ABSTRACT

Title of Dissertation: NEURAL AND COMPUTATIONAL APPROACHES TO AUDITORY SCENE ANALYSIS

Sahar Akram, Doctor of Philosophy, 2014

Dissertation directed by: Professor Shihab Shamma
Department of Electrical and Computer Engineering

Our perception of the world is highly dependent on the complex processing of the sensory inputs by the brain. Hearing is one of those seemingly effortless sensory tasks that enables us to perceive the auditory world and integrate acoustic information from the environment into cognitive experiences. The main purpose of studying auditory system is to shed light on the neural mechanisms underlying our hearing ability. Understanding the systematic approach of the brain in performing such complicated tasks is an ultimate goal with numerous clinical and intellectual applications.

In this thesis, we take advantage of various experimental and computational approaches to understand the functionality of the brain in analyzing complex auditory scenes. We first focus on investigating the behavioral and neural mechanisms underlying auditory sound segregation, also known as auditory streaming. Employing an informational masking paradigm, we explore the interaction between stimulus-driven and task-driven attentional process in the auditory cortex using magnetoencephalography (MEG) recordings from the human brain. The results

demonstrate close links between perceptual and neural consequences of the auditory stream segregation, suggesting the neural activity to be viewed as an indicator of the auditory streaming percept.

We examine more realistic auditory scenarios consisted of two speakers simultaneously present in an auditory scene and introduce a novel computational approach for decoding the attentional state of listeners in such environment. The proposed model focuses on an efficient implementation of a decoder for tracking the cognitive state of the brain, inspired from neural representation of auditory objects in the auditory cortex. The structure is based on an state-space model with the recorded MEG signal and individual speech envelopes as the input and the probability of attending to the target speaker as the output of the model. The proposed approach benefits from accurate and highly resolved estimation of attentional state in time as well as the inherent model-based dynamic denoising of the underlying state-space model, which makes it possible to reliably decode the attentional state under very low SNR conditions.

As part of this research work, we investigate the neural representation of ambiguous auditory stimuli at the level of the auditory cortex. In perceiving a typical auditory scene, we may receive incomplete or ambiguous auditory information from the environment. This can lead to multiple interpretations of the same acoustic scene and formation of an ambitious perceptual state in the brain. Here, in a series of experimental studies, we focus on a particular example of ambitious stimulus (ambitious Shepard tone pair) and investigate the neural correlates of the contextual effect and perceptual biasing using MEG. The results from psychoacoustic and neu-

ral recordings suggest a set of hypothesis about the underlying neural mechanism of short-term memory and expectation modulation in the nervous system.

NEURAL AND COMPUTATIONAL APPROACHES TO AUDITORY SCENE ANALYSIS

by

Sahar Akram

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:

Professor Shihab A. Shamma, Chair
Professor Jonathan Z. Simon
Assistant Professor Behtash Babadi
Professor Carol Espy-Wilson
Professor Yiannis Aloimonos

ACKNOWLEDGMENTS

There are many individuals without whose guidance, support and knowledge this dissertation would have been a distant dream. To those named and those inadvertently missed, I owe you my deepest and sincerest gratitude for making this dream come true.

To Shihab, my advisor and mentor, for his inspirational advice and unwavering support throughout the last several years. Thank you for providing me with numerous opportunities to explore a variety of scientific interests and joining me along the path.

To Jonathan Simon and Mounya, for the many valuable discussions, comments and advice regarding my studies. Thank you for your continued support and mentorship.

To Bernhard, a special friend and a great advisor, for providing me with encouragement and criticism when needed. Thank you for being generous with your knowledge of neuroscience, life and inspiration.

To Behtash, for his exceptional guidance, brilliant ideas and friendly support. Thank you for helping me see the light at the end of the tunnel and regaining my passion and love for science.

To Drs. Espy-Wilson and Aloimonos, who have kindly agreed to be part of my research committee. Thank you for your time and invaluable advice.

To Jonathan Fritz, for his wonderful ideas and insightful discussions. Special thoughts go also to my previous and present colleagues at NSL and CSSL, Majid

Mirbagheri, Diego Gonzales, Krishna Puvvada, Marisel Villafane and Francisco Cervantes. Thank you all for helping me in so many ways, and most importantly, thank you for your friendship.

To all my wonderful friends in the US and back home, Hamidreza, Maryam, Parastou, Azadeh, Sadaf, Naeim, Mercedeh, Payam, Shahin, Farhad, Mohamad, Alireza, Mahshid, Ladan, Bahareh, Tina, Setareh and Gilava. Thank you for sharing very many wonderful moments with me. Life would not have been as colorful without your company.

Lastly, this dissertation is certainly dedicated to my parents for a lifetime of continuous support, motivation and wisdom. Thank you for believing in every dream I've had and giving me the greatest reasons of all to succeed. To my wonderful sisters, Sara and Bita for being a great source of support, happiness and joy in my life. To you I will be forever grateful.

College Park, January 2015

*S*ahar

Contents

List of Figures

Chapter 1

Introduction

1.1 Auditory scene analysis

1.1.1 Brain, an exceptionally powerful processor

Listening to our world seems to be an effortless and undemanding activity in our daily life. We perceive and analyze our environment by taking into our brain a complex mixture of acoustic information, and suddenly we are able to localize the sound sources around us, attend selectively to different instruments in an orchestra, and reliably discriminate multiple sources in a complex auditory scene while associating related features to each auditory object. All of these magical steps should take place somewhere between the ear and the high-level processors in the brain.

Sounds in our environment are compiled together as a one-dimensional complex signal, a variable wave pressure over time, by the time it reaches the outer most part of our auditory system, the pinna. This organ, with its unique and weird

looking anatomical structure, provides preliminary cues about the direction and location of the sound. Later the sound navigates through the auditory pathways, transforming from one type of energy to another and eventually making its way through the nervous system. The nervous system has an excellent ability to put together all the necessary information and extract cues from the scene to help us answer seemingly trivial questions, such as which melody is being played, which instrument is playing it, and which singer is not singing harmony. These, in addition to so many other amazing abilities are what makes our auditory life colorful and enjoyable. Our attempts to understand the auditory mechanism of the brain, which functions as a great computational tool, has been partially successful; however, to have a comprehensive view of all the processing steps taken by the brain, we need to build a thorough map of the auditory neural circuitry, starting from physical vibrations to the perception of the sound (Yost, 1994).

1.1.2 Perceptual complexities

Segregation of an auditory scene into multiple streams is one of those highly demanding tasks which is powerfully and effortlessly resolved by the brain, even in extremely noisy and reverberant environments (Carlyon, 2004; Bregman, 1994). This procedure, also known as the cocktail party problem (Cherry, 1953) is facilitated by informative cues in the acoustic stimulus along the temporal and spectral dimensions. Although there have been intensive studies on the behavioral and neural bases of auditory stream segregation over the last decades, large aspects of this pro-

cess still remain to be explored (Griffiths and Warren, 2004a; Elhilali and Shamma, 2008b; McDermott, 2009).

While the auditory processing is a quite complicated procedure—even if all the necessary sensory information is provided by the environment—there are many situations in which our information from the environment is insufficient, limited or ambiguous. As a dynamic physical system, our brain is functioning based on a set of perceptual principles that helps it derive perceptually meaningful events out of auditory scenes with missing or limited sensory information. These perceptual principles are strongly influenced by our auditory memory, language, musical background, and many other factors that are common or specific to certain groups of people. Ambiguous sounds are one of those examples in which the perceptual realization of the sound is not necessarily a one-to-one mapping into the sensory information domain. In other words, a similar stimulus will sound differently if it is presented in different contexts. Auditory illusion is another case where universal principles override the actual information provided by the scene and makes us hear things that may be physically absent. An example of auditory illusion is hearing a missing fundamental frequency when other components of a harmonic series are present (Shepard, 1964).

1.2 Thesis outline

In this thesis, a number of novel experimental and computational approaches tackling different aspects of auditory scene analysis have been discussed. We investigate

neural mechanisms underlying complicated auditory tasks such as source segregation when multiple sound sources are present. Specifically, we focus on speech as an important and common auditory source that is routinely streamed and tracked by the human brain.

Another complex auditory scene studied here is an ambiguous stimulus for which our brain may have multiple interpretations, resulting in the exact same acoustic stimulus being perceived in one way or the other depending on various neural and environmental conditions. Our focus is mostly on understanding the neural mechanisms used by the brain to disambiguate such a stimulus, which can be internally tied to higher-level processing mechanisms in the brain such as short term memory and attention.

This dissertation is organized in six chapters. The following chapter summarizes the present knowledge on auditory pathways and the high-level neural representation of auditory stimuli in the brain, with a brief introduction to existing computational models inspired by the biology. Next, we introduce Magnetoencephalography (MEG), a non-invasive recording technique that has been used in our studies to record cortical activities of the human brain while performing an auditory task. We close this chapter by reviewing the benefits and limitations of this technique for auditory studies, and by discussing some of the common approaches used in the interpretation and analysis of MEG data.

Chapter 3 starts with a literature review over existing approaches in studying auditory stream segregation. Later in this chapter, an experimental approach for understanding the neural mechanism underlying detection of a single-frequency

target sequence embedded in spectrotemporaly randomized tones is presented as a simplified example of auditory streaming. MEG recordings collected from the auditory cortex of humans were analyzed and correlated with the behavioral performances of the listeners, while their attention was manipulated towards different aspects of the auditory scene. The results highlight the strong dependence of the behavioral performance in streaming the target sequence to the stimulus parameters which are directly reflected in the neural activity recorded from the brain, therefore suggesting a close link between the perceptual and neural consequences of the auditory streaming. This study has been published in the *PLOS ONE journal* (Akram et al. (2014a)).

In chapter 4, a novel computational approach for decoding the attentional state of listeners in a multi-speaker environment is proposed. The model is evaluated on simulation studies and used in decoding real MEG data recorded from the human brain. Later in this chapter, the main advantages of the new decoder over the existing methods is highlighted and generalization of the current model to be applied on more complex auditory scenes is discussed. The method part of this study is published in *Advances in Neural Information Processing Systems* (Akram et al. (2014b)), and the extended version is submitted to *NeuroImage journal*.

In chapter 5, the neural correlates of a compelling example of ambiguous auditory stimuli is explored, which involves the detection of pitch-direction changes in a pair of Shepard tones that are spectrally half an octave apart, and hence, equally perceived as an upward or downward step change in pitch (Deutsch, 1980). Conducting a number of psychoacoustic and MEG experiments, the neural representation of

such ambiguity in the brain is investigated and the mechanism underlying perceptual biasing of this bistable stimulus to one or the other direction is explored with the help of the surrounding context. In a final set of experiments, we examine if the effectiveness of the biasing sequence is dependent on perceptual streaming of this sequence prior to presentation of the ambitious pair.

Finally in chapter 6, a summery of the main findings of this work, along with future prospects for this field of research is provided.

Chapter 2

Background

2.1 Auditory pathway

It is appropriate to start our study of sound processing with an introduction to the auditory pathway, and to explain the structure and role of different elements along this path. The main organs in the auditory system are the ear—mainly referring to the entire peripheral auditory apparatus—and parts of the central nervous system. Here, we briefly review basic knowledge about these organs and their role in forming the perception of the sound.

2.1.1 Peripheral auditory system

The mysterious travel of sound through the auditory system starts when the one-dimensional pressure-wave hits the pinna (Figure 2.1). The function of the pinna is to perform spectral transformations on the incoming sound as well as to add directional information to the sound by applying frequency-dependent amplitude

modulation on the sound before directing it to the auditory canal.

Further along the path, the eardrum separates the external ear from the middle ear and transmits the sound from the air to the ossicles (The three smallest bones in the human body, which have a stirrup shape. These bones are located in the middle ear and transmit the sound from the air to the fluid-filled cochlea. In other words, at this stage, acoustic energy from compression waves in the air is efficiently transferred to fluid-membrane waves within the cochlea. The sound would then passes through the basilar membrane, a stiff structure element that separates two liquid-filled tubes that run along the coil of the cochlea. Changing its width and stiffness across the cochlea, the basilar membrane can capture higher frequencies at the front, and lower frequencies at the end of the membrane. It works as a tonotopically ordered frequency axis along the length of the cochlea and can be modeled as a filter bank to perform a short-term Fourier analysis on the frequency contents of the sound (Hams, 1985; Kandel et al., 2000).

Later in the pass, inner hair cells transform the fluid waves into analog nerve signals that are then relayed to the cochlear nerve fibers with a digital output. Each nerve fiber represents a particular frequency and a specific range of loudness. The encoded information carried by these fibers arrives at the first relay station in the auditory system, the cochlear nuclei. The cochlear nuclei cells encode the information provided by nerve fibers by generating action potentials with a specific rate and particular timing of individual action potentials. According to the previous studies, some features—such as the timing information of the neural patterns—are enhanced and sharpened in the cochlear nucleus prior to sending them to more

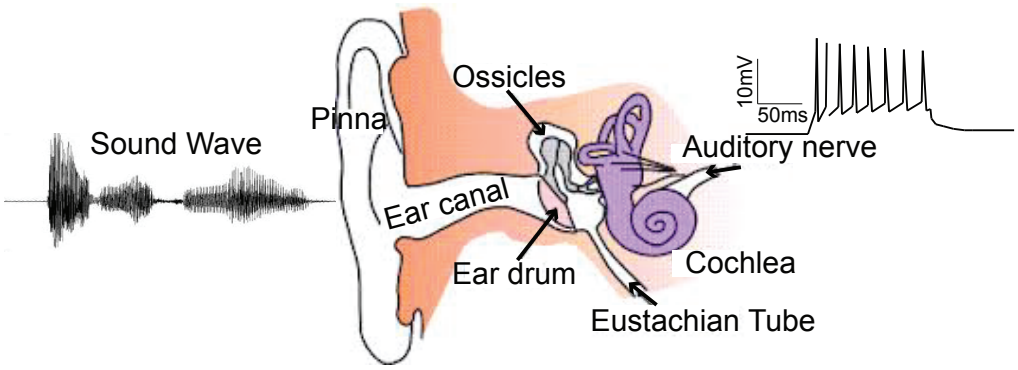

Figure 2.1: A schematic view of the periphery auditory system. Starting from the ear, this is the first stage of transferring the sound into a hearing organism, which feed directly into the nervous system by transduction of sound pressure-waves into neural action potentials.

central areas Palmer et al. (1996).

A computational and biologically inspired model to preserve the perceptual relevance and noise robustness properties we see in mammalian peripheral auditory processing is explained in Chi et al. (2005). This model is widely used in studies as one of the building blocks in auditory scene analysis and computational modeling (Elhilali and Shamma, 2007, 2008a; Mesgarani et al., 2006; Mesgarani and Shamma, 2005). Here is a brief description of the model for the early auditory processing stage.

1. To model the spectral analysis performed by the cochlear filter bank, we apply an affine wavelet transform to the input acoustic waveform $s(t)$. This filter bank, $h(x,t)$, consists of 128 overlapping constant $Q(\simeq 3)$ band-pass filters with center frequencies uniformly distributed along a logarithmic axis x over 5.3 octaves.

2. The output from the previous stage, $y_1(x,t)$, goes through a high-pass filter—non-linear compression $g(.)$—and a low-pass filter, roughly modeling the hair

cell stage and membrane leakage in which phase-locking to frequencies above 2 kHz is decreased on auditory nerves.

3. Next, the output of the previous stage, $y_2(x,t)$, is introduced to a first-order derivative with respect to the tonotopic frequency axis and then passed through a half-wave rectifier. This transformation simulates the action of the Lateral Inhibition Network (LIN), which is presumably located in the cochlear nucleus and increases the frequency selectivity of cochlear filter bank (Lyon and Shamma, 1996; Shamma, 1985).

4. The final step integrates the output of the previous stage $y_3(x,t)$ over a short window with the time constant $\tau = 8$ ms, using the function $\mu(t;\tau) = e^{(-t/\tau)}u(t)$. This stage compensates for further loss of phase-locking in the midbrain.

The mathematical formulation of the above steps is briefly described here (Chi et al., 2005) :

$$y_1(t,x) = s(t) *_t h(x,t) \tag{2.1}$$

$$y_2(t,x) = g(\delta_t y_1(t,x)) *_t w(t) \tag{2.2}$$

$$y_3(t,x) = max(\delta_x y_2(t,x), 0) \tag{2.3}$$

$$y_4(t,x) = y_3(t,x) *_t \mu(t;\tau) \tag{2.4}$$

2.1.2 Central auditory system

The anatomical complexity of the pathways and the neural morphology of this stage are far less known compared to the periphery; however, ongoing physiological and anatomical studies have gradually revealed the functionality and neural circuitry of the central auditory system. The neural information is projected to the Superior Olivary Complex (SOC) from the cochlear nucleus, and then to the Inferior Colliculi (IC). The IC works as a relay in the ascending auditory system and it is believed to integrate the information coming from the SOC and lateral lemniscus (Kandel et al., 2000), before sending it to the higher auditory areas (Figure 2.2).

Most of the necessary features from an auditory scene are extracted as the sound comes up to the level of IC and to higher auditory areas including auditory thalamus (Medical Geniculate Body). The auditory cortex is believed to play a role in discriminating or binding these features to form auditory objects (Pinto et al., 2003). Further in the auditory model proposed by (Chi et al., 2005), the central auditory processing is simulated using a multi-scale filter bank, in which each filter has a specific selectivity to the spectral and temporal modulations of sound, denoted as scale (Ω,cycle/octave) and rate (ω,cycle/time) respectively. The impulse response of these filters is defined as a two-dimensional spectrotemporally separable Gabor filter (Chi et al., 1999), which is computed as the product of spatial impulse response $h_S(x;\Omega_c,\phi_c)$, and temporal impulse response $h_T(t;\omega_c,\theta_c)$.

$$h_S(x;\Omega_c,\phi_c) = h_s(x;\phi_c)cos\phi_c + \bar{h}_s(x,\phi_c)sin\phi_c \tag{2.5}$$

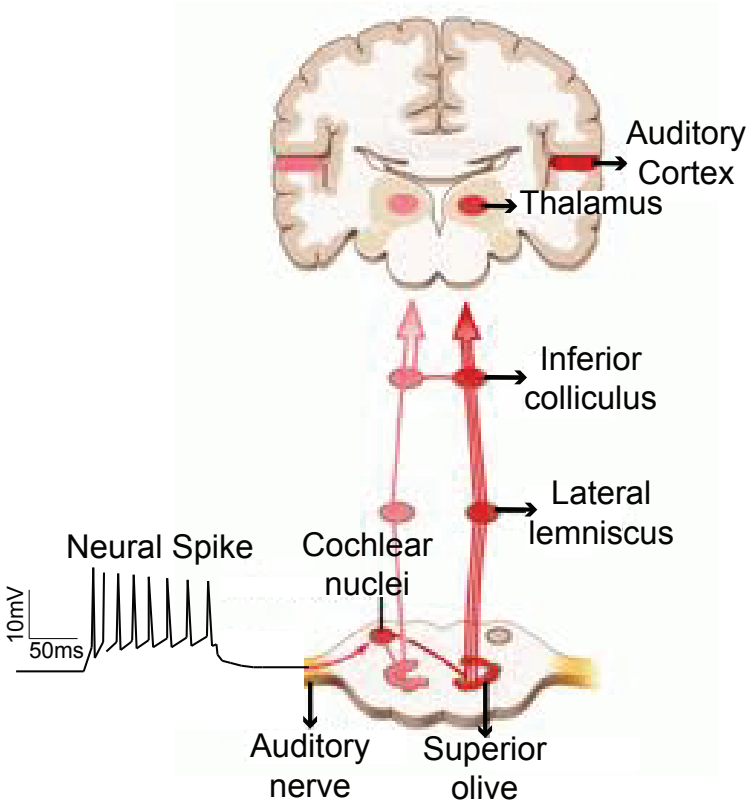

Figure 2.2: A schematic view of the central auditory system. The encoded sound information travels through intermediate stations such as cochlear nuclei and superior olivary complex of the brainstem and get further processed until eventually reaching the thalamus and from there relayed to the auditory cortex.

$$h_T(t;\omega_c,\theta_c) = h_t(t;\theta_c)cos\theta_c + \bar{h}_t(t,\theta_c)sin\theta_c \tag{2.6}$$

$$H(t,x,\Omega_c,\phi_c,\omega_c,\theta_c) = h_S(x;\Omega_c,\phi_c).h_T(t;\omega_c,\theta_c) \tag{2.7}$$

More details about h_s and h_t functions can be found in (Chi et al., 2005). The output of this model is a multi-scale, multi-rate representation of the sound in a four-dimensional complex-valued space, varying along the frequency, time, spectral

sale and temporal rate axes.

$$r(t, x, \Omega_c, \phi_c, \omega_c, \theta_c) = y(t, x) *_{tx} H(t, x, \Omega_c, \phi_c, \omega_c, \theta_c) \tag{2.8}$$

Where $y(t, x)$ is the input spectrogram derived from equations in 1.2.1, and $*_{tx}$ is a two-dimensional convolution in frequency and time. This multi-resolution representation can capture different properties of the original sound, such as upward/downward orientation and slow/fast temporal and spectral envelopes of the sound.

2.1.3 Neural representation in the auditory cortex

A complex sound can be represented in the neural domain by characterizing different neurons with respect to their specific selectivity in the spectral and temporal domains. There is a specific region for each neuron, in which the presence of the stimuli will alter the firing rate of that neuron. This region is known as the receptive field. In the auditory field, these receptive fields are a two-dimensional representation of spectral and temporal selectivity of that neuron and are known as STRFs (Spectrotemporal receptive fields) (Christopher deCharms et al., 1998; Klein et al., 2000; Theunissen et al., 2000).

The input/output relationship of a neuron can be modeled with the neurons STRF, as a spectrotemporal transfer function. The input is a two-dimensional spectrotemporal representation of the sound in the auditory cortex (described in 2.7), which is convolved with the STRF of the neuron, and the neural output is the

firing rate of the neuron as a function of time (Figure 2.3).

STRFs are providing a quantitative linear description on the neurons behavior with respect to specific stimulus patterns; however, they suffer from a number of limitations since most of the modeling approaches simplify the complicated biological pathways to make them fit certain criteria. The linearity and time invariance properties of the explained model are some examples of the shortfalls, which make this model fail in other more general cases. Recognizing these limitations, this model is employed in some of the current studies due to its inherent simplicity to explain some neural behaviors in the human auditory cortex.

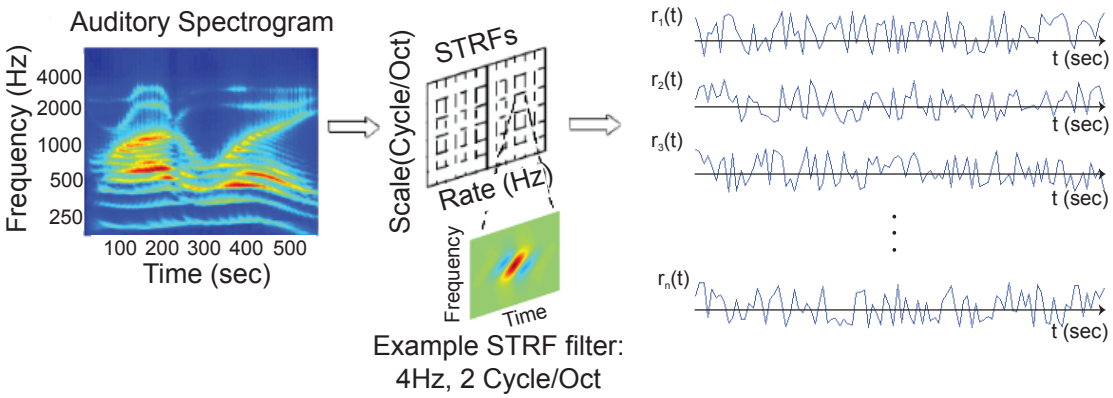

Figure 2.3: Spectrotemporal receptive fields. Assuming linearity in the system, each neuron can be modeled as having a time-varying firing rate computed from the convolution of the stimulus with the neuron's STRF. Here, the stimulus is the two-dimensional spectrao-temporal representation of the sound described in 2.1 to 2.4, which is filtered through neuron's STRFs and the outputs of the filters demonstrate the time-varying firing rates of the corresponding neurons.

2.2 Magnetoencephalography (MEG)

Investigation of the human brain is a pursuit of great intellectual interest. As Hämälainen puts it so well (Hämäläinen et al., 1993):"The whole of humanity depends on our minds". Despite extensive ongoing research on this challenging topic, the fundamental questions of the mechanisms used by the brain to store, retrieve, and process information are still largely unknown. Recent developments in the field of computational neuroscience, with many intellectual and clinical applications, are great sources of inspiration for studies of the functional principles underlying the human brain. To this end, Magnetoencephalography, a non-invasive neuroimaging technique serves us as a great tool to record ongoing neural activity from different areas of the brain.

2.2.1 Basis of MEG

Currents of all sorts deluge our body and produce electromagnetic fields. The brain also sustains ionic current flows with neurons as the strongest generators (Figure 2.4). The amount of current generated by a single cell is too small to be detected in a non-invasive measurement from outside of the head; however, if the architecture of the cell is in line with other cells and the currents are derived with a sufficient amount of synchrony, it gives rise to a large net current that is detectable centimeters away, outside the head.

Cellular currents, as explained above, are primary contributors to MEG surface signals. These current generators operate in a conductive medium and therefore

introduce a secondary type of current that circulates through the head tissues, including the skull bone, and loop back to close the electrical circuit. In certain limited conductor geometries the first and secondary currents generate equal but opposite fields, which leads to a zero net external field. Therefore, only currents with components tangential to the spherical surface are detectable by the gradiometers, and radial sources are externally silent.

Localization of the MEG sources is a crucial step in interpreting the recorded neural data. Indeterminacy of the inverse problem, even if the MEG is measured at infinitely many points around the head, leads to insufficient information for uniquely computing the distribution of the current sources within the brain (Helmholtz, 1853). Therefore, one must consider appropriate modeling constraints, such as prior knowledge of the number and approximate location of the sources, to be able to derive information about source distribution from the recorded data. In addition, discriminating between the contribution of original and secondary currents to the measurements in the proposed model further complicates the source localization procedure.

2.2.2 Instrumentation

MEG measurements are conducted externally using an extremely sensitive instrument called a superconducting quantum interference device (SQUID). The SQUID is a very low noise detector of magnetic fields that converts the magnetic flux threading a pickup coil into voltage, allowing detection of weak neuromagnetic signals. Since the SQUID relies on physical phenomena found in superconductors, it requires cryo-

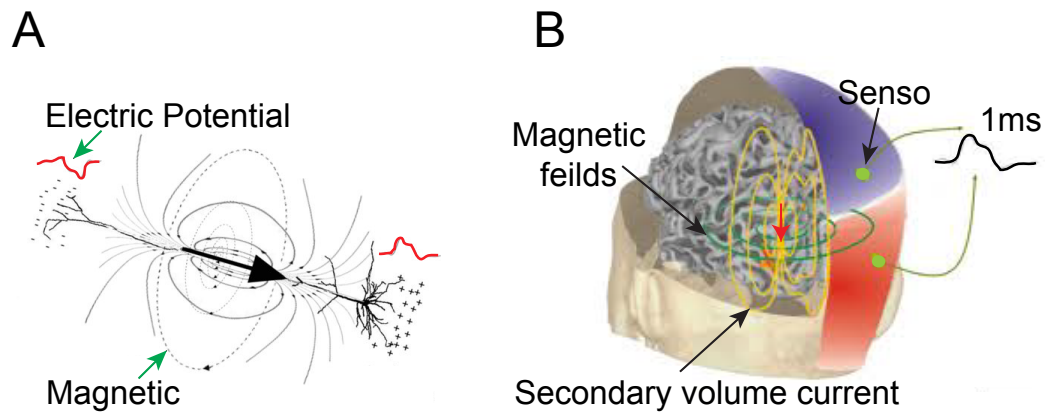

Figure 2.4: The source of MEG. (A) Neural cells drive ionic electrical currents caused by electric potential differences along the cell, which is due to the excitatory and inhibitory post-synaptic activities. The magnetic fields produced by primary currents are demonstrated in the figure (dashed black lines). The electrical circuits of currents are closed by secondary volume currents (solid black lines). (B) In a larger scale, the superposition of currents coming from synchronized activity of the neurons is shown as a red dipole surrounded by secondary volume currents shown in yellow. Resulted magnetic fields (green contours) are measurable outside of the skull with the resolution of 1 ms.

genic temperatures for operation. In a modern MEG device, an array of more than 300 SQUIDS is contained in a helmet shaped liquid helium-containing vessel called a dewar, which allows for simultaneous measurements at many points over the head (Figure 2.5–A & 2.5–B). Since the magnetic signals emitted by the brain are on the order of a few femtoteslas, shielding from external magnetic signals is necessary. The system is then operated in a Magnetically Shielded Room (MSR) that minimizes interference from external magnetic disturbances, including the earths magnetic field, noise generated by electrical equipment, and low frequency magnetic fields produced by moving magnetic objects such as elevators, cars, and trains. Appropriate magnetic shielding can be obtained by constructing rooms made of aluminum and mu-metal, thereby reducing high frequency and low frequency noise,

respectively (Figure 2.5–C).

In some MEG systems there is the possibility to record EEG from dense arrays of electrodes (≥ 60) simultaneously, thereby completing the electromagnetic signature of neural currents. Additional analog channels are usually available for miscellaneous recordings (heart monitoring (ECG), muscle activity (EMG), eye movements (EOG), respiration, skin conductance, subjects responses, etc.). Sampling rates can reach up to 5 kHz on all channels with a typical instrumental noise level limited to a few fT per square meter.

In the following studies, all MEG signals are recorded in a dimly lit magnetically shielded room (Yokogawa Electric Corporation) using a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan). Its detection coils are arranged in a uniform array on a helmet-shaped surface on the bottom of the dewar, with $\simeq 254$ mm between the centers of two adjacent 15.5 mm diameter coils. Sensors are configured as first-order axial gradiometers with a baseline of 50 mm; their field sensitivities are 5 fT/Hz or better in the white noise region. The presentation software package (Neurobehavioral Systems) is used to present stimuli to the subjects. The sounds (approximately 70 dB SPL) are delivered to the participants' ears with 50 Ω sound tubing (E-A-RTONE 3A; Etymotic Research), attached to E-A-RLINK foam plugs inserted into the ear canal. The entire acoustic delivery system is equalized to give an approximately flat transfer function from 40–3000 Hz, i.e., encompassing the range of the presently delivered stimuli. Three of the 160 channels are magnetometers separated from the others, and used as reference channels in noise-filtering methods (De Cheveigné and Simon, 2007). The magnetic

signals were filtered to the range of 1–200 Hz, notch filtered at 60 Hz, and sampled at 1 kHz.

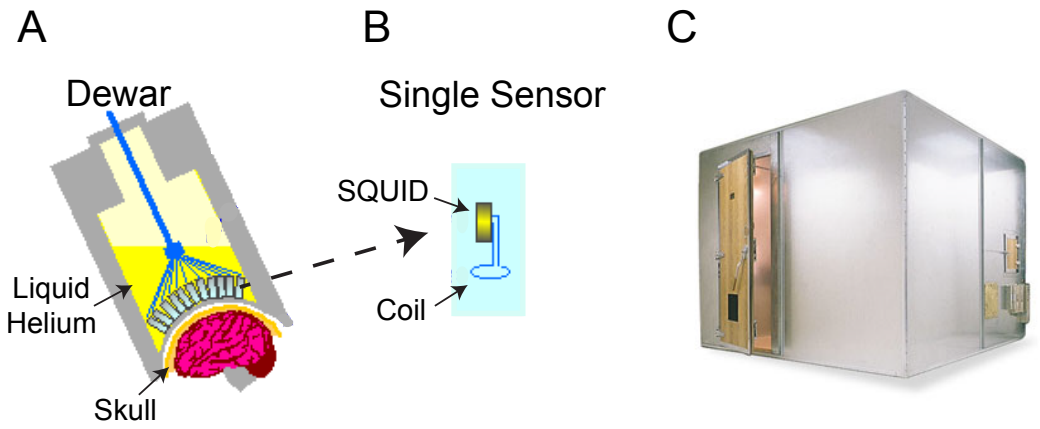

Figure 2.5: MEG system. (A) A multisensory array located in a dewar. The helmet shaped dewar, typically contain 300 sensors, covering most of the head and facilitating accumulation of the MEG data. (B) Each single sensor consists of a detection coil which is connected to a SQID. (C) A magnetic shielded room containing the MEG machine.

2.3 MEG in auditory studies

MEG has been used in numerous studies to characterize the responses to different types of acoustic stimuli, ranging from non-speech tones (Brattico et al., 2009; Gutschalk et al., 2005; Elhilali et al., 2009; Ackermann et al., 2001; Xiang et al., 2010) to components of speech sounds such as vowels and syllables (Sams et al., 1991; Phillips et al., 2000; Besle et al., 2004) to more complex speech sounds including words and sentences (Suppes and Han, 2000; Luo and Poeppel, 2007; Pulvermüller and Shtyrov, 2009; Ding and Simon, 2012a). According to the high temporal resolution of this recording technique (on the order of milliseconds), most of these studies

focus on temporal analysis and morphology of sensor measurements at the cortical level.

2.3.1 Response pattern to auditory stimuli

The earliest auditory-evoked responses in the auditory cortex peak near 20, 30 and 50 ms after the stimulus presentation, followed by a deflection at about 100 ms (M100/N1m)—a prominent, robust response across listeners and stimuli—which has been the most investigated auditory response (Näätänen and Picton, 1987; Roberts et al., 2000). According to the related studies, the neural source of a M100 response in the cortex is localized to planum temporale(Lütkenhöner and Steinsträter, 1998), whereas the earlier and smaller M50 peak possibly originates in or near the primary auditory cortex (PAC) (Rupp et al., 2002). There is evidence showing that M50 and M100 responses belong to different functional systems. Consistency of presence of the M100 response to the stimuli-like clicks, tones, and speech gives rise to the hypothesis that the M100 response reflects the process of detecting changes in sensory input, with its strength, latency, and lateralization varying with certain physical and temporal aspects of stimuli (Näätänen and Picton, 1987; Roberts et al., 2000). The amplitude and lateralization of M50, on the other hand, appears to be task independent (Chait et al., 2004), with no dependence on interaural time differences (McEvoy et al., 1994) or contralateral masking effects (Elhilali and Shamma, 2008b). The M50 response plays a key role in exploring PAC (Mäkelä et al., 1994; McEvoy et al., 1994) and early auditory system maturation (Cardy et al., 2004; Oram Cardy

et al., 2008) in humans, and is largely incriminated in neurological disorders such as schizophrenia (Adler et al., 1982; Thoma et al., 2005; Potter et al., 2006).

