1200ms following the transition), neither the ANOVA nor 2-samples t-tests revealed significant clusters of activity in the auditory or parietal cortex. Analysis for a main effect of each of the individual coherence levels revealed sources of activity in the auditory cortex as shown in figure 6.18 as well as the IPS as shown in figure 6.19. Interestingly, IPS was only activated in the coherent conditions with no activation in the control condition. The MNI coordinates of the sources for each of the above analyses are summarized in table 6.4.

Contrast	Brain areas	x	y	z	t-value	z-score
COH8	R HG	58	-6	10	5.96	4.42
	L HG	-54	-16	4	7.03*	4.87*
	R IPS	50	-56	34	5.65*	4.28*
	L IPS	-50	-52	38	3.92	3.32
COH4	R HG	50	-24	2	6.57*	4.69*
	L HG	-56	-18	2	6.49	4.66
	R IPS	56	-52	30	4.51	3.68
	L IPS	-50	-48	36	4.49	3.66
COH2	R HG	50	-24	2	7.35*	5.00*
	L HG	-50	-24	14	6.40	4.62
	R IPS	56	-52	28	4.77	3.82
	L IPS	-50	-48	36	4.66	3.76
COH0	R HG	60	-16	4	5.15	4.02
	L HG	-56	-18	2	4.88	3.88

Table 6.4: MNI coordinates for reconstruction of evoked power in the late sustained phase of the noise SFG stimulus.

Source coordinates of activity during the early phase of the transition (900-1200ms following transition) to a figure specifically in the auditory cortex and the IPS are shown for the different contrasts as indicated. Asterisk indicates statistical significant at $p < 0.05$ (FWE) whilst other results hold at $p < 0.001$ (uncorrected).

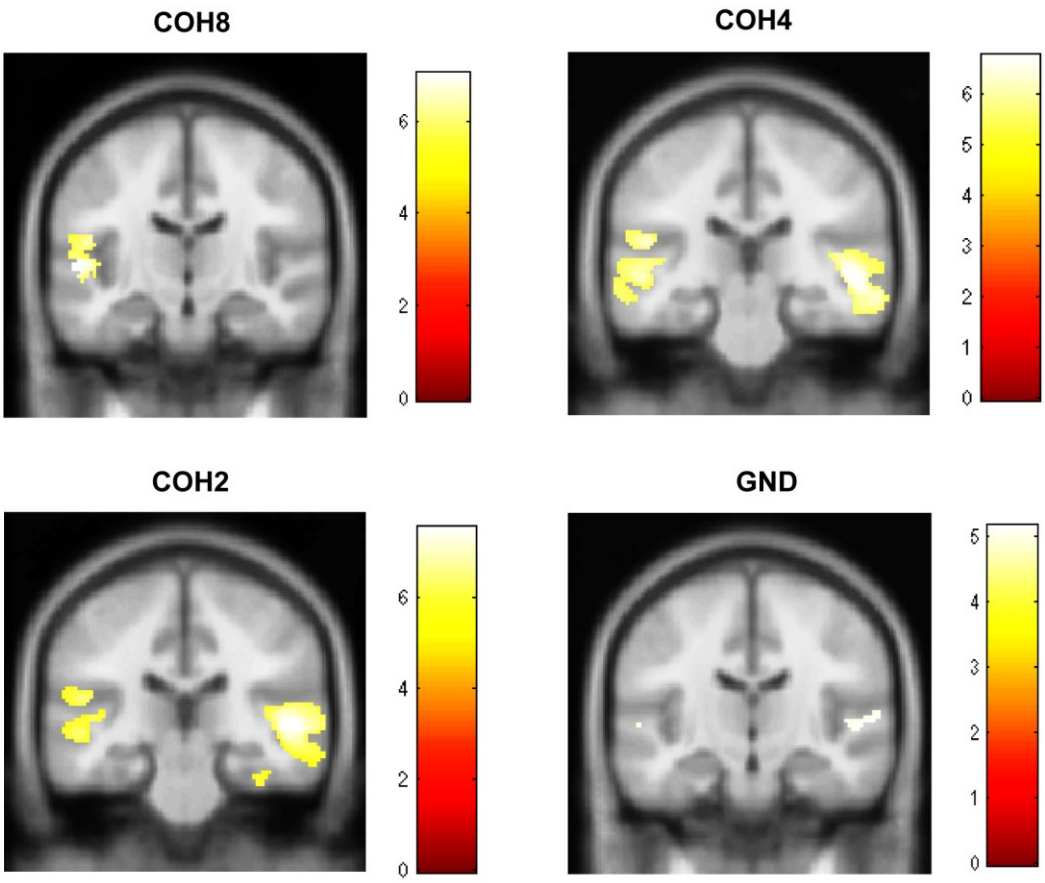

Figure 6.18: Activity in auditory cortex related to representation of figures with different coherence levels during the late phase of the noise SFG stimulus.

Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

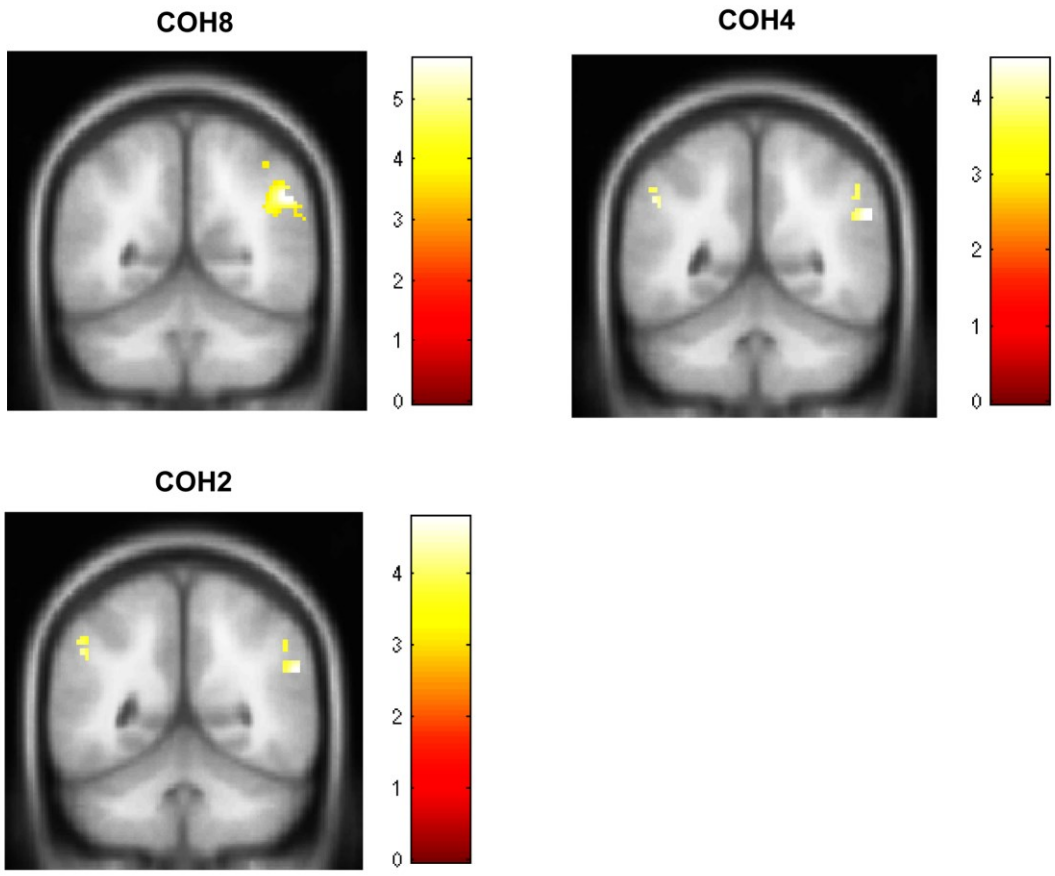

Figure 6.19: Activity in IPS related to representation of figures with different coherence levels during the late phase of the noise SFG stimulus.

Note that no activity in IPS was found for the transition to a background segment. Results are rendered on the coronal section of an average normalized brain template based on 152 T1 scans and shown at $p < 0.001$ uncorrected. A small-volume correction using a mask for IPS in the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to obtain these results. The t-values for the significant voxels for each contrast are scaled according to the heat map on the right of each image.

6.5 Discussion

The MEG experiment was designed in order to obtain a complete picture of figure-ground analysis in the SFG stimulus. The psychophysics, modeling and fMRI results suggest that detection of target signals in the SFG stimulus depends on mechanisms that cannot be explained by models based on deterministic patterns based on sequences of tones (Fishman and Steinschneider, 2001; Micheyl et al., 2005, 2007a). The temporal coherence model of auditory scene analysis, on the other hand, provided a sound explanation for segregation in the complex SFG stimulus. The fMRI results provided an intriguing hypothesis for a role for the intraparietal sulcus in mediating or representing temporal coherence associated with the salient figures (Shamma et al., 2011). These set of results provided added stimulus to examine these hypotheses in further detail using MEG. Firstly, the aim was to investigate how segregation in the SFG stimulus builds up over time and whether the auditory cortex is involved in it or not, given the fact that there was no modulation of BOLD activity in the auditory cortex. Secondly, the fMRI results provided a strong spatial prior to investigate activity in the IPS specifically related to representation of temporal coherence.

The analysis of evoked field strengths revealed an interesting picture: a strong peak developed transiently after the transition with smaller latencies for more coherent figures, and was followed by a sustained component that persisted throughout the duration of the figure. This pattern of response was observed for each of the different coherence levels but their amplitude was graded according to the coherence. The effect of coherence

was found for all coherence levels (2, 4, and 8) for transitions in the basic stimulus but only for the more coherent signals (4, and 8) in the noise SFG stimulus. Thus, for both conditions, for which there was no difference in behavioral sensitivity to the figures (see chapter 3), the MEG evoked responses also show a similar profile, suggesting a common neuronal basis for segregation of the figure in both types of signals. However, one caveat is that the post-transition segment was associated with more spectral energy than the pre-transition figure segment and the evoked responses thus cannot be considered purely perceptual responses. Still, the responses across the different coherence levels (matched for intensity) show differential scaling according to the perceptual salience of the figure segments.

These data suggest that the underlying sources are sensitive to the salience of the figure, which is not based on differences in intensity as the power associated with the post-transition figure segments was balanced. These responses were based on sensors in the auditory cortex that are most selective to sound onset (see section 6.2.4). Thus, time-locked activity in the auditory cortex shows sensitivity to the onset of coherent figures and also mediates sustained perceptual representation of the figure. The role of the auditory cortex is discussed in greater detail in the next section.

Another result that deserves attention is the correspondence between the latencies at which the evoked field strengths become significantly different from baseline and the duration of the corresponding figure that can be reliably discriminated. In the psychophysical experiments, listeners were encouraged to press a button as soon as they detected a figure and the

reaction times could thus be smaller than the duration of the figure. The analysis of the MEG evoked field latencies suggest that in passive conditions, the brain needs at least the time equal to the duration of the figure to respond to their emergence from the background. It is likely that in active task conditions, the MEG latencies might be smaller than in passive conditions and mirror behavioural reaction times more accurately.

6.5.1 Role of auditory cortex and IPS revisited

The results from source reconstruction of the early and the late phase of the transition for both the basic and the noise stimuli revealed activations in the auditory cortex including STS as well as the IPS. Different statistical tests were performed on the 3D images obtained from the inverse reconstruction to examine areas that: i) represent coherence, and ii) represent difference in coherence.