2.3.2 Information extraction

The frequency spectrum of MEG data is rich and complex, since multiple processes take place simultaneously and engage neural populations at various spatial, temporal, and frequency scales. In order to be able to study and trace the neural correlates of a specific process in the brain, we need to take some pre-processing steps on the recorded data to enhance the level of signals of interest, while attenuating noise and uninteresting signals. Filtering the data within a frequency band, removing the base-line activity, averaging over multiple trials, and other data correction techniques are among the steps to make the data ready for further analysis. Evoked and transient neural responses to auditory stimuli can be extracted from the recorded data according to both the type of information we are looking for, and how it is encoded in the neural response. In order to extract the evoked response, the same stimulus is repeated several times and the data over all the repetitions are averaged to emphasize the phase-locked response to the stimulus and cancel out the contribution of random-phase signals. Several types of MEG responses, such as the response to the slow temporal modulations, have been demonstrated to be phase-locked to the stimulus (Ding and Simon, 2009; Fuentemilla et al., 2006; Luo and Poeppel, 2007). On the other hand, the transient response results from an increase in power in response to a stimulus, similar to the evoked activity but with variability in re-

sponse time. Hence, averaging over repeated trials would not help in recovering the transient response from the noisy signal. The latency and amplitude of the transient response varies by different physical and perceptual properties of the stimulus such as loudness, signal to noise ratio, spectral pattern of the stimulus, attention, and perceptual saliency of the sound onset/offset (Näätänen and Picton, 1987; Kaplan-Neeman et al., 2006; Biermann and Heil, 2000; Poeppel et al., 1996). Different frequency and time domain analysis is employed to enable us to look at evoked and transient responses and investigate various physical and perceptual traces on neural recordings. However, the general approach for analyzing and exploring a data set should be determined specifically for each experiment based on the properties and hypothesis behind that experiment.

In all conducted experiments here, a pre-experiment consisting of 100 repetitions of a 1 kHz, 50 ms tone pip is presented before starting the real experiment. The inter-stimulus intervals (ISIs) are randomized between 0.75 ms and 1.55 s, and participants are asked to count the tone pips. The experiment is done as a control condition to check M100 responses (a prominent peak approximately 100 ms after pip onset) and verify that the location and strength of neural signals fall within a normal range.

2.3.3 Benefits and limitations

MEG is completely safe and noninvasive. The data can be collected in the seated position allowing more natural cognitive experiments than fMRI. The measurement

environment is completely silent, which facilitates auditory studies in particular. Temporal resolution in MEG is very high—on the order of milliseconds—however; spatial resolution is moderately low, on the order of centimeters.

Comparing MEG with similar non-invasive methods such as EEG, there are a number of advantages in employing MEG over EEG systems. Most importantly, while EEG is strongly degraded by the heterogeneity in conductivity within head tissues (e.g., insulating skull vs. conducting scalp), this effect is extremely limited in MEG, resulting in greater spatial discrimination of neural contributions. This has important implications for source modeling and localization of the recorded neural responses.

Additionally, subject preparation time and comfort improve in MEG, as there is no need for direct contact of the sensors to the skin. Measurements in MEG are absolute and there is no need for choosing a reference point, as opposed to EEG. Moreover, MEG is particularly useful in detecting auditory responses because the orientation of the neural sources in the auditory cortex is roughly parallel instead of perpendicular to the scalp. This orientation results in a magnetic dipole across the surface of a subjects head. In contrast, a neural current running perpendicular to the subjects head would not show any magnetic dipole.

Among the limitations of MEG technique, a major technical problem is that the localization of sources of electrical activity within the brain from magnetic measurement outside the head is complicated and does not have a unique solution. This is known as the ill-posed inverse problem and is itself the subject of research. However, feasible solutions can often be obtained by using relatively simple models. Due

to the increased distance to sources and the almost spherical symmetry of the head, it is difficult to provide reliable information about the subcortical sources of the brain activity. Also, MEG does not provide structural and anatomical information and must often be combined with MR data into a composite image of function, overlaid on anatomy, to produce activation maps.

Chapter 3

Neural correlates of auditory streaming

3.1 Introduction

Our brain is a fascinating computational tool, which makes the whole hearing process a trivial, effortless accomplishment for us; however, the underlying mechanism for different processing stages is very little understood and a lot of scientific engineering approaches to mimic the brain's functionality have fallen short in computational implementations.

Since Aristotle, many philosophers and psychologists have believed that perception is the process of using the information provided by our senses to form mental representations of the world around us, says Bregman in "A Book on Auditory Scene Analysis" (Bregman, 1990). The whole process can be divided into two main stages. The first stage involves forming the representations of the sensory input. Here, we can specifically talk about extracting informational cues and auditory features from the auditory scene. In the next stage, these features should be grouped together appropriately to form segregated and meaningful auditory objects. The latter in-

tegration stage seems to be a fairly complicated task that is optimally achieved by the brain.

Nonetheless, the separation of an auditory scene into multiple streams—also known as the Cocktail party effect (Cherry, 1953)—is a highly complex perceptual task facilitated by informative cues in the acoustic stimulus along the temporal and spectral dimensions. Although there have been intensive studies on the behavioral and neural bases of auditory stream segregation over the last decades, large aspects of this process still remain to be explored.

3.1.1 Auditory objects & grouping cues

Auditory objects are perceptually well-defined constructs, and despite their visual counterparts, it is difficult to make an intuitive sense of such a notion in auditory modality (Ahveninen et al., 2006; Alain and Arnott, 2000; DYSON, 2010; Kubovy and Van Valkenburg, 2001; Schnupp et al., 2013; Shinn-Cunningham and Best, 2008). A two-dimensional representation of an auditory object can be defined as a product of grouping mechanisms along the frequency and time dimensions (Griffiths and Warren, 2004a; Griffiths et al., 2012; Bizley and Cohen, 2013). It is a considerably challenging task to define clear perceptual boundaries between simultaneous auditory objects, and to separate the information that belongs to a specific auditory object from the rest of the auditory scene. Several grouping principles are proposed for classification of acoustic perceptual cues in a complex auditory scene (Griffiths and Warren, 2004a; Bizley and Cohen, 2013), which are mainly based

on analyzing auditory patterns in time-frequency space. Auditory grouping can be considered to have two aspects known as sequential grouping and parallel (simultaneous) grouping. Sequential procedure refers to relating the spectral components to their respective sources over time. Parallel grouping, on the other hand, requires determining which parts of the complex acoustic scene presented at the same time belong to which source, specifically. The grouping cues are presumably provided by forming connections between elements of sensory input based on the principles of similarity, continuity, proximity, and common motion. In the case of audition, auditory objects are believed to be formed based on similar features such as pitch, location cues and spectrotemporal properties, discussed in the next section.

Effective cues in auditory segregation

In a classic study of presenting a repeating sequence of temporally non-overlapping high and low-frequency tones, 'A' and 'B', the frequency separation and the presentation rates of the tones were shown to be determinant factors in perceiving the alternating tones as one or two separate steams. Small frequency separation (less than 10%) or low presentation rates, would be perceived as one coherent stream of A-B-A-B tones; however at larger frequency separations and higher presentation rates, it starts to perceptually segregate as two auditory streams of A-A-A-$\cdots$ and B-B-B-$\cdots$(van Noorden, 1975; Bregman et al., 2000). There is also an ambiguous state between the two mentioned states in which an alternating sequence can be perceived as either one or two streams depending on the attentional focus of the listeners and required behavioral tasks in that experiment (Moore and Gockel, 2012).

To address the frequency separation property, a leading theory is that if the frequency components are close enough so that they fall on the same cochlear channel in the periphery auditory system and their excitation patterns in auditory periphery overlap, they will bind together and form a single stream. This theory is also known as peripheral channeling. Hartmann and Johnson (1991) studied streaming in music using interleaved melodies, which were difficult to recognize unless they were segregated and perceived as separated streams. They got the best results for segregation task in the case for which the most peripheral channeling was expected, i.e. the successive tones differed in spectrum.

Harmonicity is another feature that is believed to play an important role in stream segregation. Frequency components with a harmonic relationship fuse together into a single pitch and give a distinct entity to a harmonic complex, which makes it easy to stream out from complexes with different fundamental frequencies (Rasch, 1978; Duifhuis et al., 1982). It is also shown that harmonic relation is not necessary for binding the tones; any regularity in spectral spacing can be sufficient for fusion of the tones into one auditory object (Roberts and Bregman, 1991).

If a sound source is on, all the features belonging to that source are present, and if it goes off, none of the corresponding features exist anymore; thus, onset and offset synchrony can play a key role in integrating those features that are believed to come from the same sound source (Rasch, 1978; Bregman, 1990). According to the reported studies, onset synchrony is a more effective integrating cue compared to the offset synchrony (Darwin and Carlyon, 1995).

Sound sources in the environment have different spatial locations and all the

sound elements belonging to one source would expectedly share the same spatial cues. While it is difficult to evaluate the role of special cues in auditory object formation independently, it is suggested that when the direction of the sound is given to the subject as a prior cue, lateralization would play a strong role in vowel identification (Darwin and Hukin, 2000). It has also been demonstrated that, in situations with more than two concurrent sound sources available, such as in a cocktail party, binaural processing plays a significant role in stream segregation (Yost, 1994).

Among other integrating cues, amplitude modulation seems to be effective, especially in the absence of spectral and temporal fine structures. In a study by Grimault et al. (2002), it is demonstrated that sequential sounds can be streamed apart based on the differences in their rate of temporal fluctuations, while no other spectral or temporal cue is available; however, the effect of frequency modulation on stream segregation seems to be more complicated and less understood.

Coherent movement in space is a powerful cue for object segregation in a visual scene, although it is not true about the analogues auditory scene in which common frequency modulation is expected to play similar role in streaming different sources (Darwin and Carlyon, 1995). According to a study by Carlyon (1991), it is very hard for the subjects to tell whether two inharmonic sounds are being frequency modulated coherently or incoherently. There is, however, some evidence that subjects can use FM coherence to separate concurrent complex sounds perceptually (McAdams, 1982).

Temporal envelope is another important cue, which involves in characterizing

timbre of the sound. In a study by Dannenbring and Bregman (1976), the alternating A-B-A paradigm was expanded to tree tone-tone, tone-noise and noise-noise paradigms, in which a narrowband noise with 1.5 semitones bandwidth around 1 kHz center frequency was used. Since the excitation pattern of the noise is very similar to that of the tone, peripheral channeling cannot play a role in streaming noise from tone sequences. According to the reported results, listeners segregation performance was better in tone-noise condition compared to the other two. One hypothesis is that different temporal envelopes (in noise and tones) characterize them with two different timbres, which leads to a better segregation.

Primitive vs. schema-based segregation

According to Bregman (1990), primitive segregation refers to bottom-up, pre-attentive processes of auditory perceptual organization, in which attention to the sound or learning parameters does not play a role. Mapping of the sound into the acoustic domain and segregating different properties of the sound mixture take place in this stage. However, a good number of our hearing experiences involve schema-based or top-down processes in which, learning, memory, and active attentional state to the sounds come into play. Top-down processes are influenced by listeners prior experiences and familiarity with the presented sound. Extensive studies in humans and animals, shows rapid, long lasting stimulus-specific changes in tuning selectivity and response magnitude of the auditory neurons associated with learning process in the brain (Recanzone et al., 1993; Menning et al., 2000; Polley et al., 2006; David et al., 2012).

There are also a good number of studies demonstrating the effect of attention on neural representation of the sounds in the brain, wherein attention-induced enhancements in the amplitude and selectivity of auditory event-related potential (ERP) components in human auditory cortex are reported (Hillyard and Picton, 1987; Näätänen, 1990). It is not very clear if a focused attention is necessary for the buildup of stream segregation (see Psychoacoustic studies).

A wide range of neurophysiology studies in animals and humans have been trying to solve and model different aspects of auditory scene analysis. Single-unit recordings from an animals brain serves us with much more detailed information about specific neural generators contributing to the recorded signal compared to non-invasive recordings from a human brain. The latter represents auditory-evoked neural activity at the population level from different sources in the brain; however, there is a possibility of performing psychoacoustic tasks in humans along with neural recordings to confirm the perceptual state of the brain at the same time, which is much harder to be done with animals. The following sections briefly describe some of the psychoacoustic, functional imaging and electrophysiology studies on behavioral and neural correlates of auditory stream segregation in humans.

3.1.2 Psychoacoustic studies

The well-known A-B-A streaming paradigm, first investigated in 1970s by Bregman and colleagues (Bregman and Campbell, 1971). As a result of these studies, several models underlying auditory grouping and segregation mechanisms were proposed

and further explored through variety of experimental techniques. Leon van Noorden examined the behavior of human listeners in response to the repeating triplets (A-B-A) of the streaming signal and investigated over perceptual characteristics involved in integration, segregation and bistability (van Noorden, 1977, 1975). Bregman studied important temporal characteristics of the auditory scene in facilitating or impeding the streaming process and highlighted the effect of presentation rates and time intervals between the successive tones in perceptual segregation of the streams (Bregman et al., 2000).

The buildup effect known as the time taken for a stream percept to emerge from an auditory mixture is another well-studied concept in auditory scene analysis. Different mechanisms underlying the gradual tendency of segregating the steams over repeated presentation of the streaming stimulus has been proposed in these studies. For example, frequency-shift detectors are hypothesized to play an important role in integrating successive tones into a single stream (Anstis and Saida, 1985). In another study, accumulation of evidence over time assumed to overwrite the default perceptual state in the brain, which is hypothesized to be perceiving all the contents in the scene as a single stream (Bregman, 1990).

Another significant facet of the streaming paradigm is the perceptual ambiguity in segregating the streams, resulting in alternation of the percept between one or two streams. Bistability in streaming is very common over a wide range of stimulus parameters (Denham and Winkler, 2006; Kashino et al., 2007). The distribution of switches in the perceptual states during streaming is quiet similar to that for visual multi-stability when measured in the same group of subjects and is shown to be

strongly modulated by the task instructions and behavioral goals (Pressnitzer and Hupé, 2006).

Finally, the role of attention has been investigated extensively in stream selection (switching between the streams) van Noorden (1975), and buildup of streaming (Carlyon et al., 2001). From an object-based point of the view, attention operates at the level of auditory streams (objects) that are already grouped in a primitive bottom-up process and it's not directly involved in the formation of the stream itself. In a study by Carlyon et al. (2003), it's demonstrated that tendency to report streaming is reduced by the absence of attention or switch in attention. Also, reseting in the buildup process when attention is briefly diverted away from the streaming signal, reported in a study by Cusack et al. (2004), indicates the prominent role of attention in the buildup process.

3.1.3 Functional imaging studies

EEG/MEG and fMRI techniques have been widely used as non-invasive methods to study the neural correlates of auditory stream segregation in human (Denham and Winkler, 2006; Micheyl et al., 2007a; Melcher et al., 2009; Gutschalk and Dykstra, 2014). To explore the underlying mechanism of stream formation in the brain, specifically the bottom-up, pre-attentive grouping mechanism and a higher order attention-dependent buildup mechanism, both proposed by Bregman, ERP component has been investigated in a variety of experiment designs (Winkler et al., 2005; Snyder et al., 2006). They have found that P2 and N1c responses to the streaming

sequence increased in amplitude correlated with behavioral measures of streaming.

Moreover, the mismatch negativity (MMN) was combined with streaming paradigms to investigate the effect of attention in streaming process (Sussman et al., 1999, 2007; Winkler et al., 2003). In a study by Sussman et al. (2007), ERPs evoked by short trains of tones were recorded while the listeners attention was driven away to a difficult noise intensity change detection task, simultaneously. Segregation was promoted by increasing the frequency separation between subsets of the tones, in which a number of probe tones with different intensity were present. The stimulus design was such that when the whole sequence was heard as a single stream, no regular pattern of iso-intensity existed because of the presence of probe tones with variable intensity. On the other hand, when the sequence was perceived as two separate streams, a regularity of intensity emerges in one of the streams, such that the probe tones with different intensity from the standard tones would pop out and elicit miss-match negativity (MMN) in neural response. According to the diversion of attention away from the tone sequences, the results reported in this study suggest that sustained attention is not required to initiate the formation of the auditory stream. The paradigm for most of the MMN-involved streaming experiments is such that the presence or absence of MMN is used as an indicator to check if the brain has pre-attentively parsed a sequence of sounds into separate auditory streams (Snyder et al., 2006; Winkler et al., 2003; Sussman, 2007). However, these studies do not reveal any information about underlying mechanism of auditory streaming (Snyder and Alain, 2007).

In another study by Gutschalk et al. (2005), they measured the auditory evoked

neuromagnetic field in response to the streaming stimulus. The results demonstrate a strong correlation between the magnitude of the auditory response affected by frequency intervals and inter-stimulus interval (ISI) and perceptual detectability of the streams. In addition, the results for dipole fitting indicates activation of the non-primary auditory cortex in majority of subjects while performing the streaming experiment.

Functional MRI has been also used to investigate the neural bases of streaming. Despite very poor temporal resolution which fails this technique in following the dynamics of behavioral and perceptual states, high spatial resolution allows a highly resolved exploration of the brain regions involved in the streaming process. In a study by Deike et al. (2004), it has been shown that using a variant of the streaming paradigm, the left auditory cortex has an increased activity in segregation of sounds based on spectral cues. In another fMRI study by Cusack (2005), they could not find any activity difference in the auditory cortex while subjects reported the percept of two vs. one stream in the ambiguous conditions; however, they found significant activity difference in the intraparietal sulcus (IPS), for the two perceptual conditions. It is not yet discovered whether the IPS activity is a cause or consequence of the perceptual change in the bistable state (Shamma and Micheyl, 2010), but the results keep up with the role of IPS in visual binding (Xu and Chun, 2009; Hill and Miller, 2009), in which there are evidences of IPS activation during the perceptual state switches. In other studies (Kondo and Kashino, 2009; Kashino and Kondo, 2012) they used an event-related design to investigate the temporal dynamics of brain activity when the perceptual state switches between one and two-stream percepts.

In their studies they confirmed the role of Medial Geniculate Body (MGB) and the primary auditory cortex in perceptual switching during streaming and found that activations in these regions are correlated with the individual differences in perceptual pre-dominance in streaming.

3.1.4 Neurophysiology studies

Direct recordings from human auditory cortex has been performed in studies by Bidet-Caulet et al. (2007); Bidet-Caulet and Bertrand (2009), in which depth electrodes were inserted in the temporal cortex of epileptic patients listening to the streaming paradigms for which the onset asynchrony was manipulated to induce either streaming or grouping. Electrophysiological responses to the identical stimuli and for different percepts (one or two streams) demonstrated a larger transient and steady state responses as well as induced gamma band oscillations for onset synchrony of the two concurrent sounds than in the case of onset asynchrony. A more recent study via intracranical EEG recordings from temporal, frontal and parietal cortex indicates the involvement of areas spread across the superior temporal and perirolandic cortex, middle temporal gyrus, inferior and middle frontal gyrus in the auditory streaming process (Dykstra et al., 2011). These studies provide evidence in favor of the role of higher order non-auditory areas in sound segregation procedure and auditory scene analysis in general.

3.2 Neural correlates of streaming in an informational masking paradigm

As discussed in the previous section, a commonly experienced paradigm for studying auditory perceptual organization is a sequence of two pure tones alternating in time that can be perceived as a single or two segregated auditory objects under certain conditions (Bregman, 1990; Elhilali and Shamma, 2008b; McDermott, 2009; Shamma et al., 2011a). Here, we have used a more spectrally rich paradigm developed by Kidd Jr et al. (1994), known as *informational masking* paradigm, consisted of a target tone sequence, embedded in a cloud of masker tones that are randomly desynchronized (Figure 5.2–A) and has been shown to yield streaming percepts analogous to those of the simpler two-tone sequences (Kidd Jr et al., 1994, 1995, 2011), both in their systematic dependence on stimulus parameters, as well as the improvement of detection over the time course of few seconds (the so-called buildup of streaming, (Carlyon et al., 2001; Micheyl et al., 2005; Näätänen et al., 2001)). We consider this analogy to be the working hypothesis for the current study. Alternative hypotheses regarding differences in streaming mechanisms between the two paradigms are considered in detail in the discussion section.

Previous studies demonstrated critical dependence of target detection ability on the attentional focus of the listeners as well as the density of the masker tones and spectral separation between the target and masker tones (Elhilali et al., 2009; Gutschalk et al., 2008; Micheyl et al., 2007b). The present study expands earlier studies by further manipulating the effect of temporal and spectral components in

the informational masking paradigm and exploring different perceptions of the scene that can be used as an experimental tool for improving or impeding streaming ability of the listeners (Akram et al. (2014a)).

The motivation for this study follows from two earlier studies by Elhilali et al. (2009) and Xiang et al. (2010). In the latter, the interaction between task-driven and stimulus-driven attentional processes for two competing rhythmic sequences at two different rates (4 and 7 Hz) was explored and despite many similarities in neural responses to the two presentation rates, the faster sequence (7 Hz) was quite different from the slower (4 Hz) behaviorally, especially with regard to buildup over the time course of each trial. Modulation rates in the range of 2–10 Hz, are known to be crucially important in grouping the physical and perceptual cues in a complex acoustic scene and stream formation (Kowalski et al., 1996; Miller et al., 2002; Moore and Gockel, 2002). Since the presentation rates studied by Xiang et al. (2010), fall in the range of the slow temporal modulations mentioned above, we decided to use a paradigm similar to the earlier study by Elhilali et al. (2009), and replace the slow target rate (4 Hz) with a faster rate (7 Hz) to explore the neural and behavioral responses to the faster presentation rate more independently as well as conducting a richer behavioral study on the buildup of target detectability as a function of target sequence presentation rate. Moreover, we were interested in testing the hypothesis that magnetoencephalographic (MEG) neural recordings obtained during the psychoacoustic experiments directly reflects the behavioral dependence of streaming on stimulus parameters, under different behavioral conditions. Specifically, the informational masking paradigm allowed us to explore stream formation in a single tone

sequence (target) as (1) a function of target/masker separation, (2) target tone frequency and (3) target repetition rates. The task-specific design of the experiment enabled us to manipulate the attentional focus of the subjects to different features in the scene towards or away from the target sequence (Figure 5.2–A). In the target task subjects detected a frequency-shifted deviant in the target sequence; in the masker task subjects detected a sudden elongation of the masker tones in time. Both tasks required focused attention, but to spectrally and temporally different features of the auditory scene. Since the stimuli presented in both tasks were identical, any differences in the neural representation of the sound are expected to be a result of attentional modulation. The main hypothesis behind this study is to examine the possibility of viewing the neural activity as an indicator of the streaming percept.

3.2.1 Experimental design & procedures

Stimuli

The paradigm is related closely to previous stream segregation experiments in terms of the parameters governing performance (Fishman et al., 2001; Gutschalk et al., 2005; Micheyl et al., 2005; Snyder et al., 2006), but used the informational masking stimulus, a regular foreground embedded within an irregular background. Specifically, each stimulus consisted of two concurrent streams, a narrow-band, temporally regular target tone sequence and a wide-band cloud of tones that were temporally irregular - the masker stimulus. Subjects were asked to detect either a deviation in frequency of one tone in the target sequence (target task), or a deviation in duration

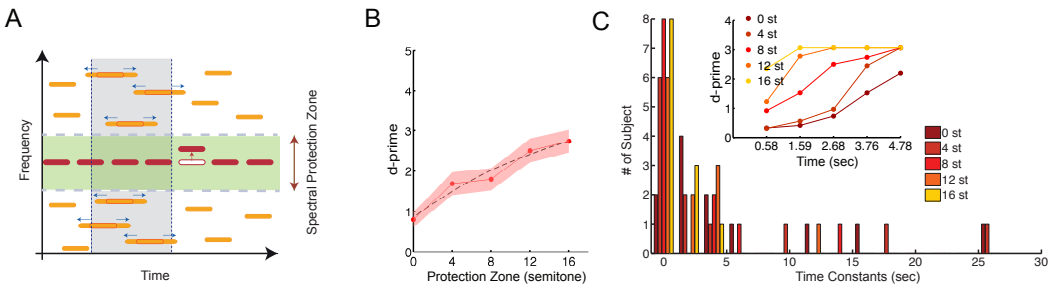

Figure 3.1: Stimulus paradigm and behavioral performance. (A) Schematic representation of the stimulus design. A rhythmic sequence of pure tones (target sequence, red) is placed within a background of randomly distributed (in time and frequency) tones (maskers, yellow) and protected by a spectral zone with no stimulus energy (green region). In the target task, subjects detected a randomly occurring frequency-shifted tone (red arrow). In the masker task, subjects were instructed to detect an elongation of all constituent tones of the masker in a 0.5 s time window (blue arrows). Each trial contained only one type of deviant, both, or none. Subjects performed the tasks in separate blocks, with the order counterbalanced across subjects. (B) Behavioral performance in the target task as a function of protection zone in a range from 0 to 16 semitones (Psychoacoustic experiment A, N=14) (C) Behavioral build-up of detection in the target task. Histogram of time constants obtained from exponential fitting to the buildup curves of the behavioral responses as a function of the size of the protection zone (0–16 semitones). The inset shows behavioral buildup of target task detection for a sample subject illustrating the changes in the buildup speed as a function of different protection zone sizes.

in one of the masker tones (masker task), in separate blocks. Identical stimuli were thus presented in the two different tasks, but with the subjects attention guided to different aspects of the stimuli in the two tasks.

The target sequence was a sequence of identical pure tones with frequency chosen randomly in the range of 250–500 Hz (in 2 semitone steps). The pure tones were presented at a fixed rate in the range of 2–10 Hz, according to the experimental condition. The masker stimulus formed a complex acoustic background consisting of pure tones placed at randomized temporal and spectral positions. The temporal positions of the tones were uniformly distributed over time at a density of 50 tones/s

and over frequency at a spectral resolution of 2 semitones. The spectral positions were chosen uniformly in logarithmic frequency in a range of 5 octaves centered at 353 Hz (ranging from approximately 62 Hz to 1997 Hz), excluding a spectral protection zone, i.e. a frequency band around the target sequence with no masker tone allowed in it. The random sampling of frequencies kept the probability of harmonically related maskers at a minimum. The protection zone on either side of the target sequence varied in width, ranging from 0–16 semitones in 4 semitone steps. The duration of the target and masker tones was 75 ms with 10 ms onset and offset cosine ramps. For the target and masker tasks, deviations were introduced at a randomly chosen constituent tone for either sequence, introducing a frequency, or a duration change, respectively. There were 4 types of trials: (i) null condition (no deviant); (ii) target condition (one target deviant per stimulus); (iii) masker condition (one masker deviant per stimulus); and (iv) combined condition (one target and one masker deviant independently, per stimulus). A target deviant was an upward or downward displacement, of a randomly chosen target note, from the target frequency by 2 semitones. A masker deviant was a single 500 ms window in which all masker tones starting in this window were elongated from 75 ms to 400 ms. For each condition 15 exemplars were generated, differing in the position/tone which was modified. The stimuli were generated using MATLAB (The MathWorks). Each trial stimulus was 5.5 s long and sampled at 44.1 kHz.

Experiments

The effect of manipulation of the paradigms parameters and attentional modulation on task performance and neural responses was explored in 4 different experimental blocks, 2 psychoacoustic (experiments A, B), and 2 MEG (experiments C, D). In the psychoacoustic experiment A and MEG experiment C, the effect of different spectral protection zone width was studied. In psychoacoustic experiment B, the effect of changing target sequence rates with a fixed protection zone was investigated. In MEG experiment D the dependence of the responses on target frequency was measured, as well as changes in buildup and lateralization in the different tasks, all in the context of a fixed protection zone and target tone rate.

Psychoacoustic studies

In the psychoacoustic studies participants performed the tasks at a computer in a soundproof room. They were asked to control the computer using a Graphical User Interface (GUI) and they were allowed to adjust the volume to a comfortable level before starting the experiment. No change of stimulus intensity was allowed after starting the experiment. A complete explanation of the required task, as well as the basic instructions on using the GUI, was given in advance.

Psychoacoustic Experiment A In psychoacoustic experiment A, the effect of different protection zone widths (0, 4, 8, 12 and 16 semitones) with a fixed target rate of 7 Hz was examined. A block of 200 stimuli consisting of 5 protection zones $\times$ 4 conditions $\times$ 10 exemplars were presented to the subjects. Participants could

proceed from one trial to the next, by pressing a button when they were ready. For this experiment, participants were required to do the target task only.

A training block of 15 trials in a decreasing order with respect to the protection zone width was played to the subjects prior to the actual experiment. Participants could listen to each sound as many times as desired and after each trial, they were asked about the presence of deviants in that trial with the correct answer displayed on the screen afterwards. For the real experiment, each stimulus was presented only once, and no feedback was given after each trial. This part lasted approximately 1 hr.

Psychoacoustic Experiment B In psychoacoustic experiment B, the protection zone was fixed to 8 semitones and the rates were varied from 2–10 Hz in steps of 2 Hz. A block of 200 trials consisted of 5 target sequence rates $\times$ 4 conditions $\times$10 exemplars with fixed protection zone was presented to the subjects. For each block participants were required to do the target task only. Training sets of 20 stimuli were provided for each section and they were allowed to listen to each sound as many times as they needed to be able to perform the task. The training block was presented with rates increasing from 2 to 10 Hz in steps of 2 Hz. For the real experiment, stimulus was presented only once, and no feedback was given after each trial. This part lasted approximately 1 hr. Participants performed experiments A and B (psychoacoustics) on 2 different days.

Subjects

In psychoacoustic experiment A, a total of 14 subjects participated (6 male; mean age, 26 y; range 19–33 y), in psychoacoustic experiment B, 12 subjects (5 male; mean age 25 y; range 19–30 y) participated. For MEG experiment C, 12 (7 male; mean age, 25 y, range 18–33 y) participated in the study, whereas in MEG experiment D, 12 subjects (6 male; mean age, 23 y, range 18–33 y) participated. Six subjects took part in all experiments. Psychoacoustic and MEG experiments were conducted over a period exceeding 15 months and therefore subjects participating in different experiments were partly non-overlapping. All participants were right handed (Oldfield, 1971), and had no history of hearing problems or neurological disorders. The University of Maryland Institutional Review Board approved the experiments, and written informed consent was obtained from each participant.

MEG recordings

MEG experiments C & D In the MEG experiment C, 3 identical blocks of 72 trials (3 protection regions $\times$ 4 conditions $\times$ 6 exemplars) presented for each task (totaling 432 trials), whereas in MEG experiment D, only the 8 semitones protection zone stimuli were used and more trials from the same condition were collected. Three identical blocks of 60 stimuli (1 protection region $\times$ 4 conditions $\times$ 15 exemplars) were presented for each task (totaling 360 trials). For both parts, the inter-trial intervals were randomly chosen to be 1.8, 1.9, and 2.0 s. Participants were allowed to rest after each block, but otherwise required to stay still. For both

target and masker tasks, an identical stimulus ensemble (including identical inter-trial intervals) was presented for all subjects and the participants were asked to listen for the presence of a frequency deviant in the target rhythm (target task), or duration deviant in the masker (masker task), based on the task order. Each task deviant was present in exactly half of the trials.

A training block of 20 trials was presented before each task and for each experiment. For the target task, training trials were played in a decreasing order with respect to the protection zone width. For the masker task, the order was increasing. For the real experiment, each participant performed both the masker and the target task, with block order counterbalanced across participants. Each stimulus was presented only once, and no feedback was given after each trial. The entire session of both tasks lasted approximately 2 hrs. Details on MEG signal recording can be found in chapter 2.

A pre-experiment consisting of 200 repetitions of a 1 kHz, 50 ms tone pip was presented before starting the real experiment. The inter-trial intervals were randomized between 0.75 ms and 1.55 s, and participants were asked to count the tone pips. The experiment was done as a control condition to check the M100 response (a prominent peak approximately 100 ms after pip onset) and verify that the location and strength of neural signals fell within a normal range. Next, a training block of 20 trials was presented before each task inside the MEG recorder, during which the sounds could be repeated. Participants verbally indicated the existence of the deviants and the correct answer was given afterwards by the investigator. In the main experiment, participants were presented with three blocks of stimuli

described above. They performed both the masker and the target tasks, with task orders counterbalanced across participants, and were instructed to press a button whenever they heard the appropriate deviant. The button controller was held in the right hand, far away from the sensors. The entire session of both tasks in each MEG experiments (C & D) took around 1 hr.

3.3 Data analysis

3.3.1 Behavioral performance analysis

To evaluate the ability of the participants to perform each task, a d' measure of performance was calculated (Kay, 1993). The hit rate and false alarm probabilities corresponding to deviant detection for each requested task were calculated and converted to z-scores to compute the d' value. To investigate the effect of the pure tone frequency of the target sequences on the behavioral responses in psychoacoustic experiment A, the stimuli were divided into two spectral groups (low-frequency target and high-frequency target), depending on whether the target tone was lower or higher than the middle frequency 353 Hz (those with target frequency of 353 Hz itself were randomly assigned to low- or high-frequency classes). Then we derived a d' measure for each frequency class and across different tasks. To study the build-up of detectability of the target deviant in psychoacoustic part A, we divided the deviant trials into 5 groups according to the deviants location in time, such that each group covered two possible temporal locations for the deviant throughout

the stimulus sequence (out of 10 possible temporal locations for deviants). The hit probability was measured for each group and the false alarm rate was averaged over all 5 groups, independent of its occurrence time because of the uncertainty in false alarm trials. The specific hit rate for each time segment and the averaged false alarm were used to calculate the d' value for corresponding segments. Only one participant gave non-positive d' value of -0.7 due to her high false alarm rate and low hit rate, whose data was excluded from the analysis of build-up. For the psychoacoustic experiment B, d' values were computed as a function of different target sequence rates (Figure 3.2–A). We repeated the same build-up analysis as above for different rate conditions to investigate the interaction between target sequence rate and build-up of target detectability (Figure 3.2–B)

3.3.2 Neural data analysis

To analyze recordings from MEG experiments C and D, in each trial the temporal range from 1.21 to 5.5 s was selected, to exclude onset effects. All shortened responses were concatenated then to make an extended response with duration T = 4.29 s $\times$ number of trials $\times$ number of blocks, for each channel and for each task block. Each extended response was translated to the frequency domain using a discrete Fourier transform (DFT), yielding a frequency spectrum from 0 to 500 Hz at a resolution of $1/T$ Hz. The complex magnetic field strength was obtained by the product of the DFT and the sampling interval ($1/fs$). Power spectral densities were computed by squaring the complex magnetic field strength and normalizing it by

T, the signal duration. Then we calculated the square magnitude of the frequency component at 7 Hz, divided by the average square magnitude of the frequency components in a window around 7 Hz (1 Hz on each side), excluding the component at 7 Hz. The resulting quantity will be referred to as the normalized neural response at 7 Hz and we averaged this quantity over the 20 channels with the strongest normalized responses for each participant. For channel selection, we pooled all trials together regardless of the performed task and 20 best channels with strongest response to the target sequence were chosen. The average square magnitude of the frequency components in the mentioned window (excluding the 7 Hz frequency bin) did not show any significant difference across tasks, so the normalization was task independent, i.e. it was not biased by one of the two tasks. To explore the effect of protection zone width on neural response strength normalized response amplitude for target and masker tasks per protection zone width were calculated for each participant, and averaged over all 12 participants in MEG experiment C (Figure 3.4–A). We investigated the effect of attention on neural response strength, by taking the ratio between the normalized responses to the target vs. masker tasks per participant in MEG experiment D. As described above, the effect of the pure tone frequency of the target sequences on the behavioral responses in MEG experiment D was explored, by dividing the stimuli into two spectral groups (low-frequency target and high-frequency target), depending on whether the target tone was lower or higher than the middle frequency 353 Hz. Then the normalized neural responses were calculated for each frequency class and across different tasks. To study the effect of attention across different frequencies in MEG experiment D, the difference

of normalized responses for two tasks was calculated at 7 Hz and 5 other frequencies: two adjacent bins ($7Hz - df$ and $7Hz + df$), with $df = 7/30$ Hz and 3 other frequencies in theta, alpha and low beta frequency bands, that were multiple integers of df ($21df \simeq 4.9, 43df \simeq 10$ and $64df \simeq 15$). Calculated differences did not show a significant task-dependent effect, since there was not a significant difference over average squared magnitude of the frequency components between 6 Hz and 8 Hz except for the 7 Hz. To analyze the effect of attention on the synchronization between two distinct neural populations in MEG experiment D, phase coherence between two channels m and n, was calculated using all $Q = 180$ trials (Srinivasan et al., 1999):

$$\gamma^2_{mn}(f) = \frac{X^2_{mn}(f)}{< X_{mm}(f) >< X_{nn}(f) >} \tag{3.1}$$

Here, $X_{mn}(f)$ is the average cross spectrum between channel m and channel n, and $X_{mm}(f)$ is the average power spectrum of the individual channel m:

$$X_{mn}(f) = \frac{1}{Q}\sum_{q=1}^{Q} F_{mq}(f)F^*_{nq}(f) \tag{3.2}$$

$F_{mq}(f)$ is the Fourier transform of the q^{th} trial of channel m at frequency f. If two channels keep the same phase difference on every trial, the coherence value would be one, while a random phase difference across trials leads to a coherence value near zero. The coherence difference between target task and masker task was computed for every channel pair. The standard error of the mean (SEM) was constructed to

identify robust coherence change (Bendat and Piersol, 1986; Srinivasan et al., 1999):

$$\epsilon_{mn} = \sqrt{\frac{2}{Q}} \frac{(1 - \gamma_{mn}^2)}{|\gamma_{mn}|} \tag{3.3}$$

To investigate phase modulation in auditory cortex, 20 channels with the strongest normalized neural responses were chosen from 157 neural channels. To further exclude the artificial coherence caused by volume conduction effect on extracranial magnetic filed only distant channel pairs ($>$ 100 mm) were used in the analysis (Srinivasan et al., 1999). In this way, artificial coherence caused by volume conduction was eliminated. The difference between number of channel pairs with robust increased coherence and channel pairs with decreased coherence was normalized over the total number of long-range channel pairs for each participant.

To analyze the possibility of hemispheric difference in response to stimuli in MEG experiment D, the 20 best channels i.e. with the strongest normalized neural response at the target sequence rate, were chosen from each hemisphere separately. The hemispheric normalized neural responses showed no significant lateralization in either task, in contrast to analogous results with the present paradigm at 4 Hz (Elhilali et al., 2009), and hence hemispheric differences were not further analyzed.

The build-up of detectability was studied in MEG experiment D by dividing the entire responses into five temporal segments of approximately 714 ms duration since shorter segments did not show any buildup effect. Corresponding segments extracted from all trials were concatenated to form single extended responses with duration $T \simeq 0.714$ s $\times$ 60 trials $\times$ 3 blocks for each channel. Then the discrete

Fourier transform (DFT) of each single response was computed, resulting in a single Fourier response in the range of 0 to 500 Hz with a frequency resolution of $1/T$ Hz. Different segment durations were used to find the time-scale on which the build-up can be best resolved. Segment lengths were chosen to span an integer number of periods at 7 Hz since we expect to see the build-up in detectability over time windows corresponding to the target sequence rate of 7 Hz.

Behavioral versus neural correlation and bootstrap analysis

The effect of high versus low target frequencies on the behavioral and normalized neural responses from MEG experiment B were correlated using the psychometric and neurometric measures for each subject. Concretely, we computed

$$\delta = \frac{arctan(aNNR(HF) - aNNR(LF))}{(d'(HF) - d'(LF))} \tag{3.4}$$

Where NNR stands for averaged normalized neural response at High (HF) and low frequencies (LF). This angle represents the relationship between the effects of target frequency and neurometric/psychometric measures. It has the virtue of keeping within-subject correlations including their sign relation, but discarding absolute co-scaling of the two measures. The across-participant angles were then combined using circular statistics to yield an angular mean for each task (Fisher, 1995). As a pre-processing step, the neural data (the normalized responses to target) was scaled by a factor of two in order to match the absolute ranges of both neural and behavioral values. A bootstrap procedure was then performed in order to confirm the posi-

tive (respectively, negative) correlation between the neurometric and psychometric functions in the target, respectively, masker task. A balanced bootstrap sampling procedure (Efron and Tibshirani, 1994) was employed by randomly selecting 12 participants with replacement and computing their angular sample mean and repeating this process 1000 times. The procedure was controlled to ensure that all participants appeared the same number of times over all 1000 bootstrap samplings. Confidence measures were then derived from the bootstrap statistics.

Neural source localization

In order to localize the source regions in the brain underlying the magnetic response in all MEG experiments, we used equivalent current dipole analysis. A limited set of complex equivalent current dipoles, best fitting the complex magnetic field configuration at 7 Hz peak in each hemisphere, were computed (Simon and Wang, 2005). Only cortical sources were considered since MEG is not sensitive to subcortical neural sources. The same localization process was done for the M100 neural responses obtained in an auditory test prior to the experiment, in which pure 1 kHz tones were presented to the subjects. Significance of the relative displacement between the target and M100 dipole sources were determined by a two-tailed paired t-test in each of the three dimensions: lateral/medial, anterior/posterior, and superior/inferior. Goodness of fit was computed as the residual variance ratio, as a function of the complex current-equivalent dipole (Simon and Wang, 2005). Only channels with $SNR > 4$ were used in the fitting.

3.3.3 Statistical analysis

Non-parametric tests were used throughout the study to avoid assumptions regarding distributional shape. Single group medians were assessed with the Wilcoxon signed rank test, two group median comparisons with the Mann-Whitney U-test, and multiple groups with the Friedman test, all available in the Matlab Statistics toolbox (The MathWorks, Natick).

3.4 Results

3.4.1 Psychoacoustic results

Wider protection zones facilitated the target task and increase build-up speed

The protection zone—i.e., gap of spectral energy around the target sequence—partially controled the difficulty of segregating the target from the competing maskers background and could potentially induce varying degrees of stream formation. Here, the effect of protection zone width was investigated in a range of 0 to 16 semitones in steps of 4 semitones in psychoacoustic experiment part A (Figure 5.2–B). Participants were asked to perform the target task only. A positive correlation between the protection zone width and behavioral performance of the target task was measured using bootstrap across participants ($p < 0.001$). An exponential recovery curve was fitted to the performance curve, yielding a decay constant of 9.2 semitones and a positive asymptote of 4.4 starting at 0.8. This indicated a progression of the behavioral performance over a large range of protection zones. Notably, even with no

protection zone (PZ = 0), performance remained above chance ($d' = 0.8$, signed ranks test, $p < 0.001$, d' value for chance level was 0), since in this case a frequency change in the target sequence was cued by a disappearance of the target tone at its expected frequency.

We next investigated the build-up of streaming by considering the progression of behavioral performance when the deviants were placed at different times in the target sequence. In the target task the detection performance followed roughly an exponential time course and improved with the width of the protection zone (Figure 5.2–C, inset shows data from a sample subject to reveal the trend for buildup speed as a function of protection zone). This was quantified by the asymptotic values of the exponential fits being positive, a necessary condition to demonstrate build-up (bootstrap across participants, $p < 0.001$).

Time constants of the fitted exponentials decreased significantly from 0 semitones to 4 semitones (6.2 to 5.1 s, bootstrap across participants, $p < 0.001$) and from 4 semitones to 8 semitones (5.1 to 2.6 s, bootstrap across participants, $p < 0.001$), but did not change significantly from 8 to 12 semitones (2.6 to 2.4 s, $p > 0.05$). It also had a significant drop from 12 semitones to 16 semitones (2.4 to 1 s, bootstrap across participants, $p < 0.001$). To better demonstrate the distribution of time constants as a function of different protection zones, a histogram of the time constants for fitted exponential curves to the behavioral buildup curves of individual subjects is plotted for different protection zone widths in Figure 5.2–C. The inset shows example buildup curves of an individual subject. These results suggested that the detection of the target task was easier for larger protection zones, while more

buildup time was required for smaller protection zones.

Faster presentation rates facilitated the target task

Facilitation of task performance in the context of stream segregation has been studied in a number of previous studies (Miller et al., 2002; Shamma et al., 2011a; Vliegen et al., 1999). Here, the effect of the presentation rate of a sequence of stimuli was studies for its known influence on stream formation in the well-known ABA two-tone paradigm (Moore and Gockel, 2002). We investigated this dependence in an informational masking paradigm with targets at different rates in psychoacoustic experiment B (Figure 3.2–A). Using a fixed protection zone width (8 semitones), the rate was varied between 2–10 Hz in steps of 2 Hz. The trials were presented in 5 consecutive blocks corresponding to 5 different rates. Over the range of tested rates, the performance showed significant variation, with higher rates leading to improved detection performance (Figure 3.2–A, signed rank test, $p < 0.0001$). Behavioral performance increased over 2 and 4 Hz presentation rates and hit the maximum level at 6, 8 and 10 Hz (Figure 3.2–A). Looking at the build-up of task detectability as a function of presentation rate, faster build-up was observed for higher presentation rates (Figure 3.2–B, $d' = 2.65$, for the 40 trials in each condition).

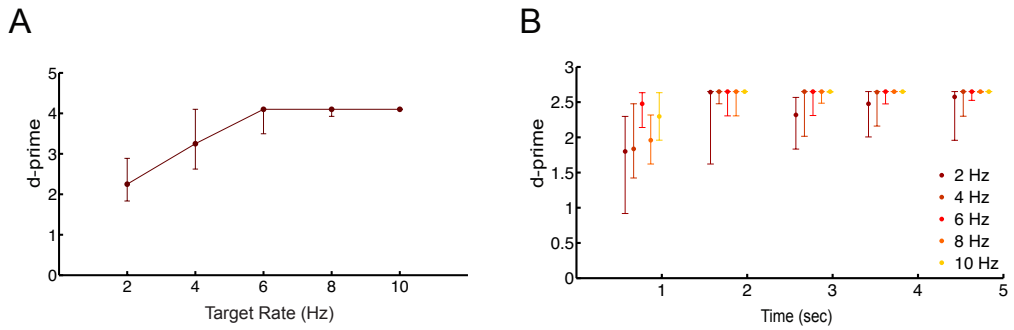

Figure 3.2: Behavioral performance improvement with target sequence rate reflected in neural build-up curve. (A) Behavioral performance (psychoacoustic experiment B, N = 12) as a function of target sequence rate for an expanded range from 2 to 10 Hz, in steps of 2 Hz. Overall performance increased with presentation rate, eventually reaching the ceiling value of $d' = 4.1$ (for 200 trials) (B) Build-up of the behavioral performance as a function of presentation rate. The time for achieving ceiling detection performance is reduced for faster presentation rates. Results are depicted as median and [25,75] % percentiles.

3.4.2 MEG results

Magnetic field distribution showed a stereotypical pattern for neural activity

The magnetic field distributions of the target sequence rate response component revealed the stereotypical pattern for neural activity originating separately in the left and right auditory cortex. The neural sources of all target rhythm response components with high signal-to-noise ratio ($SNR > 4$) originated in auditory cortex (Simon and Wang, 2005). The mean displacement of the neural source from the source of the auditory M100 response was calculated for each hemisphere. The displacement was significantly different in the anterior direction for both right (11.5± 5.8) and left hemisphere (10.8±4.3), using a two-tailed t-test ($t = 3.1$, $p < 0.05$ in the right and $t = 2.4$, $p < 0.05$ in the left hemisphere), but no statistically significant

displacement was observed in other directions. Goodness of fit for these sources was 0.6 ± 0.18 (artificially reduced in accordance with Simon and Wang (2005)). Assuming a M100 origin of planum temporale, this is consistent with an origin for the neural response to the target rhythm in Heschls gyrus, the site of the core auditory cortex, a region known for its good phase-locking to most naturally occurring rates (< 40 Hz) (Liégeois-Chauvel et al., 2004; Miller et al., 2002; Steinschneider et al., 2013).

Attentional modulation of response power and phase coherence

Neural responses to the acoustic stimuli were expected to reflect the physical attributes of the stimulus, but also aspects of the subject's attentional state. Since the stimuli were acoustically identical for the two tasks, differences in the neural responses during the two tasks have to relate to differences in the attentional state. The phase-locked response to the target sequence rate at 7 Hz in MEG experiment C was used as an indicator for the strength of representation of the target stream (Elhilali et al., 2009). As expected, this phase-locked response was stronger during the target task than during the masker task, as indicated by the amplitude of the response power spectrum at 7 Hz (Figure 3.3–A). The individual normalized neural responses at a target rate of 7 Hz showed a larger average power gain than for 4 Hz, which was studied before by Elhilali et al. (2009). Power gain was defined as the ratio of normalized neural response to the target sequence in target vs. masker task. For the present case of 7 Hz the power gain was 3.86 (SEM = 0.87, 3.3–B, red and blue error bars represent target and masker task, respectively), while it had

only been 2.1 at 4 Hz in the previous study by Elhilali et al. (2009). At the same time, the overall amplitudes in 7 Hz experiment were almost a factor 10 smaller than those reported by Elhilali et al. (2009). This overall reduction in amplitudes was likely a consequence of the known low-pass property of auditory cortical responses (Eggermont, 1991; Kilgard and Merzenich, 1999; Phillips et al., 1989; Schreiner and Raggio, 1996).

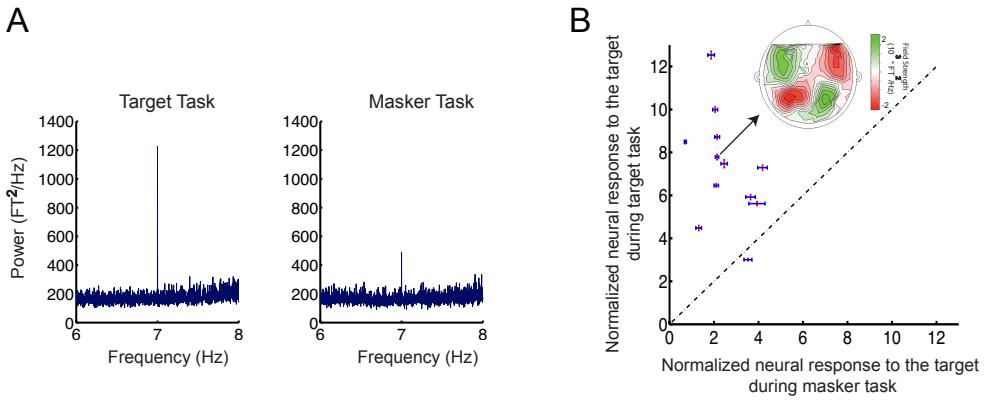

Figure 3.3: Attention modulated the normalized neural response. (A) The power at the target sequence rate was larger in the target task compared to the masker task (MEG experiment D, N = 12, 20 best channels selected for each participant, see Methods for details). (B) Normalized neural response to the target sequence is plotted in target-masker normalized response space for each participant. The normalized neural response is computed as the ratio of the neural response power at the target sequence rate (7 Hz) to the average power of the background neural activity (from 6–8 Hz). Error bars represent the standard error for the target task (red, orthogonal bars) and the masker task (blue, horizontal bars). Inset: the MEG magnetic field distributions of the target rhythm response component for a single participant, with red and green representing the target magnetic field strength projected onto a line with constant phase.

Wider protection zones facilitated the target task, but not the masker task

To get a better understanding of the neural mechanism underlying performance increase as a function of wider protection zone, we used 4, 8, and 12 semitone protection zones to perform MEG experiment C. Behavioral and neural results for both target and masker tasks are shown in Figure 3.4. For the target task, widening the protection zone facilitated the segregation of the target tones, and hence the detection of the frequency deviant (Figure 3.4–A, right panel, signed rank test, $p < 0.001$; significantly positive slope, bootstrap across participants, $p < 0.001$) in agreement with the results obtained in psychoacoustic experiment A. A corresponding increase in the normalized neural response to the target sequence as a function of protection zone was consistent with the changes in the behavioral results (Figure 3.4–A, right panel, signed rank test, $p < 0.001$; significantly positive slope, bootstrap across participants, $p < 0.001$). However, for the masker task, increasing the protection zone did not have a significant effect on behavioral performance (Figure 3.4–A, left panel, signed rank test, p = 0.21). Consistently, there was no significant change in neural activity recorded during the same task (signed rank test, p = 0.1).

Faster rates facilitated the buildup of target detectability

As an extension to the psychoacoustic study part B, it was examined whether the normalized neural responses reflected a similarly rapid build-up of performance for higher target presentation rates To this end, neural and behavioral responses from MEG experiment D were compared with those of 4 Hz target rate studied by Elhilali

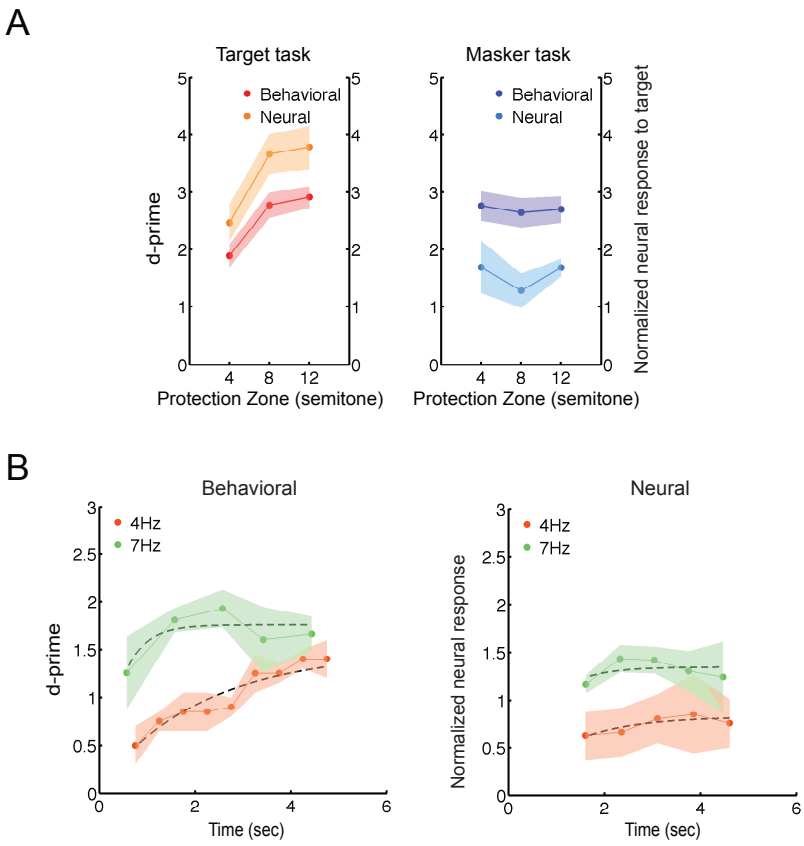

Figure 3.4: Larger protection zones ease the target task, but not the masker task.(A) Behavioral performance and neural results (MEG experiment C, N = 12) for the target task (left panel) and the masker task (right panel), as a function of protection zone. (B) Analysis of neural and behavioral build-up over time for the target task. Behavioral performance (left panel) and neural responses, normalized with respect to the masker task neural response power (right panel) are plotted as a function of time for both the 4 and 7 Hz target sequence rate (orange and green curves, respectively), averaged over participants. Data shown for the 4 Hz target rate is obtained from the study by Elhilali et al. 2009. Neural responses and corresponding behavioral performances are acquired only for the 8-semitone protection zone.

et al. (2009). In the current study, behavioral detectability of the target deviant was calculated for each of the 5 time segments corresponding to the target deviants location (Figure 3.4–B, left panel, green). The build-up of the normalized neural response was measured over the duration of the trial by separating the response into

non-overlapping segments and computing the 7 Hz contributions in each segment. No build-up was observed for window sizes less than 5 cycles, likely due to lack of sufficient statistical power. A weak build-up as a flattened curve was obtained for segment length approximately 714 ms (5 cycles) (Figure 3.4–B, right panel, green), consistent with the progression speed of behavioral response (in MEG recording session, left panel, green). Time constants given by fitted exponentials to both neural and behavioral curves in 7 Hz target, were significantly positive but small (0.63 s for neural curve and 0.1 s for behavioral curve, signed rank test, p = 0.03). Given the fast build-up obtained psychoacoustically for the 8 semitones protection zone of the 7 Hz target (Figure 5.2–C, 3^{rd} panel), we conjecture that a fast neural build-up was occurring at the beginning of trials, but early enough that it could not be resolved using the current analysis. To further validate this analysis, the 4 Hz target data from Elhilali et al. (2009) was reanalyzed. For better comparison of the neural build-up curves, the normalized neural responses for both 4 and 7 Hz target sequence rates were further normalized by the average power of the normalized neural responses in the corresponding masker tasks (Figure 3.4–B, right panel). A significant build-up was obtained using a 750 ms time window (three periods) for 4 Hz. The time constants obtained from exponential curve fittings were significantly positive (1.17 ms for the neural data curve and 11.8 ms for the behavioral data, signed rank test, $p < 0.01$) and larger than the ones for the 7 Hz curves (rank sum test; $p < 0.003$ for behavioral curves and $p < 0.02$ for neural curves), suggesting that the detection task for the 4 Hz target sequence embedded in a 8 semitones protection region was harder than the detection task for 7 Hz target sequence under

similar conditions.

High frequency targets facilitated the target task

Acoustic stimulus parameters influence the saliency of a streaming percept. One such parameter is the frequency of the target tone sequence, which influenced the results both for behavioral and neural data. In MEG experiment D target sequences were divided into high and low frequency tones (above or below 353 Hz). Both behavioral and neural data showed a significantly positive slope (bootstrap across participants, $p < 0.001$) as a function of target frequency in the target detection task (dark/light red line, Figure 3.5–A, left panel), indicating that high frequency tones facilitated target detection. Neither slope was significantly non-zero in the masker task for the average behavioral and normalized neural responses (Figure 3.5–A; right panel); however, the individual behavioral and normalized neural response trends showed a significant negative correlation as explained below. To better demonstrate the correspondence between the normalized neural response and behavioral measures, we computed the correlation between the two indicators during both tasks as a function of target frequency. As described in methods, we computed an angular measure relating neurometric and psychometric changes as a function of frequency. The resulting average angle over subjects was positive (42.4°) for the target task and negative (−29.8°) for the masker task (Figure 3.5–B, yellow line). Bootstrap analysis was performed across participants and the estimated angle is plotted as a green line with the corresponding 95% confidence interval as the pink / blue backgrounds for the target / masker tasks. The positive and negative correlations

obtained for target and masker task respectively, confirmed that behavioral performance in the target task was better for higher frequency targets (> 350 Hz) than for lower frequencies (sum rank test, $p < 0.01$). An increase to the neural response of the target is correlated with this trend. Conversely, the masker task showed a trend of being oppositely affected by the physical saliency of the target task despite the independence of the two tasks.

Attention to the target stream lead to selective power and phase enhancement at target rate

The normalized neural response to the 7 Hz rhythms obtained from MEG experiment D, showed a significant increase in the target vs. masker task (Figure 3.6–A, signed rank test, $p < 0.0001$). In contrast, no significant change in the normalized neural response to the nearby or distant frequencies was obtained, suggesting that the sustained attention to the target stream leads to a feature-selective modulation of the cortical response, but has no significant impact on responses to the other nearby or distant frequencies (signed rank test, p-values $= 0.15$). The perceptual saliency of the attended stimulus may also be the consequence of a more widespread activation by the target in the brain. To explore this possibility, we examined phase coherence between long-distance (> 100 mm) channel pairs. The results revealed a significant enhancement at the target sequence rate (Figure 3.6–B) expressed as the mean number of channels with increased or decreased coherence, in percent of the total number of channel pairs analyzed. The phase enhancement was significantly positive only at 7 Hz (signed ranks test, $p < 0.001$), while changing attentional state of the

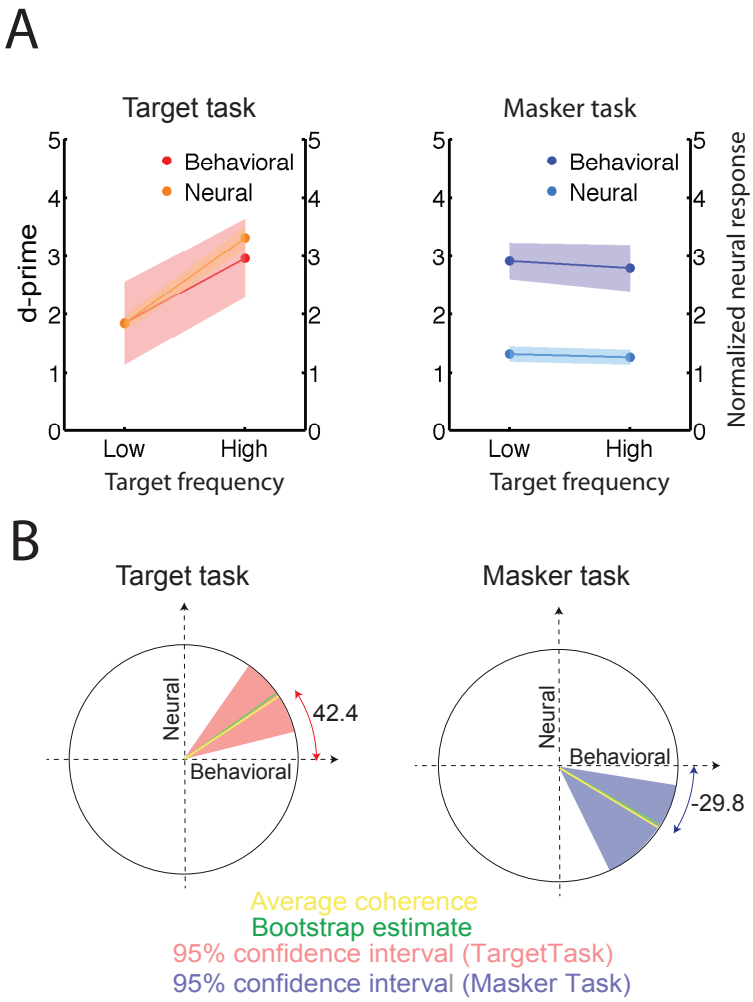

Figure 3.5: Bottom-up saliency of the target sequence increased for higher target frequencies. (A) Behavioral and neural responses (MEG experiment C, N = 12) as a function of target frequency. In the left panel, the red/orange line corresponds to behavioral and neural responses for the target task with respect to the low and high target frequency. In the left panel dark/light blue corresponds to behavioral and neural responses for the masker task. Error hull represent 1 SEM. (B) Correlation of the behavioral and neural responses as a function of target frequency. The ratio of the neural to behavioral response differences as a function of target frequency was averaged across participants. A mean slope angle of 42.4° for target (left plot) task and −29.8° for masker (right plot) task (yellow line) were obtained in this analysis. As detailed in methods, the slope angle corresponds to the strength of correlation between neural and behavioral data. Bootstrap estimates (overlying green lines) and their 95% confidence intervals (pink and blue background for the target and masker task, respectively) confirmed the positive/negative correlations for target/masker task.

subjects did not enhance other rates, and was widely distributed indicating a phase coherence increase within and across hemispheres for the target task relative to the masker task. Therefore, these results suggest an attention or stream formation dependent spatial spread of activity beyond the primary auditory areas.

3.5 Discussion

Stream formation is a central process in parsing the acoustic environment. Perceptual cues and attentional focus modify this segmentation, both qualitatively and quantitatively. In the present experiments, we were pursuing two goals: first, to examine the correspondence between the mechanisms and percepts of stream segregation versus those of the informational masking paradigm, and especially their dependence on the spectrotemporal properties of the stimuli; second, to investigate the potential mechanisms of stream segregation and their interaction with selective attention, for which we employed neuromagnetic imaging. Specifically, while holding the stimulus fixed, we investigated the changes in the neural responses as attention was directed to different components of the acoustic scene. This was repeated under different spectrotemporal stimulus conditions so as to explain the integration of perceptual features of a complex acoustic scene mediated by the processes of attention.

We have based our experimental paradigm and analysis on the hypothesis that detection of the target sequence in an IM stimulus employs similar neural mechanisms used for detection of a target sequence in a classic 2-tone paradigm.

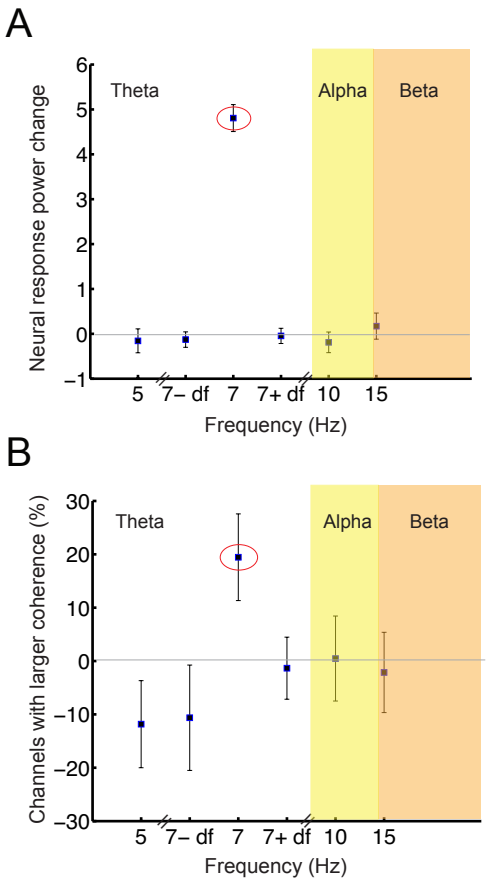

Figure 3.6: Attention to the target stream lead to selective power and phase enhancement at target rate. (A) Power enhancement during target task. The difference between the normalized neural responses in the target task versus the masker task (MEG experiment D, N = 12) showed a significant and highly precise enhancement at the frequency of the target sequence (7 Hz, circled in red). (B) Phase coherence enhancement between distant MEG channels of target relative to masker task. The difference between the number of long-range channel pairs with robust increased coherence in target task, and channel pairs with decreased coherence, was normalized over the total number of long-range channel pairs and converted to %. Error bars represent 1 SEM in each graph.

This hypothesis is supported by a number of earlier studies in which, the similarities between the two paradigms are discussed. First, in line with the arguments discussed in Micheyl et al. (2007b), we believe the systematic dependence of performance on the size of the protection zone is analogous to the frequency separation parameter in streaming experiments using the classic two tone paradigm (Bregman, 1990). Nevertheless, this dependence can rely on the frequency selectivity of neurons in the central auditory system, and hence a low-level neural mechanism (Fishman et al., 2001; Micheyl et al., 2007b). Also in the same study by Micheyl et al. (2007b), it has been shown that performance in detection of the target task decreases if the target tone is presented every other burst. This is also consistent with the results found by Bregman et al. (2000) that the degree of streaming shown to be related to the gaps between successive tones in stream. Moreover, regarding the buildup effect, in the classic streaming paradigm, it has been shown that detection of the target task improves with increasing the number of tone bursts in the target sequence. This is similar to what we see for the IM paradigm. The underlying mechanism in both cases might be explained via sensory-evidence-accumulation, with a causal relationship to the build-up of stream segregation.