For the early phase of the basic stimulus, the right auditory cortex was found to represent the difference in the coherence between figures with coherence of 8 vs. each of the lower coherence levels respectively. However, the auditory cortex was activated bilaterally for the representation of coherence as a main effect for each of the coherence levels including the ground condition. Similar effects were found for the same contrast in the IPS as well, following a small-volume correction at a significance threshold of $p < 0.001$ (uncorrected). In the later phase, similar results were obtained as a function of both contrasts in the auditory cortex but interestingly, the IPS was additionally activated for the representation of the perceptual salience of the figures, i.e. it was found to represent the difference in

coherence between figures. Overall, these data suggest that the auditory cortex is sensitive to the onset of coherent segments and also represents coherence (or temporal coherence). The IPS on the other hand, shows a differential pattern of response: it represents the difference in coherence only in the later sustained phase of the transition (together with the auditory cortex). This evidence points towards a possible hierarchy of processing the transition from background to figure: the auditory cortex may be involved in initial encoding of the figures and the IPS may be additionally recruited in the perceptual representation of the figure.

The source reconstruction results for the early phase of the noise stimulus revealed modulation of the left inferior frontal gyrus (IFG) and IPS as a function of difference in coherence. Contrary to the results from the early phase of the basic stimulus, no activation in the auditory cortex was found. The activity in the fronto-parietal cortex may reflect greater top-down drive to segregate the figure chords interspersed by white noise. On the other hand, auditory cortex along with the IPS was activated as a function of discrete coherence levels. For the source modeling of the late phase, however, although auditory cortex and IPS were involved in representing the coherence levels (2, 4, and 8) there was no activity associated with IPS for the ground condition. These set of results suggest a mechanism based in the temporal and fronto-parietal cortex that is related to encoding the figures. However, the role of the auditory cortex in perceptual representation of the figures in the noise stimulus was not substantiated to the same extent as in the case for the basic stimulus. This could be due to

the masking effects of the noise segments whereby detection of the embedded figures relies more on top-down mechanisms in the fronto-parietal cortex.

The activation of auditory cortex in the MEG compared to the lack of activation in the fMRI study may be due to differences in the stimulus paradigm, background acoustic environment (continuous scanning in MRI vs. quiet conditions in MEG) or the temporal resolution of the measurement technique. The BOLD response may not have adequately capture time-locked activity in the auditory cortex.

In other MEG paradigms based on IM stimuli, activity in the auditory cortex has been demonstrated (Gutschalk et al., 2008; Elhilali et al., 2009b; Wiegand and Gustschalk, 2012). Gutschalk and colleagues (2008) uncovered a response termed as the awareness related negativity (ARN) specifically for detected target tones with no activation for the undetected targets in the auditory cortex. In a more recent IM experiment, Wiegand and Gutschalk (2012) found BOLD activity in the medial Heschl's gyrus as a function of detected vs. undetected targets. They also carried out MEG recordings and found an ARN response similar to the previous study (Gutschalk et al., 2008). The analysis of the fMRI data, however, was limited only to HG, PT and STG and did not focus on areas outside the temporal lobe. These results offer impetus for an active figure-detection task based on the SFG stimulus and to examine whether ARN is elicited in the auditory cortex as well as the IPS.

6.5.2 Temporal coherence

As demonstrated in chapters 3 and 4, the onset of the figure in the SFG stimulus is associated with an increase in temporal coherence. According to the model of Shamma and colleagues (2011), coherence between different frequency channels can temporally bind these channels together and assign them as belonging to one source or representing a separate object distinct from the background with uncorrelated elements. In the context of this model, the present stimulus design results in an increase in temporal coherence after the transition to the coherent figure segment. Modeling results (see section 4.4) suggest temporal coherence as a plausible mechanism in the detection of the figure based on coherent elements in a background of incoherent elements. Another aspect of the model is that it is based on stimulus-driven, phase-locked activity between distinct set of neuronal populations that code for the feature of interest (here, frequency). Thus, source reconstruction based on evoked power in the post-transition segment may reflect sources that compute or represent temporal coherence.

As described in the previous section, the auditory cortex and the IPS were found to be activated during the figure segment and encoded coherence as well as difference in coherence. In the basic stimulus, however, the IPS was found to be activated during the sustained phase rather than the early encoding phase suggesting that it may be involved exclusively in the representation of temporal coherence that may be processed in the auditory cortex and fed forward to higher centres in the parietal (or frontal) cortices. These results provide a basis to consider a

hierarchical network in the processing of novel salient sounds in the acoustic environment based on predictive coding mechanisms. This account holds that the brain is constantly trying to predict sensory input and generates prediction errors when the incoming input does not match predictions based on long-term templates formed on the basis of exposure to the environment (Friston, 2005). In this model, regions that are placed lower in the hierarchy are specifically involved in processing prediction errors generated by the mismatch between the predictions derived from higher centres and the incoming sensory information. In this context, one may speculate that the IPS represents a higher node in this hierarchical processing system that signals (and represents) a change in the acoustic environment based on low-level stimulus processing in the auditory cortex. Although predictive coding has been shown to be relevant for mediating the MMN response (Garrido et al., 2009) and multistability in the streaming signal (Winkler et al., 2012; Mill et al., 2013), whether it applies to detection of changes in complex acoustic scenes and its relationship with the temporal coherence model remains to be investigated.

6.5.3 Limitations

The purpose of the MEG study was to examine stimulus-driven or bottom-up segregation in the absence of directed attention to the stimulus. Listeners' were kept naïve regarding the SFG stimuli and focused on a visual task. However, the visual task was quite easy to perform and may not have taxed their attentional resources much. It is also possible that listeners may briefly focus on the sounds whilst performing the incidental visual task.

These potential confounds limit the explanation of a purely passive account of segregation. However, listeners did not report any particular interest in the sound stimulation at the end of the experiment.

These results suggest the basis for a pre-attentive mechanism that is sensitive to temporal correlations across frequency channels in accordance with the temporal coherence model. Further work is required to elucidate the specific role of attention in an active task paradigm, for instance, by manipulating the attentional load or difficulty of an unrelated visual or auditory task. It is also possible that non time-locked activity in auditory and parietal cortices is crucial and ongoing frequency-time analysis and localization of induced power may shed further insights into the processing of salient figures in a random background.

Chapter 7. General Discussion

This thesis examined the brain bases of auditory segregation based on a novel stochastic signal, referred to as the stochastic figure-ground (SFG) stimulus. The problem of how a natural scene is parsed into its constituent components, i.e. individual objects for subsequent processing and perceptual representation is a fundamental problem in neuroscience. Visual and auditory scenes comprise multiple objects, and objects of interest need to be encoded as a coherent whole that is distinct from other objects in the background. Encoding of an object may be based on a number of grouping mechanisms that operate on certain attributes of the object, for instance, grouping based on luminance in vision or grouping on the basis of frequency in audition. The Gestalt psychologists examined a number of such principles of binding such as common fate, collinearity, good continuation, symmetry and convexity which were originally examined in the context of visual binding but have also constructively influenced the principles of grouping in audition (Bregman, 1990; Denham and Winkler, 2013).

Bregman proposed principles of auditory perceptual organization based on the analysis of simple deterministic sequences of alternating low and high frequency tones, known as the streaming signal (Bregman and Campbell, 1971; van Noorden, 1975). A number of fundamental principles of auditory scene analysis have been uncovered based on psychophysical and physiological examination of responses to these stimuli (Bregman, 1990; Carlyon, 2004; Fishman et al., 2001; Micheyl et al., 2007a; Denham and Winkler, 2013). In the following section, the relative merits and

limitations of the different signals used to study segregation are discussed, leading to the motivation for the development of stochastic stimuli like the SFG stimulus.

7.1 Stimuli for studying auditory segregation

A typical biological entity has to deal with a plethora of sensory inputs that occur in a random, unpredictable manner. Moreover, such signals are often contaminated with noise which renders the problem of accurate perception more challenging, necessitating robust neural encoding mechanisms. Another aspect of the nature of sensory stimulation can be understood in the context of information theory: deterministic signals convey less information, whilst stochastic signals convey more information and more effectively engage the neural machinery underlying perception. Signals used in auditory scene analysis research can thus be classified accordingly as deterministic or stochastic stimuli.

Deterministic signals used to investigate auditory segregation include streaming signals (van Noorden, 1975; Bregman, 1990), sequences of tones based on the phenomenon of informational masking (Neff and Green, 1987), or oddball stimulus patterns (Näätänen et al., 2007). These stimuli are discussed in detail in section 1.4. In spite of their usefulness, they are characterized by certain features that constrain their utility in understanding auditory perception as occurs in the natural world. These signals share a few limitations: they have a deterministic temporal structure; contain narrowband target signals; the foreground and background streams are usually non-overlapping and out of phase; and, the target signals are often

separated by a band-stop region with little energy. These features are not representative of signals in the natural environment characterized by multiple, overlapping channels with stochastic temporal structures. These signals, to their advantage, offer a simplistic approach with flexible control of stimulus parameters that enables systematic analysis of neural responses to specific acoustic attributes.

The aim of this thesis was to understand segregation mechanisms that operate in realistic auditory environments. To overcome the limitations posed by the conventional stimulus paradigms as discussed above, a novel signal with a stochastic spectrotemporal structure was developed. The SFG stimulus is conceptually similar to the visual coherent dots motion paradigm (Shadlen and Newsome, 1996) which involves manipulation of the direction of motion of certain dots that form the "figure" against the backdrop of other dots which move in random directions that comprise the "ground". The SFG signal is based on a similar approach: the figure was based on coherence in time, i.e., it was comprised of a few channels that repeated synchronously whilst the remaining frequency channels were characterized with random fluctuation patterns. Unlike other tonal sequences, the figure in the SFG stimulus was indistinguishable at each moment in time and could only be extracted by integrating across both frequency and time. This also negated the use of selective attention to follow the target stream and extract it from the background. Moreover, the pattern of the figure, i.e., its spectrotemporal properties varied from trial to trial and required robust integration across both spectral and temporal dimensions to detect the

figure. These differences highlight the significance of the SFG stimulus for characterizing naturalistic auditory segregation behaviour.

Chapter 3 reports a number of psychophysical experiments that examined listener's segregation performance in the SFG stimulus. The basic experimental paradigm required listeners to detect brief figures (duration ranging from 100 – 350ms) with varying number of temporally correlated channels (defined as the 'coherence': 1, 2, 4, 6, or 8). Sensitivity or d' was measured as a function of both these factors and was found to increase monotonically with increasing duration and the coherence of the signal. Furthermore, figure-ground discrimination abilities were quite robust: listeners could successfully detect coherent patterns as brief as a few hundred milliseconds (maximum duration of figure was 350ms). Segregation in streaming stimuli, on the other hand, builds up over a couple of seconds (Anstis and Saida, 1985). This suggests the existence of a highly tuned bottom-up segregation mechanism that is sensitive to the salience of brief figures.

Additional experiments reported in chapter 3 manipulated the spectrotemporal structure of the figure and examined the sensitivity of the listeners to the modified figure patterns. In spite of changes in the temporal structure (speeding up of the stimulus in experiment 3), spectral shape of the figures (ramped vs. linear figure patterns in experiments 4a and 4b), changes in the stimulus pattern (isolated presentation of figures without the preceding and succeeding chords in experiment 5), and introduction of masking noise between successive chords (experiments 6a and 6b), figure-

detection performance was mostly unaffected in comparison to experiment 1 (except for a slight drop in performance for experiments 4a and 4b). These data suggest that the "pop-out" of these salient figures may be mediated by a robust, bottom-up, stimulus-driven mechanism. The figure patterns are characterized by correlations in frequency and time and implicate a mechanism that computes such correlations in complex stimulus patterns.

Irrespective of the nature of stimuli used in auditory scene analysis, the goal of the auditory neuroscientist is to answer the question – *"What in neural terms, corresponds to the final representation of what we hear?"*

Treisman (1999) explored the nature of this question in vision and proposed an important role for temporal coding mechanisms. In the following section, this proposal is examined in the case of vision as well as audition, with specific examination of the temporal coherence model of auditory scene analysis.