Comparing behavioral and neural measures, we have confirmed that attending to one stream significantly modulates the neural response to the attended stimulus. Despite the known transient effects of attention on auditory signals (Näätänen, 1990; Tiitinen et al., 1997), a sustained increase in the normalized neural response was found to correlate with sustained attention. This enhancement is consistent with the behavioral improvement in target detection for individual subjects, which supports

the hypothesis that attentional manipulation can lead to increased responses to the attended features, and suppression of the response to the background or unattended features (Bidet-Caulet et al., 2007; Ding and Simon, 2012a; Mesgarani and Chang, 2012a; Elhilali et al., 2009; Paltoglou et al., 2009; Somers et al., 1999; Zion Golumbic et al., 2013). However, given the stimulus design of the current study, it is also possible that the neural response enhancement during target task is a consequence of having more number of trials in which listeners are aware of the presence of the target sequence vs. the masker task. Therefore, it is quiet difficult to establish whether the enhanced neural power is a cause or an effect of selective attention and a different experimental design would be needed to dissociate between the two interpretations of the current results. Recent studies suggest that oscillatory entrainment in various frequency bands can be enhanced by attention (Kim et al., 2006; Lakatos et al., 2008; Morgan et al., 1996); however, the results of this experiment are unlikely to be a consequence of entrainment since the normalized neural responses show a significant power change only at frequency of the target presentation rate and no other frequencies, even nearby frequency bins (See results and Figure 3.6–A). We conclude accordingly that attentional modulation underlying our results is feature selective engendering plain evoked MEG responses. This power enhancement was in addition accompanied by a significant long-distance coherence of the responses exactly at the rate of the target sequence, indicating an increase in synchronization among distributed populations of responsive neurons.

We also observed a systematic dependence of performance and normalized neural response strength on the width of protection zone for the target task. This is

analogous to the increase in frequency separation between two-tone sequences studied in the more traditional ABA streaming paradigms (Bregman, 1990). According to our findings, increasing the spectral separation improves behavioral detection and, notably, also causes an increase in the normalized neural response. This can be speculatively attributed to the well-known lateral inhibitory interactions, which may occur among tones as much as an octave or more apart (Bartlett et al., 2011; Fishman et al., 2013). In this case, the boundary between energetic and informational masking cannot be defined sharply since very close spectral distances between target and masker tones, will eventually inhabit the same frequency band (dependent on the tuning width of auditory neurons), and thus could be considered as energetic masking rather than informational masking. But it is also possible that these suppressive interactions between the two temporally incoherent streams (target and masker) are inherently due to the desynchronized activation of these respective frequency channels, and not simply to pre-existing inhibitory connections (Shamma et al., 2011a). If so, we would predict that this enhancement of response amplitude with increasing frequency separation would also occur for two alternating tones despite the fact that they are not simultaneous. Given our observation of no significant change of the neural response for the masker task, we postulate attention to be a required component to instantiate the increase of neural responses in the target task.

Higher target tone frequencies produced stronger normalized neural responses, and their deviants were easier to detect. This effect may be due to an enhanced bottom-up saliency that increases as a function of frequency, i.e. tones at higher frequencies (350 Hz to 500 Hz) are perceived to be louder compared to low frequency

tones (250 Hz to 350 Hz) at the same amplitude (ISO 226:2003). Interestingly, the strong, positive correlation between the behavioral and normalized neural response for the target task was complemented by a significant negative correlation in the masker condition. Thus, subjects with a positive/negative behavioral trend as a function of sound frequency, showed a decrease/increase in their corresponding normalized neural response, respectively. This could be explained by the competitive nature of the tasks, i.e. better detection in the masker task requires more effective suppression (decrease) of the competing target sequence. This finding is also significant as it confirms that the difficulty of the tasks was sufficient to manipulate the listeners attention towards or away from the target sequence.

Similarly, increasing the presentation rate of the target sequence had a significant, positive effect on behavioral performance (2–10 Hz range) and its neural correlates in the target detection task (4 vs. 7 Hz), presumably because of the more rapid buildup of target/masker segregation (streaming). This is consistent with previously measured effects of temporal rates in auditory scene analysis in which faster rates induced stronger streaming effects (Bregman, 1990). An earlier study by Xiang et al. (2010) found conflicting results when presenting competing pairs of different temporal rate sequences (4 and 7 Hz) to the subjects, and instructed them to attend to one of the two rhythms and detect in it a deviant temporal jitter. This psychoacoustic study found a streaming advantage for the 4 Hz rhythms relative to the 7 Hz, inconsistent with our current findings and classical streaming studies (Bregman, 1990). We conjecture that this may simply have been a consequence of the reliance on temporal jitter as the deviant, which is more difficult to detect

with faster rates, leading to a decrease in the detection scores of the 7 Hz sequence. Finally, average temporal alignment (termed coherence) has recently been suggested as a dominant contributor to stream formation (Elhilali et al., 2009; Shamma et al., 2011a; Teki et al., 2013). According to the temporal coherence hypothesis, distinct neural populations with temporally correlated responses are grouped together representing one single stream, whereas neural populations with uncorrelated temporal responses are segregated representing different streams. In the present study, temporal alignment did not play an important role for binding across multiple frequency channels, since target and masker streams were temporally uncorrelated (due to the random nature of the masker). However, this lack of coherence may have been used as a discriminating factor, strengthening the perceptual and neural activity independence between the target and masker components of the stimulus, leading to a better target/masker segregation ability for listeners. Using both behavioral and neural measures we have shown that conditions, which facilitate target detection, are paralleled by enhancements in neural activity. This suggests that the neural sources of the MEG signal associated with the target sequence are already affected by the conditions that give rise to the streaming percepts, and are in fact good indicators of the perceptual state of the subjects in perceiving the presence of informational masking in the auditory scene, and perhaps other scenes too.

Chapter 4

A state-space model for decoding auditory attentional modulation in a competing speaker environment

4.1 Introduction

One of the hallmarks of brain function is the ability to segregate and perceive an auditory object in a complex auditory scene composed of several independent sources. From a mathematical perspective, this is a highly ill-posed problem; however, our brain is able to solve this problem in a seemingly effortless fashion. It has been hypothesized that after entering the auditory system, the complex auditory signal resulting from coincident sound sources in a crowded environment is decomposed into acoustic features at different stages of the auditory pathway. Then, a rich representation of spectrotemporal features reaches the auditory cortex, where an appropriate binding of the relevant features and discounting of others leads to the perception of an auditory object (Bergman, 1994; Griffiths and Warren, 2004b; Fishman and Steinschneider, 2010; Shamma et al., 2011b). A compelling example is the Cocktail Party effect (Cherry, 1953; Brungart, 2001; McDermott, 2009), in which a

listener is able to attend to an individual speaker in the presence of other competing speakers and to segregate the attended speech from all other sound sources in the environment.

The neural representation of speech as a distinct auditory object has been extensively studied using auditory scenes consisting of pairs of concurrent speech streams mixed into a single acoustic channel with no spatial cues provided (Ding and Simon, 2012b,c; Mesgarani and Chang, 2012b; Pasley et al., 2012; O'Sullivan et al., 2014). Any neural representation of a single stream of speech (considered as an auditory object) involves complex segregation and grouping processes, given the substantial overlaps in spectral and temporal domains. As reported by these studies, concurrent auditory objects—even those with highly overlapping spectrotemporal features—are neurally encoded as a distinct object in the auditory cortex and emerge as fundamental representational units for high-level cognitive processing. In the case of listening to speech, it has recently been demonstrated that the auditory response manifested in magnetoencephalographic recordings is strongly modulated by the spectrotemporal features of the speech (Ding and Simon, 2012c; Pasley et al., 2012). In the presence of two speakers, this modulation appears to be strongly phase-locked to the spectrotemporal features of the attended speaker as opposed to the unattended speaker (See Figure 4.1–A) (Ding and Simon, 2012b; Mesgarani and Chang, 2012b).

The complexity of this process becomes apparent when one tries to emulate the underlying mechanism via computer algorithms. Mimicking the brain's ability to separate different sound sources when multiple sources are mixed into a single or

small number of channels is not possible without imposing specific constraints on the structure of the signal components. Some existing mathematical models have proven effective in imposing such additional constraints required to achieve single-channel source separation or scene analysis (Wang and Brown, 2006; Ellis, 2006; Shao and Wang, 2008; Li and Huang, 2010; Wang et al., 2013). In conjunction with statistical methods, these models form powerful domains for integrating different information and constraints as well as for probing various hypotheses regarding the underlying neural processes.

Common techniques, which are mainly based on the reverse correlation methodology, create the best approximation of the stimulus from the response of the neural population, which can be compared with the original stimulus to reveal preserved or dismissed features in the population response (Bialek et al., 1991; Gielen et al., 1988; Hesselmans and Johannesma, 1989). Although useful for evaluating data from multiple neurons using electrocorticography(ECoG) (Mesgarani et al., 2009; Mesgarani and Chang, 2012b) or MEG (Ding and Simon, 2012b), these methods have a number of limitations. First, the achievable temporal resolution of the current techniques is of the order of several minutes. In a real-world scenario, attention of the listener can switch dynamically from one speaker to another; therefore, an appropriate decoder needs to have a dynamic estimation framework with high temporal resolution in order to capture attention switches in real-time, especially in light of the emergence and rapid growth of brain-computer interface systems. Second, the full spectrotemporal features of speech are employed for decoding neural responses. It is not clear whether the decoding can be carried out with a more parsimonious set of spec-

trotemporal features. Finally, these decoders often rely on ad-hoc assumptions and simplifications, which in turn overshadow a reliable statistical interpretation of the data.

In this chapter, the aforementioned limitations are addressed by introducing a biologically-inspired state-space model that accounts for the dynamicity of the attentional state as well as its correlation with MEG observation in a competing-speaker scenario (Akram et al., 2014b). To this end, a forward model relating the auditory MEG activity to the envelopes of the two speech streams is utilized by employing the sparse structure of the auditory response. The attentional state of the listener is then modeled using a non-stationary Bernoulli process. Finally, von Mises-Fisher circular statistics are employed to form a robust inverse model that accounts for the correlation of the observed MEG data with respect to the two speech streams. The Maximum *a posteriori* (MAP) estimation framework is used to infer the state-space parameters from the observed data. In particular, a novel application of two nested Expectation-Maximization (EM) algorithms is devised to efficiently solve the MAP problem.

The proposed model has several advantages over existing methods. First, theoretically speaking, the proposed state-space model is able to preserve dynamics as fast as the sampling resolution. Simulation studies as well as application to experimental data reveal that this model is indeed capable of predicting the attentional state of the listener with a temporal resolution of several *seconds*, which is a significant improvement over the state-of-the-art temporal resolution of several *minutes*. Second, This model only requires the envelopes of the two speech streams as co-

variates, which is a substantial reduction in the dimension of the spectrotemporal feature set used for decoding auditory attention. Finally, the current state-space framework provides confidence bounds on the state parameters as a by-product of this estimation method, which can in turn be used for precise statistical inference precedes such as hypothesis testing. Further simulation studies as well as applications of this method on experimentally acquired MEG data is provided in this chapter. The analyses reveal the superior performance of the proposed decoder in tracking the attentional state of a listener in a competing-speaker environment, as compared to existing techniques.

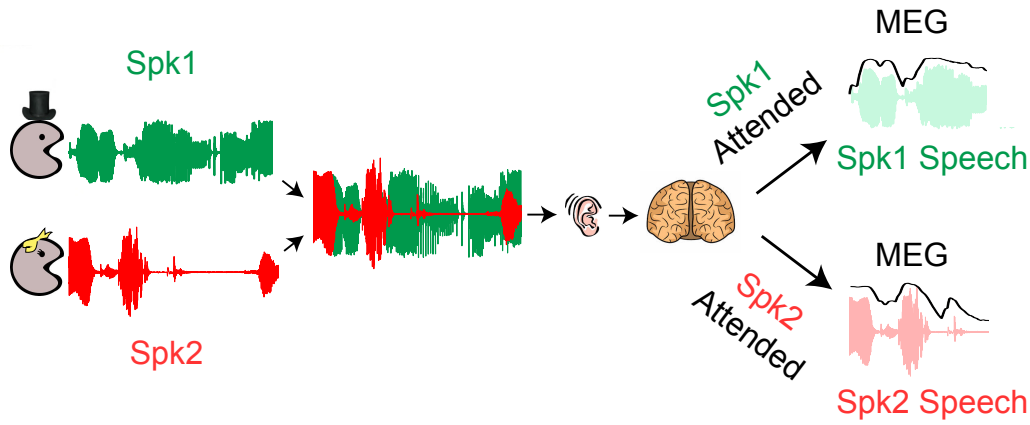

Figure 4.1: Schematic depiction of auditory object encoding in the auditory cortex. Here, the auditory scene consists of the mixture of two concurrent speech streams. Recent studies show that cortical activity (solid thick traces) is selectively phased-locked to the temporal envelope of the attended speaker (solid thin traces) as opposed to the unattended speaker's envelope.

4.2 Computational modeling

Current modeling framework is devided into three stages: the forward problem of relating the MEG observations to the temporal features of the attended and unattended speech streams; the attention model which takes into account the dynamics of selective attention; and the inverse problem of decoding the attentional state of the listener given the MEG observations and the temporal features of the two speech streams.

4.2.1 The forward problem: estimating temporal receptive fields

Consider a task where the subject is passively listening to a speech stream. Let the discrete-time MEG observation at time t, sensor j, and trial r be denoted by $x_{t,j,r}$, for $t = 1, 2, \cdots, T$, $j = 1, 2, \cdots, M$ and $r = 1, 2, \cdots, R$. Let the time series $y_{1,r}, y_{2,r}, \cdots, y_{T,r}$ denote the auditory component of the MEG observations. This component can be obtained through source localization techniques or sensor-space source separation algorithms, and will be denoted hereafter by MEG data (See Section 4.2.3). Also, let E_t be the speech envelope of the speaker at time t in dB scale. In a linear model, the MEG data is linearly related to the envelope of speech as:

$$y_{t,r} = \tau_t * E_t + v_{t,r}, \tag{4.1}$$

where τ_t is a linear filter of length L denoted by the temporal response function (TRF), $*$ denotes the convolution operator, and $v_{t,r}$ is a nuisance component accounting for trial-dependent and stimulus-independent components manifested in

the MEG data. It is known that the TRF is a sparse filter, with significant components analogous to the M50 and M100 auditory responses (Ding and Simon, 2012c,b)(See Figure 4.3–B). A commonly used technique for estimating the TRF is known as Boosting (David et al., 2007; Ding and Simon, 2012c), where the components of the TRF are greedily selected to decrease the mean square error (MSE) of the fit to the MEG data. Here, an alternative estimation framework based on ℓ_1-regularization is employed. Let $\boldsymbol{\tau} := [\tau_L, \tau_{L-1}, \cdots, \tau_1]'$ be the time-reversed version of the TRF filter in vector form, and let $\mathbf{E}_t := [E_t, E_{t-1}, \cdots, E_{t-L+1}]'$. In order to obtain a sparse estimate of the TRF, we seek the ℓ_1-regularized estimate:

$$\widehat{\boldsymbol{\tau}} = \underset{\boldsymbol{\tau}}{\operatorname{argmin}} \sum_{r,t=1}^{R,T} \|y_{t,r} - \boldsymbol{\tau}'\mathbf{E}_t\|_2^2 + \gamma\|\boldsymbol{\tau}\|_1, \tag{4.2}$$

where γ is the regularization parameter. The above problem can be solved using standard optimization software. Here, a fast solver based on iteratively re-weighted least squares is employed (Ba et al., 2014). The parameter γ is chosen by two-fold cross-validation, where the first half of the data is used for estimating $\boldsymbol{\tau}$ and the second half is used to evaluate the goodness-of-fit in the MSE sense. In a competing-speaker environment, where the subjects are only attending to one of the two speakers, the linear model takes the form:

$$y_{t,r} = \tau_t^a * E_t^a + \tau_t^u * E_t^u + v_{t,r}, \tag{4.3}$$

with τ_t^a, E_t^a, τ_t^u, and E_t^u, denoting the TRF and envelope of the attended and unattended speakers, respectively. The above estimation framework can be generalized to the two-speaker case by replacing the regressor $\boldsymbol{\tau}'\mathbf{E}_t$ with $\boldsymbol{\tau}^{a\prime}\mathbf{E}_t^a + \boldsymbol{\tau}^{u\prime}\mathbf{E}_t^u$, where $\boldsymbol{\tau}^a$, $\mathbf{E}_t^a$, $\boldsymbol{\tau}^u$, and $\mathbf{E}_t^u$ are defined in a fashion similar to the single-speaker case. Similarly, the regularization $\gamma\|\boldsymbol{\tau}\|_1$ is replaced by $\gamma^a\|\boldsymbol{\tau}^a\|_1 + \gamma^u\|\boldsymbol{\tau}^u\|_1$.

Selective attention: non-stationary Bernoulli process Suppose that at each window of observation, the subject is attending to either of the two speakers. Let $n_{k,r}$ be a binary variable denoting the attention state of the subject at window k and trial r:

$$n_{k,r} = \begin{cases} 1 & \text{attending to speaker 1} \\ 0 & \text{attending to speaker 2} \end{cases} \tag{4.4}$$

The subjective experience of attending to a specific speech stream among a number of competing speeches reveals that the attention often switches to the competing speakers, although not intended so by the listener. Therefore, we model the statistics of $n_{k,r}$ by a Bernoulli process with a success probability of p_k:

$$P(n_{k,r}|p_k) = p_k^{n_{k,r}}(1-p_k)^{1-n_{k,r}}. \tag{4.5}$$

A value of p_k close to 1 (vis-á-vis 0) implies attention to speaker 1 (vis-á-vis 2). The process $\{p_k\}_{k=1}^K$ is assumed to be common among different trials. In order to model the dynamics of p_k, we define a variable z_k such that

$$p_k = \mathsf{logit}^{-1}(z_k) := \frac{\exp(z_k)}{1+\exp(z_k)}. \tag{4.6}$$

When z_k tends to $+\infty$ (vis-á-vis $-\infty$), p_k tends to 1 (vis-á-vis 0). We assume that z_k obeys first-order autoregressive dynamics of the form:

$$z_k = z_{k-1} + w_k, \tag{4.7}$$

where w_k is a zero-mean i.i.d. Gaussian random variable with a variance of η_k. We further assume that η_k are distributed according to the conjugate prior given by the inverse-Gamma distribution with hyper-parameters α (shape) and β (scale).

4.2.2 The inverse problem: decoding attentional modulation

Let $y_{1,r}, y_{2,r}, \cdots, y_{T,r}$ denote the MEG data time series at trial r, for $r = 1, 2, \cdots, R$ during an observation period of length T. For a window length W, let

$$\mathbf{y}_{k,r} := \left[y_{(k-1)W+1,r}, y_{(k-1)W+2,r}, \cdots, y_{kW,r}\right], \tag{4.8}$$

for $k = 1, 2, \cdots, K := \lfloor T/W \rfloor$. Also, let $E_{i,t}$ be the speech envelope of speaker i at time t in dB scale, $i = 1, 2$. We extract the envelope of the speech signal by taking the absolute value of its analytic extension (Hilbert Transform). In order to eliminate ringing and obtaining a smoothed envelope, the result was subjected to a low-pass filter with a cut-off frequency of 20 Hz. Let τ_t^a and τ_t^u denote the TRFs of the attended and unattended speakers, respectively. The MEG predictors in the

linear model are given by:

$$\begin{cases} e_{1,t} := \tau_t^a * E_{1,t} + \tau_t^u * E_{2,t}, & \text{attending to speaker 1} \\ e_{2,t} := \tau_t^a * E_{2,t} + \tau_t^u * E_{1,t}, & \text{attending to speaker 2} \end{cases}, \quad t = 1, 2, \cdots, T. \quad (4.9)$$

Let

$$\mathbf{e}_{i,k} := \left[e_{i,(k-1)W+1}, e_{i,(k-1)W+2}, \cdots, e_{i,kW} \right], \quad (4.10)$$

for $i = 1, 2$ and $k = 1, 2, \cdots, K$. Recent work by Ding and Simon (2012b) suggests that the MEG data $\mathbf{y}_k$ is more correlated with the predictor $\mathbf{e}_{i,k}$ when the subject is attending to the ith speaker at window k. Let

$$\theta_{i,k,r} := \arccos\left(\left\langle \frac{\mathbf{y}_{k,r}}{\|\mathbf{y}_{k,r}\|_2}, \frac{\mathbf{e}_{i,k}}{\|\mathbf{e}_{i,k}\|_2} \right\rangle \right) \quad (4.11)$$

denote the empirical correlation between the observed MEG data and the model prediction when attending to speaker i at window k and trial r. When $\theta_{i,k,r}$ is close to 0 (vis-á-vis π), the MEG data and its predicted value are highly (vis-á-vis poorly) correlated. Inspired by the findings of Ding and Simon (2012b), we model the statistics of $\theta_{i,k,r}$ by the von Mises-Fisher distribution (Fisher, 1993):

$$p\left(\theta_{i,k,r}\right) = \frac{\kappa_i^{W/2-1}}{2\pi^{W/2} I_{W/2-1}(\kappa_i)} \exp\left(\kappa_i \cos\left(\theta_{i,k,r}\right)\right), \qquad \theta_{i,k,r} \in [0\ \pi], \quad i = 1, 2 \quad (4.12)$$

where $I_W(\cdot)$ is the W^{th} order modified Bessel function of the first kind, and κ_i denotes the spread parameter of the von Mises-Fisher distribution for $i = 1, 2$. The von Mises-Fisher distribution gives more (vis-á-vis less) weight to higher (vis-á-vis

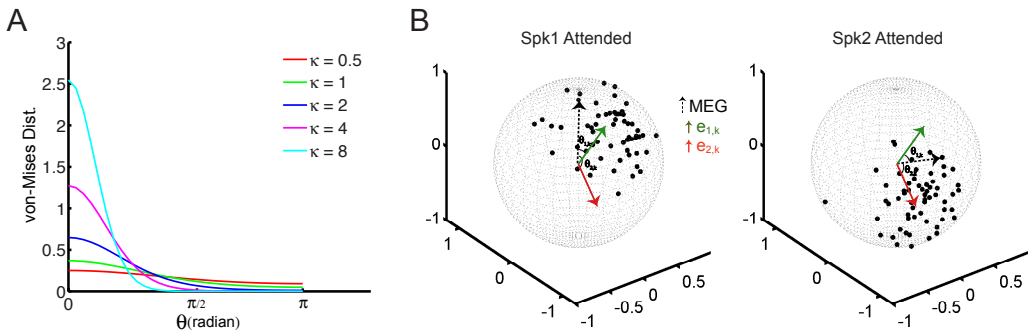

Figure 4.2: (A) von Mises–Fisher probability density for different κ parameters. (B) Schematic view of von Mises–Fisher statistics on a three dimensional sphere: normalized MEG data points are shown by black dots on the unit sphere. Red and green arrows indicate the vectors of predicted MEG based on attending to speaker 1 or speaker 2, respectively. The angles between the MEG data at window k and each of the predictors are shown as $\theta_{1,k}$ and $\theta_{2,k}$, for the case of attending to speaker 1 (right plot) and speaker 2 (left plot), respectively. The point cloud formed by the MEG data is aligned with the direction of the predictor vector corresponding to the attention state.

lower) values of correlation between the MEG data and its predictor. The spread parameter κ_i accounts for the concentration of $\theta_{i,k,r}$ around 0. Figure 4.2 shows a schematic depiction of the von Mises–Fisher statistics in modeling the correlation of MEG data with its predictors based on speech envelopes. We assume a conjugate prior of the form $p(\kappa_i) \propto {\kappa_i}^{d(W/2-1)} \frac{\exp(c_0 d \kappa_i)}{I_{W/2-1}(\kappa_i)^d}$ over κ_i, for some hyper-parameters c_0 and d.

Parameter estimation: a novel EM-based decoder Let

$$\Omega := \left\{ \kappa_1, \kappa_2, \{z_k\}_{k=1}^K, \{\eta_k\}_{k=1}^K \right\} \tag{4.13}$$

be the set of state-space parameters. In principle, these parameters can be estimated through maximum *a posteriori* (MAP) estimation. However, due to the

involved functional form of the log-likelihood and particularly temporal coupling of the state parameters, direct maximization of the log-posterior requires solving a high dimensional convex optimization problem. Instead, we use a novel form of the Expectation-Maximization (EM) algorithm to efficiently estimate the state parameters (Dempster et al., 1977). Taking $\{n_{k,r}\}_{k=1,r=1}^{K,R}$ as the unobserved data, the complete data log-posterior can lead to a feasible MAP estimate of the parameters, due to its tractable functional form for optimization purposes (Appendix A).

The overall estimation procedure consists of two nested EM algorithms and is outlined in Algorithm 1. In the outer EM, the E-step involves computing $\mathbb{E}\left\{n_{k,r}^{(\ell+1)}\middle|\Omega^{(\ell)},\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\right\}$, using Bayes' rule, and the M-step updates $\kappa_1^{(\ell+1)}$, $\kappa_2^{(\ell+1)}$, $\{\eta_k\}_{k=1}^K$ and $\{z_k\}_{k=1}^K$. As for the last two sets of parameters, the maximization in the M-step itself is computed using the inner EM algorithm. In the inner EM algorithm, the E-step corresponds to a Bernoulli smoothing algorithm (Smith and Brown, 2003; Smith et al., 2004) and the M-step updates the state variance sequence (Shumway and Stoffer, 1982). The detailed derivations of the estimation procedure are provided in Appendices A, B and C. Confidence intervals for $p_k^{(\ell)}$ can be obtained by mapping the confidence intervals for the Gaussian variable $z_k^{(\ell)}$ via the inverse logit mapping. In summary, the decoder inputs the MEG observations and the envelopes of the two speech streams, and outputs the Bernoulli success probability sequence corresponding to attending to speaker 1. The choice of the hyper-parameters will be discussed in Section 4.4.

Algorithm 1: Estimation of the State-Space Parameters

input : MEG observations $\{\mathbf{y}_{k,r}\}_{k,r=1}^{K,R}$, tolerance $\mathsf{tol} \in (0, 0.001)$ and maximum number of iterations for outer and inner EM algorithms $L_{\max}$ *and* $M_{\max} \in N^{+}$, respectively.

Initialization: initial guess of state variables $z_k^{(0)}$ and state-noise variances $\eta_k^{(0)}$ for $k = 1, 2, \cdots, K$, initial conditions $z_{0|0}$ and $\sigma_{0|0}$, Initial values for von Mises-Fisher distribution parameters $\kappa_1^{(0)}$ and $\kappa_2^{(0)}$. Initialize iteration numbers to $l = 1$ *and* $m = 1$;

Outer EM iteration:

while $l \leq L_{\max}$ *or relative change in log-posterior* $\geq \mathsf{tol}$ **do**

E-step: Compute $\mathbb{E}^{(\ell)}\{n_{k,r}\} := \mathbb{E}\left\{n_{k,r} \middle| \{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}, \Omega^{(\ell)}\right\}$, *for all* $k = 1, 2, \cdots, K$ *and* $r = 1, 2, \cdots, R$. (A.2).

M-step: Update $\kappa_1^{(\ell+1)}$ *and* $\kappa_2^{(\ell+1)}$ (A.3).

Inner EM iteration:

while $m \leq M_{\max}$ *or relative change in log-posterior* $\geq \mathsf{tol}$ **do**

E-step: Compute $\bar{z}_{k|K}^{(\ell+1,m)} := \mathbb{E}\left\{z_k \middle| \{\mathbb{E}^{(\ell)}\{n_{k,r}\}\}_{k,r=1}^{K,R}\right\}$ *for all* $k = 1, 2, \cdots, K$ *using Bernoulli smoothing* (A.5,A.6).

M-step: Update $\eta_k^{(\ell+1,m)}$, *for all* $k = 1, 2, \cdots, K$. (A.7).

end

$z_k^{(\ell+1)} := \bar{z}_{k|K}^{(\ell+1,m)}$ *and* $\eta_k^{(\ell+1)} := \eta_k^{(\ell+1,m)}$, *for all* $k = 1, 2, \cdots, K$.

end

output: $\widehat{\kappa}_1 := \kappa_1^{(L+1)}, \widehat{\kappa}_2 := \kappa_2^{(L+1)}$, $\widehat{p}_k := \mathsf{logit}^{-1}\left(z_k^{(L+1)}\right)$ and $\widehat{\eta}_k := \eta_k^{(L+1)}$ for all $k = 1, 2, \cdots, K$, where $L \leq L_{\max}$ is the final counter value of the outer EM.

4.2.3 Experimental procedure & data analysis

Subjects, stimuli, and procedures Eleven normal-hearing, right-handed young adults (ages between 20 and 31) participated in this study, consisting of two experiments: constant-attention experiment (eight subjects, three female) and attention-switch (seven subjects, four female). Four subjects (three female) participated in both experiments. All subjects were compensated for their participation. The experimental procedures were approved by the University of Maryland Institutional

Review Board. Written, informed consent was obtained from each subject before the experiment.

The stimuli consist of segments from the book *A Child's History of England* by Charles Dickens, narrated by two different readers (of opposite genders). Four speech segments (one target and one masker segment for each speaker) were used to generate three speech mixtures. Each speech mixture was constructed by mixing two speech segments digitally in a single channel with duration of 1 minute, as described next. The first mixture was generated using the male target segment and the female masker segment, whereas the second mixture was generated using the female target segment and the male masker segment. The third mixture was generated using male and female target segments. Periods of silence longer than 300 ms were shortened to 300 ms to keep the speech streams flowing continuously. All stimuli were low-pass-filtered below 4 kHz and delivered diotically at both ears using tube phones plugged into the ear canals. In all trials, the stimuli were mixtures with equal root-mean-square values of sound amplitude, presented roughly at a 65 dB sound pressure level (SPL).

In the constant-attention experiment, subjects were asked to focus on one speaker (speaker 1, male; speaker 2, female) through the entire trial. In the attention-switch experiment, subjects were instructed to focus on one speaker in the first 28 seconds of the trial, switch their attention to the other speaker after hearing a 2 second pause (28^{th} to 30^{th} seconds), and maintain their focus on the latter speaker through the end of that trial. Consequently, there were four conditions: 1) attending to speaker 1 for the entire trial duration, 2) attending to speaker 2 for the entire trial

duration, 3) attending to speaker 1 and switching to speaker 2 halfway through the trial, and 4) attending to speaker 2 and switching to speaker 1 halfway through the trial. The first mixture was used as the stimulus for condition 1, second mixture for condition 2 and third mixture for conditions 3 and 4. Each mixture was repeated three times during each experimental condition. The first second of each section was replaced by the clean recording from the target speaker to help the listener attend to the target speaker. After each condition was presented, subjects answered comprehensive questions related to the passage on which they focused, as a way to keep them motivated on attending to the target speaker. Eighty percent of the questions were correctly answered on average. The order of presentation for the constant-attention experiment (conditions 1 and 2), and the attention switch (conditions 3 and 4) was counterbalanced across subjects participating in that experiment.

A pilot recording from subjects listening to single speaker trials was performed prior to the actual study. In this experiment, 6 trials (3 repetitions of each male and female target segments) were presented to the subjects and recordings were used for estimating the Temporal Receptive Fields (TRFs) in the forward model.

A pre-experiment consisting of 100 repetitions of a 1 kHz, 50 ms tone pip was presented to all subjects after entering the MEG machine. The results were used as a control condition to check the M100 response (a prominent peak in auditory response, approximately 100 ms after pip onset) and verify that the location and strength of neural signals fell within a normal range (Lütkenhöner and Steinsträter, 1998). The inter-trial intervals were randomized between 0.75 ms and 1.55 s, and participants were asked to count the tone pips.

Data recording MEG signals were recorded in a dimly lit magnetically shielded room (Yokogawa Electric Corporation) using a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan), and with a sampling rate of 1 kHz. Detection coils were arranged in a uniform array on a helmet-shaped surface on the bottom of the dewar, with 25 mm between the centers of two adjacent 15.5-mm-diameter coils. Sensors are configured as first-order axial gradiometers with a baseline of 50 mm; their field sensitivities are 5 fT/Hz or better in the white noise region.

The presentation software package from Neurobehavioral Systems was used to present stimuli to the subjects. The sounds (approximately 65 dB SPL) were delivered to the participants ears with 50 sound tubing (E-A-RTONE 3A; Etymotic Research), attached to E-A-RLINK foam plugs inserted into the ear canal. The entire acoustic delivery system was equalized to give an approximately flat transfer function from 40 to 3000 Hz, thereby encompassing the range of the presently delivered stimuli.

A 200 Hz low-pass filter and a notch filter at 60Hz were applied to the magnetic signal online. Three of the 160 channels were magnetometers separated from the others and used as reference channels in measuring and canceling environmental noise (de Cheveigné and Simon, 2007). Five electromagnetic coils were used to measure each subject's head position inside the MEG machine. The head position was measured twice during the experiment, once before and once after to quantify the head movement.

MEG processing and neural source localization Recorded MEG signals contained both stimulus-driven responses and stimulus-irrelevant background neural activity. In order to extract components that were phase-locked to the stimulus and consistent over trials, as opposed to the random irrelevant activities, we employed the Denoising Source Separation (DSS) algorithm (de Cheveigné and Simon, 2008). This algorithm is a blind source separation method that decomposes the data into temporally uncorrelated components by removing inconsistent temporal components that are not phased-locked to the stimulus. In other words, DSS suppresses the components of the data that are noise-like and enhances those that are consistent across trials, with no knowledge of the stimulus or the timing of the task. The recorded neural response during each 60s was band-pass filtered between 1–8 Hz and down sampled to 200 Hz before submission to the DSS analysis. We found that only the first DSS component contains a significant amount of stimulus information, so analysis was restricted to this component, which we denote by the auditory MEG component throughout this paper. The spatial magnetic field distribution pattern of the auditory MEG component was used for neural source localization. In all subjects, the magnetic field corresponding to the auditory MEG component showed a stereotypical bilateral dipolar pattern (See Figure 4.3–A).

Statistical analysis All trials were tested for significance of the correlation values between MEG data and the model prediction—attended speaker's speech envelope—to remove noise-contaminated trials from further analysis. In order to test the uncorrelatedness hypothesis, the Fisher transformation of the Pearson's correlation

ρ between the MEG data and the model prediction was computed:

$$r = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{4.14}$$

The empirical Fisher transformed correlation values are assumed to be approximately normally distributed with a mean value of r and a standard deviation of $\frac{1}{\sqrt{N-3}}$, where N indicates the number of samples within each trial. All trials were evaluated on a two-tailed test of the population mean with the null hypothesis of $r = 0$. Poorly correlated trials (at a confidence level of 70%) were omitted, resulting in rejection of about 0.6% of the trials (1 out of 180).