7.2 Role of temporal structure in binding

Standard models of segmentation in vision and audition place great emphasis on the role of the spatial and spectral structure respectively. Differences in spatial location and spectral profiles provide strong cues to parsing the visual and acoustic scene respectively. However, recent work has explored the role of temporal structure in binding in vision (Treisman, 1999; Blake and Lee, 2005) and audition (Elhilali et al., 2009a; Shamma et al., 2011, 2013) which is discussed in this section.

In vision, grouping principles focus on features that are defined in terms of spatial discontinuities in luminance, colour, or texture that constitute what is referred to as *spatial structure* (Blake and Lee, 2005). Spatial cues help define edges and borders between objects in the visual scene that represent static cues for segregation. However, it is evident that the visual world is highly dynamic, characterized by the movement of objects and observers. Thus, it is useful to consider visual segmentation as a more complicated process that must also integrate dynamic cues for parsing. This was conceptualized by the Gestalt psychologists as grouping by *common fate.* This principle encapsulates the importance of *temporal structure* and has been demonstrated to be an important grouping factor in several studies of visual segmentation (reviewed by Blake and Lee, 2005). The role of temporal structure in visual grouping is considered to be complementary to the role of spatial structure, and when these two structures are in conflict, the relative salience of the two cues determines the final outcome of the grouping process (Blake and Lee, 2005).

Similarly, the Gestalt principle of common fate has also been employed by auditory scientists in the perceptual analysis of acoustic scenes (van Noorden, 1975; Bregman, 1990). Sounds that start and stop together are said to share a common fate and can be attributed to the same acoustic source. A source of sound is associated with several spectrotemporal properties such as pitch, and intensity that co-vary together in time, and this temporal feature can be exploited for segregation. The most commonly accepted models of stream segregation, however, attribute a predominant

role to spectral structure: differences in frequency between two streams promote the activation of distinct populations of neurons in the auditory cortex that corresponds to the perceptual representation of the streams (Fishman et al., 2001; Micheyl et al., 2005, 2007b). Studies in macaques, songbirds, guinea pigs and humans demonstrate this phenomenon at different levels of the auditory pathway from the cochlea to the cortex (Fishman et al., 2001, 2004; Bee and Klump, 2004, 2005; Micheyl et al., 2005; Gutschalk et al., 2005, 2007; Wilson et al., 2007; Pressnitzer et al., 2008). This 'population-separation' model of segregation can explain the classical streaming effect but is not sufficient to explain segregation in more complicated stimulus patterns with multiple, overlapping frequency components.

Elhilali and colleagues (2009a) demonstrated the shortcomings of the population-separation model by showing that alternating and synchronous patterns of tones produce the same response profiles in the auditory cortex although they have different perceptual signatures: the alternating sequence of tones is perceived as two streams whilst the synchronous sequence of tones is perceived as a single stream. These findings suggest the importance of temporal structure in auditory segregation: sound elements with high temporal coherence may be grouped as one stream whilst elements with low coherence are perceived as separate streams (Shamma et al., 2011). This model of segregation is known as the 'temporal coherence' model and stresses the importance of temporal features in addition to spectral features in determining the perceptual representation of sound scenes. The features

of the model are discussed in detail in section 4.2: its main advantages being that it can be applied even in the case of complex stimuli with multiple overlapping frequencies such as the SFG stimuli. Indeed, modeling of the SFG stimuli revealed patterns of temporal coherence that mirrored the behavioural figure-detection responses (see Figure 4.4). The modeling simulations suggest that the SFG figure patterns show strong temporal coherence which may drive the segregation of these figures.

The temporal coherence model presents a strong case for segregation of stimuli that share temporal dependencies. However, whether the model applies in case of stimuli associated with bistable perception remains an open question. Furthermore, although the experimental results provide a counter-argument against population-separation models of segregation, the neurophysiological bases of temporal coherence computations are yet to be determined. How does the brain compute temporal relationship amongst spatially distributed neuronal ensembles that encode different acoustic features? Until these questions are resolved, the temporal coherence model remains incomplete. The following section discusses the neural substrates of auditory segregation and temporal coherence in more detail.

7.3 Neural substrates of auditory segregation

Models of stream segregation suggest that physiological properties of auditory cortical neurons such as frequency selectivity (and tonotopic organization of the auditory system), adaptation and forward masking result in the activation of spatially segregated populations of neurons that encode different streams. Recordings from the auditory cortex in macaques

(Fishman et al., 2001, 2004) first revealed these features which were later confirmed in songbirds as well (Bee and Klump, 2004, 2005). Several lines of evidence including functional imaging studies in humans showed that auditory segregation occurs along distributed centres of the ascending auditory pathway including the cochlea (Pressnitzer et al., 2008), the thalamus (Kondo and Kashino, 2009, 2012), the primary and non-primary auditory cortex (Deike et al., 2004, 2010; Gutschalk et al., 2005, 2007; Wilson et al., 2007; Dysktra et al., 2011; Ding and Simon, 2012; Mesgarani and Chang, 2012; Zion-Golumbic et al., 2013) as well as the parietal cortex (Cusack, 2005; Hill et al., 2011).

A majority of these studies were based on stimuli with no temporal correlations amongst its components and thus the implicated structures cannot be said to reflect processing or representation of temporal coherence. Temporal coherence may be encoded at the level of the auditory cortex although Elhilali and colleagues (2009a) did not find single-unit evidence supporting this hypothesis. It is possible that ensembles of neurons compute coherence within the cortex or in higher-order auditory-related areas such as the parietal or frontal cortex. The data from the fMRI study reported in chapter 5 suggest that the IPS may represent the locus of temporal coherence analysis. BOLD activity in the IPS most strongly co-varied with the salience of the figure which was shown to be correlated to temporal coherence in the modeling simulations shown in chapter 4. The fMRI data, however, cannot specifically resolve whether coherence is encoded at the level of the IPS directly or after encoding at the level of the auditory cortex.

The latter represents a significant possibility as the IPS is in receipt of anatomical projects from the auditory cortical areas (Cohen, 2009) and is also a locus of selective attention (Cusack, 2005; Fritz et al., 2007) as well as auditory spatial attention (Lee et al., 2013).

The MEG data, reported in chapter 6, provide evidence for a role for the sustained representation of temporal coherence in the IPS after initial encoding in the auditory cortex. Temporal coherence reflects a phase-locked operation and reconstruction of phase-locked evoked power after the transition to a figure segment revealed clusters of activity in the IPS along with the auditory cortex including STS. Furthermore, the representation of temporal coherence was found to scale with the coherence of the figure in a manner that is consistent with the effects of coherence on behaviour as well as the modeling response curves. Activity in the parietal (as well as frontal) cortex appeared to be more relevant for encoding of figures that were embedded in alternating white noise segments. Together, the fMRI and MEG data suggest that both the auditory and parietal cortices may be involved in the analysis of temporal coherence. Although the MEG data suggest that activation of auditory and parietal cortex becomes stronger over time, the exact temporal relationship and causal interplay between these areas remains to be determined.

In terms of the underlying neural coding schemes, two basic mechanisms are relevant: rate coding (Barlow, 1972; Shadlen and Newsome, 1994) and temporal correlation (Abeles, 1982; Mainen and Sejnowski, 1995). Rate coding schemes propose that information is

conveyed by the average firing rates of neurons. It is unlikely, however, that the firing of single neurons can capture the fine temporal structure of dynamic stimuli, as information about the timings of individual spikes is not retained in the average rate code. However, the firing rate of an ensemble of neurons may fluctuate in a time-locked manner to time-varying stimuli with high precision (Shadlen and Newsome, 1994). The temporal correlation hypothesis, on the other hand, suggests that information is conveyed in the timing of individual spikes. The spiking activity between separate groups of neurons may be synchronized in a stimulus-locked fashion. In human vision, both coding schemes are considered to be able to explain its temporal acuity (Blake and Lee, 2005) but whether the same applies in the case of audition remains to be investigated in greater detail.

7.4 Future directions for research

A feature of the functional imaging studies examined in this thesis is that they examined figure-ground discrimination in passive listening paradigms. Although this allowed an examination of bottom-up, stimulus-driven mechanisms that mediate the pop-out of the salient figures, the influence of attention on behaviour and the underlying neural responses could not be assessed. Attentional state modulates sensory responses in the auditory cortex (Fritz et al., 2007) and can also bias selection of particular acoustic features that bind other temporally correlated features belonging to the same object (Shamma et al., 2011). Analysis of active figure-ground discrimination may further help to disentangle the specific roles of higher-order areas such as the parietal or prefrontal cortex in auditory segregation.

Although both the fMRI and MEG experiments revealed activations in IPS related to figure-ground processing, a causal role for IPS is yet to be established. This may be achieved by the use of transcranial magnetic stimulation (TMS) to selectively disrupt neuronal processing in the area and examine segregation performance before and after creating virtual lesions in the IPS (e.g., Kanai et al., 2008).

A prominent feature of the streaming paradigm is the phenomenon of bistability that allows an investigation of perceptual responses in the absence of corresponding physical changes. The frequency separation and tone presentation rate can be modified to produce a bistable percept that switches between that of one stream to two streams (van Noorden, 1975; Pressnitzer and Hupe, 2006). It remains to be explored whether similar bistability can be achieved in the SFG stimulus: alternating chords could comprise different repeating frequencies to form two sets of competing figures that vie for perceptual dominance.

Another unresolved question that merits attention is the brain basis of temporal coherence analysis as discussed in the previous section. Unraveling the precise structural and functional bases of temporal coherence computations would help to elucidate the neural architecture of auditory segregation in complex scenes as investigated here. The use of SFG stimulus in neurophysiological experiments in behaving animals such as ferrets represents a promising approach in this vein (Shamma et al., 2013).

Another question that deserves further investigation in the auditory perception literature is the relationship between perceptual performance on an auditory task and structural features (e.g. surface area, or grey matter volume) of relevant areas such as the auditory (or parietal) cortex. It has been demonstrated the surface area of the primary visual cortex predicts variability in conscious visual experience (Schwarzkopf et al., 2011). In multistability, perceptual fluctuation rates may be constant for a certain individual but vary considerably across individuals. In vision, neuroanatomical substrates for individual variability in spontaneous switching behaviour have been identified in the parietal cortex (Kleinschmidt et al., 2012), but similar assessment of individual differences in auditory multistable phenomena is lacking. Thus, the structural basis of inter-individual differences in target detection for the streaming as well as SFG paradigms may offer new insights linking behaviour and cognition to anatomy (Kanai and Rees, 2011).

Recently, a theory of brain function based on predictive coding (Rao and Ballard, 1999; Friston, 2005) has become a prominent model of human behaviour and cognition. In the specific case of audition, predictive coding accounts for the MMN response (Garrido et al., 2009), pitch perception (Kumar et al., 2011), as well as multistability in auditory stream segregation (Winkler et al., 2012) have been proposed. A recent study on bistable visual perception demonstrated modulation of top-down connectivity from the fronto-parietal to visual cortex during perceptual transitions using dynamic causal modeling of BOLD data (Weilnhammer et al., 2013). This work

represents another direction, embedded in the predictive coding framework, to investigate functional interactions and causal interplay between fronto-parietal cortex and auditory cortex in auditory bistable phenomena in particular, and auditory perceptual organization, in general.

BIBLIOGRAPHY

1. Abeles M (1982) Role of the cortical neuron: integrator or coincidence detector? Isr J Med Sci 18: 83–92.
2. Aertsen AM, Johannesma PI (1981a) A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. Biol Cybern 42: 145–156.
3. Aertsen AM, Johannesma PI (1981b) The spectro-temporal receptive field. A functional characteristic of auditory neurons. Biol Cybern 42: 133–143.
4. Alain C (2007) Breaking the wave: effects of attention and learning on concurrent sound perception. Hear. Res 229: 225-236.
5. Alain C, Woods DL (1997) Attention modulates auditory pattern memory as indexed by event-related brain potentials. Psychophysiology 34: 534-546.
6. Alais D, Blake R., Lee SH (1998) Visual features that vary together over time group together over space. Nat Neurosci 1: 160–164.
7. Anderson LA, Christianson GB, Linden JF (2009) Stimulus-specific adaptation occurs in the auditory thalamus. J Neurosci 29: 7359–7363.
8. Andersen RA, Asanuma C, Cowan WM. 1985. Callosal and prefrontal associational projecting cell populations in area 7A of the macaque monkey: a study using retrogradely transported fluorescent dyes. J Comp Neurol 232: 443–455.

9. Andersson JL, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling geometric deformations in EPI time series. Neuroimage 13: 903–919.

10. Andreou L-V, Kashino M, Chait M (2011) The role of temporal regularity in auditory segregation. Hear Res 280: 228–235.

11. Anstis S, Saida S (1985) Adaptation to auditory streaming of frequency-modulated tones. J Exp Psychol: Human Percept Perf 11: 257-271.

12. Antunes FM, Nelken I, Covey E, Malmierca MS (2010) Stimulus-specific adaptation in the auditory thalamus of the anesthetized rat. PLoS ONE 5: e14071.

13. Arnott SR, Alain C (2002) Stepping out of the spotlight: MMN attenuation as a function of distance from the attented location. Neuroreport 13: 2209-2212.

14. Ayala YA, Pérez-González D, Duque D, Nelken I, Malmierca MS (2012) Frequency discrimination and stimulus deviance in the inferior colliculus and cochlear nucleus. Front Neural Circuits 6: 119.

15. Baillet S, Friston K, Oostenveld R (2011) Academic software applications for electromagnetic brain mapping using MEG and EEG. Comput Intell Neurosci 2011: 972050.

16. Bandettini PA, Wong EC (1998) Echo-planar magnetic resonance imaging of human brain activation. In: Schmitt F et al. (eds.), Echo-Planar Imaging: Theory, Technique and Applcation, 2nd Ed., Springer Verlag, New York, pp. 493-530.

17. Barbas H, Mesulam M (1981) Organization of afferent input to subdivisions of area 8 in the rhesus monkey. J Comp Neurol 200: 407-431.

18. Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1: 371–394.

19. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical microcircuits for predictive coding. Neuron 76: 695–711.

20. Baumann S, Griffiths TD, Rees A, Hunter D, Sun L, Thiele A (2010) Characterisation of the BOLD response time course at different levels of the auditory pathway in non-human primates. Neuroimage 50: 1099–1108.

21. Baumann S, Petkov CI, Griffiths TD (2013) A unified framework for the organization of the primate auditory cortex. Front Syst Neurosci 7: 11.

22. Bays P, Husain M (2008) Dynamic shifts of limited working memory resources in human vision. Science 321(5890): 81-85.

23. Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41: 809-823.

24. Beauvois MW (1998) The effect of tone duration on auditory stream formation. Percept Psychophys 60: 852–861.

25. Beauvois MW, Meddis R (1991) A computer model of auditory stream segregation. Q J Exp Psychol A 43(3): 517-541.

26. Beauvois MW, Meddis R (1996) Computer simulation of auditory stream segregation in alternating-tone sequences. J Acoust Soc Am 99:2270–2280.

27. Bee MA, Klump GM (2004) Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. J Neurophysiol 92: 1088-1104.

28. Bee MA, Klump GM (2005) Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. Brain Behav Evol 66: 197-214.

29. Behrens TEJ, Johansen-Berg H (2005) Relating connectional architecture to grey matter function using diffusion imaging. Philos Trans R Soc Lond, B, Biol Sci 360: 903–911.

30. Behrens TEJ, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott CAM, Boulby PA, Barker GJ, Sillery EL, Sheehan K, Ciccarelli O, Thompson AJ, Brady JM, Matthews PM (2003) Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nat Neurosci 6: 750–757.

31. Belin P, Zatorre RJ, Hoge R, Evans AC, Pike B (1999) Event-related fMRI of the auditory cortex. Neuroimage 10:417–429.

32. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403: 309-312.

33. Bendixen A, Denham SL, Gyimesi K, Winkler I (2010) Regular patterns stabilize auditory streams. J Acoust Soc Am 128: 3658–3666.

34. Bendor D, Wang X (2008) Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. J Neurophysiol 100: 888–906.

35. Bidet-Caulet A, Bertrand O (2009) Neurophysiological mechanisms involved in auditory perceptual organization. Front Neurosci 3: 182-191.

36. Bidet-Caulet A, Fischer C, Besle J, Aguera P, Giard M, Bertrand O (2007) Effects of Selective Attention on the Electrophysiological Representation of Concurrent Sounds in the Human Auditory Cortex. J Neurosci 27: 9252-9261.

37. Billig AJ, Davis MH, Deeks JM, Monstrey J, Carlyon RP (2013) Lexical Influences on Auditory Streaming. Curr Biol 23(16): 1585-1589.

38. Bisley JW, Goldberg M.E (2010) Attention, intention, and priority in the parietal lobe. Ann Rev Neurosci 33: 1–21.

39. Bizley JK, Cohen YE (2013) The what, where and how of auditory-object perception. Nat Rev Neurosci 14: 693-707.

40. Blake R, Lee SH (2005) The role of temporal structure in human vision. Behav Cogn Neurosci Rev 4: 21-42.

41. Boly M, Garrido MI, Gosseries O, Bruno M-A, Boveroux P, Schnakers C, Massimini M, Litvak V, Laureys S, Friston KJ (2011) Preserved feedforward but impaired top-down processes in the vegetative state. Science 332(6031): 858-862.

42. Bregman AS, Campbell J (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J Exp Psychol 89:244–249.

43. Bregman AS (1978) Auditory streaming is cumulative. J Exp Psychol: Human Percept Perf 11: 257-271.

44. Bregman AS (1990) Auditory scene analysis: The Perceptual Organization of Sound. Cambridge (Massachusetts): MIT Press.

45. Bregman AS, Ahad PA, Crum PA, O'Reilly J (2000) Effects of time intervals and tone durations on auditory stream segregation. Percept Psychophys 62(3): 626-636.

46. Bregman AS (2008) Auditory scene analysis. In Squire, L.R. (Editor-in-Chief.) Encyclopedia of Neuroscience. Oxford, UK: Academic Press.

47. Bremmer F, Schlack A, Shah NJ, Zafiris O, Kubischik M, Hoffmann K, Zilles K, Fink GR (2001) Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. Neuron 29: 287-296.

48. Brodmann K (1909) Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues. Barth, Leipzig, Germany.

49. Brosch M, Schreiner CE (1997) Time course of forward masking tuning curves in cat primary auditory cortex. J Neurophysiol 77: 923–943.

50. Brugge JF, Volkov IO, Garell PC, Reale RA, Howard MA 3rd (2003) Functional connections between auditory cortex on Heschl's gyrus and on the lateral superior temporal gyrus in humans. J Neurophysiol 90: 3750–3763.

51. Brugge JF, Volkov IO, Oya H, Kawasaki H, Reale RA, Fenoy A, Steinschneider M, Howard MA 3rd (2008) Functional localization of auditory cortical fields of human: click-train stimulation. Hear Res 238: 12–24.

52. Buelte D, Meister IG, Staedtgen M, Dambeck N, Sparing R, Grefkes C, Boroojerdi B (2008) The role of the anterior intraparietal sulcus in crossmodal processing of object features in humans: an rTMS study. Brain Res 1217: 110-118.

53. Buxton RB, Wong EC, Frank LR (1998) Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn Reson Med 39: 855–864.

54. Calford MB, Semple MN (1995) Monaural inhibition in cat auditory cortex. J Neurophysiol 73: 1876–1891.

55. Calvert GA (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex 11: 1110-1123.

56. Campbell AW (1905) Histological studies on the Localization of Cerebral Function. Cambridge University Press, Cambridge, UK.

57. Carlyon RP, Cusack R, Foxton JM, Robertson IH (2001) Effects of attention and unilateral neglect on auditory stream segregation. J Exp Psychol: Human Percept Perf 27: 115–127.

58. Carlyon RP, Plack CJ, Fantini DA, Cusack R (2003) Cross-modal and non-sensory influences on auditory streaming. Perception 32: 1393–1402.

59. Carlyon RP (2004) How the brain separates sounds. Trends Cog Sci 8: 465-471.

60. Chait M, Poeppel D, de Cheveigné A, Simon JZ (2007) Processing asymmetry of transitions between order and disorder in human auditory cortex. J Neurosci 27: 5207–5214.

61. Chait M, Poeppel D, Simon JZ (2008) Auditory temporal edge detection in human auditory cortex. Brain Res 1213: 78–90.

62. Chambers J, Akeroyd MA, Summerfield AQ, Palmer AR (2001) Active control of the volume acquisition noise in functional magnetic resonance imaging: method and psychoacoustical evaluation. J Acoust Soc Am 110: 3041–3054.

63. Cherry EC (1953) Some experiments on the recognition of speech, with one and two ears. J Acoust Soc Am 25: 975-979.

64. Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118: 887-906.

65. Clarke S, Rivier F (1998) Compartments within human primary auditory cortex: evidence from cytochrome oxidase and acetylcholinesterase staining. Eur J Neurosci 10: 741–745.

66. Clarke S, Morosan P (2012) Architecture, Connectivity, and Transmitter Receptors of Human Auditory Cortex. In: Human Auditory Cortex (Poeppel D, Overath T, Popper AN, Fay RR, eds). New York: Springer Science+Business Media.

67. Cohen D (1968) Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. Science 161: 784–786.

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

69. Cohen D, Halgren E (2009) Magnetoencephalography. In: Squire LR (ed.) Encyclopedia of Neuroscience 5: 615-622.

70. Cohen MS (1996) Rapid MRI and functional applications. In: Toga AW, Mazziotta JC (eds), Brain Mapping: The Methods, Academic Press: San Diego, pp. 223-255.

71. Cohen MS (1999) Echo-planar imaging and functional MRI. In: Moonen CTW et al. (eds), Functional MRI, Springer Verlag: New York, pp. 137-148.

72. Cohen YE (2009) Multimodal activity in the parietal cortex. Hear Res 258: 100-105.

73. Constantino FC, Pinggera L, Paranamana S, Kashino M, Chait M (2012) Detection of appearing and disappearing objects in complex acoustic scenes. PLoS ONE 7: e46167.

74. Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3: 201–215.

75. Crottaz-Herbette S, Menon V (2006) Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence. J Cogn Neurosci 18: 766–780.

76. Cusack R, Brett M, Osswald K (2003) An evaluation of the use of magnetic field maps to undistort echo-planar images. Neuroimage 18: 127–142.

77. Cusack R, Deeks J, Aikman G, Carlyon RP (2004) Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J Exp Psychol: Human Percept Perf 30: 643–656.

78. Cusack R (2005) The intraparietal sulcus and perceptual organization. J Cogn Neurosci 17: 641-651.

79. Cusack R, Mitchell DJ, Duncan J (2010) Discrete Object Representation, Attention Switching, and Task Difficulty in the Parietal Lobe. J Cogn Neurosci 22: 32-47.

80. Da Costa S, van der Zwaag W, Marques JP, Frackowiak RSJ, Clarke S, Saenz M (2011) Human primary auditory cortex follows the shape of Heschl's gyrus. J Neurosci 31: 14067–14075.

81. Damadian R, Minkoff L, Goldsmith M, Stanford M, Koutcher J (1976) Field focusing nuclear magnetic resonance (FONAR): visualization of a tumor in a live animal. Science 194: 1430–1432.

82. Darwin CJ, Carlyon RP (1995) Auditory Grouping. In: Moore BCJ (ed) Hearing. Orlando, FL, Academic, pp. 387–424.

83. Darwin CJ, Hukin RW (1999) Auditory objects of attention: the role of interaural time differences. J Exp Psychol: Human Percept Perf 25: 617-629.