In order to verify the significance of difference in correlation values between the MEG data and the attended (ρ_1) vs. unattended (ρ_2) speakers, a similar procedure was carried out for each trial. A two-tailed test of the population mean with the null hypothesis of $r_1 = r_2$ —with r_1 and r_2 denoting the Fisher transformation of ρ_1 and ρ_2, respectively —was evaluated, and all those trials for which the resulted normally distributed correlation values were not significantly different (at a confidence level of 70%) were rejected. The Benjamini-Hochberg False Discovery Rate (FDR) correction (Hochberg and Benjamini, 1990) at an FDR rate of 30% was applied to the last two conditions with attention-switch to correct for multiple comparisons. Three percent of the trials (4 out of 179) were rejected as a result; all further analysis was carried out using the remaining 175 trials.

4.3 Results

In order to evaluate the performance of the state-space model in decoding the attentional state of listeners and to illustrate the effectiveness of this model in various stimulus conditions, a number of realistic simulations and experimental data sets were employed. We first present our results on the robust estimation of TRF, which forms the basis of the forward models used in both simulations and experimental data analysis. We will then present simulation results which highlight the capability of our proposed estimation framework in tracking the attentional state under a wide range of SNR values as well as dynamics. Finally, we will apply the proposed attentional decoding framework to real MEG data from several subjects which chimes in accordance to our simulation studies.

4.3.1 TRF estimation

The spatial magnetic field distribution pattern of the auditory MEG component is shown for all subjects in Figure 4.3–A. As expected for auditory evoked fields, the field maps show a stereotypical bilateral dipolar pattern. Estimated TRFs using the auditory MEG component for all subjects are shown in Figure 4.3–B. TRFs corresponding to the attended speaker were estimated from the pilot conditions where only single speech streams were presented to the subjects. Separate TRFs were obtained for male and female speakers, using 3 repeated trials for each and the TRF with smaller fitting normalized least square error was picked and used through the rest of analysis. The TRF corresponding to the unattended speaker

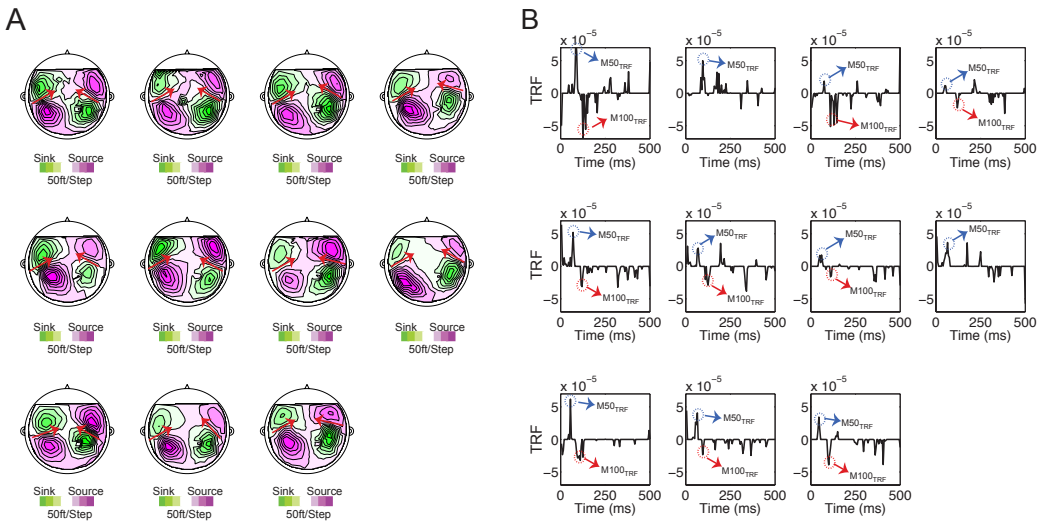

Figure 4.3: TRF estimation. (A) MEG magnetic field distributions of the first DSS components for all (11) subjects, show a stereotypical pattern of neural activity originating separately in the left and right auditory cortices. Purple and green contours represent the magnetic field strength. Red arrows schematically represent the locations of the dipole currents, generating the measured magnetic field. (B) TRFs estimated from MEG data for all (11) subjects. Most TRFs have significant components analogous to the well-known M50 and M100 auditory responses as marked in blue (M50) and red (M100) for each individual subject.

was approximated by truncating the attended TRF beyond a lag of 90ms, on the grounds of the recent findings of Ding and Simon (2012b) which show that the components of the unattended TRF are significantly suppressed beyond the M50 evoked field.

4.3.2 Decoding auditory attention from MEG: a simulation study

In order to simulate MEG data modulated by attention, first a binary sequence $\{n_{k,r}\}_{k=1,r=1}^{240,3}$ was generated as realizations of a Bernoulli process with success prob-

ability $p_k = 0.95$ or 0.05, corresponding to attention to the first or second speakers, respectively. The total observation time was 60s with a sampling rate of $F_s = 200$Hz ($T = 12000$ samples) and the processing window length was chosen to be 250ms ($W = 50$ samples). Using a TRF template of length 0.5s estimated from experimental data (See Section4.3.1), we generated 3 trials for various SNR values and with multiple number of attention switches throughout each trial.

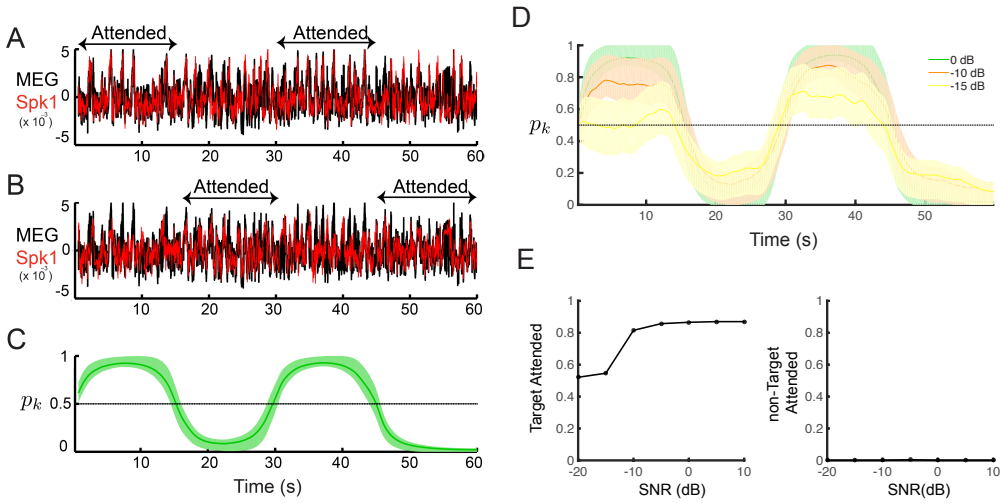

Figure 4.4: MEG data simulation. Simulated MEG data (black traces) and model prediction (red traces) of (A) speaker one and (B) speaker two at SNR = 10 dB. Regions highlighted in yellow indicate the attention of the listener to each of the speakers. (C) Estimated values of $\{p_k\}$ with 95% confidence intervals. (D) Estimated values of $\{p_k\}$ from simulated MEG data vs. SNR = 0, -10 and -15 dB. Error hulls indicate 95% confidence intervals. The MEG units are in pT/m. (E) Behavioral results of the simulated MEG data vs. SNR values ranging from -20 to 10 dB. The time fraction for which the estimated attentional sate follows the target speaker (the opposite speaker) as a function of different SNRs is shown in the left panel (right panel).

Figures 4.4–A and 4.4–B show the simulated MEG signal (black traces) and predictors of attending to speaker one and two (red traces) at an SNR of 10 dB. Regions indicated by arrows in panels A and B demonstrate the time intervals, in

which listeners are supposed to attend to either of the two speakers.

The hyper-parameters for the von Mises-Fisher distribution were chosen as $d = 100KR/2$ and $c_0 = 0.01$, as the resulting correlation between the simulated MEG data and the model prediction is in the range of $\approx$ [0–0.2]. The choice of $d = 100KR/2$ gives more weight to the prior than the empirical estimate of κ_i. The hyper-parameters α and β for the inverse-Gamma prior on the state variance were chosen as $\alpha = 2.01$ and $\beta = 0.5$. This choice of α close to 2 results in a non-informative prior, as the variance of the prior is given by $\beta^2/[(\alpha-1)^2(\alpha-2)] \approx 245$, while the mean is given by $\beta/(\alpha-1) \approx 0.5$.

Estimated values of $\{p_k\}_{k=1}^{240}$ (green trace) and the corresponding confidence intervals (green hull) are shown in Figure 4.4–C. The estimated p_k values reliably track the attentional state, and the transitions are captured with high accuracy. MEG data recorded from the brain is usually contaminated with environmental noise as well as nuisance sources of neural activity, which can considerably decrease the SNR of the measured signal. In order to test the robustness of the decoder with respect to observation noise, we repeated the above simulation with SNR values ranging from -20 to 10 dB. As demonstrated in Figure 4.4–D, the confidence intervals and the estimated transition width widen gracefully as the SNR decreases. The dynamic denoising feature of the proposed state-space model results in a desirable decoding performance for SNR values above -15 dB (Figure 4.4–E).

4.3.3 Decoding auditory attention from MEG: application to real MEG data

We assessed our proposed state-space model and decoder on experimental MEG data recorded from 11 human subjects who listened to one of the two competing speakers in constant-attention and attention-switch experiments (see Methods). All hyper-parameters in the model were chosen similar to those of the simulation studies in the previous section, except for the prior parameter c_0 for the von Mises-Fisher distribution which was conservatively chosen as $c_0 = 0.01$, since the observed correlation values between real MEG data and their predictors are typically in the range $[0 - 0.2]$.

The predicted p_k values resulted from single and multi-trial analysis are shown in Figure 4.5 for three sample subjects. For multi-trial analysis (3^{rd} panel of each plot) 90% confidence intervals are shown by the shaded hulls around the estimated values. In the first and second conditions subjects were instructed to maintain their attention through the entire experiment to the male and female speakers, respectively (Figures 4.5–A and 4.5–B). The decoding results demonstrate the decoder's reliable recovery of the attention modulation by estimating $\{p_k\}$ close to 1 for the first condition and values close to 0 for the second condition. For the third and fourth conditions, subjects were instructed to switch their attention after hearing a 2 s pause, in the middle of each trial, from the male to the female speaker (Figure 4.5–C) and from the female to the male speaker (Figure 4.5–D). Using multiple-trial analysis, the decoder was able to capture the attentional switch occurring roughly

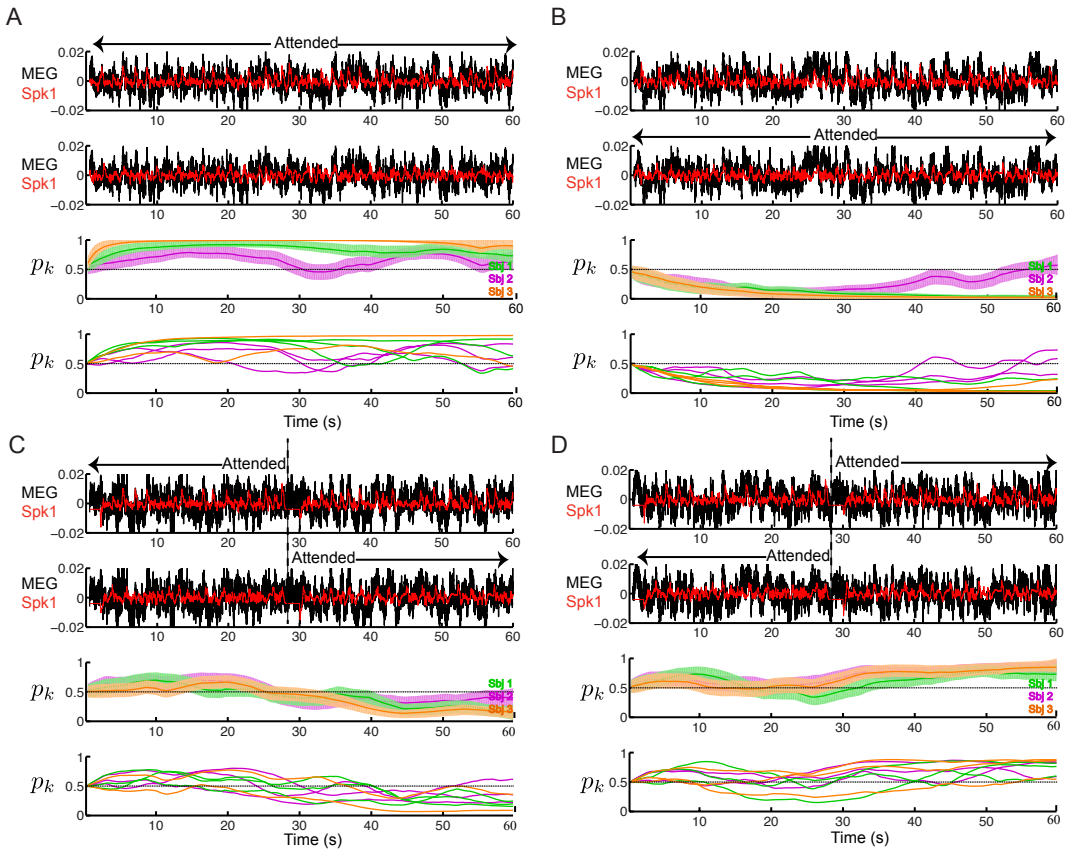

Figure 4.5: Decoding auditory attentional modulation in experimental MEG data. In each subplot, the MEG data (black traces) and the model prediction (red traces) for attending to speaker 1 (male) and speaker 2 (female) are shown in the first and second panels, respectively, for one sample subject. The third panel shows the estimated values of $\{p_k\}$ and the corresponding confidence intervals using multi-trial analysis for three sample subjects. The fourth panel shows the estimated $\{p_k\}$ values for single trials. A) Condition one: attending to the speaker 1 through the entire trial. B) Condition two: attending to the speaker 2 through the entire trial. C) Condition three: attending to the speaker 1 until $t = 28s$ and switching attention to the speaker 2 after the 2 s pause. D) Condition four: attending to the speaker 2 until $t = 28s$ and switching attention to the speaker 1 after the 2 s pause. Dashed lines in subplots C and D indicate the start of the 2s silence cue for attentional switch. Error hulls indicate 90% confidence intervals. The MEG units are in pT/m.

halfway through the trial. The decoding of individual trials in the forth panel of Figure 4.5–C & 4.5–D suggest that the exact switching times were not consistent across different trials, as the attentional switch might have occurred slightly earlier or later than the presented cue.

The performance of individual subjects were evaluated by computing time fractions in which the target speaker or the alternative speaker were followed according to the estimated results from the state-space decoder. All computations were done within the confidence interval of 90% for multi-trial and 70% for single-trial analysis. An illustrative example of the time intervals in which a sample subject is in target, alternative target (Alt-target) or unfollowed attentional sate is shown in Figure 4.6, for a sample trial in male-female attention-switch condition (condition 3). The evaluated target and Alt-target attentional time fractions for single trials are plotted in Figure 4.7–A1 and 4.7–A2, for the constant-attention and the attention-switch experiments, respectively. As shown in these figures, most of the data points fall above the identity line, indicating larger time fractions in which the target speakers were attended vs. the alternative targets. The behavioral results from multi-trial analysis were significantly improved compared to the single-trial estimations (one way ANOVA, $P < 0.01$). This is indeed expected from the the state-space formulation, as the variance of the state variable z_k is inversely proportional to the number of trials R (See Eq. (A.5)). The results of multi-trial estimations are shown in Figure 4.7–B1 & 4.7–B2 for each individual subject and two experimental conditions. The median, 25% and 75% quartile values are shown in separate box plots for target and Alt-target attended time fractions and for each

individual experiment. In addition, individual subject performances averaged over condition pairs within constant-attention experiment (conditions one & two) and attention-switch experiment (conditions three & four) are plotted in blue on top of the corresponding box plots. Evaluated performances for the decoded attentional states show that time fractions in which the target speakers were attended to, were significantly larger than the Alt-target attended time fractions (one way ANOVA, $P < 0.001$), highlighting the successful decoding of the attentional states via the state-space model.

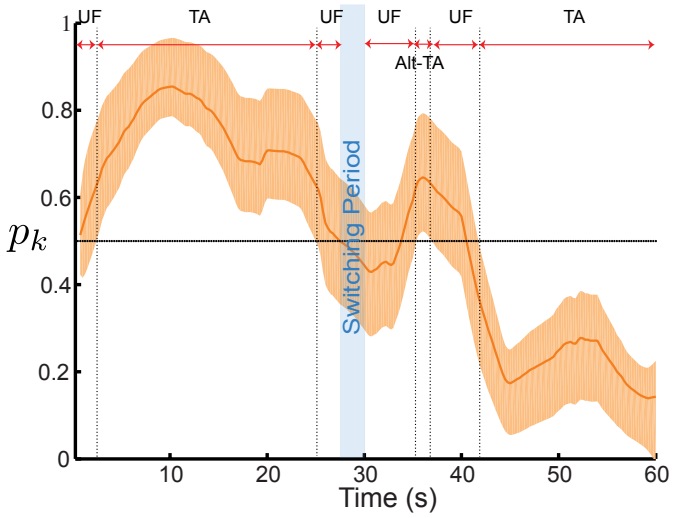

Figure 4.6: Schematic illustration of attentional states for a sample subject, in an example of male-female attention-switch condition experiment. The estimated attentional probability at each time point can be in one of the following states: Target Attended (TA), Alternative Target Attended (Alt-TA), and the Unfollowed state (UF).

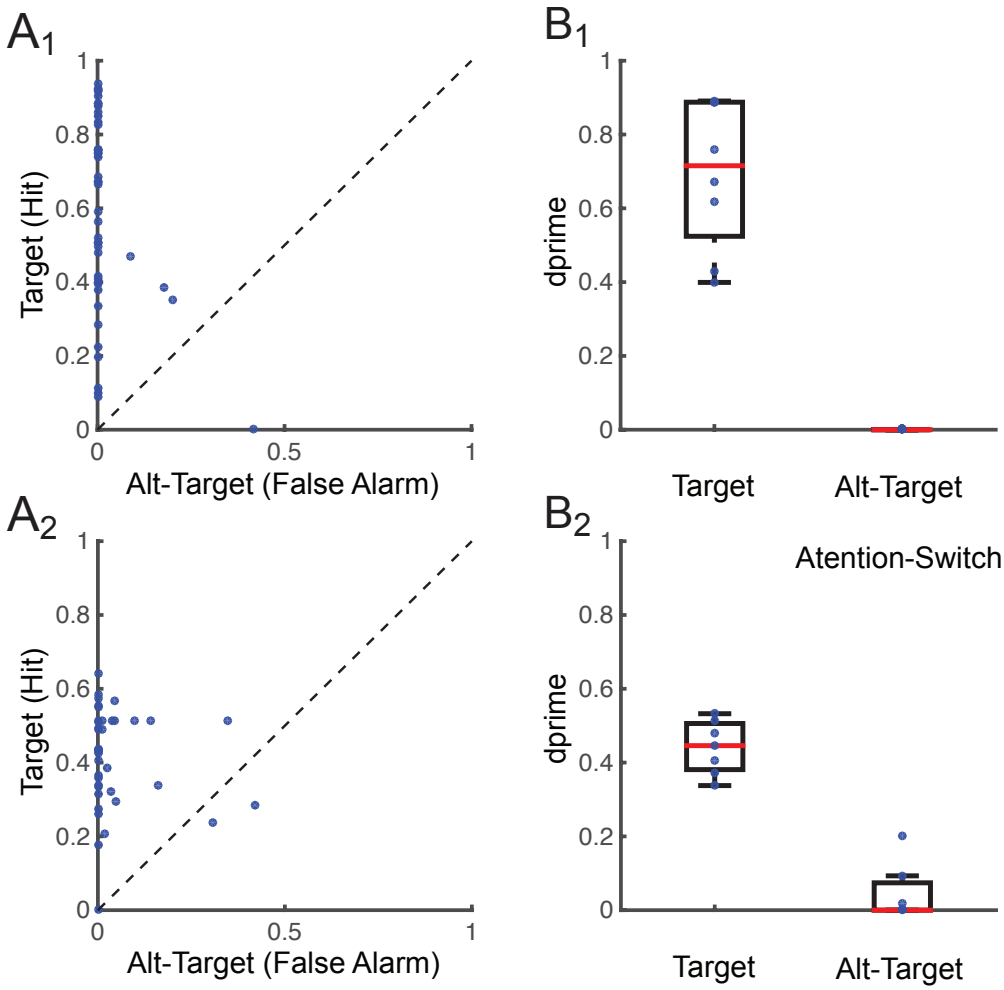

Figure 4.7: Behavioral analysis of experimental MEG data. (A1) Time fractions in which the target speaker attended is plotted with respect to the time fractions in which the Alt-target speaker is attended for individual subjects in constant-attention experiment. (A2) Same analysis results for attention-switch experiment. (B1) Target and Alt-target attended time fractions are computed via multi-trial analysis. The first and second box plots indicate the median and quartile percentages for target attended and non-target attended respectively. (B2) Same analysis for attention switch experiment. Individual subject performances, shown in blue markers, are plotted on top of the box plots.

4.4 Discussion

In this study, we developed a neurobehaviorally inspired state-space model that provides an estimation framework for decoding the attentional state of a listener in

a competing-speaker environment. The proposed algorithm takes advantage of the temporal continuity in the attentional state, resulting in a decoding performance, which is highly accurate and resolved in time. Parameter estimation of this model is carried out using the EM algorithm, which is tied to the efficient computation of the Bernoulli process smoothing, resulting in a very low overall computational complexity. The output of the state-space model at each EM iteration is plotted in Figure 4.8 for a sample subject and all four experimental conditions. These plots illustrate the convergence path of the EM iterations in estimating the attention probability values p_k, starting from values at chance level (0.5) and converging to values near 0 or 1 depending on the targeted speaker.

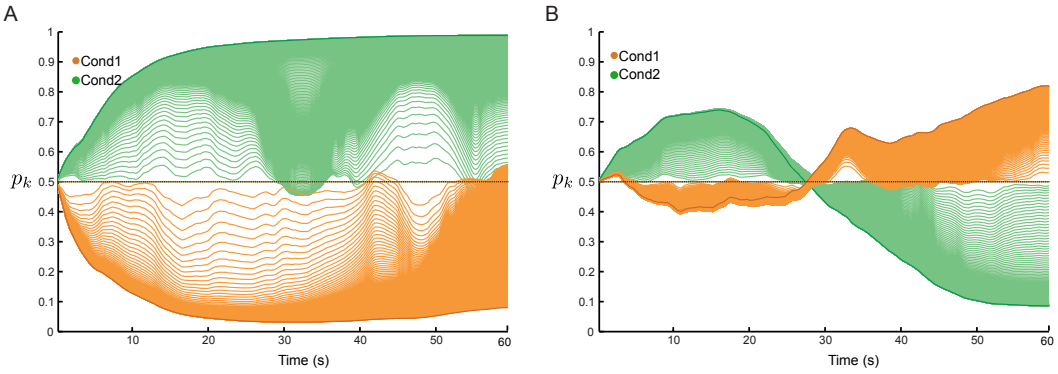

Figure 4.8: A step-wise illustration of the EM convergence. The output of the state-space decoder is plotted after each EM iteration for sample trials of a subject in Constant-Attention (A) and Attention-Switch (B) experiments. Estimated values from final iterations are shown in bold green/orange lines.

The novel state space model proposed in this study is supported by performance evaluation of the model on realistic simulated data, as well as evoked neural activity from the auditory cortex of humans, recorded via MEG. These studies divulge three main advantages in the current model over the state of the art methods

such as the reverse correlation technique (Bialek et al., 1991; Gielen et al., 1988; Hesselmans and Johannesma, 1989).

First, in the proposed model, temporal resolution of the estimated state of attention is in the order of multi-seconds. This resolution is comparable to empirically estimated speed of attention switching in humans; therefore the proposed model provides a dynamic framework for tracking the attentional state of a listener in real world scenarios.This is a considerable improvement over the commonly used methods based on reverse correlation, in which the recovery of the stimulus paradigm from the corresponding neural response results is a poor reconstruction of the stimulus using short processing time windows, and therefore fails in tracking the attentional state in a precise fashion (Ding and Simon, 2012b; Mesgarani and Chang, 2012b).

Second, the proposed algorithm employs only the envelopes of the two speech streams as the stimulus covariates. This is a substantial reduction in the dimension of the spectro-temporal feature set used for decoding compared to those used in previous studies. Reducing the dimensionality of the feature space is not only an extremely important issue in real-world applications, but also can hint to the features of speech which are encoded by the brain as markers of auditory attention, and thereby lead to a better understanding of the neural basis of auditory stream segregation.

Third, the principled statistical framework used in constructing the decoder allows us to obtain confidence bounds on the estimated attentional state. This feature is crucial to obtaining a statistically-principled framework for assessing the

validity of the algorithm output. Moreover, the proposed approach benefits from the inherent model-based dynamic denosing of the underlying state-space model, and is able to reliably decode the attentional state under very low SNR conditions.

A potential application of this analysis framework is to be used as a real-time cocktail party analyzer, tracking the attentional state of a listener in a complex auditory environment. The state space model provides estimation of the probability of attending to either one of the speakers at each time point t based on the recorded MEG data at all other time points before (via non-linear filtering) and after (via backward smoothing) t. Assuming that the cognitive state of attention is a continuous process in time, this continuity is appropriately accounted for in the proposed model; however, for real-time Brain-Computer Interface (BCI) applications, the smoothing step can be omitted and estimation of the attentional state can be causally carried out via the proposed non-linear filter.

Future work includes generalization of the proposed model to more realistic and complex auditory environments with more diverse sources such as mixtures of speech, music and structured background noise. Nevertheless, the promising performance of the proposed algorithm for MEG recordings makes it an appealing candidate for EEG applications.

Chapter 5

Perceptual mechanism of contextual effect on ambiguous stimuli in the auditory cortex

5.1 Introduction

Human perception in real-world scenarios is typically based on incomplete or ambiguous information. In order to arrive at a consistent interpretation of a continuous stimulus, personal knowledge about the world is integrated with the current stimulus to embody perceptual expectations. Such knowledge arises over a wide range of time frames, from lifelong experiences to immediately preceding or even simultaneously occurring stimuli. These experiences form the context within which the current stimulus is interpreted and more generally represent the constructive neural process that generates a coherent and unified representation of the world.

The effect of context on perception can be studied precisely in the laboratory with the use of stimuli that are perceptually ambiguous, i.e., they can have multiple perceptual states, while their physics are unchanged. These are often referred to as multi-stable stimuli. Well-known visual examples include the Necker cube (Necker,

1832) and bistable apparent motion (Ramachandran and Anstis, 1985), whereas some auditory examples include the verbal transformation effect (Warren and Gregory, 1958) and auditory stream segregation (A-B-A paradigm; (Bregman, 1990)). Multi-stability can also take place in combined audio-visual modalities (Pressnitzer and Hupé, 2006; Hupé et al., 2008; Schwartz et al., 2012). In all these examples, one or more percepts emerge over time, while the stimulus itself is kept unchanged. In some cases, this perceptual switching can be induced by preceding biasing stimuli. Examples of this can be found in well-known effects of prior spectral bias on phonemic percepts (Holt, 2006; Holt et al., 2005; Stephens and Holt, 2003), where prior acoustic stimuli composed of sine-wave tones that are drawn from spectral distributions with different mean frequencies robustly affect speech categorization.

Another example of perceptual switching induced by preceding stimuli is the tritone paradox, an auditory illusion first introduced by Deutsch (1980), in which a sequentially played pair of Shepard tones, separated by a semi-octave, is heard as ascending by some people and as descending by others. A Shepard tone, named after Roger Shepard (Shepard, 1964), is a complex sound consisting of a superposition of octave spaced tones. A schematic view of a Shepard tone is shown in Figure 5.1–A (with all the constituent tones enclosed within a dashed green rectangle). Since the Shepard tone complex theoretically covers the entire spectrum, it can be parameterized by a circular property in pitch, such that shifting it up or down by an octave (12 semitones) over the frequency axis results in the exact same Shepard tone (Figure 5.1–B).

When presenting two Shepard tones sequentially, with the second tone shifted

up or down on the frequency axis (e.g., by ± 2 semitones with respect to the first tone), one can clearly perceives a pitch-change in the corresponding direction. In Figure 5.1–A, the spectral transition between the first and second Shepard tones is heard as a clear downward step in pitch and for the second and third tones is heard as a clear upward step in pitch. By contrast, in the case of the tritone paradox, where the pair of Shepard tones are separated by half octaves (± 6 semitones), the spectral transition between the two Shepard tones becomes ambiguous in that it can be equally viewed as an upward or downward step in pitch (Figure 5.1–A, third and forth Shepard tones).

Earlier studies by Chambers and Pressnitzer (2011); Chambers et al. (2012) showed that presentation of a spectrally defined acoustic context prior to presentation of the ambiguous Shepard pair can reliably influence the perceived direction of pitch-change in the pair. The acoustic context, also known as the biasing sequence, consists of a sequence of Shepard tones that are spectrally confined to the lower or upper half-octave spectral regions with respect to the first Shepard tone in the ambiguous pair (Figure 5.2–A, to simplify the diagram, only one octave is shown). The ascending or descending percept is strongly dependent on the spectral region in which the biasing tones are placed. If the biasing sequence is located in the upper or lower half-octave with respect to the first Shepard tone in the ambiguous pair, it is highly probable that an ascending or descending pitch-change is perceived by the listeners. An analogous case of such a priming effect in visual modality has been described by Zhang et al. (2012), in which a brief exposure to a vertical or horizontal motion increased the probability of perceiving the motion in the corresponding

direction. Thus there exist both auditory and visual examples of percepts that are strongly influenced by the preceding sequence of stimuli.

There have been several number of human psychoacoustic experiments and animal neurophysiological studies that investigated the neural correlates of the contextual effect on perception in both the auditory and visual modalities. However, the main limitation of all these studies is that where there is access to perceptual states of humans, there is no corresponding neural information available, and when there is access to the neural information from the animal's brain, it is very challenging to verify the perceptual state. In this chapter, several psychoacoustic experiments are described in which neural recordings from the auditory cortex of humans—using MEG—were used to record both neural and behavioral data simultaneously, and hence to investigate the underlying mechanisms of perceptual manipulation caused by contextual biasing in the tritone paradox example.

In a first set of experiments, the neural correlates of the contextual influence of the biasing sequence on the directional percept of the ambiguous Shepard tone pairs were investigated. Specifically, we examined the auditory cortical responses immediately following the biasing tones so as to assess the responsiveness of the neurons at various frequency-specific channels along the tonotopic axis, where they induced to be in a sensitized state as reported in the vision literature by Wissig and Kohn (2012) or in an adapted (suppressed) state prior to the final percepts, as suggested by Linke et al. (2011) & Englitz et al. (2013).

In the following two experiments, the spectrotemporal properties of the contextual trace (biasing sequence) were further characterized with slightly modified

paradigms to improve SNR and resolution of the neural measurements in the spectral and temporal domains.

Finally, a series of psychoacoustic and neural experiments were conducted to determine whether the biasing mechanism is merely dependent on the bottom-up primary feature extraction procedures or whether it involves high-level top-down mechanisms such as attentional modulation. Specifically, potential links between the biasing effect and steaming of the biasing sequence is explored and discussed in this section.

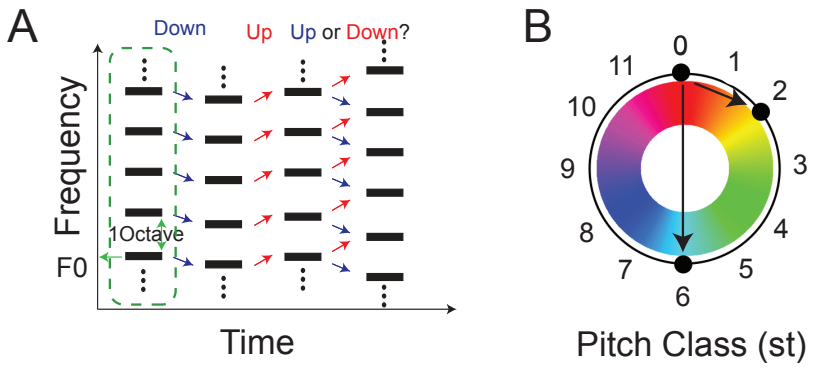

Figure 5.1: Shepard tones and circular property in pitch space. (A) Shepard tones are a stack of octave-spaced tones that cover the entire spectrum from low to high frequencies. From left to right a Shepard tone was translated slightly up or down relative to the previous tone, resulting in a clear perception of up and down shift in pitch, respectively. However in the last pair on the right, the second tone is shifted by 6 semitones, causing an ambiguous percept. (B) Perceived pitch of a Shepard tone has a circular property in the spectral domain, i.e., shifting the Shepard tone by an octave leads to the same Shepard tone. Examples of ambiguous (0 & 6 st) and unambiguous (0 & 2 st) pairs are shown on the circle in black arrows.

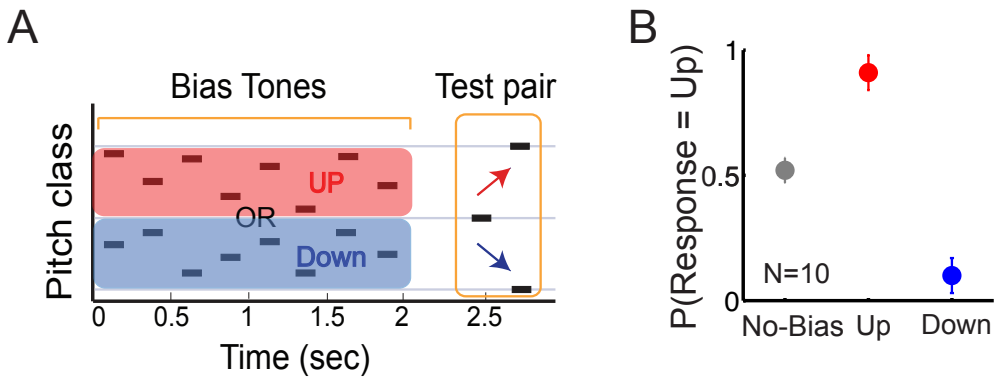

Figure 5.2: A psychoacoustic paradigm to study contextual influences on perception of steps in pitch.(A) Stimulus paradigm of previous psychoacoustic studies; biasing tones are presented in the upper (red) or lower (blue) spectral region with respect to the first tone in the test pair followed by an ambiguous pair. (B) Psychoacoustic results comparing the percept of the ambiguous test pairs without a preceding biasing sequence (gray dot) to the percept when there is a preceding sequence (UP = red, DOWN = blue, $N = 10$, error bars = 1 SEM)

5.2 Neural correlates of the ambiguous Shepard tone pairs percept

To investigate neural correlates of the ambiguous percept of the Shepard tone pairs, MEG recordings of the neural activity in response to the biasing context were analyzed and correlated with the perceptual responses of human listeners. The behavioral data and neural recordings were collected from multiple subjects performing a 2 Alternative Forced Choice (2AFC) task, in which subjects were asked to decide whether a pair of Shepard tones ascended or descended in pitch, after listening to the biasing sequence. In order to estimate the activation state of the auditory cortex from the MEG recordings a rhythmic sequence of Shepard tones, known as the probe sequence, was interspersed between the contextual sequence and the ambiguous pair

(Figure 5.4–A). This otherwise silent period is considered to be the maintenance period during which the memory of the biasing tones persists in the auditory cortex and subsequently interacts with and leads to the percepts of the final test tones.