84. David SV, Fritz JB, Shamma SA (2012) Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proc Natl Acad Sci USA 109: 2144–2149.

85. deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280: 1439–1443.

86. de Cheveigné A (2001) The auditory system as a separation machine. In: Physiological and Psychophysical Bases of Auditory Function (Houtsma AJM, Kohlrausch A, Prijs VF, Schoonhoven R, editors), pp 453-460. Maastricht, The Netherlands: Shaker Publishing BV.

87. de Cheveigné A, Parra LC (2013) A versatile tool for multichannel data analysis. (Manuscript under submission).

88. De la Mothe LA, Blumell S, Kajikawa Y, Hackett TA (2006) Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. J Comp Neurol 496: 27–71.

89. Deichmann R, Schwarzbauer C, Turner R (2004) Optimisation of the 3D MDEFT sequence for anatomical brain imaging: technical implications at 1.5 and 3 T. NeuroImage 21: 757-767.

90. Deike S, Scheich H, Brechmann A (2010) Active stream segregation specifically involves the left human auditory cortex. Hear Res 265: 30-37.

91. Deike S, Gaschler-Markefski B, Brechmann A, Scheich H (2004) Auditory stream segregation relying on timbre involves left auditory cortex. Neuroreport 15: 1511–1514.

92. Denham SL, Winkler I (2006) The role of predictive models in the formation of auditory streams. J. Physiol. Paris 100: 154-170.

93. Denham SL, Winkler I (2013) Auditory perceptual organization. Oxford Handbook of Perceptual Organization (ed: J Wagemans). Oxford University Press: Oxford.

94. Dick F, Tierney AT, Lutti A, Josephs O, Sereno MI, Weiskopf N (2012) In vivo functional and myeloarchitectonic mapping of human primary auditory areas. J Neurosci 32: 16095–16105.

95. Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci USA 109: 11854–11859.

96. Di Salle F, Formisano E, Seifritz E, Linden DE, Scheffler K, Saulino C, Tedeschi G, Zanella FE, Pepino A, Goebel R, Marciano E (2001) Functional fields in human auditory cortex revealed by time-resolved fMRI without interference of EPI noise. Neuroimage 13: 328–338.

97. Divac I, Lavail JH, Rakic P, Winston KR (1977) Heterogeneous afferents to the inferior parietal lobule of the rhesus monkey revealed by the retrograde transport method. Brain Res 123: 197-207.

98. Donner TH, Kettermann A, Diesch E, Ostendorf F, Villringer A, Brandt SA (2002) Visual feature and conjunction searches of equal difficulty engage only partially overlapping frontoparietal networks. Neuroimage 15: 16-25.

99. Durlach NI, Mason CR, Kidd G Jr, Arbogast TL, Colburn HS, Shinn-Cunningham BG (2003) Note on informational masking. J Acoust Soc Am 113(6): 2984-2987.

100. Dykstra AR, Halgren E, Thesen T, Carlson CE, Doyle W, Madsen JR, Eskandar EN, Cash SS (2011) Widespread Brain Areas Engaged during a Classical Auditory Streaming Task Revealed by Intracranial EEG. Front Hum Neurosci 5: 74.

101. Edmister WB, Talavage TM, Ledden PJ, Weisskoff RM (1999) Improved auditory cortex imaging using clustered volume acquisitions. Hum Brain Mapp 7: 89–97.

102. Eickhoff S, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage 25(4): 1325-1335.

103. Elhilali M, Fritz JB, Chi T-S, Shamma SA (2007) Auditory cortical receptive fields: stable entities with plastic abilities. J Neurosci 27: 10372–10382.

104. Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA (2009a) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron 61: 317-329.

105. Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J Acoust Soc Am 124: 3751-3771.

106. Elhilali M, Xiang J, Shamma SA, Simon JZ (2009b) Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol 7: e1000129.

107. Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM (1993) 3D statistical neuroanatomical models from 305 MRI volumes. In: Nuclear Science Symposium and Medical Imaging Conference, 1993 IEEE Conference Record., pp 1813–1817.

108. Fahle M (1993) Figure-ground discrimination from temporal information. Proc Biol Sci 254: 199–203.

109. Fastl H, Zwicker E (2007) Psychoacoustics: Facts and Models. Springer: Berlin.

110. Fay RR (1998) Auditory stream segregation in goldfish (Carassius auratus). Hear Res 120: 69–76.

111. Fishman YI, Arezzo JC, Steinschneider M (2004) Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. J Acoust Soc Am 116: 1656-1670.

112. Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear Res 151: 167-187.

113. Fishman YI, Steinschneider M (2010a) Formation of auditory streams. In The Oxford Handbook of Auditory Science, eds Rees A, Palmer AR (Oxford University Press), pp 215-246.

114. Fishman YI, Steinschneider M (2010b) Neural correlates of auditory scene analysis based on inharmonicity in the monkey primary auditory cortex. J Neurosci 30(37): 12480-12494.

115. Fleschig P (1908) Bemerkungen über die Hörsphäre des enschlichen Gehirns. Neurol. Zentralbl 27 (2-7): 50-57.

116. Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, Goebel R (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40: 859–869.

117. Fox PT, Raichle ME (1986) Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. Proc Natl Acad Sci USA 83: 1140–1144.

118. Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RS (1995a) Spatial registration and normalisation of images. Hum Brain Mapp 2: 165–189.

119. Friston KJ, Harrison L, Daunizeau J, Kiebel SJ, Phillips C, Trujillo-Barreto N, Henson RNA, Flandin G, Mattout J (2008) Multiple sparse priors for the M/EEG inverse problem. Neuroimage 39: 1104-1120.

120. Friston KJ, Henson RNA, Phillips C, Mattout J (2005) Bayesian estimation of evoked and induced responses. Human Brain Mapp 27: 722-735.

121. Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RS (1995b) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189 –210.

122. Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond, B, Biol Sci 360: 815–836.

123. Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138.

124. Friston KJ, Harrison L, Penny W (2003) Dynamic causal modeling. Neuroimage 19: 1273–1302.

125. Friston KJ, Holmes AP, Worsley KJ (1999) How many subjects constitute a study? Neuroimage 10:1–5.

126. Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R (1996) Movement-related effects in fMRI time-series. Magn Reson Med 35:346–355.

127. Fritz J, Elhilali M, Shamma S (2005) Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. Hear Res 206: 159–176.

128. Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6: 1216–1223.

129. Fritz JB, David SV, Radtke-Schuller S, Yin P, Shamma SA (2010) Adaptive, behaviourally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. Nat Neurosci 13: 1011–1019.

130. Fritz JB, Elhilali M, David SV, Shamma SA (2007) Auditory attention--focusing the searchlight on sound. Curr Opin Neurobiol 17: 437–455.

131. Fuchs P (2010) The Oxford Handbook of Auditory Science: The Ear. Oxford University Press, Oxford, UK.

132. Gaab N, Gabrieli JDE, Glover GH (2007) Assessing the influence of scanner background noise on auditory processing. II. An fMRI study comparing auditory processing in the absence and presence of recorded scanner noise using a sparse design. Hum Brain Mapp 28: 721–732.

133. Galaburda AM, Pandya DN (1983) The intrinsic architectonic and connectional organization of the superior temporal region of the rhesus monkey. J Comp Neurol 221: 169–184.

134. Geng JJ, Mangun GR (2009) Anterior intraparietal sulcus is sensitive to bottom-up attention driven by stimulus salience. J Cogn Neurosci 21: 1584–1601.

135. Giraud A-L, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H (2007) Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron 56: 1127–1134.

136. Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15: 511–517.

137. Goldinger SD (1998) Echoes of echoes? An episodic theory of lexical access. Psych Reviews 105(2): 251-279.

138. Gottlieb JP, Kusunoki M, Goldberg M.E (1998) The representation of visual salience in monkey parietal cortex. Nature 391: 481–484.

139. Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338: 334–337.

140. Griffiths TD, Warren JD (2004) What is an auditory object? Nat Rev Neurosci 5: 887–892.

141. Griffiths TD, Micheyl C, Overath T (2012) Identification tasks I: Auditory object analysis. In: Human Auditory Cortex (Poeppel D, Overath T, Popper AN, Fay RR, eds). New York: Springer Science+Business Media.

142. Grimault N, Bacon SP, Micheyl C (2002) Auditory stream segregation on the basis of amplitude-modulation rate. J Acoust Soc Am 111: 1340-1348.

143. Gross J, Baillet S, Barnes GR, Henson RN, Hillebrand A, Jensen O, Jerbi K, Litvak V, Maess B, Oostenveld R, Parkkonen L, Taylor JR, van Wassenhove V, Wibral M, Schoffelen J-M (2013) Good practice for conducting and reporting MEG research. Neuroimage 65: 349–363.

144. Gutschalk A, Micheyl C, Melcher JR, Rupp A, Scherg M, Oxenham AJ (2005) Neuromagnetic correlates of streaming in human auditory cortex. J Neurosci 25: 5382-5388.

145. Gutschalk A, Oxenham AJ, Micheyl C, Wilson EC, Melcher JR (2007) Human Cortical Activity during Streaming without Spectral Cues Suggests a General Neural Substrate for Auditory Stream Segregation. J Neurosci 27: 13074–13081.

146. Gutschalk A, Micheyl C, Oxenham AJ (2008) Neural correlates of auditory perceptual awareness under informational masking. PLoS Biol 6: e138.

147. Gutschalk A, Dykstra A (2013) Functional imaging of auditory scene analysis. Hear Res pii: S0378-5955(13): 00194-9.

148. Hackett TA, Stepniewska I, Kaas JH (1998a) Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. J Comp Neurol 394: 475–495.

149. Hackett TA, Stepniewska I, Kaas JH (1998b) Thalamocortical connections of the parabelt auditory cortex in macaque monkeys. J Comp Neurol 400: 271–286.

150. Hackett TA, Preuss TM, Kaas JH (2001) Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. J Comp Neurol 441: 197–222.

151. Hahn EL (1950) Spin Echoes. Phys Rev 80: 580–594.

152. Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW (1999) “Sparse” temporal sampling in auditory fMRI. Hum Brain Mapp 7: 213–223.

153. Hämäläinen M, Hari R, Ilmoniemi R, Knuutila J, Lounasmaa OV (1993) Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev. Mod. Phys. 65: 413–497.

154. Hari R (1990) The neuromagnetic method in the study of the human auditory cortex. In: Grandori, F., Hoke, M., Romani, G. (Eds.), Auditory Evoked Magnetic Fields and Potentials. Advances in Audiology, Vol 6. Karger, Basel, pp. 222–282.

155. Hari R, Salmelin R (2012) Magnetoencephalography: From SQUIDs to neuroscience. Neuroimage 20th anniversary special edition. Neuroimage 61: 386–396.

156. Hartmann WM, Johnson D (1991) Stream Segregation and Peripheral Channeling. Music Perception 9: 155–183.

157. Hauk O (2004) Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data. Neuroimage 21(4): 1612-1621.

158. Herdener M, Esposito F, Scheffler K, Schneider P, Logothetis NK, Uludag K, Kayser C (2013) Spatial representations of temporal and spectral sound cues in human auditory cortex. Cortex pii: S0010-9452(13)00110-X.

159. Hill KT, Bishop CW, Miller LM (2012) Auditory grouping mechanisms reflect a sound's relative position in a sequence. Front Hum Neurosci 6: 158.

160. Hill KT, Bishop CW, Yadav D, Miller LM (2011) Pattern of BOLD signal in auditory cortex relates acoustic response to perceptual streaming. BMC Neurosci 12: 85.

161. Hill KT, Miller LM (2010) Auditory attentional control and selection during cocktail party listening. Cereb Cortex 20: 583-590.