5.2.1 Experimental design & procedures

Stimuli The acoustic stimuli were digitally generated using MATLAB (MathWorks, Natick) at a sampling rate of 44.1 kHz. Each trial consisted of three parts: 1) a biasing sequence, followed by 2) a probing sequence and 3) an ambiguous test pair at the end. The constituent tones of all these sequences were Shepard tones (Shepard, 1964), which varied in spectral location and duration.

An individual Shepard tone was generated as the sum of octave-spaced pure tones with a flat spectral envelope (Figure 5.2–A, tones within the green dashed rectangle) and randomized phases. A cosine shaped gating function with a half-period of 5 ms were applied to all Shepard tones. The spectral location of a Shepard tone is given by the position of each of its constituent tones within an octave relative to a base-frequency. The assigned spectral location of a Shepard tone, also known as pitch class, is conventionally provided in semitones, in the range 0–11 st. Here, a base-frequency of 440 Hz is assigned as pitch-class 0. The number of constituent tones in a Shepard tone varied to cover the full range of hearing by the subject. If the Shepard tone is spectrally shifted by a full octave, it maps to the same stimulus, which is the circular property of the space of Shepard tones mentioned earlier (Figure 5.1–B).

The biasing sequence was composed of a series of 8 Shepard tones with a tone duration of 125 ms and inter-stimulus intervals of 125 ms, and hence a total duration of 2 s. Spectral locations of the biasing tones were randomly selected from the upper or lower semi-octave relative to the spectral location of the first tone in the ambiguous pair (Figure 5.2–A). These spectral locations lead to strong upward or downward biasing effects, respectively (as seen in Figure 5.2–B).

The probe sequence was also composed of Shepard tones with a constant pitch-class and shorter tone lengths compared to the tones in the bias sequence (Figure 5.4–A). The characteristic properties of the probe sequence, such as tone length, sequence duration and presentation rate were determined in a separate psychoacoustic experiment as described later.

Finally, the test pair consisted of a sequence of two Shepard tones that were spectrally separated by half an octave. Such pairs of Shepard tones are ambiguous with respect to the perceived direction of change in pitch. The test pair was presented 315 ms after the end of the probe sequence and with an inter-stimulus gap of 125 ms.

Subjects Ten participants (four males; mean age 26 y, range 24–34 y) took part in an initial series of psychoacoustic experiments to determine the parameters of the probing sequence. Seventeen normal-hearing right-handed adults participated in the MEG studies. One MEG participant was excluded from further analysis due to an excess of electrical artifacts. Nine subjects (four males; mean age 24 y, range 18–29 y) participated in the first MEG experiment. Seven participants (three males; mean

age 26 y, range 18–32 y) participated in the second and third MEG experiments.

5.2.2 Psychoacoustic studies

In the psychoacoustic experiments, participants performed the tasks at a computer in a sound-attenuated room (J.W. Manny Inc., Eckel Sound Rooms). They were asked to control the computer using a Graphical User Interface (GUI) and were allowed to adjust the volume to a comfortable level before starting the experiment. A complete explanation of the required task, as well as the basic instructions for using the GUI, were given in advance. For each presented sequence, subjects were asked to judge the direction of pitch-changes for the last two stimuli, i.e., the ambiguous test pair. They were instructed to press a designated button for each choice. By design, no correct answer existed for an ambiguous pair and, and hence no feedback was provided.

In an initial psychoacoustic study, a simplified paradigm consisting of a biasing sequence followed by an ambiguous test pair was used to provide a baseline for behavioral performance of the listeners (Figure 5.2–B, N=10). As a result, $92 \pm 8\%$ of the UP-biases and $88 \pm 10\%$ of DOWN-biases led to ascending and descending responses, respectively. In the absence of the bias sequence, the probability of UP responses was at chance level ($52 \pm 6\%$) averaged over all subjects (Figure 5.2–B, the gray marker).

In a second psychoacoustic study, the presentation rates and durations of the probe tone sequences were determined. The probe sequence consisted of tones sim-

ilar in structure to the bias sequence, except for having a fixed pitch-class and a much shorter duration. To ensure that the probe sequence did not itself induce a perceptual bias, psychoacoustic tests were conducted to determine the optimal parameters (tone length, presentation rate, and overall sequence length) that minimized its biasing potential. Subjects were tested in 9 experimental conditions: 3 different presentation rates (4, 7 and 10 Hz) and 3 different tone lengths (35, 65, 95 ms) for the probe sequence. Each condition contained 240 trials (12 pitch class $\times$ 2 biasing direction $\times$ 10 repetitions). The results are depicted in Figure 5.3–A. It is found that the presentation rate did not have a significant influence on the biasing strength (one-way ANOVA, $p > .05$ for all pair comparisons within each column in Figure 5.3–A). Therefore, the lowest tested rate (4 Hz) was chosen to maximize the SNR of the neural recordings. This arises from a so-called $1/f$ power law relationship, found in EEG/MEG activities in the following form:

$$S_x(f) = \frac{\text{Constant}}{|f|^\gamma} \tag{5.1}$$

Here $S_x(f)$ is the power spectral density, f is the frequency and γ is some spectral parameter which is usually close to 1 but can lie in the range $0 < \gamma < 2$, and can be greater than 2 in the presence of the noise sources (Keshner, 1982).

Tone duration was found to be critical, with shorter tones producing a negligible biasing effect (down from $92 \pm 5\%$ for 75 ms to $50 \pm 7\%$ at 35 ms). The effect of the original bias on the ambiguous pair was also evaluated as a function of different probe sequence lengths, in a range of 0.5–4 s (Figure 5.3–B). Since longer durations

of probe sequence showed a decrease in the original biasing effect, the length of the sequence was limited to 2 s, to minimize any incidental biasing by the probes.

Overall, a direct measurement of the correlation between the spectral location of the probe sequences and the test tone percepts was shown to be negligible for the selected parameters above (probe sequence of 4 Hz presentation rate, 35 ms tone length, and 2 s duration, Figure 5.3–C)), compared to the effect of the bias sequence. Therefore, these parameters were used for all subsequent MEG experiments.

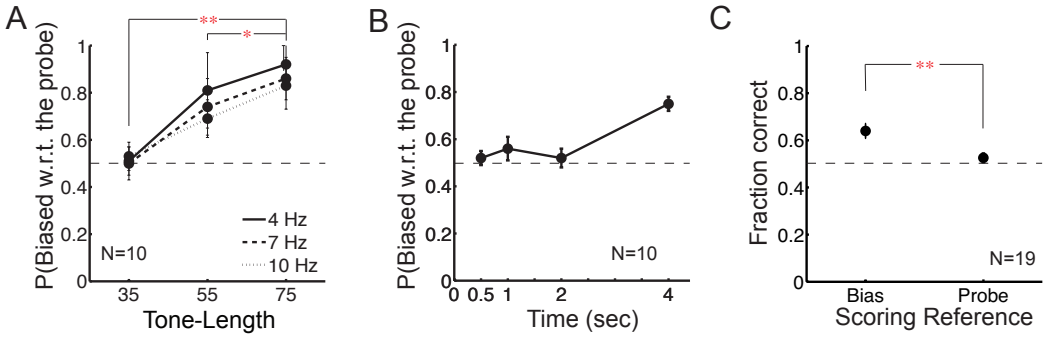

Figure 5.3: A spectrally local MEG probe sequence with short tone durations has little influence on the directional percept. (A) The biasing effect of the probe sequence is reduced with a shorter tone length, and is not different from chance at a tone length of 35 ms. Repetition rate had only a little effect on the percept, and we thus chose 4 Hz for its maximal response size in MEG recordings (N=10). (B) Behavioral performance with respect to the probing sequence duration. Probe sequences with a length of up to 2 s (corresponding to 8 tones) preserved the bias effect of the bias sequence. (C) Subjects' percepts were not influenced by the probe sequence, as shown here by scoring the behavioral performances either based on the spectral locations of the bias (strong separation) or the probe (chance performance) sequence. These results are derived from the recording in the MEG machine (N=19).

5.2.3 MEG studies

In MEG experiments, participants listened to sequences of Shepard tones consisting of bias, probe, and test pairs as described above. At the end of each trial, subjects were required to indicate whether they heard an ascending or descending change in pitch. A short training block of 20–30 trials was conducted at the start of each MEG recording session, which typically lasted about 1 hour.

The most influential pairs of Shepard tones were determined for each subject prior to the MEG studies. Unlike in Repp (1997), we deployed a flat spectral envelope on the amplitudes of the constituent tones of a Shepard tone, and hence no specific frequency region controlled the level of ambiguity. Subjects listened to a block of 120 trials (12 pitch classes $\times$ 10 repetitions), each trial consisting of a bias sequence followed by an ambiguous test pair positioned at different pitch classes (0–11). The three most perceptually ambiguous pitch classes, and hence maximally influenced by the preceding bias tones were then selected and used in the subsequent MEG experiments.

In the first MEG study, 3 identical blocks of 120 trials consisting of four conditions (2 biasing locations $\times$ 2 probe locations) and 30 trials in each condition (3 pitch classes $\times$ 10 repetitions) were presented in random order within each block.

Earlier psychoacoustic experiments by Chambers and Pressnitzer (2011) suggested that the biasing effect on the perceived pitch direction lasted for a comparatively long period of time (up to multiple seconds). A second MEG experiment was designed to verify whether the suppression of the neural response in specific channels

following the bias was also maintained for the whole period after presentation of the bias sequence and before presentation of the test pair. Since presentation of the tones in the probe sequence itself caused adaptation in the neural response to the probe sequence, our measurements in the previous experiment became unreliable after they hit the noise level (∼1 s). In this experiment, low SNR measurements were avoided by introducing a delay in the presentation onset of the probe sequence with respect to the last tone in the bias sequence. The delays varied between 1.5 and 3 s, and the amount of suppression was estimated from the responses to the early tones in the sequence as depicted in Figure 5.6–A. Three identical blocks of 96 trials consisting of 8 conditions (2 pause durations × 2 bias locations × 2 probe locations) with 12 trials in each condition (3 pitch classes × 4 exemplars) were presented to the subjects in a randomized order within each block.

In the first MEG study the activation state of the cortex was only measured at two frequency locations: (1) at the center of the biasing sequence, and (2) in the opposite location, a half-octave away (Figures 5.4–A & 5.6–A). To refine the spectral estimate of the suppression pattern, the activation state at 6 additional locations were measured in a third MEG study, covering the full octave with a resolution of 2 st (Figure 5.7–A), providing a total of 6 measurements for each octave, i.e., at the locations 2, 4, 6, 8, 10 and 12 st, relative to the first tone in the test pair (Figure 5.7–A). Three identical blocks of 108 trials, consisting of 12 experimental conditions (2 biasing locations × 6 probing locations) and 9 trials in each condition (3 pitch classes × 3 exemplars) were presented to the participants with randomized trial orders within each block.

5.3 Data analysis

To analyze recordings from all MEG experiments, the 2 s temporal responses to the probe sequences were extracted and divided into 4 time windows of 500 ms each. Neural responses of all trials corresponding to each time window were then concatenated to obtain an extended response with duration $T =[$ 500 ms $\times$ number of trials $]$ within a conditional number of blocks, for each time window and for each channel. Concatenated responses were represented in the frequency domain using a Discrete Fourier Transform (DFT), yielding a frequency spectrum from 0 to 500 Hz at a resolution of $1/T$ Hz. The complex magnetic field strength was obtained by using the product of the DFT and the sampling interval (1/SR). Power spectral densities were computed by squaring the complex magnetic field strength, normalized by the duration T of the signal. We then calculated the square magnitude of the frequency component at 4 Hz, averaged over the 50 channels with the strongest normalized responses for each participant.

Neural recordings of all trials were first separated into two groups: (1) *Congruent* trials, for which the percept of pitch-change is located on the same side as the Bias, and (2) *Incongruent* trials, for which the percept is located on the opposite side of the Bias. Within each group, trials were further separated into *Biased* (Probe sequence in the same spectral region as the Bias tones) and *Unbiased* (Probe sequence in the opposite region as the Bias tones) trials. Due to the modulation rate of 4 Hz of the probe sequence, only 4 Hz response power was analyzed and then averaged within these four groups (*Congruent-Biased, Congruent-Unbiased,*

Incongruent-Biased, Incongruent-Unbiased).

In order to have a fair cross-conditional comparison of the results, an equal number of trials for each condition were analyzed. This addressed the dependence of the SNR on the number of trials, and the variation of the number of trials for each *Congruent* and *Incongruent* group per participant (based on individual performance). Trial equalization was done by performing the analysis on the minimum number of trials available for conditions, and then bootstrap resampling was applied for those conditions with a higher number of trials and averaged over the results of bootstrapping to obtain the final neural response power.

Within-subject normalization was required to make the final results comparable across subjects and independent of individual differences in MEG signal size. Neural responses were first normalized by the mean value of all conditions for a given subject, and then averaged over all subjects to obtain the population means and standard errors shown in Figures 5.5, 5.6 and 5.7.

Noise level at the 4 Hz presentation rate for each MEG experiment was computed by extracting and concatenating the inter-trial pauses over the whole experimental session. The extended response was then analyzed in the same way as the neural response to the probing sequence. In order to account for the variability of noise level in different experiments, noise levels were computed independently for all MEG experiments and scaling factors of the noise levels for later experiments with respect to the first experiment were computed. All neural responses for the later experiment were then normalized with the corresponding scaling factors.

The strength of the neural suppression of the probe sequence at 4 Hz presen-

tation rate was correlated with the percentage correct behavioral results obtained from subjects on their perceived direction of pitch-change in the ambiguous pair. In the *Congruent* group, the correlation between the neural strength of the *Biased* sequences for the first analyzed time window and the behavioral scores (calculated with respect to the biasing direction) were computed. We used Pearson correlation coefficients in this analysis, e.g., for the *Congruent* group we evaluated:

$$CC[N,B] = \frac{\sum_{i=1}^{K}[N(i) \times B(i)]}{\sqrt{\sum_{i=1}^{K} N(i)^2}\sqrt{\sum_{i=1}^{K} B(i)^2}} \tag{5.2}$$

Here, K is the number of subjects, N is the normalized neural response to the probe sequence for the first time window, and B is the behavioral performance. Bootstrap procedures were then performed in order to confirm the negative correlation between the neurometric and psychometric functions in each of the two groups. A balanced bootstrap sampling procedure (Efron and Tibshirani, 1994) was done by randomly selecting 9 participants with replacement and computing the corresponding correlation coefficient. This process was repeated 1,000 times. The procedure was controlled to ensure that all participants appeared the same number of times over all 1,000 bootstrap samplings. Confidence measures were then derived from the bootstrap statistics.

To localize the source regions in the brain underlying the magnetic responses, subjects from all MEG experiments were pooled together. The localization procedure was done with respect to the source of the M100 response (a prominent peak

$\sim$ 100 ms after the onset of a pure tone (Näätänen and Picton, 1987)), for which all subjects participated in a preliminary experiment listening to 100 repetitions of a 1 kHz pure tone (see methods). In response to these short tones, a transient auditory evoked response was observed in the MEG recordings of every participant. The source of the M100 was then localized using an equivalent-current dipole model best fitting the magnetic field configuration at the M100 peak, in each hemisphere. The MEG laboratory software program v.2.001M (Yokogawa Electric, Eagle Technology, Kanazawa Institute of Technology) was used for this procedure. A similar localization procedure with complex equivalent-current dipole model was used to estimate the source for probing sequence best fitting the complex magnetic field configuration at 4 Hz peak, in each hemisphere (Simon and Wang, 2005). Only channels with SNR > 4 were used in the fitting. Also, since MEG is not sensitive to subcortical neural sources, only cortical sources were considered. Goodness of fit was computed as a function of the complex equivalent-current dipole and was given by one minus the residual variance ratio (Simon and Wang, 2005). Significance of the relative displacement between the estimated dipole locations for the 4 Hz and M100 neural responses were determined by two-tailed paired t-tests in each of the three dimensions: lateral/medial, anterior/posterior, and superior/inferior.

5.4 Results

To clarify the relationship between the response strength to the probe sequence and the corresponding perceptual pitch-change of the ambiguous test pair, neural

responses were separated into the *Congruent* and *Incongruent* groups. All the analyses were performed in parallel on these two groups and compared across other experimental conditions.

5.4.1 Local suppression and the percept

To explore how the effects of the biasing sequence persisted over the maintenance period, the sensitivity of auditory cortical frequency-specific (tuned) regions following the presentation of biasing sequence were explored. Specifically, we probed response strength at the spectral centers of the bias sequences, i.e., $\pm$ 3 st relative to the first tone in the test pair, and starting 375 ms after the last tone in the biasing sequence as depicted in Figure 5.4–A. Responses in these two channels were expected to be maximally different since they corresponded to the *Biased* and *Unbiased* conditions, i.e., preceded or not preceded by the bias tones, respectively.

Considering all the trials together, there was no significant difference in the neural responses from *Biased* and *Unbiased* conditions (Figure 5.4–B); however, when we grouped the trials based on each subjects performance, neural activity in the *Biased* and *Unbiased* regions showed significant difference, especially in the first two time windows of the probe sequences. This difference demonstrated opposite patterns for the *Congruent* and *Incongruent* trials. In the *Congruent* trials, neural responses during the *Biased* condition were significantly suppressed compared to the *Unbiased* condition during the first and second 500 ms analysis windows (Figure 5.5–A; $F_{1,16} = 5.8$, $p < 0.05$ and $F_{1,16} = 5.23$, $p < 0.05$ for the first and

second time-windows respectively). This difference was not tractable after the first two time-windows ($\sim$ 1 s), because the probe tone responses themselves adapted rapidly, decreasing to low SNR's close to the noise level. This limitation on the ability to monitor the persistence of the suppression over longer time intervals was circumvented in a subsequent experiment as described below. The pattern was exactly the opposite for the *Incongruent* trials—in which the subjects perceived the pitch direction incorrectly or opposite to the location of the bias sequence. Here, the normalized neural responses to the probe sequence in the *Biased* versus *Unbiased* conditions were also differentiated during the first and second time windows, but with a significantly higher suppression for neural activity in the *Unbiased* condition (Figure 5.5–B) ($F_{1,16} = 4.54$, $p < 0.05$ and $F_{1,16} = 11.49$, $p < 0.005$ for the first and second time-windows, respectively).

To demonstrate the relationship between subjects behavioral performances (behavioral scores with respect to the biasing direction) and the relative amount of *Biased* suppression, the correlation coefficients between the two measures were computed in the *Congruent* trials. The outcome was a significantly negative correlation of -0.65 ± 0.19 ($p < 0.05$, N = 9, indicated confidence bounds were obtained by bootstrapping across participants). This suggests that higher performance is closely tied to stronger suppression of the *Biased* regions during the *Congruent* trials (Figure 5.5–C).

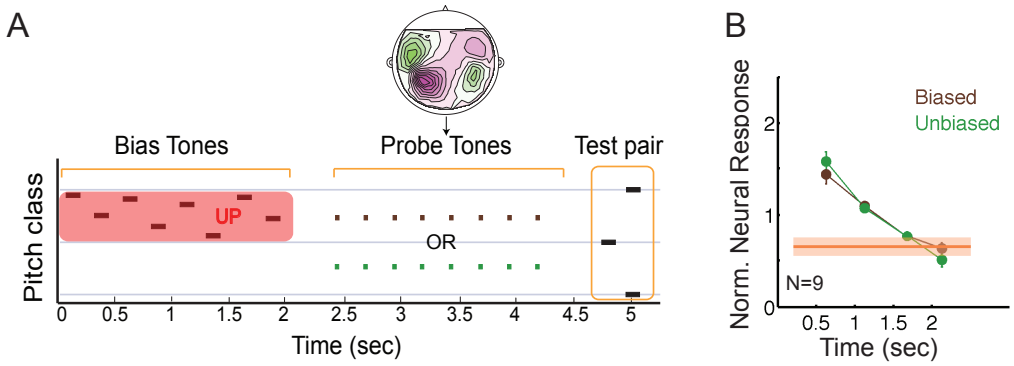

Figure 5.4: Schematic view of the stimulus paradigm. (A) The activation state for different spectral locations was measured by placing a probe sequence in one of the two spectral locations: in the center of the probe sequence or the opposite location half an octave away. The power of the response at 4 Hz (as a function of time) was used as an indicator of the recruitable response at the respective spectral location. (B) No significant difference in the neural responses of the *Biased* and *Unbiased* conditions was observed, averaged over all trials.

5.4.2 Neural suppression persistence over multiple seconds

Suppression in the neural response to the probe was found to last over multiple seconds, consistent with the earlier results on the time scale of the biasing effect (Chambers and Pressnitzer, 2011; Chambers et al., 2012). Specifically, time constants for recovery from suppression were estimated by fitting an exponential to the normalized neural responses, computed over the first 500 ms time window of the probing sequence, at 0.375 s (obtained from the first experiment), 1.5 s and 3 s delays after presentation of the bias sequence. The results shown in Figure 5.6–B & 5.6–C illustrated that suppression in the *Biased* region was deep and lasted over the time course of multiple seconds. Furthermore, the relative response suppression in the *Biased* and *Unbiased* conditions for the *Congruent* versus *Incongruent* groups was consistent with the earlier results described in Figure 5.4–B. The suppression

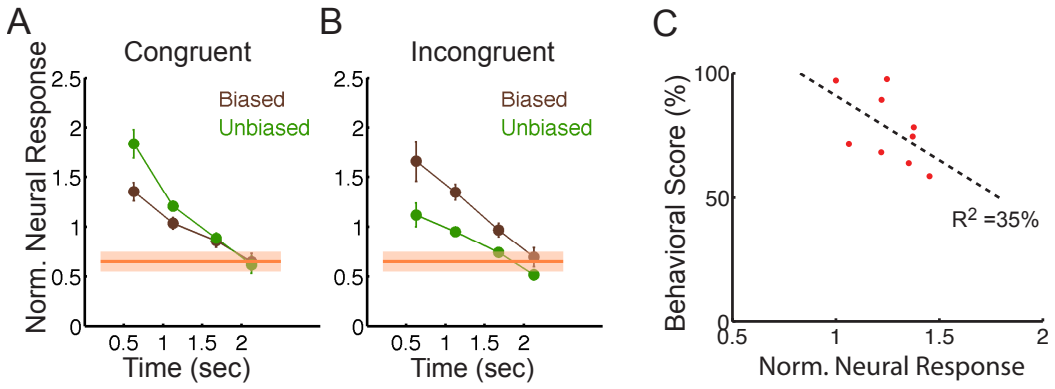

Figure 5.5: Reduction in the auditory cortex response predicts the directional percept. (A) In the *Congruent* trials, activation of the auditory cortex is found to be more reduced at the location of the bias than at the opposite location, measured as the normalized neural power to the 4 Hz probe sequence presented after the biasing sequence (N=9). This suggests a suppressive trace after the bias, rather than a hypothetical positive priming effect. *Biased* and *Unbiased* conditions are color coded according to the probe locations (brown and green for *Biased* and *Unbiased* conditions, respectively) consistent with colors used in the stimulus cartoon for probe locations. Averaged noise level is shown as a flat line within each plot. Error bars and error hulls represent 1 SEM. (B) In the *Incongruent* trials, activation of the auditory cortex was found to be opposite to the *Congruent* trials, i.e., with weaker suppression on the side of the bias sequence. Hence, stronger reduction was found in agreement with the subsequent percept, suggesting a causal relation with the percept. (C) Behavioral responses are plotted as a function of normalized neural response for the first 500 ms window of the *Biased* condition in the *Congruent* trials. A linear fit to the data points from 9 subjects reveals a negative slope, suggesting stronger adaptation in the *Biased* neural responses for better performance scores ($R^2 = 35\%$).

weakened exponentially over the time course of the maintenance period with a time constant of 10.83 ± 4.12 and 4.71 ± 0.4 s (bootstrapping across participants, $p < 0.01$) in the *Congruent* and *Incongruent* trials, respectively. Correlated with the reduction in the strength of the perceived contextual influence over time (0.375 s: $67 \pm 4.2\%$; 1.5 s:$60 \pm 4.8\%$; 3 s: $55 \pm 5.1\%$; chance is at 50%), the difference in suppression strength between the *Biased* and *Unbiased* neural responses is also reduced (See

Figure 5.6–B & 5.6–C).

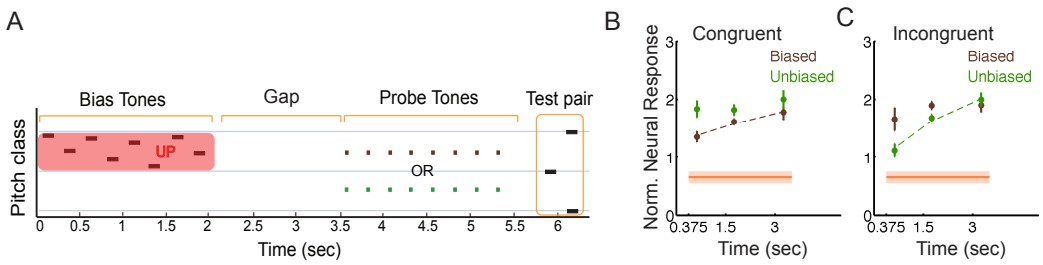

Figure 5.6: The suppressive trace persists on the order of seconds after the bias sequence. (A) We measured the strength of the adaptive trace at different times after the bias sequence by introducing a pause of variable length between the bias and the probe sequence. (B) In the *Congruent* trials the average neural response (N=7) remained more reduced on the *Biased* side than the *Unbiased* side, over the measured period up to 3 s. *Biased* and *Unbiased* conditions are color coded according to the probing locations (brown for *Biased* and green for *Unbiased* conditions) consistent with colors used in (A) for probe locations. (C) In the *Incongruent* trials, only the *Unbiased* responses show a significant reduction, consistent with the hypothesis of a causal influence of suppression on the subsequent percept. Average noise level is shown as a flat line within each plot. Error bars and error hulls represent 1 SEM.

5.4.3 Frequency profile of net suppression

In this experiment, the normalized neural responses to the first 500 ms time window of each probing sequence were extracted and analyzed. To emphasize the spectral shape differences between the *Congruent* and *Incongruent* conditions, the neural responses were subtracted in the *Incongruent* condition from the *Congruent* condition at each frequency location and for each subject. The spectral shape of the neural response differences averaged over 6 subjects was found to be significantly different as a function of spectral location (Figure 5.7–B; $F_{1,54} = 3.29, p < 10^{-5}$), with its trough near the middle of the biasing frequency region (at +3 st). Further

implication of these experiments and the potential neural mechanisms underlying the biasing effect are provided in the Discussion.

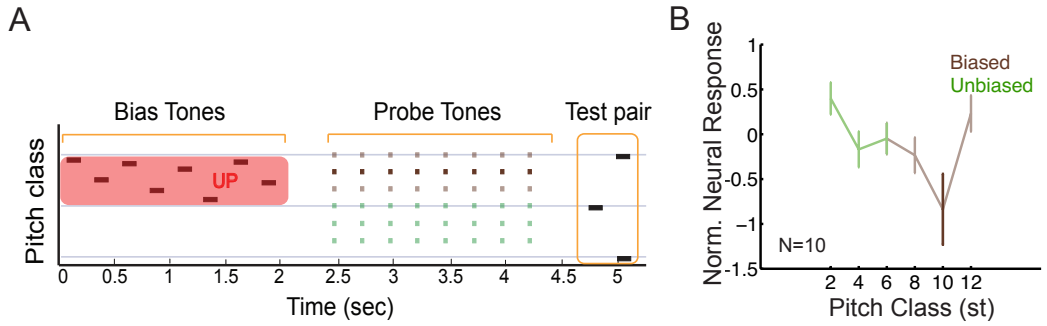

Figure 5.7: The suppressive trace left by the bias varies simply with distance to the center of the contextual sequence. (A) To better resolve the shape of the adaptive trace proceeding the biasing sequence, we probed frequency regions in a finer resolution. Probed locations were chosen at 2 semitone intervals starting at 2 st relative to the first tone in the test pair. In a given trial, only one location was probed (B) Normalized neural response difference of the suppressive trace between the *Congruent* and *Incongruent* trials revealed significant change ($p < 10^{-5}$) as a function of pitch class with a through roughly in the middle of the biased region. Results are plotted in brown (*Biased* condition) and green (*Unbiased* condition) colors and averaged over 6 subjects. Error bars show 1 SEM.

5.4.4 Magnetic field distribution & source localization

Magnetic field distribution of the probe sequence response component revealed the stereotypical pattern for neural activity originating separately in the left and right auditory cortex. A phasor map of the neural response spatial pattern is shown in Figure 5.4–A(inset), as a graph of the complex (magnitude and phase) magnetic field on all channels. Red and green contours represent the magnetic field strength projected onto the lines with constant phase that maximized the projected fields variance (Simon and Wang, 2005). Goodness of fit for these sources was

0.7 ± 0.12 (artificially reduced in accordance with (Simon and Wang, 2005)). The topography of the special magnetic field of all probing rhythm response components with sufficiently high signal-to-noise ratio was similar to those of the M100 response. The mean displacement of the neural source from the source of the auditory M100 response (Näätänen and Picton, 1987) was then calculated for each hemisphere. The displacement was significantly different in the anterior direction for both right (9.4 ± 3.8) and left hemispheres (10.1 ± 2.3), using a two-tailed t-test ($t = 4.2, p = 0.002$ in the right and t = 3.4, $p < 0.01$ in the left hemispheres), but no statistically significant displacement was observed in other directions. Assuming a M100 origin of planum temporale, this is consistent with an origin for the neural response to the probe rhythm in the Heschls gyrus, the site of the core auditory cortex, a region known for its good phase-locking to most naturally occurring rates (< 40 Hz) (Liégeois-Chauvel et al., 2004; Miller et al., 2002).

5.5 Biasing effect: a case of streaming?

Earlier experiments have revealed that listening to a biasing sequence proceding an ambiguous test pair can induce an ascending or descending percept, depending on the spectral properties of the bias sequence. It is nevertheless unclear whether the underlying mechanism of the biasing effect is dependent purely on bottom-up feature-extraction processes or if it involves higher-level top-down mechanisms such as attentional modulation. The experiments described here aimed to investigate the role of attention in contextual biasing, using a slightly modified task-dependent

paradigm. Specifically, they explored whether hearing out the biasing tones as an auditory stream has a significant impact on perceptual manipulation of the ambiguous test pair. That is, whether or not one needs to hear the biasing tones as a "stream", in order to get effective biasing. Discovery of a potential connection between the two well-explored mechanisms in auditory scene analysis—auditory streaming and auditory contextual biasing—enables us to explore each mechanism in the light of the other, as an objective tool.

5.5.1 MEG & psychoacoustics (I)

As discussed in chapter 3, one of the important cues facilitating auditory streaming is the onset/offset synchrony of the constituent components of an auditory sequence (Rasch, 1978; Bregman, 1990). The synchronization feature and its influence on the biasing sequences was the key focus of the design of the following experiments. Bias sequences, composed of synchronous or asynchronous Shepard tones and preceded an ambiguous test pair. They were first played to the subjects in separate blocks to evaluate the effectiveness of each sequence in manipulating the percept. Then, the two sequences were presented simultaneously in each trial to compete with each other in biasing listeners' percept to opposite directions (up and down) at the same time. According to the earlier studies on the effect of temporal synchrony on the streaming ability of listeners (Rasch, 1978; Bregman, 1990), it is hypothesized that if steamability of the biasing sequence is important in creating the biasing experience, then synchronous sequence, compared to the asynchronous one, should be easier to

stream and are therefore more effective in manipulating the percept of pitch-change direction. Although obtaining such a result is not sufficient to establish a causal relationship between the streaming of the biasing sequence and experiencing the biasing effect, it can reveal some potential links between the two auditory experiences.

Experimental design

Stimuli & procedure In the psychoacoustic experiments, the stimulus was a 3 s long sequence, consisting of 8 Shepard tones with 125 ms tone length, and 125 ms inter-tone interval. The ambiguous test pair was presented 156 ms after the last biasing tone with tone duration and inter-tone interval similar to those of the biasing sequence. For the synchronous biasing tones all the components in each Shepard tone were synchronized and had the same onset (as in the previous experiments), whereas in the asynchronous case the onsets of the constituent components in each Shepard tone were randomly jittered in the range of $\pm$ 8 ms (Figure 5.8–A).

The experiment consisted of a presentation of 3 blocks with 80 trials each (4 pitch classes $\times$ 2 biasing direction $\times$ 10 exemplars). In the first block, trials contained only the synchronous biasing tones and in the second block, trials contained only the asynchronous biasing tones (Figure 5.8–A, first and second panels). In the third block, synchronous and asynchronous biasing tones were both present at the same time and competed with each other in biasing the perceptual state of the listeners by having their tones in spectrally opposite regions (Figure 5.8–A, third panel).

In the MEG part, only the paradigm with competing synchronous and asyn-

chronous tones was presented simultaneously was presented to the subjects. The degree of asynchrony in the sequence was varied in 3 experimental conditions (±0–30%, ±30–60% and ±60–100% jittering of the onsets of the constituent components in a Shepard tone). To increase the chance of seeing neural response differences via MEG, the degree of asynchrony in the sequence was made considerably larger compared to those in the psychoacoustic experiment. Biasing tones were then followed by a probing sequence and an ambiguous test pair, as shown schematically in Figure 5.9–A. Three identical blocks of 108 trials (2 biasing directions × 2 probe locations × 3 asynchrony degrees × 3 pitch classes × 3 exemplars) were presented to the subjects, while being inside the MEG machine.

In both experiments, subjects listened passively to the biasing sequence with no task that manipulated their attention toward one or the other biasing sequence.

Subjects Eleven normal-hearing right-handed adults with a mean age of 27 (range 18–35 y; five males) participated in this study. One subject was excluded from additional analysis because of inability to perform the task. Seven subjects (range 18–35 y; three males) participated in the MEG experiment.

Results

Psychoacoustic results for synchronous, asynchronous and mixed bias sequences are shown in Figure 5.8–B. Subjects were clearly biased in the separate synchronous and asynchronous conditions with relatively high biasing scores, computed with respect to the biasing sequences' spectral locations (see Figure 5.8–B, bar plots for

synch/asynch conditions). The difference between the behavioral scores for synchronous and asynchronous conditions was quite small but significantly positive ($p < 0.05$, $mean = 1\%$), indicating that synchronous tones are slightly more effective than asynchronous tones when presented separately in each trial.

The results for the mixed condition is shown in Figure 5.8–B (third panel). The behavioral scores computed based on the synchronous and asynchronous bias tones' spectral locations are shown in gray and yellow bars, respectively. In most trials listeners were biased with respect to the synchronous sequence (one-way ANOVA, $p < 0.001$), showing that the synchronous biasing sequence is more powerful in manipulating the subject's percept in comparison with the simultaneously presented asynchronous sequence. Again, although this result is not sufficient to deduce a causal relationship between streaming of the biasing sequence and contextual biasing effect on the test pair, it motivates us to conduct further experiments to explore the potential links between the two auditory tasks.

In the MEG study, the behavioral performance of the subjects is shown in Figure 5.9–B. For different degrees of asynchrony, there was a slight but significant improvement in the behavioral scores with respect to the synchronous sequence as the asynchronous sequence became more and more jittered (one-way ANOVA, $p < 0.01$). The neural responses to the probe sequences following the synchronous and asynchronous biasing tones were analyzed similarly to the procedure discussed earlier in the MEG experiments in this chapter. Trials in each condition were separated into the *Congruent* (perceptually biased w.r.t. the synchronous biasing sequence) and *Incongruent* (perceptually biased w.r.t. the asynchronous biasing sequence) groups.

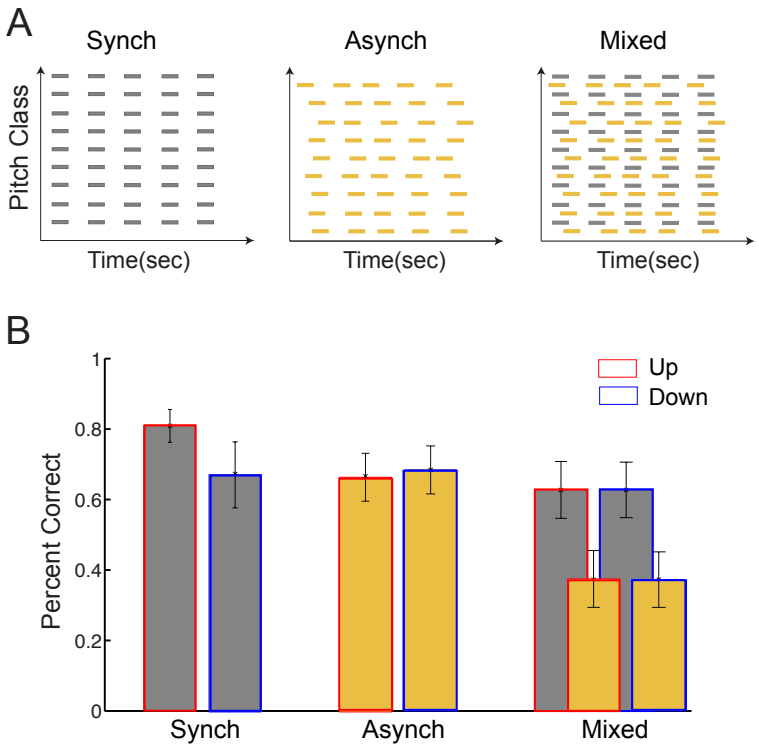

Figure 5.8: Psychoacoustics (I). (A) Schematic representation of the bias sequences for synchronous, asynchronous and mixed conditions. (B) Averaged behavioral performances, evaluating the biasing effect in synchronous, asynchronous and mixed conditions. Behavioral results scored with respect to the spectral location of the synchronous and asynchronous sequences are shown in gray and yellow bars, respectively. Colored borders indicate the expected perceived pitch-change direction (blue: up and red: down), in each experimental condition. The results reveal significantly higher averaged scores for biasing effect with respect to the synchronous sequence as opposed to the asynchronous sequences in the Mixed condition (one-way ANOVA, $p < 0.001$)

Medians, 25% and 75% quartiles plotted for the two *Congruent* and *Incongruent* groups (Figure 5.10–A & 5.10–B, respectively) indicated stronger suppression in the spectral regions corresponding to the perceptual biasing direction; however, the difference is weaker compared to the earlier experiments in which only one biasing sequence was present during each trial. This is presumably due to the simultaneous presence of both biasing sequences and their interaction.

Finally, neural responses to the probe tones in the spectral region of the asynchronous tones did not show significant change with respect to different degrees of jitter, perhaps because the amount of asynchrony was already sufficiently large to dominate the effects even at its smallest amount. It would be useful in future experiments to investigate perceptual and neural differences for smaller jitter, i.e., < 30 ms.

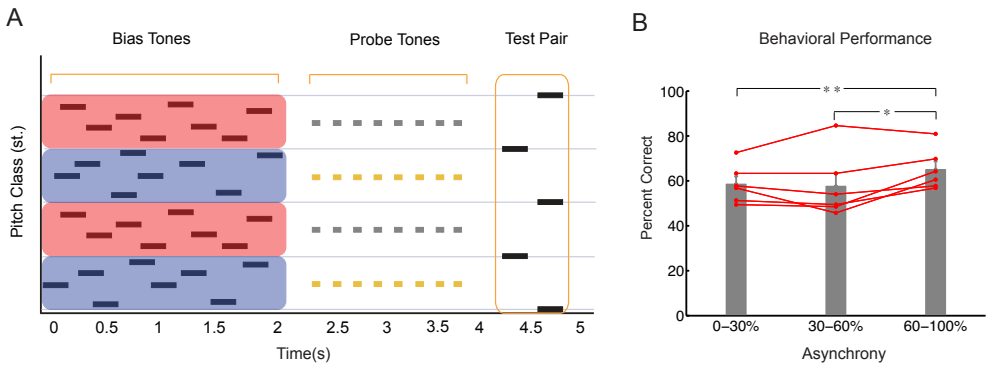

Figure 5.9: Stimulus description and behavioral performances. (A) Cartoon spectrogram of the stimulus in MEG experiment. (B) Averaged behavioral performance of the subjects while performing the pitch-change direction task inside the MEG machine, for different degrees of asynchrony. Individual performances are shown in red. The averaged performance reveals a significant improvement in the third versus the first asynchrony condition (one-way ANOVA, $p < 0.001$), and in the third versus the second asynchrony condition (one-way ANOVA, $p < 0.01$), indicating that synchronous biasing sequences become more effective as the amount of asynchrony in the asynchronous sequence increases.

5.5.2 Psychoacoustics II

In order to evaluate the role of attention in the biasing mechanism, a second set of psychoacoustic experiments were conducted in which subjects were asked to maintain their attention on one of the synchronous or asynchronous biasing sequences in

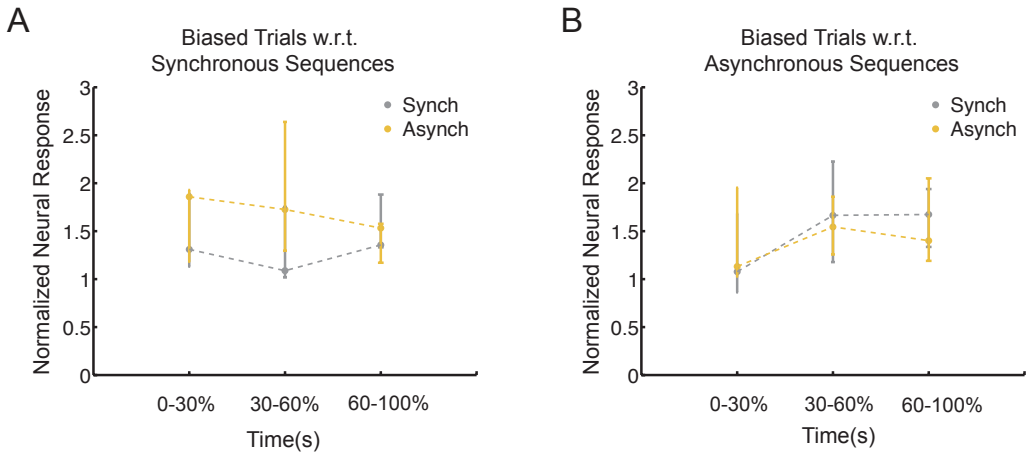

Figure 5.10: MEG neural responses in different asynchrony conditions. (A) and (B) represent Median, 25% and 75% quartiles of the normalized neural responses to the probing sequences in the *Congruent* and *Incongruent* groups across all subjects, respectively. Responses to the probe sequences that precede by synchronous and asynchronous biasing sequences are shown in gray and yellow, respectively.

the presence of the competing biasing sequence. To ensure attention to the targeted biasing tones, subjects were asked to perform an attention task while listening to the biasing sequences. They were then presented with an ambiguous test pair and asked to report their perceived pitch-change direction as in the previous experiments. The cross-effect of streaming the biasing tones and the strength of perceptual biasing was then studied by evaluating listeners' conditional performances from the two behavioral tasks (See Results).

Experimental design

Stimuli & procedure The stimulus was a 5 s long sequence, consisting of 2 superimposed biasing sequences, each with 8 Shepard tones (125 ms tone length, and 125 ms inter-tone interval). The ambiguous test pair was presented 156 ms

after the biasing period with a tone duration and inter-tone interval of 125 ms. For the synchronous biasing tones all of the components in each Shepard tone were synchronized and had the same onset (as in the previous experiments), whereas in the asynchronous case the onsets of the constituent components in each Shepard tone were randomly jittered in the range of $\pm$[42–84ms] (Figure 5.11–A).

A training block of 20 trials was presented to the subjects before the start of the experiment to make sure they were able to identify and follow each biasing sequence, independently. Subjects were then presented with 2 blocks of 60 trials each (2 biasing directions $\times$ 2 task conditions (presence or absence of the deviant) $\times$ 5 deviant locations$\times$ 3 pitch classes). They were asked to maintain their attention on one of the biasing sequences (synchronous or asynchronous) through the entire biasing period of each trial and for all trials in a block. The order of the task (targeted biasing sequence) was counterbalanced across subjects. After listening to the biasing tones in each trial, subjects had to decide if they heard a deviant on any of the targeted biasing tone sequence. They were then presented with an ambiguous Shepard pair and were asked to report their perceived direction of pitch-change in that pair. Subjects reported the answers to both tasks after each trial.

Subjects Seven subjects (mean age, 25 y; four males) participated in this experiment.

Results

In order to evaluate subjects' performances in the attention task (deviant detection), d' scores were computed for individual subjects. The results indicated that the detection performance was higher overall in the synchronous task compared to the asynchronous task (one-way ANOVA, $p < 0.0001$). Trials were then separated into 5 groups based on the temporal locations of the deviants. Computed d' values of the five time windows demonstrated a temporal buildup of deviant detection in both synchronous and asynchronous attention tasks (Figure 5.11–B). This is a similar result to the temporal build-up of the target task detection in the informational masking paradigm discussed in Chapter 3. It was interpreted then as listeners' improved ability in streaming the target sequence over the time. This observation along with the earlier results on the significant improvement of the behavioral performance of listeners in getting biased with respect to the synchronous sequence as the amount of asynchrony in the competing asynchronous sequence increased, support this hypothesis that subjects are able to stream the targeted biasing sequence in this paradigm.

Next, the effect of the streaming performance on the biasing strength was evaluated through computation of the marginal and conditional probabilities of subjects' performances in the second task (perceived pitch-change direction). The probability of reporting the pitch-change direction according to the targeted biasing sequence spectral location, $P(T2)$, was evaluated for each subject, as well as $P(T2|T1)$, the conditional probability of reporting $T2$ correctly given that the deviant on the tar-

geted biasing sequence is correctly detected. In earlier experiments we have found that presentation of only one or two biasing tones can still have a significant biasing effect on subject's percept; therefore, only those trials with their deviants located on the fourth and fifth time-window were included in the analysis of conditional probability, so as to exclude all other trials for which the deviants occurred earlier in the sequence and listeners may have had the chance to switch their attention between the two basing sequences afterward.

The results for the targeted synchronous and asynchronous sequences are plotted in gray and yellow lines in Figure 5.11–C. The difference between the conditional and marginal probabilities is significantly positive across all 7 subjects for both attentional conditions (Figure 5.11–C, bar plots, one-way ANOVA, $p < 0.01$ & $p < 0.001$, for synch and asynch attentional conditions, respectively). These results indicate a positive impact of the first task on the second task, meaning that subjects were biased more strongly in those trials, in which they were able to detect the deviant on the targeted biasing sequence.

Differences between the marginal and conditional distributions for synchronous conditions are significantly smaller than the asynchronous conditions (one-way ANOVA, $p < 0.01$). This is probably due to the task effort being higher in following the asynchronous biasing sequences compared to the synchronous tones. Note that it was shown earlier that listeners were more strongly biased by the synchronous sequences while passively listening to trials with mixed biasing condition.

Overall, the results suggest that although selectively attending to the sequences is not a necessary component in perceptually biasing listeners (subjects

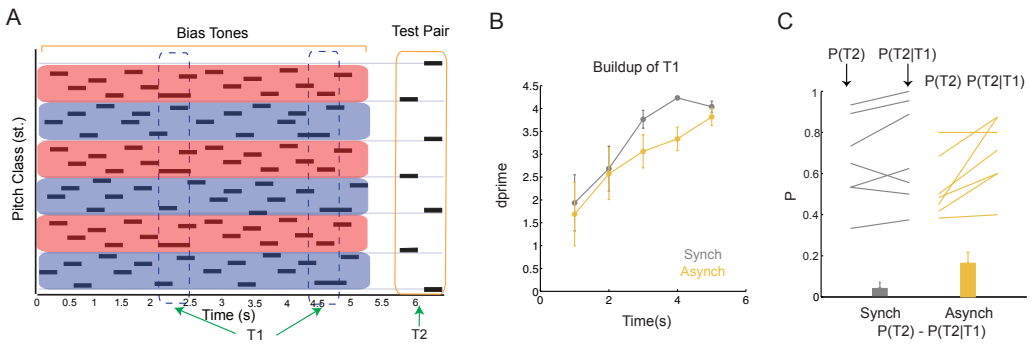

Figure 5.11: Psychoacoustics (II). (A) Schematic view of a typical stimulus. During the presentation of the bias sequence, subjects are asked to stay focused on the synchronous or asynchronous sequence and detect the presence of a deviant (elongation of the constituent tones in a Shepard tone) along the presentation of the biasing sequences ($T1$). The deviant occurs randomly in 50% of the trials and can be placed on any Shepard tone in the sequence. Subjects are then presented with an ambiguous test pair and asked to report the perceived pitch-change direction ($T2$). (B) Behavioral buildup of the detection task ($T1$) is plotted for synchronous and asynchronous sequences. (C) The difference between the conditional and marginal probabilities, $P(T2)$ and $P(T2|T1)$, averaged over all subjects is significantly positive (one-way ANOVA, $p < 0.01$ & $p < 0.001$, for synch and asynch attentional conditions). It is plotted in gray and yellow bar plots for synch and asynch conditions, respectively. Differences between the marginal and conditional distributions for synchronous condition are significantly smaller than the asynchronous condition (one-way ANOVA, $p < 0.01$). Gray and yellow lines indicate the individual performance probabilities $P(T2)$ and $P(T2|T1)$, for each subject.

were able to get the biasing effect by listening passively to the biasing tones), it nevertheless enables the listeners to change their perceptual bias by actively focusing on certain streams and ignoring others.

5.6 Discussion

In this chapter, the neural correlates of contextual influences in auditory perception were investigated, and it was found that a stimulus-specific suppression of the auditory cortex activity predicts a listeners subsequent percept. Further, the results

indicated that the strength of suppression was predictive of the subjects performance. The suppression lasted for multiple seconds consistent with the time-scales of human performance in psychoacoustic studies. While in the present paradigm stimulus-specificity refers to a spectral location, we hypothesize that such effects would occur for more general contextual stimuli, and that the suppression thus reflects a form of early sensory memory, represented already in the auditory cortex (e.g., in Asari and Zador (2009)).

Properties of the neural manifestation of the contextual influence The representation of the contextual influence could have differed in multiple ways with respect to brain stage, shape and direction. First, the contextual influence could have lacked a representation in the auditory cortex and instead only have been represented in higher order auditory or non-sensory areas, as the effect could have been predicted to be closer to decision-related areas, such as the prefrontal cortex (Fritz et al., 2010). Based on source localization (using dipole fitting) and the frequency-specificity of the response, we conclude that the responses come from the auditory cortex instead. Second, the shape of the contextual representation could have been very broad, based on the broad spectral content of the Shepard tones. Instead, the specificity of the representation—in the sense of being specific to a limited range of frequencies, and hence neighboring Shepard tone stimuli—indicates that the auditory cortex represents different Shepard tones, and hence the restricted range of Shepard tones in the context in a locally distinguishable manner. More basic mechanisms of local representation such as SSA (stimulus-specific adaptation discussed

below, (Ulanovsky et al., 2003)) may contribute to this shape, but do not account for the relation to perception in their current from. The locality suggests that a certain amount of contextual content is retained in the representation. Third, the contextual influence could have been expected to be in the form of an enhancement or sensitization, rather than a suppression as was described in the primary visual cortex of the awake mouse (Zhang et al., 2012). Instead, a representation of the contextual information in terms of suppression is found here, a finding that potentially has profound consequences for theories on how prior expectations in a Bayesian framework are encoded in the brain (e.g., (Huys et al., 2007)). Traditionally, priors have been represented as an increase in neuronal firing rates or neuronal sensitivity. However, physiological measurements in the auditory cortex have often displayed suppression or reductions in firing rates during various behavioral states, such as in preparation for task performance (Otazu et al., 2009) and memory maintenance (Linke et al., 2011), and may thus represent a more typical strategy in the auditory system.

Suppression reflects external context and internal state If this representation of context is general, one may hypothesize that it is not only generated by external stimuli but is also modulated by internal expectations or just fluctuations of brain state. Moreover, regardless of how this suppressive representation is generated, it is likely that it influences perception.

The findings in the current studies demonstrated this latter point by the fact that contextual suppression was stronger when trials were grouped according to

performance (*Congruent* versus *Incongruent*; Figure 5.5–A & 5.5–B) rather than by simply the biasing spectral location. There was no significant separation based on the latter; therefore, no data are shown. This suggests that strong suppression induced either externally by biasing sequences or internally by expectations and other top-down factors, played a causal role in inducing the reported percepts. Further support for this hypothesis is provided by the correlation between the strength of suppression and the average performance of the subjects (Figure 5.5–C) . Naturally, a follow-up question of the present study would be to test the manifestation of top-down influences directly, e.g., if a cross-modal cue for directional perception manifests itself in the activity of the auditory cortex as a local suppression.

Relation to neuronal recordings of contextual effects The ensemble potentials obtained in MEG/EEG recordings provide high temporal resolution, but cannot provide insights into the representation at the level of individual cells. Previous neurophysiological studies have addressed the question of contextual tone-tone interactions in the auditory cortex (Brosch and Schreiner, 1997; Brosch et al., 1999; Asari and Zador, 2009); however, their focus has not been on the influence on the representation of perceived directionality of pitch-change. Also, most studies have been conducted in anesthetized animals, making the transition to human perception a complicated affair. Adaptation of sensory neurons to a maintained sensory stimulation has historically been considered merely to reflect fatigue in spike generation or synaptic transmission. However, numerous studies have demonstrated that adaptation/suppression is a much more general phenomenon, both on the per-

ceptual and neural levels. Several forms of activity dependent reduction of neural sensitivity have previously been described in the auditory cortex, most notably forward suppression (Brosch and Schreiner, 1997; Wehr and Zador, 2005; Scholl et al., 2010) and stimulus-specific adaptation (SSA) (Ulanovsky et al., 2003, 2004; Condon and Weinberger, 1991). While these forms of adaptation have been discussed separately, commonalities and interrelationships are likely to exist among them, and with the present results. For example, in forward suppression, the interactions between subsequent tones can be facilitatory or suppressive, with suppressive interactions dominating especially at the frequency location of the first stimulus much like what we see relative to the effects of the biasing tone sequences (Brosch and Schreiner, 1997; Brosch et al., 1999). In stimulus-specific adaptation, a surprising level of locality in suppression has been demonstrated for sequences of two more and less stimuli, with a precision on the order of a few semitones (Condon and Weinberger, 1991), and thus below the tuning width of most neurons. Furthermore, SSA has been demonstrated to last for multiple seconds (Ulanovsky et al., 2004), to be most prominent for the onset response (von der Behrens et al., 2009) and to be of mixed thalamocrtical and cortical origin (Taaseh et al., 2011). Almost all studies of SSA have been performed in anesthetized animals, leaving open the question of top-down control of SSA. Given the similarity in time-scale and frequency specificity between the current findings and SSA, it is likely that the two phenomena share similar origins or underlying mechanisms. In a parallel study (Englitz et al., 2013), neural activity in the auditory cortex of awake, non-behaving ferrets, in response to Shepard biasing and test tone sequences (but not probe sequences) was recorded. As in

the present studies, a localized suppression was measured in the frequency channels stimulated by the bias sequences. It persisted for over a second following the end of the bias sequence, and decayed monotonically away from the center frequency of the bias. These findings led to the proposal that the pitch shift percepts can be accounted for accurately by a decoding scheme based on differential suppression of local frequency-shift-detector cells, as previously advocated by (Demany et al., 2009).

Suppression and sensory memory Memory formation is thought to be a multi-step process. Some computations relating to the working memory content may most efficiently be executed close to the periphery. A putative function of the suppression demonstrated here could be a form of sensory memory located already in primary sensory cortex, as has been proposed by Jääskeläinen et al. (2007). Retaining the memory of previous stimuli can be useful for comparison tasks, which are based on a stimulus property explicitly represented in sensory cortex. An example of this kind for frequency comparison was demonstrated by Linke et al. (2011), where a locally reduced activity was demonstrated in the auditory cortex during the maintenance phase between reference and probe stimulus. The authors interpreted their results as a mechanism for the brain to prevent new information during this phase from overwriting the memory. Similarly, Nelken and colleagues (Nelken and Ulanovsky, 2007) have argued that the local reduction observed in SSA could be interpreted as the memory of the standard sequence, and the relatively increased response to the deviant stimulus as the mismatch signal, commonly observed in mismatch negativ-

ity. While this view has been challenged on more general grounds (Farley et al., 2010), it still remains a building block for the memory involved in mismatch negativity with first order properties like frequency. In the present studies, the local spectral suppression due to the biasing tones may mark the most likely connection between the two subsequent test tones in the ambiguous pair, as has been proposed more generally by Huys et al. (2007). In contrast to the aforementioned study, the representation of this memory is, however, in the form of a reduction rather than an increase in activity (Akram et al., 2013).

Potential controls for the probe stimulus paradigm Using MEG, it is challenging, if not impossible, to measure the activation state of the auditory cortex during the maintenance period after the biasing context without introducing additional stimuli. In these studies a set of such probing stimuli is chosen to perform this measurement, while ensuring that they did not influence the bias induced by the preceding sequences. For example, in a series of psychoacoustic experiments it has been demonstrated that the probe sequences did not overwrite the perceptual effect of the biasing sequence. Therefore, response differences measured with the same probe sequences can be attributed to the aftereffect of the bias sequences on the activation state and not to the probe sequences themselves.

In a further control experiment, it was examined whether the directional percepts reported by the listeners were dependent on the spectral locations of the probe by evaluating the behavioral scores based on the locations of the probe sequences. The behavioral performances were expected to be around chance level since the

probes caused no perceptual change and had no functional relationship to the percepts. The lack of such a relationship was shown in Figure 5.2–A, where the probe's location did not have an influence on the perceived direction (at least in the presence of a preceding bias). Finally, it should be noted that even if the probe sequences had a minor influence on the subsequent percept, the present results would remain valid, since every response was evaluated at the location of the probe (by the design of the measurement), and hence the presence of the probe could not explain differences in the response.

Contextual effects for more general stimuli The presently used stimulus paradigm has been simplified to allow specific conclusions in a laboratory setting. Natural stimulus contexts will rarely present such simplicity. Also, the type of ambiguity studied here, the pitch shift between the Shepard test pair, does not occur much in natural sounds. Nevertheless, realistic scenes are rich in ambiguities of various types, most of which are automatically resolved based on contexts at different levels, ranging from immediately preceding stimuli to higher-level knowledge about the current scene. In a series of audio-visual experiments, it has been shown that videos can moderate phonetic contextual effect and disambiguate an acoustically ambiguous precursor syllable (Fowler et al., 2000; Holt et al., 2005). According to these studies, the presence of auditory-visual interactions cannot necessarily be distinguished among theories of speech perception. In most of the existing theories, the information for recognition of speech sound can be provided from multiple sources, e.g., lexical or visual, but the extent to which these sources are important can differ.

Lexical and statistical information has been shown to result in second-order interactions (Elman and McClelland, 1988; Magnuson et al., 2003; Samuel and Pitt, 2003), whereas for visual sources, there is no evidence that the visual information produces such indirect interactions (Huys et al., 2007). A rich body of behavioral studies in humans report aftereffects on the perception of higher-level stimulus properties such as identity, gender and emotion of faces (Rhodes et al., 2003; Webster et al., 2004), for review see Clifford et al. (2007)), which are consistent with adaptive/suppressive effects: they have been rationalized as a repulsion caused by the suppressed representation of the adaptor. These higher level effects suggest that adaptive mechanisms are not only present in early sensory stage, but are in fact employed throughout multiple levels of perceptual processing. Another kind of generality of stimulus contexts is presented by complex scenes, such as crowded auditory scenes, e.g., the proverbial cocktail party situation (Cherry, 1953), where multiple speakers and other acoustic sources are simultaneously present. Humans outperform man-made systems in their ability to recognize speech under these circumstances, likely by their superior use of contextual cues, which allow better separation of sources and more accurate predictions of subsequent input. Although the complexity of these scenes is much higher than in the present paradigm, similar mechanisms could be applied to enhance stimulus recognition. A local suppression of activity would have differential effects on attended and unattended information. Non-attended background information is predicted on a statistical level only and leads to a cancellation with the actual stimulus. Attended foreground information would lead to suppression on the basis of multi-level expectations and, consequently, the difference from the actual

stimulus can become informative as it signifies the part of the information one is unable to predict.

Contextual effect and Streaming Although an auditory stream is not a well-defined concept in the field of auditory science, the terminology has been extensively used among auditory scientists over the last decades. It is not yet clear what properties an auditory stream should possess or what the neural representation of an auditory stream might be in the brain; however, some general principles are proposed to be used as a basis for object analysis studies in any sensory domain including auditory (see the review by Griffiths and Warren (2004a)). According to these principles, auditory streaming involves the separation of information related to the auditory stream and information related to the rest of the auditory world. It should also involve the abstraction of information that is independent of the auditory representation, i.e, the extraction of invariant characteristics that define a stream in the auditory domain.

Several models in auditory scene analysis are based on the analysis of spectrotemporal features that form a basis for the grouping of features on which stream formation depends. One of the important features known to facilitate auditory streaming procedure is onset synchrony(Rasch, 1978; Bregman, 1990; Shamma et al., 2011a). In the study of contextual effect on the ambitious Shepard tone in this chapter, it has been shown that perceptual biasing strength is also modulated by the onset synchrony of the biasing Shepard tones. Therefore, it was investigated whether there is a potential connection between the streamability of the biasing

sequence and perceptual biasing of the ambiguous test pair by conducting a number of psychoacoustic and MEG experiments.

Dependance of the two auditory experiences above, can establish an objective neurobehavioral measure for auditory streaming by evaluating the perceptual biasing effect strength in a joint experimental paradigm. The results obtained in the current studies suggests that listeners can successfully perform the streaming related tasks during the basing sequence presentation, which specifically has a positive correlation with their ability to get biased according to the attended Shepard sequence. However, given the current information, one still cannot argue if streaming the biasing sequence is a necessary condition for perceptual biasing of the listeners. More psychoacoustic and neurophysiological experiments, with modified or novel experimental paradigms should be used to follow up on the current studies and shed light on the underlying connections between the two auditory experiences.

Chapter 6

Conclusion

6.1 Thesis overview

Hearing problems and different aspects of approaching them have been extensively explored over the last decades; however, there is still a lot to discover in this field of research. Analyzing a complex auditory scene is a seemingly effortless accomplishment for a normal-hearing listener, but it is extremely difficult to simulate such a procedure in an artificial system. The main goal of the studies in this thesis is to first investigate the neural mechanisms underlying auditory perception of complex sounds at cortical level, and second, to explore neuro-biologically inspired computational models for these brain's cognitive functions.

At an experimental level, we are most interested in unraveling the computational strategies that explains cortical activities and their underlying neural mechanisms. In chapter 3, we focused on investigating the neural correlates of auditory streaming in a simplified stream segregation scenario, using an informational mask-

ing paradigm. Evidence shown in this work argues for an interesting interaction between bottom-up feature extraction mechanisms and top-down attentional modulations, leading to perceptual segregation of a single frequency tone-sequence from a background of spectrotemporaly randomized tone cloud. While earlier studies focused on the role of attention as a high-level mechanism in binding relevant features and forming a segregated stream, here we present evidences for strong dependence of the behavioral performance to the stimulus parameters, which are directly reflected in the neural activity recorded from the brain. The results suggest that the recorded neural activity from the brain can be employed as an objective indicator of the streaming percept in listeners.

Taking advantage of the recent findings on close links between perceptual and neural correlates of streaming, we were able to implement a computational model to track the attentional state of listeners. In chapter 4, the proposed model capitalizes on prior observations to build a statistically-principled decoder of human auditory attentional states under competing-speaker auditory stimulation. The underlying measurements represent the problem as one of Bayesian inference and propose an EM-based procedure for estimating the latent variables. The performance of this algorithm is demonstrated on simulated data, and then tested on real MEG data where subject's attention manipulated to either speaker in the stimulus mixture.

We also explored the neural mechanism underlying the contextual effect on the perceptual organization of ambiguous sounds in the auditory cortex. The experimental approaches, described in chapter 5 employed a variation of tritone paradox as the stimulus and characterized the spectrotemporal properties of the contextual

trace using MEG technique. The results reveal a tonotapically localized suppression trace that persists for multiple seconds after presentation of the context. Most interestingly, the strength of the suppression was predictive of the decisions of the listeners even when the stimulus context was not. The findings support the notion that cortical responses to an auditory scene are not static, but are rapidly and locally modified by the stimulus to leave a persistent effect which influences subsequent stimuli and determines perceptual states.

In summery, this thesis is an attempt to explore some of the behavioral and neural mechanisms employed by the brain to analyze complex acoustic scenes and unravel the ambiguous auditory scenarios with the help of the contextual effect. It would be of great interest to combine all these experimental and modeling approaches to conjecture generalized theories of auditory perception and bridge the gap between theoretical and experimental neuroscience, hence the development of practical engineering tools for sound processing.

6.2 Future prospects

In the last few years, there has been considerable interest in the neuroscience community in decoding high-level latent brain states from noisy neural recordings. One such state variable is attention: primates and other high-level organisms can preferentially distribute resources to encode and process a selective set of incoming stimuli in a way that is typically not externally visible. In the case of auditory selective attention, empirical studies have identified a set of neural variables that can be mea-

sured to provide information about sound to which a subject is currently directing attention. Some of these variables can be measured via magnetoencephalography (MEG), as described in chapter 4.

An interesting future direction for this study is to use the proposed attentional state decoder in developing an intelligent hearing aid device to help hearing-impaired listeners in analyzing the complex auditory scenes around them. Current hearing aids fail to recover users' ability to hear out the speech signal in a multi-speaker or noisy environment. Using an intelligent hearing-aid, the attended target can be detected via the EEG/MEG recordings of the brain and enhanced through guiding a directional microphone array or other methods that enables us to amplify and track the target of attention. Another future prospect is to develop a brain-computer-interface (BCI) system, which has potential applications in helping paralyzed patients. Discovering the direction of attention for these patients can be used as a cue, combined with other sources of information, to understand their intention in the first place and help them facilitate executing their intended task in the next step.

Appendix A

Parameter Estimation of the Inverse Problem

Let

$$\Omega := \left\{ \kappa_1, \kappa_2, \{z_k\}_{k=1}^K, \{\eta_k\}_{k=1}^K \right\} \tag{A.1}$$

be the set of parameters.