162. Hromádka T, Deweese MR, Zador AM (2008) Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol 6: e16.

163. Humphries C, Liebenthal E, Binder JR (2010) Tonotopic organization of human auditory cortex. Neuroimage 50: 1202–1211.

164. Hutton C, Bork A, Josephs O, Deichmann R, Ashburner J, Turner R (2002) Image Distortion Correction in fMRI: A Quantitative Evaluation. NeuroImage 16: 217-240.

165. Hyvärinen J (1982) The Parietal Cortex of Monkey and Man, Springer-Verlag, Berlin.

166. Itti L, Koch C (2001) Computational modeling of visual attention. Nat Rev Neurosci. 2: 194-203.

167. Iverson P (1995) Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes. J Exp Psychol: Human Percept Perf 21: 751-763.

168. Izumi A (2002) Auditory stream segregation in Japanese monkeys. Cognition 82(3): B113-122.

169. Jäncke L, Shah NJ, Posse S, Grosse-Ryuken M, Müller-Gärtner HW (1998) Intensity coding of auditory stimuli: an fMRI study. Neuropsychologia 36: 875–883.

170. Jones EG, Dell'Anna ME, Molinari M, Rausell E, Hashikawa T (1995) Subdivisions of macaque monkey auditory cortex revealed by calcium-binding protein immunoreactivity. J Comp Neurol 362: 153–170.

171. Julesz B (1962) Visual pattern discrimination. IRE Trans. Inf. Theory, IT-8: 84-92.

172. Kaas JH, Hackett TA (1998) Subdivisions of auditory cortex and levels of processing in primates. Audiol Neurootol 3: 73–85.

173. Kaas JH, Hackett TA (1999) "What" and "where" processing in auditory cortex. Nat Neurosci 2: 1045–1047.

174. Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci USA 97: 11793–11799.

175. Kanai R, Muggleton NG, Walsh V (2008) TMS over the intraparietal sulcus induces perceptual fading. J Neurophysiol 100(6): 3342-3350.

176. Kanai R, Rees G (2011) The structural basis of inter-individual differences in human behaviour and cognition. Nat Rev Neurosci 12(4): 231-242.

177. Kashino M, Okada M, Mizutani S, Davis P, Kondo HM (2007) The dynamics of auditory streaming: psychophysics, neuroimaging, and modeling. In Hearing—from sensory processing to perception (eds B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Upperkamp & J. Verhey), pp. 275– 283. Berlin, Germany: Springer.

178. Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of on-going cortical speech representations in a "cocktail party." J Neurosci 30: 620–628.

179. Kidd G, Mason CR (2003) Multiple bursts, multiple looks, and stream coherence in the release from informational masking. J Acoust Soc Am 114: 2835-2845.

180. Kidd G, Mason CR, Dai H (1995) Discriminating coherence in spectro-temporal patterns. J Acoust Soc Am 97: 3782-3790.

181. Kidd G, Mason CR, Deliwala PS, Woods WS, Colburn HS (1994) Reducing informational masking by sound segregation. J Acoust Soc Am 95: 3475-3480.

182. Kidd G, Richards VM, Streeter T, Mason CR (2011) Contextual effects in the identification of nonspeech auditory patterns. J Acoust Soc Am 130: 3926-3938.

183. Kiebel SJ, Daunizeau J, Phillips C, Friston KJ (2008) Variational Bayesian inversion of the equivalent current dipole in EEG/MEG. Neuroimage 39(2): 728-741.

184. Kiebel SJ, Garrido MI, Moran R, Chen C-C, Friston KJ (2009) Dynamic causal modeling for EEG and MEG. Hum Brain Mapp 30: 1866–1876.

185. King AJ, Nelken I (2009) Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat Neurosci 12: 698–701.

186. Kitada R, Kochiyama T, Hashimoto T, Naito E, Matsumura M (2003) Moving tactile stimuli of fingers are integrated in the intraparietal and inferior parietal cortices. Neuroreport 14: 719-724.

187. Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci 9: 85–111.

188. Kleinschmidt A, Sterzer P, Rees G (2012) Variability of perceptual multistability: from brain state to individual trait. Philos Trans R Soc Lond B Biol Sci 367(1591): 988-1000.

189. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4: 219-227.

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

191. Kondo HM, Kashino M (2009) Involvement of the Thalamocortical Loop in the Spontaneous Switching of Percepts in Auditory Streaming. J Neurosci 29: 12695-12701.

192. Kashino M, Kondo HM (2012) Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. Philos Trans R Soc Lond, B, Biol Sci 367: 977–987.

193. Kriegstein KV, Giraud A (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22: 948-955.

194. Koffka K (1935) Principles of Gestalt Psychology. Harcourt: New York.

195. Kowalski N, Depireux DA, Shamma SA (1996a) Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J Neurophysiol 76: 3503–3523.

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

197. Kumar S, Sedley W, Nourski KV, Kawasaki H, Oya H, Patterson RD, Howard MA 3rd, Friston KJ, Griffiths TD (2011) Predictive coding and pitch processing in the auditory cortex. J Cogn Neurosci 23: 3084–3094.

198. Kusmierek P, Rauschecker JP (2009) Functional specialization of medial auditory belt cortex in the alert rhesus monkey. J Neurophysiol 102: 1606–1622.

199. Langers DRM, Backes WH, van Dijk P (2007) Representation of lateralization and tonotopy in primary versus secondary human auditory cortex. Neuroimage 34: 264–273.

200. Langers DRM, van Dijk P (2012) Mapping the tonotopic organization in human auditory cortex with minimally salient acoustic stimulation. Cereb Cortex 22:2024–2038.

201. Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. Science 320:110–113.

202. Lee AKC, Larson E, Maddox RK, Shinn-Cunningham BG (2013) Using neuroimaging to understand the cortical mechanisms of auditory selective attention. Hear Res pii: S0378-5955(13)00170-6.

203. Lee CC, Winer JA (2008a) Connections of cat auditory cortex: I. Thalamocortical system. J Comp Neurol 507: 1879–1900.

204. Lee CC, Winer JA (2008b) Connections of cat auditory cortex: II. Commissural system. J Comp Neurol 507: 1901–1919.

205. Leek MR, Brown ME, Dorman MF (1991) Informational masking and auditory attention. Percept Psychophys 50: 205–214.

206. Leff AP, Iverson P, Schofield TM, Kilner JM, Crinion JT, Friston KJ, Price CJ (2009) Vowel-specific mismatch responses in the anterior superior temporal gyrus: an fMRI study. Cortex 45: 517–526.

207. Leopold DA, Logothetis NK (1999) Multistable phenomena: changing views in perception. Trends Cogn Sci 3: 254–264.

208. Levitt H (1971) Transformed up-down methods in psychoacoustics. J Acoust Soc Am 49:Suppl 2: 467+.

209. Liegeois-Chauvel C, Musolino A, Chauvel P (1991) Localization of the primary auditory area in man. Brain 114: 139–151.

210. Lipp R, Kitterick P, Summerfield Q, Bailey PJ, Paul-Jordanov I (2010) Concurrent sound segregation based on inharmonicity and onset asynchrony. Neuropsychologia 48: 1417-1425.

211. Litvak V, Mattout J, Kiebel S, Phillips C, Henson R, Kilner J, Barnes G, Oostenveld R, Daunizeau J, Flandin G, Penny W, Friston K (2011) EEG and MEG data analysis in SPM8. Comput Intell Neurosci 2011: 852961.

212. Logothetis NK (2002) The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. Philos Trans R Soc Lond, B, Biol Sci 357: 1003–1037.

213. Logothetis NK (2003) The underpinnings of the BOLD functional magnetic resonance imaging signal. J Neurosci 23: 3963–3971.

214. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412: 150–157.

215. Logothetis NK (2012) Intracortical recordings and fMRI: an attempt to study operational modules and networks simultaneously. Neuroimage 62: 962–969.

216. Lütkenhöner B, Krumbholz K, Lammertmann C, Seither-Preisler A, Steinsträter O, Patterson RD (2003) Localization of primary auditory cortex in humans by magnetoencephalography. Neuroimage 18: 58–66.

217. Lutti A, Dick F, Sereno MI, Weiskopf N (2013) Using high-resolution quantitative mapping of R1 as an index of cortical myelination. Neuroimage pii: S1053-8119(13)00642-3

218. MacDougall-Shackleton SA, Hulse SH, Gentner TQ, White W (1998) Auditory scene analysis by European starlings (Sturnus vulgaris): perceptual segregation of tone sequences. J Acoust Soc Am 103: 3581–3587.

219. Macmillan NA, Creelman CD (2005) Detection Theory: A User's Guide (2nd ed.). Mahwah , N.J. : Lawrence Erlbaum Associates

220. Mainen ZF, Sejnowski TJ (1995) Reliability of spike timing in neocortical neurons. Science 268: 1503–1506.

221. Malonek D, Grinvald A (1996) Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. Science 272: 551–554.

222. Mansfield P (1977) Multi-planar image formation using NMR spin echoes. J Phys[C] 10: L55-L58.

223. Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J (1995) A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). Neuroimage 2: 89–101.

224. McCabe SL, Denham MJ (1997) A model of auditory streaming. J Acoust Soc Am 101: 1611-1621.

225. McDermott JH (2009) The cocktail party problem. Curr Biol 19: 1024–1027.

226. McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nat Neurosci 16(4): 493-498.

227. McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71: 926-940.

228. McDonald KL, Alain C (2005) Contribution of harmonicity and location to auditory object formation in free field: Evidence from event-related brain potentials. J. Acoust. Soc. Am. 118: 1593-1604.

229. Melcher JR (2009) Auditory Evoked Potentials. In: Encyclopedia of Neuroscience (Larry R. Squire, ed), pp 715–719. Oxford: Academic Press.

230. Micheyl C, Carlyon RP, Gutschalk A, Melcher JR, Oxenham AJ, Rauschecker JP, Tian B, Courtenay Wilson E (2007a) The role of auditory cortex in the formation of auditory streams. Hear. Res 229: 116-131.

231. Micheyl C, Hanson C, Demany L, Shamma S, Oxenham AJ (2013a) Auditory Stream Segregation for Alternating and Synchronous Tones. J Exp Psychol Human Percept Perform: [Epub ahead of print].

232. Micheyl C, Kreft H, Shamma S, Oxenham AJ (2013b) Temporal coherence versus harmonicity in auditory stream formation. J Acoust Am Soc 133(3): 188-194.

233. Micheyl C, Oxenham AJ (2010) Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. Hear Res 266: 36-51.

234. Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques. Neuron 48: 139-148.

235. Micheyl C. Shamma S, Oxenham AJ (2007a) Hearing Out Repeating Elements in Randomly Varying Multitone Sequences: A Case of Streaming? In: Kollmeier B, Klump G, Hohmann V, Langemann U, Mauermann M, et al., editors. Hearing – from basic research to application. Berlin: Springer: pp. 267-274.

236. Mill RW, Bőhm TM, Bendixen A, Winkler I, Denham SL (2013) Modeling the emergence and dynamics of perceptual organisation in auditory streaming. PLoS Comput Biol 9: e1002925.

237. Miller CT, Mandel K, Wang X (2010). The communicative content of the common marmoset phee call during antiphonal calling. Am J Primatol. 72(11): 974-980.

238. Miller LM, D'Esposito M (2005) Perceptual Fusion and Stimulus Coincidence in the Cross-Modal Integration of Speech. J Neurosci 25: 5884-5893.

239. Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87: 516–527.

240. Mishkin M (1979) Analogous neural models for tactual and visual learning. Neuropsychologia 17: 139–151.

241. Moelker A, Pattynama PMT (2003) Acoustic noise concerns in functional magnetic resonance imaging. Hum Brain Mapp 20: 123–141.

242. Moerel M, De Martino F, Formisano E (2012) Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. J Neurosci 32: 14205–14216.

243. Molholm S, Martinez A, Ritter W, Javitt DC, Foxe JJ (2005) The neural circuitry of pre-attentive auditory change-detection: an fMRI study of pitch and duration mismatch negativity generators. Cereb Cortex 15: 545-551.

244. Moore BCJ, Glasberg BR (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J Acoust Soc Am 74: 750-753.

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

246. Moore BCJ, Gockel HE (2012) Properties of auditory stream formation. Phil Trans R Soc 367 (1591): 919-931.

247. Morel A, Garraghty PE, Kaas JH (1993) Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys. J Comp Neurol 335: 437–459.

248. Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K (2001) Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13(4): 684-701.

249. Näätänen R, Gaillard AW, Mäntysalo S (1978) Early selective-attention effect on evoked potential reinterpreted. Acta Psychol (Amst) 42: 313–329.

250. Näätänen R (1992) Attention and brain function. Hillsdale, NJ: Lawrence Erlbaum.

251. Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clin Neurophys 118: 2544-2590.

252. Nagarajan S, Gabriel RA, Herman A (2010) Magnetoencephalograpy. In: Human Auditory Cortex (Poeppel D, Overath T, Popper AN, Fay RR, eds). New York: Springer Science+Business Media.

253. Necker LA (1832) Observations on some remarkable optical phenomena seen in Switzerland; and on an optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. Lond Edinb Phil Mag J Sci 1: 329–337.

254. Neff DL, Green DM (1987) Masking produced by spectral uncertainty with multicomponent maskers. Percept Psychophys 41: 409–415.

255. Nelken I (2004) Processing of complex stimuli and natural scenes in the auditory cortex. Curr Opin Neurobiol 14: 474–480.

256. Ogawa S (2012) Finding the BOLD effect in brain images. Neuroimage 62: 608–609.

257. Ogawa S, Lee TM, Kay AR, Tank DW (1990a) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci USA 87: 9868–9872.

258. Ogawa S, Lee TM, Nayak AS, Glynn P (1990b) Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magn Reson Med 14: 68–78.

259. Ogawa S, Tank DW, Menon R, Ellermann JM, Kim SG, Merkle H, Ugurbil K (1992) Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proc Natl Acad Sci USA 89: 5951–5955.

260. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–609.

261. Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011: 156869.

262. Opitz B, Rinne T, Mecklinger A, von Cramon DY, Schröger E (2002) Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. Neuroimage 15: 167–174.

263. Overath T, Kumar S, Stewart L, von Kriegstein K, Cusack R, Rees A, Griffiths TD (2010) Cortical mechanisms for the segregation and representation of acoustic textures. J. Neurosci 30: 2070-2076.

264. Overath T, Kumar S, von Kreigstein K, Griffiths TD (2008) Encoding of spectral correlation over time in auditory cortex. J Neurosci 28(49): 13268-13273.

265. Pandya DN (1995) Anatomy of the auditory cortex. Rev Neurol (Paris) 151: 486–494.

266. Pandya DN, Kuypers HGJM (1969) Cortico-cortical connections in the rhesus monkey. Brain Res 13: 13-36.

267. Pauling L, Coryell CD (1936) The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin. Proc Natl Acad Sci USA 22: 210–216.

268. Pelleg-Toiba R, Wollberg Z (1989) Tuning properties of auditory cortex cells in the awake squirrel monkey. Exp Brain Res 74:353–364.

269. Penny W, Holmes AP (2004) Random-effects analysis. In: Frackowiak RS, Friston KJ, Frith C, Dolan RJ, Price CJ, editors. Human brain function. San Diego: Academic. pp. 843-850.

270. Petkov CI, Kayser C, Augath M, Logothetis NK (2006) Functional imaging reveals numerous fields in the monkey auditory cortex. PLoS Biol 4: e215.

271. Petrides M, Pandya DN (1984) Projections to the frontal cortex from the posterior parietal region in the rhesus monkey. J Comp Neurol 228: 105-116.

272. Plack CJ, Moore BCJ (1990) Temporal window shape as a function of frequency and level. J Acoust Soc Am 87: 2178-2187.

273. Polich J (2007) Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 118(10): 2128-2148.

274. Pollack I (1975) Auditory informational masking. J Acoust Soc Am Suppl 1, 57: S5.

275. Pressnitzer D, Hupé J-M (2006) Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. Curr Biol 16: 1351–1357.

276. Pressnitzer D, Sayles M, Micheyl C, Winter IM (2008) Perceptual organization of sound begins in the auditory periphery. Curr Biol 18: 1124–1128.

277. Pressnitzer D, Suied C, Shamma SA (2011) Auditory scene analysis: the sweet music of ambiguity. Front Hum Neurosci 5: 158.

278. Price DL, De Wilde JP, Papadaki AM, Curran JS, Kitney RI (2001) Investigation of acoustic noise on 15 MRI scanners from 0.2 T to 3 T. J Magn Reson Imaging 13: 288–293.

279. Pulvermuller F, Shtyrov Y (2006) Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes. Prog Neurobiol 2006: 79: 49-71.

280. Rademacher J, Caviness VS Jr, Steinmetz H, Galaburda AM (1993) Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology. Cereb Cortex 3: 313–329.

281. Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund HJ, Zilles K (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage 13: 669–683.

282. Rajendran VG, Harper NS, Willmore BD, Hartmann WM, Schnupp JWH (2013) Temporal predictability as a grouping cue in the perception of auditory streams. J Acoust Soc Am 134: EL98–EL104.

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

284. Rauschecker JP (1998) Cortical processing of complex sounds. Curr Opin Neurobiol 8: 516–521.

285. Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12: 718–724.

286. Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268: 111–114.

287. Rauschecker JP, Tian B, Pons T, Mishkin M (1997) Serial and parallel processing in rhesus monkey auditory cortex. J Comp Neurol 382: 89–103.

288. Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of “what” and “where” in auditory cortex. Proc Natl Acad Sci USA 97: 11800–11806.

289. Ravicz ME, Melcher JR, Kiang NY (2000) Acoustic noise during functional magnetic resonance imaging. J Acoust Soc Am 108: 1683–1696.

290. Read HL, Winer JA, Schreiner CE (2001) Modular organization of intrinsic connections associated with spectral tuning in cat auditory cortex. Proc Natl Acad Sci USA 98: 8042–8047.

291. Recanzone GH (2000) Response profiles of auditory cortical neurons to tones and noise in behaving macaque monkeys. Hear Res 150: 104–118.

292. Rice NJ, Tunik E, Grafton ST (2006) The anterior intraparietal sulcus mediates grasp execution, independent of requirement to update: new insights from transcranial magnetic stimulation. J Neurosci 26: 8176-8182.

293. Rivier F, Clarke S (1997) Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. Neuroimage 6: 288–304.

294. Roberts B, Glasberg B, Moore BCJ (2002) Primitive stream segregation of tone sequences without differences in F0 or passband. J Acoust Soc Am 112: 2074-2085.

295. Rodrigues-Dagaeff C, Simm G, De Ribaupierre Y, Villa A, De Ribaupierre F, Rouiller EM (1989) Functional organization of the ventral division of the medial geniculate body of the cat: evidence for a rostro-caudal gradient of response properties and cortical projections. Hear Res 39: 103–125.

296. Roland PE, Zilles K (1994) Brain atlases--a new research tool. Trends Neurosci 17: 458–467.

297. Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 2: 1131–1136.

298. Rouiller EM, Rodrigues-Dagaeff C, Simm G, De Ribaupierre Y, Villa A, De Ribaupierre F (1989) Functional organization of the medial division of the medial geniculate body of the cat: tonotopic organization, spatial distribution of response properties and cortical connections. Hear Res 39: 127–142.

299. Rushworth MF, Ellison A, Walsh V (2001a) Complementary localization and lateralization of orienting and motor attention. Nat. Neurosci 4: 656-661.

300. Rushworth MF, Krams M, Passingham RE (2001b) The attentional role of the left parietal cortex: the distinct lateralization and localization of motor attention in the human brain. J Cogn Neurosci 13: 698-710.

301. Saenz M, Langers DRM (2013) Tonotopic mapping of human auditory cortex. Hear Res pii: S0378-5955(13)00187-1.

302. Salmi J, Rinne T, Koistinen S, Salonen O, Alho K (2009) Brain networks of bottom-up triggered and top-down controlled shifting of auditory attention. Brain Res 1286: 155-164.

303. Sander K, Brechmann A, Scheich H (2003) Audition of laughing and crying leads to right amygdala activation in a low-noise fMRI setting. Brain Res Brain Res Protoc 11: 81–91.

304. Sanders LD, Joh AS, Keen RE, Freyman RL (2008) One sound or two? Object-related negativity indexes echo perception. Percept Psychophys 70: 1558-1570.

305. Schadwinkel S, Gutschalk A (2010) Activity Associated with Stream Segregation in Human Auditory Cortex is Similar for Spatial and Pitch Cues. Cereb Cortex 20(12): 2863-2873.

306. Schadwinkel S, Gutschalk A (2011) Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus. J Neurophysiol 105: 1977–1983.

307. Scheich H, Baumgart F, Gaschler, Markefski B, Tegeler C, Tempelmann C, Heinze HJ, Schindler F, Stiller D (1998) Functional magnetic resonance imaging of a human auditory cortex area involved in foreground-background decomposition. Eur J Neurosci 10: 803-809.

308. Schneider DM, Woolley SMN (2013) Sparse and background-invariant coding of vocalizations in auditory scenes. Neuron 79: 141–152.

309. Schofield TM, Iverson P, Kiebel SJ, Stephan KE, Kilner JM, Friston KJ, Crinion JT, Price CJ, Leff AP (2009) Changing meaning causes coupling changes within higher levels of the cortical hierarchy. Proc Natl Acad Sci USA 106: 11765–11770.

310. Shofner WP, Niemiec AJ (2010) Comparative psychoacoustics. In The Oxford University Press Handbook of Auditory Science: Auditory Perception, (ed.) Christopher Plack, pp. 145-176.

311. Schönwiesner M, von Cramon DY, Rübsamen R (2002) Is it tonotopy after all? Neuroimage 17: 1144–1161.

312. Schwarzkopf DS, Song C, Rees G (2011) The surface area of human V1 predicts the subjective experience of object size. Nat Neurosci 14(1): 28-30.

313. Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. Curr Opin Neurobiol 4: 569–579.

314. Shadlen MN, Newsome WT (1996) Motion perception: seeing and deciding. Proc Natl Acad Sci USA 93: 628–633.

315. Shafritz KM, Gore JC, Marois R (2002) The role of the parietal cortex in visual feature binding. Proc Natl Acad Sci USA 99: 10917-10922.

316. Shamma SA, Symmes D (1985) Patterns of inhibition in auditory cortical cells in awake squirrel monkeys. Hear Res 19: 1–13.

317. Shamma S, Elhilali M, Ma L, Micheyl C, Oxenham AJ, Pressnitzer D, Yin P, Xu Y (2013) Temporal coherence and the streaming of complex sounds. Adv Exp Med Biol 787: 535–543.

318. Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34: 114-123.

319. Shamma SA, Micheyl C (2010) Behind the scenes of auditory perception. Curr Opin Neurobiol 20: 361-366.

320. Sheft S, Yost WA (2008) Method-of-adjustment measures of informational masking between auditory streams. J Acoust Soc Am 124: EL1-7.

321. Sigalovsky IS, Fischl B, Melcher JR (2006) Mapping an intrinsic MR property of gray matter in auditory cortex of living humans: a possible marker for primary cortex and hemispheric differences. Neuroimage 32: 1524–1537.

322. Silver AH, Zimmerman JE (1965) Quantum transitions and loss in multiply connected superconductors. Phys Rev Lett 15: 888–891.

323. Snyder JS, Alain C (2007) Toward a neurophysiological theory of auditory stream segregation. Psychol Bull 133: 780-799.

324. Snyder JS, Alain C, Picton TW (2006) Effects of attention on neuroelectric correlates of auditory stream segregation. J Cogn Neurosci 18: 1-13.

325. Snyder JS, Gregg MK, Weintraub DM, Alain C (2012). Attention, awareness, and the perception of auditory scenes. Front Psychol 3: 15.

326. Sporns O, Tononi G, Edelman GM (1991) Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. Proc Natl Acad Sci USA 88(1): 129-133.

327. Stainsby TH, Fullgrabe C, Flanagan HJ, Waldman S, Moore BCJ (2011) Sequential streaming due to manipulation of interaural time differences. J Acoust Soc Am 130: 904-914.

328. Stanton GB, Bruce CJ, Goldberg ME (1995) Topography of projections to posterior cortical areas from the macaque frontal eye fields. J Comp Neurol 353: 291-305.

329. Striem-Amit E, Hertz U, Amedi A (2011) Extensive cochleotopic mapping of human auditory cortical fields obtained with phase-encoding FMRI. PLoS ONE 6: e17832.

330. Sussman ES (2005) Integration and segregation in auditory scene analysis. J Acoust Soc Am 117: 1285–1298.

331. Sussman ES, Bregman AS, Wang WJ, Khan FJ (2005) Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. Cogn Affect Behav Neurosci 5: 93–110.

332. Sussman ES, Ceponiene R, Shestakova A, Näätänen R, Winkler I (2001) Auditory stream segregation processes operate similarly in school-aged children and adults. Hear Res 153:108–114.

333. Sussman ES, Horváth J, Winkler I, Orr M (2007) The role of attention in the formation of auditory streams. Percept Psychophys 69: 136–152.

334. Sussman ES, Ritter W, Vaughan HG Jr (1999) An investigation of the auditory streaming effect using event-related brain potentials. Psychophysiology 36: 22–34.

335. Sussman ES, Winkler I, Schröger E (2003) Top-down control over involuntary attention switching in the auditory modality. Psychon Bull Rev 10: 630–637.

336. Taaseh N, Yaron A, Nelken I (2011) Stimulus-specific adaptation and deviance detection in the rat auditory cortex. PLoS ONE 6: e23369.

337. Talairach P and Tournoux J (1988) A Stereotactic Coplanar Atlas of the Human Brain. Stuttgart: Thieme.

338. Talavage TM, Edmister WB, Ledden PJ, Weisskoff RM (1999) Quantitative assessment of auditory cortex responses induced by imager acoustic noise. Hum Brain Mapp 7: 79–88.

339. Talavage TM, Edmister WB (2004) Nonlinearity of FMRI responses in human auditory cortex. Hum Brain Mapp 22: 216–228.

340. Talavage TM, Sereno MI, Melcher JR, Ledden PJ, Rosen BR, Dale AM (2004) Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. J Neurophysiol 91: 1282–1296.

341. Talavage TM, Hall DA (2012) How challenges in auditory fMRI led to general advancements for the field. Neuroimage 62: 641–647.

342. Tallon-Baudry, Bertrand (1999) Oscillatory gamma activity in humans and its role in object representation. Trends Cogn Sci 3: 151–162.

343. Tanji K, Leopold DA, Ye FQ, Zhu C, Malloy M, Saunders RC, Mishkin M (2010) Effect of sound intensity on tonotopic fMRI maps in the unanesthetized monkey. Neuroimage 49: 150–157.

344. Teki S, Barnes GR, Penny WD, Iverson P, Woodhead ZVJ, Griffiths TD, Leff AP (2013) The right hemisphere supports but does not replace left hemisphere auditory function in patients with persisting aphasia. Brain 136: 1901–1912.

345. Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292: 290–293.

346. Toga AW, Ambach K, Quinn B, Hutchin M, Burton JS (1994) Postmortem anatomy from cryosectioned whole human brain. J Neurosci Methods 54: 239–252.

347. Treisman A (1999) Solutions to the binding problem: progress through controversy and convergence. Neuron 24: 105–125.

348. Ulanovsky N, Las L, Farkas D, Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. J Neurosci 24: 10440–10453.

349. Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. Nat Neurosci 6: 391–398.

350. Ungerleider LG, Haxby JV (1994) “What” and “where” in the human brain. Curr Opin Neurobiol 4: 157–165.

351. Upadhyay J, Ducros M, Knaus TA, Lindgren KA, Silver A, Tager-Flusberg H, Kim D-S (2007) Function and connectivity in human primary auditory cortex: a combined fMRI and DTI study at 3 Tesla. Cereb Cortex 17: 2420–2432.

352. Upadhyay J, Silver A, Knaus TA, Lindgren KA, Ducros M, Kim D-S, Tager-Flusberg H (2008) Effective and structural connectivity in the human auditory cortex. J Neurosci 28: 3341–3349.

353. van Noorden LPAS (1975) Temporal coherence in the perception of tone sequences. Eindhoven: University of Technology.

354. Van Veen BD, van Drongelen W, Yuchtman M, Suzuki A (1997) Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans Biomed Engg 44(9): 867-880.

355. Vélez A, Bee MA (2011) Dip listening and the cocktail party problem in grey treefrogs: Signal recognition in temporally fluctuating noise. Anim Behav 82: 1319–1327.

356. Vliegen J, Oxenham AJ (1999) Sequential stream segregation in the absence of spectral cues. J Acoust Soc Am 105: 339-346.

357. Von Békésy, G (1960) Experiments in Hearing. (EG Wever, ed.) New York: McGraw-Hill.

358. Von Békésy G (1970) Travelling waves as frequency analysers in the cochlea. Nature 225: 1207–1209.

359. von der Malsburg C (1981) The Correlation Theory of Brain Function (Department of Neurobiology, Max Planck Institute for Biophysics and Chemistry, Gottingen Internal Report 81-2.

360. Von der Malsburg C, Schneider W (1986) A neural cocktail-party processor. Biol Cybern 54: 29–40.

361. von Economo C, Koskinas GN (1925) Die Cytoarchitektonik der Grosshirnrinde des erwachsenen Menschen. Spinger, Berlin.

362. von Economo C, Horn L (1930) Über Windungsrelief, Maße und Rindenarchitektonik der Supratemporalfläche, ihre individuellen und ihre Seitenunterschiede. Z Neurol Psychiatr 130: 678-757.

363. Wallace MN, Johnston PW, Palmer AR (2002) Histochemical identification of cortical areas in the auditory region of the human brain. Exp Brain Res 143: 499–508.

364. Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Netw 19: 1395–1407.

365. Walther DB, Koch C (2007) Attention in hierarchical models of object recognition. Prog Brain Res 165: 57-78.

366. Wang D, Chang P (2008) An oscillatory correlation model of auditory streaming. Cogn Neurodyn 2: 7–19.

367. Warren JD, Jennings AR, Griffiths TD (2005) Analysis of the spectral envelope of sounds by the human brain. Neuroimage 24: 1052-1057.

368. Wasserthal C, Brechmann A, Stadler J, Fischl B, Engel K (2013) Localizing the human primary auditory cortex in vivo using structural MRI. Neuroimage pii: S1053-8119(13): 00813-6.

369. Watkins S, Dalton P, Lavie N, Rees G (2007) Brain mechanisms mediating auditory attentional capture in humans. Cereb Cortex 17: 1694-1700.

370. Werner S, Noppeney U (2010) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. J Neurosci 30: 2662-2675.

371. Wiegand K, Gutschalk A (2012) Correlates of perceptual awareness in human primary auditory cortex revealed by an informational masking experiment. Neuroimage 61: 62–69.

372. Wielnhammer VA, Ludwig K, Hesselmann G, Sterzer P (2013) Frontoparietal cortex mediates perceptual transitions in bistable perception. J Neurosci 33(40): 16009-16015.

373. Wilson EC, Melcher JR, Micheyl C, Gutschalk A, Oxenham AJ (2007) Cortical FMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming. J Neurophysiol 97: 2230-2238.

374. Winer JA, Lee CC (2007) The distributed auditory cortex. Hear Res 229: 3–13.

375. Winkler I (2007) Interpreting the Mismatch Negativity. Journal of Psychophysiology 21: 147–163.

376. Winkler I, Kushnerenko E, Horváth J, Ceponiene R, Fellman V, Huotilainen M, Näätänen R, Sussman E (2003b) Newborn infants can organize the auditory world. Proc Natl Acad Sci USA 100: 11812–11815.

377. Winkler I, Sussman E, Tervaniemi M, Horváth J, Ritter W, Näätänen R (2003a) Preattentive auditory context effects. Cogn Affect Behav Neurosci 3: 57–77.

378. Winkler I, Takegata R, Sussman E (2005) Event-related brain potentials reveal multiple stages in the perceptual organization of sound. Brain Res Cogn Brain Res 25: 291–299.

379. Winkler I, Teder-Sälejärvi WA, Horváth J, Näätänen R, Sussman E (2003c) Human auditory cortex tracks task-irrelevant sound sources. Neuroreport 14: 2053–2056.

380. Winkler I, Denham S, Mill R, Bohm TM, Bendixen A (2012) Multistability in auditory stream segregation: a predictive coding view. Philos Trans R Soc Lond, B, Biol Sci 367: 1001–1012.

381. Winkler I, Denham SL, Nelken I (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn Sci (Regul Ed) 13: 532–540.

382. Woods DL, Herron TJ, Cate AD, Yund EW, Stecker GC, Rinne T, Kang X (2010) Functional properties of human auditory cortical fields. Front Syst Neurosci 4: 155.

383. Xu Y, Chun MM (2009) Selecting and perceiving multiple visual objects. Trends Cogn Sci 13: 167-174.

384. Yokoi I, Komatsu H (2009) Relationship between neural responses and visual grouping in the monkey parietal cortex. J. Neurosci 29: 13210-13221.

385. Zion-Golumbic EM, Cogan GB, Schroeder CE, Poeppel D (2013a) Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party.” J Neurosci 33:1417–1426.

386. Zion-Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013b) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron 77:980–991.

APPENDIX I: PUBLICATIONS ARISING FROM THIS THESIS

1. Teki S, Chait M, Kumar S, von Kriegstein K, Griffiths TD (2011). Brain bases for auditory stimulus-driven figure-ground segregation**.** J Neurosci 31(1): 164-171.
2. Teki S, Chait M, Kumar S, Shamma S, Griffiths TD (2013). Segregation of complex acoustic scenes based on temporal coherence**.** eLife 2: e00699.
3. Teki S, Payne C, Kumar S, Griffiths TD, Chait M. Figure-ground segregation in complex acoustic scenes: an MEG study (in preparation).

APPENDIX II: AUTHOR CONTRIBUTIONS

Chapter 3 (Psychophysics): The author was involved in conception and design, acquisition of data, analysis and interpretation of data, and writing the article for publication. Maria Chait was involved in conception and design, analysis and interpretation of data and writing the manuscript. Sukhbinder Kumar was involved in the design of the experimental stimulus. Tim Griffiths was involved in conception and design, interpretation of data and writing the manuscript.

Chapter 4 (Temporal Coherence Modeling): The author was involved in performing the modeling, interpretation of modeling results and writing the results for publication. Maria Chait, Shihab Shamma and Tim Griffiths were involved in interpretation of the modeling results and writing the manuscript.

Chapter 5 (Functional Magnetic Resonance Imaging): The author was involved in analysis and interpretation of data, and writing the article for publication. Maria Chait was involved in conception and design, acquisition of data, and writing the manuscript. Katharina von Kreigstein was involved in design and acquisition of data. Sukhbinder Kumar was involved in conception and design and acquisition of data. Tim Griffiths was involved in conception and design, and writing the manuscript.

Chapter 6 (Magnetoencephalography): The author was involved in conception and design, acquisition of data, analysis and interpretation of data, and writing the manuscript. Chris Payne was involved in acquisition of data. Tim Griffiths was involved in conception and design, interpretation of data and writing the manuscript. Maria Chait was involved in conception and design, analysis and interpretation of data, and writing the manuscript.