The log-posterior of the parameter set Ω given the observed data $\left\{\theta_{i,k,r}\right\}_{i,k,r=1}^{2,T,R}$ is given by:

$$\begin{aligned}
\log p\Big(\Omega\Big|\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\Big) = &\sum_{r,k=1}^{R,K} \log\left[\frac{\kappa_1^{W/2-1} p_k}{2\pi^{W/2} I_{W/2-1}(\kappa_1)} \exp\left(\kappa_1 \cos\left(\theta_{1,k,r}\right)\right) + \frac{\kappa_2^{W/2-1}(1-p_k)}{2\pi^{W/2} I_{W/2-1}(\kappa_2)} \exp\left(\kappa_2 \cos\left(\theta_{2,k,r}\right)\right)\right] \\
&+ \left[(\kappa_1+\kappa_2)c_0 d + d(W/2-1)(\log\kappa_1 + \log\kappa_2) - d\big(\log I_{W/2}(\kappa_1) + \log I_{W/2}(\kappa_2)\big)\right] \\
&- \sum_{r,k=1}^{R,K} \left\{ \frac{1}{2\eta_k}\left(z_k - z_{k-1}\right)^2 + \frac{1}{2}\log\eta_k + (\alpha+1)\log\eta_k + \frac{\beta}{\eta_k} \right\} + \mathsf{cst}.
\end{aligned}$$

where $\mathsf{cst}.$ denotes terms that are not functions of Ω. The MAP estimate of the parameters is difficult to obtain given the involved functional form of the log-posterior. However, the complete data log-posterior, where the unobservable sequence $\{n_{k,r}\}_{k=1,r=1}^{K,R}$ is given, takes the form:

$$\begin{aligned}
\log p\Big(\Omega\Big|\{\theta_{i,k,r}, n_{k,r}\}_{i,k,r=1}^{2,K,R}\Big) = &\sum_{r,k=1}^{R,K} n_{k,r}\left[(W/2-1)\log(\kappa_1) + \kappa_1\cos\left(\theta_{1,k,r}\right) - \log I_{W/2-1}(\kappa_1)\right] \\
&+ \sum_{r,k=1}^{R,K} (1-n_{k,r})\left[(W/2-1)\log(\kappa_2) + \kappa_2\cos\left(\theta_{2,k,r}\right) - \log I_{W/2-1}(\kappa_2)\right] \\
&+ \left[(\kappa_1+\kappa_2)c_0 d + d(W/2-1)(\log\kappa_1 + \log\kappa_2) - d\big(\log I_{W/2}(\kappa_1) + \log I_{W/2}(\kappa_2)\big)\right] \\
&+ \sum_{r,k=1}^{R,K} \left[n_{k,r}\log p_k + (1-n_{k,r})\log(1-p_k)\right] \\
&- \sum_{r,k=1}^{R,K} \left\{ \frac{1}{2\eta_k}\left(z_k - z_{k-1}\right)^2 + \frac{1}{2}\log\eta_k + (\alpha+1)\log\eta_k + \frac{\beta}{\eta_k} \right\} + \mathsf{cst}.
\end{aligned}$$

The log-posterior of the parameters given the complete data has a tractable

functional form for optimization purposes. Therefore, by taking $\{n_{k,r}\}_{k=1,r=1}^{K,R}$ as the unobserved data, we can estimate Ω via the EM algorithm (Dempster et al., 1977). Using Bayes' rule, the expectation of $n_{k,r}$, given $\left\{\theta_{i,k,r}\right\}_{i,k,r=1}^{2,K,R}$ and current estimates of the parameters $\Omega^{(\ell)} := \left\{\kappa_1^{(\ell)}, \kappa_2^{(\ell)}, \left\{z_k^{(\ell)}\right\}_{k=1}^{K}, \left\{\eta_k^{(\ell)}\right\}_{k=1}^{K}\right\}$ is given by:

$$\mathbb{E}\left\{n_{k,r}\middle|\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}, \Omega^{(\ell)}\right\} = \frac{\frac{{\kappa_1^{(\ell)}}^{W/2-1} p_k^{(\ell)}}{2\pi^{W/2} I_{W/2-1}(\kappa_1^{(\ell)})} \exp\left(\kappa_1^{(\ell)} \cos\left(\theta_{1,k,r}\right)\right)}{\frac{{\kappa_1^{(\ell)}}^{W/2-1} p_k^{(\ell)}}{2\pi^{W/2} I_{W/2-1}(\kappa_1^{(\ell)})} \exp\left(\kappa_1^{(\ell)} \cos\left(\theta_{1,k,r}\right)\right) + \frac{{\kappa_2^{(\ell)}}^{W/2-1} (1-p_k^{(\ell)})}{2\pi^{W/2} I_{W/2-1}(\kappa_2^{(\ell)})} \exp\left(\kappa_2^{(\ell)} \cos\left(\theta_{2,k,r}\right)\right)}. \tag{A.2}$$

Denoting the above expectation by the shorthand $\mathbb{E}^{(\ell)}\{n_{k,r}\}$, the M-step of the EM algorithm for $\kappa_1^{(\ell+1)}$ and $\kappa_2^{(\ell+1)}$ gives:

$$\kappa_i^{(\ell+1)} = A^{-1}\left(\frac{\sum_{r,k=1}^{R,K} \varepsilon_{i,k,r}^{(\ell)} \cos\left(\theta_{i,k,r}\right) + c_0 d}{\left(d + \sum_{r,k=1}^{R,K} \varepsilon_{i,k,r}^{(\ell)}\right)}\right), \quad \varepsilon_{i,k,r}^{(\ell)} = \begin{cases} \mathbb{E}^{(\ell)}\{n_{k,r}\} & i = 1 \\ 1 - \mathbb{E}^{(\ell)}\{n_{k,r}\} & i = 2 \end{cases} \tag{A.3}$$

where $A(x) := -\frac{W/2-1}{x} + \frac{0.5(I_{W/2-2}(x) + I_{W/2}(x))}{I_{W/2-1}(x)}$, with $I_W(\cdot)$ denoting the W^{th} order modified Bessel function of the first kind. Inversion of $A(\cdot)$ can be carried out numerically in order to find $\kappa_1^{(\ell+1)}$ and $\kappa_2^{(\ell+1)}$. The M-step for $\{\eta_k\}_{k=1}^K$ and $\{z_k\}_{k=1}^K$ corresponds to the following maximization problem:

$$\underset{\{z_k,\eta_k\}_{k=1}^K}{\operatorname{argmax}} \sum_{r,k=1}^{R,K} \left[\mathbb{E}^{(\ell)}\{n_{k,r}\} z_k - \log(1+\exp(z_k)) - \frac{1}{2\eta_k}\left[(z_k - z_{k-1})^2 + 2\beta\right] - \frac{1+2(\alpha+1)}{2}\log\eta_k\right]. \tag{A.4}$$

An efficient approximate solution to this maximization problem is given by another EM algorithm, where the E-step is the point process smoothing algorithm (Smith and Brown, 2003; Smith et al., 2004) and the M-step updates the state variance sequence (Shumway and Stoffer, 1982). At iteration m, given an estimate of $\eta_k^{(\ell+1)}$, denoted by $\eta_k^{(\ell+1,m)}$, the forward pass of the E-step for $k = 1, 2, \cdots, K$ is given by:

$$\begin{cases} \bar{z}_{k|k-1}^{(\ell+1,m)} = \bar{z}_{k-1|k-1}^{(\ell+1,m)} \\ \sigma_{k|k-1}^{(\ell+1,m)} = \sigma_{k-1|k-1}^{(\ell+1,m)} + \frac{\eta_k^{(\ell+1,m)}}{R} \\ \bar{z}_{k|k}^{(\ell+1,m)} = \bar{z}_{k|k-1}^{(\ell+1,m)} + \sigma_{k|k-1}^{(\ell+1,m)}\left[\sum_{r=1}^{R} \mathbb{E}^{(\ell)}\{n_{k,r}\} - R\frac{\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)}{1+\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)}\right] \\ \sigma_{k|k}^{(\ell+1,m)} = \left[\frac{1}{\sigma_{k|k-1}^{(\ell+1,m)}} + R\frac{\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)}{\left(1+\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)\right)^2}\right]^{-1} \end{cases} \tag{A.5}$$

Note that the third equation in the forward filter is non-linear in $\bar{z}_{k|k}^{(\ell+1,m)}$, and can be solved using standard techniques (e.g., Newton's method). More details on derivation of the non-linear forward filter can be found in Appendix B. For

$k = K-1, K-2, \cdots, 1$, the backward pass of the E-step is given by:

$$\begin{cases} s_k^{(\ell+1,m)} = \sigma_{k|k}^{(\ell+1,m)} / \sigma_{k+1|k}^{(\ell+1,m)} \\ \bar{z}_{k|K}^{(\ell+1,m)} = \bar{z}_{k|k}^{(\ell+1,m)} + s_k^{(\ell+1,m)} \left(\bar{z}_{k+1|K}^{(\ell+1,m)} - \bar{z}_{k+1|k}^{(\ell+1,m)} \right) \\ \sigma_{k|K}^{(\ell+1,m)} = \sigma_{k|k}^{(\ell+1,m)} + s_k^{(\ell+1,m)} \left(\sigma_{k+1|K}^{(\ell+1,m)} - \sigma_{k+1|k}^{(\ell+1,m)} \right) s_k^{(\ell+1,m)} \end{cases} \tag{A.6}$$

The M-step gives the updated value of $\eta_k^{(\ell+1,m+1)}$ as:

$$\begin{aligned} \eta_k^{(\ell+1,m+1)} &= \frac{\mathbb{E}\left(z_k^2 \middle| \Omega^{(\ell)}, \{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R} \right) + \mathbb{E}\left(z_{k-1}^2 \middle| \Omega^{(\ell)}, \{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R} \right) - 2\mathbb{E}\left(z_k, z_k - 1 \middle| \Omega^{(\ell)}, \{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R} \right) + 2\beta}{1 + 2(\alpha+1)} \\ &= \frac{\left(\bar{z}_{k|K}^{(\ell+1,m)} - \bar{z}_{k-1|K}^{(\ell+1,m)} \right)^2 + \sigma_{k|K}^{(\ell+1,m)} + \sigma_{k-1|K}^{(\ell+1,m)} - 2\sigma_{k|K}^{(\ell+1,m)} s_{k-1}^{(\ell+1,m)} + 2\beta}{1 + 2(\alpha+1)}. \end{aligned} \tag{A.7}$$

Appendix B

Recursive non-linear filter algorithm

Assume that at time $(k-1)$, $z_{k-1|k-1}$ and $\sigma^2_{k|k-1}$ are given under the gaussian continuity assumption on z_k. The distribution of z_k given $z_{k-1|k-1}$ is $N(z_{k-1|k-1}, \sigma^2_{k|k-1})$, where $\sigma^2_{k|k-1} = \eta_k + \sigma^2_{k-1|k-1}$. To derive the non-linear recursive filter, we keep track of the parameters of the posterior distribution $p(z_k|\Omega)$.

$$\log(p(z_k|\Omega)) = \left\{ -\frac{(z_k - z_{k-1|k-1})^2}{\sigma^2_{k|k-1}} \right\} \left\{ \mathbb{E}^{(\ell)}\{n_k\} z_k - \log(1+\exp(z_k)) - \frac{\beta}{\eta_k} + \frac{1+2(\alpha+1)}{2} \log \eta_k \right\}. \tag{B.1}$$

To find the optimal estimate of z_k, we apply a gaussian approximation to posterior prediction equation. The approximation is based on recursively computing the posterior mode $z_{k|k}$ and computing its variance $\sigma^2_{k|k}$ as the negative inverse Hessian of the log posterior probability density (Tanner). Differentiating equation (B.1) w.r.t. z_k gives

$$-\frac{z_k - z_{k-1|k-1}}{\sigma^2_{k|k-1}} + \mathbb{E}^{(\ell)}\{n_k\} - \frac{\exp(z_k)}{1+\exp(z_k)} = 0 \tag{B.2}$$

and solving for z yields

$$z_k = z_{k-1|k-1} + \sigma^2_{k|k-1} \left\{ \mathbb{E}^{(\ell)}\{n_k\} - \frac{\exp z_k}{1+\exp(z_k)} \right\}. \tag{B.3}$$

This equation is non-linear w.r.t. z_k and can be solved using Newton's method. The Hessian of equation (B.1) is given by:

$$\frac{-1}{\sigma^2_{k|k-1}} - \frac{\exp(z_k)(1+\exp(z_k)) - \exp^2(z_k)}{(1+\exp(z_k))^2} \tag{B.4}$$

and hence, the variance of z_k, under the Gaussian approximation is given by:

$$\sigma^2_{k|k} = \left(\frac{1}{\sigma^2_k k - 1} + \frac{\exp z_{k|k}}{(1+\exp(z_{k|k}))^2} \right)^{-1}. \tag{B.5}$$

Appendix C

State-space covariance algorithm

The lagged covariance estimate $\sigma_{k,u|K}$ can be computed from the state-space covariance smoothing algorithm (De Jong and Mackinnon, 1988) given by the following equation:

$$\sigma_{k,u|K} = \sigma^2_{k|k}(\sigma^2_{k+1|k})^{-1}\sigma_{k+1,u|K} \tag{C.1}$$

for $1 \leq k \leq u \leq K$. It follows that the covariance terms required for the E-step in the state-space model are

$$\mathbb{E}\{z_k, z_{k-1}\} = \sigma_{k,k+1|K} + \bar{z}_{k|K}\bar{z}_{k+1|K} \tag{C.2}$$

and

$$\mathbb{E}\{z_k^2\} = \sigma^2_{k|K} + \bar{z}^2_{k|K} \tag{C.3}$$

Bibliography

Ackermann, H., Hertrich, I., Mathiak, K., Lutzenberger, W., 2001. Contralaterality of cortical auditory processing at the level of the m50/m100 complex and the mismatch field: A whole-head magnetoencephalography study. Neuroreport 12, 1683–1687.

Adler, L.E., Pachtman, E., Franks, R., Pecevich, M., Waldo, M., Freedman, R., 1982. Neurophysiological evidence for a defect in neuronal mechanisms involved in sensory gating in schizophrenia. Biological psychiatry .

Ahveninen, J., Jääskeläinen, I.P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., Levänen, S., Lin, F.H., Sams, M., Shinn-Cunningham, B.G., et al., 2006. Task-modulated what and where pathways in human auditory cortex. Proceedings of the National Academy of Sciences 103, 14608–14613.

Akram, S., Englitz, B., Chambers, C., Pressnitzer, D., Simon, J., Shamma, A., 2013. Neural correlates of sensory memory in auditory cortex. Psychological Science 19, 143–148.

Akram, S., Englitz, B., Elhilali, M., Simon, J.Z., Shamma, S.A., 2014a. Investigating the neural correlates of a streaming percept in an informational-masking paradigm. PloS one 9, e114427.

Akram, S., Simon, J.Z., Shamma, S.A., Babadi, B., 2014b. A state-space model for decoding auditory attentional modulation from meg in a competing-speaker environment, in: Advances in Neural Information Processing Systems, pp. 460–468.

Alain, C., Arnott, S.R., 2000. Selectively attending to auditory objects. Front. Biosci 5, D202–D212.

Anstis, S.M., Saida, S., 1985. Adaptation to auditory streaming of frequency-modulated tones. Journal of Experimental Psychology: Human Perception and Performance 11, 257.

Asari, H., Zador, A.M., 2009. Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. Journal of neurophysiology 102, 2638–2656.

Ba, D., Babadi, B., Purdon, P.L., Brown, E.N., 2014. Convergence and stability of iteratively re-weighted least squares algorithms. IEEE Trans. on Signal Processing 62, 183–195.

Bartlett, E.L., Sadagopan, S., Wang, X., 2011. Fine frequency tuning in monkey auditory cortex and thalamus. Journal of neurophysiology 106, 849–859.

von der Behrens, W., Bäuerle, P., Kössl, M., Gaese, B.H., 2009. Correlating stimulus-specific adaptation of cortical neurons and local field potentials in the awake rat. The Journal of Neuroscience 29, 13837–13849.

Bendat, J.S., Piersol, A.G., 1986. Random data: measurement and analysis procedures.

Bergman, A.S., 1994. MIT Press, Cambridge, MA.

Besle, J., Fort, A., Delpuech, C., Giard, M.H., 2004. Bimodal speech: early suppressive visual effects in human auditory cortex. European Journal of Neuroscience 20, 2225–2234.

Bialek, W., Rieke, F., Van Steveninck, R.d.R., Warland, D., 1991. Reading a neural code. Science 252, 1854–1857.

Bidet-Caulet, A., Bertrand, O., 2009. Neurophysiological mechanisms involved in auditory perceptual organization. Frontiers in neuroscience 3, 182.

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

Biermann, S., Heil, P., 2000. Parallels between timing of onset responses of single neurons in cat and of evoked magnetic fields in human auditory cortex. Journal of Neurophysiology 84, 2426–2439.

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

Brattico, E., Pallesen, K.J., Varyagina, O., Bailey, C., Anourova, I., Järvenpää, M., Eerola, T., Tervaniemi, M., 2009. Neural discrimination of nonprototypical chords in music experts and laymen: an meg study. Journal of Cognitive Neuroscience 21, 2230–2244.

Bregman, A., 1990. Auditory scene analysis: The perceptual organization of sound. 1990.

Bregman, A.S., 1994. Auditory scene analysis: The perceptual organization of sound. MIT press.

Bregman, A.S., Ahad, P.A., Crum, P.A., OReilly, J., 2000. Effects of time intervals and tone durations on auditory stream segregation. Perception & psychophysics 62, 626–636.

Bregman, A.S., Campbell, J., 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. Journal of experimental psychology 89, 244.

Brosch, M., Schreiner, C.E., 1997. Time course of forward masking tuning curves in cat primary auditory cortex. Journal of Neurophysiology 77, 923–943.

Brosch, M., Schulz, A., Scheich, H., 1999. Processing of sound sequences in macaque auditory cortex: response enhancement. Journal of Neurophysiology 82, 1542–1559.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. The Journal of the Acoustical Society of America 109, 1101–1109.

Cardy, J.E.O., Ferrari, P., Flagg, E.J., Roberts, W., Roberts, T.P., 2004. Prominence of m50 auditory evoked response over m100 in childhood and autism. Neuroreport 15, 1867–1870.

Carlyon, R.P., 1991. Discriminating between coherent and incoherent frequency modulation of complex tones. The Journal of the Acoustical Society of America 89, 329–340.

Carlyon, R.P., 2004. How the brain separates sounds. Trends in cognitive sciences 8, 465–471.

Carlyon, R.P., Cusack, R., Foxton, J.M., Robertson, I.H., 2001. Effects of attention and unilateral neglect on auditory stream segregation. Journal of Experimental Psychology: Human Perception and Performance 27, 115.

Carlyon, R.P., Plack, C.J., Fantini, D.A., Cusack, R., 2003. Cross-modal and non-sensory influences on auditory streaming. PERCEPTION-LONDON- 32, 1393–1402.

Chait, M., Simon, J.Z., Poeppel, D., 2004. Auditory m50 and m100 responses to broadband noise: functional implications. Neuroreport 15, 2455–2458.

Chambers, C., Akram, S., Shamma, S., Pressnitzer, D., 2012. The influence of perceptual organisation on an auditory context effect. The Journal of the Acoustical Society of America 131, 3230–3230.

Chambers, C., Pressnitzer, D., 2011. The effect of context in the perception of an ambiguous pitch stimulus. In Association for Research in Otolaryngology , 1025.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America 25, 975–979.

de Cheveigné, A., Simon, J.Z., 2007. Denoising based on time-shift pca. Journal of Neuroscience Methods 165, 297–305.

de Cheveigné, A., Simon, J.Z., 2008. Denoising based on spatial filtering. Journal of Neuroscience Methods 171, 331–339.

Chi, T., Gao, Y., Guyton, M.C., Ru, P., Shamma, S., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. The Journal of the Acoustical Society of America 106, 2719–2732.

Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America 118, 887–906.

Christopher deCharms, R., Blake, D.T., Merzenich, M.M., 1998. Optimizing sound features for cortical neurons. science 280, 1439–1444.

Clifford, C.W., Webster, M.A., Stanley, G.B., Stocker, A.A., Kohn, A., Sharpee, T.O., Schwartz, O., 2007. Visual adaptation: neural, psychological and computational aspects. Vision research 47, 3125–3131.

Condon, C.D., Weinberger, N.M., 1991. Habituation produces frequency-specific plasticity of receptive fields in the auditory cortex. Behavioral neuroscience 105, 416.

Cusack, R., 2005. The intraparietal sulcus and perceptual organization. Journal of cognitive neuroscience 17, 641–651.

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

Dannenbring, G.L., Bregman, A.S., 1976. Effect of silence between tones on auditory stream segregation. The Journal of the Acoustical Society of America 59, 987–989.

Darwin, C., Carlyon, R.P., 1995. Auditory grouping, in: The handbook of perception and cognition. Hearing, pp. 387–424.

Darwin, C., Hukin, R., 2000. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. The Journal of the Acoustical Society of America 107, 970–977.

David, S.V., Fritz, J.B., Shamma, S.A., 2012. Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proceedings of the National Academy of Sciences 109, 2144–2149.

David, S.V., Mesgarani, N., Shamma., S.A., 2007. Selective cortical representation of attended speaker in multi-talker speech perception. Network: Computation in Neural Systems 18, 191–221.

De Cheveigné, A., Simon, J.Z., 2007. Denoising based on time-shift pca. Journal of neuroscience methods 165, 297–305.

De Jong, P., Mackinnon, M.J., 1988. Covariances for smoothed estimates in state space models. Biometrika 75, 601–602.

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

Demany, L., Pressnitzer, D., Semal, C., 2009. Tuning properties of the auditory frequency-shift detectors. The Journal of the Acoustical Society of America 126, 1342–1348.

Dempster, A.P., Laird, N.M., , Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society 39, 1–38.

Denham, S., Winkler, I., 2006. The role of predictive models in the formation of auditory streams. Journal of Physiology-Paris 100, 154–170.

Deutsch, D., 1980. The processing of structured and unstructured tonal sequences. Perception & Psychophysics 28, 381–389.

Ding, N., Simon, J.Z., 2009. Neural representations of complex temporal modulations in the human auditory cortex. Journal of neurophysiology 102, 2731–2743.

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

Ding, N., Simon, J.Z., 2012b. Emergence of neural encoding of auditory objects while listening to competing speakers. PNAS 109, 11854–11859.

Ding, N., Simon, J.Z., 2012c. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. Journal of Neurophisiology 107, 78–89.

Duifhuis, H., Willems, L.F., Sluyter, R., 1982. Measurement of pitch in speech: An implementation of goldsteins theory of pitch perception. The Journal of the Acoustical Society of America 71, 1568–1580.

Dykstra, A.R., Halgren, E., Thesen, T., Carlson, C.E., Doyle, W., Madsen, J.R., Eskandar, E.N., Cash, S.S., 2011. Widespread brain areas engaged during a classical auditory streaming task revealed by intracranial eeg. Frontiers in human neuroscience 5.

DYSON, J., 2010. Auditory organization. Oxford Handbook of Auditory Science 3, 177–206.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. volume 57. CRC press.

Eggermont, J.J., 1991. Rate and synchronization measures of periodicity coding in cat primary auditory cortex. Hearing research 56, 153–167.

Elhilali, M., Shamma, S., 2007. A cortical view on auditory scene analysis: A physiological and computational approach, in: 19th International Congress on Acoustics.

Elhilali, M., Shamma, S., 2008a. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. Journal of the Acoustical Society of America 124, 3751–3771.

Elhilali, M., Shamma, S.A., 2008b. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. The Journal of the Acoustical Society of America 124, 3751–3771.

Elhilali, M., Xiang, J., Shamma, S.A., Simon, J.Z., 2009. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS biology 7, e1000129.

Ellis, D.P., 2006. Model-based scene analysis. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications , 115–147.

Elman, J.L., McClelland, J.L., 1988. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. Journal of Memory and Language 27, 143–165.

Englitz, B., Akram, S., David, S., Chambers, C., Pressnitzer, D., Depireux, D., Fritz, J., Shamma, S.A., 2013. Putting the tritone paradox into context: Insights from neural population decoding and human psychophysics, in: Basic Aspects of Hearing. Springer, pp. 157–164.

Farley, B.J., Quirk, M.C., Doherty, J.J., Christian, E.P., 2010. Stimulus-specific adaptation in auditory cortex is an nmda-independent process distinct from the sensory novelty encoded by the mismatch negativity. The Journal of Neuroscience 30, 16475–16484.

Fisher, N.I., 1993. Statistical analysis of spherical data. Cambridge University Press.

Fisher, N.I., 1995. Statistical analysis of circular data. Cambridge University Press.

Fishman, Y.I., Micheyl, C., Steinschneider, M., 2013. Neural representation of harmonic complex tones in primary auditory cortex of the awake monkey. The Journal of Neuroscience 33, 10312–10323.

Fishman, Y.I., Reser, D.H., Arezzo, J.C., Steinschneider, M., 2001. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hearing research 151, 167–187.

Fishman, Y.I., Steinschneider, M., 2010. Neural correlates of auditory scene analysis based on inharmonicity in monkey primary auditory cortex. The Journal of Neuroscience 30, 12480–12494.

Fowler, C.A., Brown, J.M., Mann, V.A., 2000. Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. Journal of Experimental Psychology: Human perception and performance 26, 877.

Fritz, J.B., David, S.V., Radtke-Schuller, S., Yin, P., Shamma, S.A., 2010. Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. Nature neuroscience 13, 1011–1019.

Fuentemilla, L., Marco-Pallarés, J., Grau, C., 2006. Modulation of spectral power and of phase resetting of eeg contributes differentially to the generation of auditory event-related potentials. Neuroimage 30, 909–916.

Gielen, C., Hesselmans, G., Johannesma, P., 1988. Sensory interpretation of neural activity patterns. Mathematical Biosciences 88, 15–35.

Griffiths, T.D., Micheyl, C., Overath, T., 2012. Auditory object analysis, in: The Human Auditory Cortex. Springer, pp. 199–223.

Griffiths, T.D., Warren, J.D., 2004a. What is an auditory object? Nature Reviews Neuroscience 5, 887–892.

Griffiths, T.D., Warren, J.D., 2004b. What is an auditory object? Nature Reviews Neuroscience 5, 887–892.

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

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

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

Gutschalk, A., Micheyl, C., Oxenham, A.J., 2008. Neural correlates of auditory perceptual awareness under informational masking. PLoS biology 6, e138.

Hämäläinen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J., Lounasmaa, O.V., 1993. Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of modern Physics 65, 413.

Hams, J.D., 1985. Bases of hearing science. Ear and Hearing 6, 117–118.

Hartmann, W.M., Johnson, D., 1991. Stream segregation and peripheral channeling. Music perception , 155–183.

Helmholtz, H.v., 1853. Ueber einige gesetze der vertheilung elektrischer ströme in körperlichen leitern mit anwendung auf die thierisch-elektrischen versuche. Annalen der Physik 165, 211–233.

Hesselmans, G.H., Johannesma, P.I., 1989. Spectro-temporal interpretation of activity patterns of auditory neurons. Mathematical biosciences 93, 31–51.

Hill, K.T., Miller, L.M., 2009. Auditory attentional control and selection during cocktail party listening. Cerebral cortex , bhp124.

Hillyard, S.A., Picton, T.W., 1987. Electrophysiology of cognition. Comprehensive Physiology .

Hochberg, Y., Benjamini, Y., 1990. More powerful procedures for multiple significance testing. Statistics in medicine 9, 811–818.

Holt, L.L., 2006. Speech categorization in context: Joint effects of nonspeech and speech precursors. The Journal of the Acoustical Society of America 119, 4016–4026.

Holt, L.L., Stephens, J.D., Lotto, A.J., 2005. A critical evaluation of visually moderated phonetic context effects. Perception & psychophysics 67, 1102–1112.

Hupé, J.M., Joffo, L.M., Pressnitzer, D., 2008. Bistability for audiovisual stimuli: perceptual decision is modality specific. Journal of Vision 8, 1.

Huys, Q.J., Zemel, R.S., Natarajan, R., Dayan, P., 2007. Fast population coding. Neural Computation 19, 404–441.

Jääskeläinen, I.P., Ahveninen, J., Belliveau, J.W., Raij, T., Sams, M., 2007. Short-term plasticity in auditory cognition. Trends in neurosciences 30, 653–661.

Kandel, E.R., Schwartz, J.H., Jessell, T.M., et al., 2000. Principles of neural science. volume 4. McGraw-Hill New York.

Kaplan-Neeman, R., Kishon-Rabin, L., Henkin, Y., Muchnik, C., 2006. Identification of syllables in noise: electrophysiological and behavioral correlates. The Journal of the Acoustical Society of America 120, 926–933.

Kashino, M., Kondo, H.M., 2012. Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 977–987.

Kashino, M., Okada, M., Mizutani, S., Davis, P., Kondo, H.M., 2007. The dynamics of auditory streaming: psychophysics, neuroimaging, and modeling, in: Hearing–From Sensory Processing to Perception. Springer, pp. 275–283.

Kay, S.A., 1993. Fundamentals of Statistical Signal Processing: Estimation Theory. Technometrics.

Keshner, M.S., 1982. 1/f noise. Proceedings of the IEEE 70, 212–218.

Kidd Jr, G., Mason, C.R., Deliwala, P.S., Woods, W.S., Colburn, H.S., 1994. Reducing informational masking by sound segregation. The Journal of the Acoustical Society of America 95, 3475–3480.

Kidd Jr, G., Mason, C.R., Rohtla, T.L., 1995. Binaural advantage for sound pattern identification. The Journal of the Acoustical Society of America 98, 1977–1986.

Kidd Jr, G., Richards, V.M., Streeter, T., Mason, C.R., Huang, R., 2011. Contextual effects in the identification of nonspeech auditory patterns. The Journal of the Acoustical Society of America 130, 3926–3938.

Kilgard, M.P., Merzenich, M.M., 1999. Distributed representation of spectral and temporal information in rat primary auditory cortex. Hearing research 134, 16–28.

Kim, Y.J., Grabowecky, M., Paller, K.A., Muthu, K., Suzuki, S., 2006. Attention induces synchronization-based response gain in steady-state visual evoked potentials. Nature neuroscience 10, 117–125.

Klein, D.J., Depireux, D.A., Simon, J.Z., Shamma, S.A., 2000. Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. Journal of computational neuroscience 9, 85–111.

Kondo, H.M., Kashino, M., 2009. Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. The Journal of Neuroscience 29, 12695–12701.

Kowalski, N., Depireux, D., Shamma, S., 1996. Analysis of dynamic spectra in ferret primary auditory cortex: Characteristics of single unit responses to moving ripple spectra. J. Neurophysiology 76.

Kubovy, M., Van Valkenburg, D., 2001. Auditory and visual objects. Cognition 80, 97–126.

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

Li, Q., Huang, Y., 2010. Robust speaker identification using an auditory-based feature, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE. pp. 4514–4517.

Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J., Chauvel, P., 2004. Temporal envelope processing in the human left and right auditory cortices. Cerebral Cortex 14, 731–740.

Linke, A.C., Vicente-Grabovetsky, A., Cusack, R., 2011. Stimulus-specific suppression preserves information in auditory short-term memory. Proceedings of the National Academy of Sciences 108, 12961–12966.

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

Lütkenhöner, B., Steinsträter, O., 1998. High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiology and Neurotology 3, 191–213.

Lyon, R., Shamma, S., 1996. Auditory representations of timbre and pitch, in: Auditory computation. Springer, pp. 221–270.

Magnuson, J.S., McMurray, B., Tanenhaus, M.K., Aslin, R.N., 2003. Lexical effects on compensation for coarticulation: The ghost of christmash past. Cognitive Science 27, 285–298.

Mäkelä, J., Hämäläinen, M., Hari, R., McEvoy, L., 1994. Whole-head mapping of middle-latency auditory evoked magnetic fields. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 92, 414–421.

McAdams, S., 1982. Spectral fusion and the creation of auditory images, in: Music, mind, and brain. Springer, pp. 279–298.

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

McEvoy, L., Mäkelä, J., Hämäläinen, M., Hari, R., 1994. Effect of interaural time differences on middle-latency and late auditory evoked magnetic fields. Hearing research 78, 249–257.

Melcher, J.R., Levine, R.A., Bergevin, C., Norris, B., 2009. The auditory midbrain of people with tinnitus: abnormal sound-evoked activity revisited. Hearing research 257, 63–74.

Menning, H., Roberts, L.E., Pantev, C., 2000. Plastic changes in the auditory cortex induced by intensive frequency discrimination training. Neuroreport 11, 817–822.

Mesgarani, N., Chang, E.F., 2012a. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.

Mesgarani, N., Chang, E.F., 2012b. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2009. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. Journal of neurophysiology 102, 3329.

Mesgarani, N., Shamma, S., 2005. Speech enhancement based on filtering the spectrotemporal modulations, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, IEEE. pp. 1105–1108.

Mesgarani, N., Slaney, M., Shamma, S., 2006. Content-based audio classification based on multiscale spectro-temporal features. IEEE Transactions on Speech and Audio processing 14, 920–930.

Micheyl, C., Carlyon, R.P., Gutschalk, A., Melcher, J.R., Oxenham, A.J., Rauschecker, J.P., Tian, B., Courtenay Wilson, E., 2007a. The role of auditory cortex in the formation of auditory streams. Hearing research 229, 116–131.

Micheyl, C., Shamma, S.A., Oxenham, A.J., 2007b. Hearing out repeating elements in randomly varying multitone sequences: a case of streaming?, in: Hearing–From Sensory Processing to Perception. Springer, pp. 267–274.

Micheyl, C., Tian, B., Carlyon, R.P., Rauschecker, J.P., 2005. Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48, 139–148.

Miller, L.M., Escabí, M.A., Read, H.L., Schreiner, C.E., 2002. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. Journal of neurophysiology 87, 516–527.

Moore, B.C., Gockel, H., 2002. Factors influencing sequential stream segregation. Acta Acustica United with Acustica 88, 320–333.

Moore, B.C., Gockel, H.E., 2012. Properties of auditory stream formation. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 919–931.

Morgan, S., Hansen, J., Hillyard, S., 1996. Selective attention to stimulus location modulates the steady-state visual evoked potential. Proceedings of the National Academy of Sciences 93, 4770–4774.

Näätänen, R., 1990. The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. Behavioral and Brain Sciences 13, 201–233.

Näätänen, R., Picton, T., 1987. The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24, 375–425.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., Winkler, I., 2001. primitive intelligencein the auditory cortex. Trends in neurosciences 24, 283–288.

Necker, L.A., 1832. Lxi. observations on some remarkable optical phænomena seen in switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid .

Nelken, I., Ulanovsky, N., 2007. Mismatch negativity and stimulus-specific adaptation in animal models. Journal of Psychophysiology 21, 214.

van Noorden, L.P., 1977. Minimum differences of level and frequency for perceptual fission of tone sequences abab. The Journal of the Acoustical Society of America 61, 1041–1045.

van Noorden, L.P.A.S., 1975. Temporal coherence in the perception of tone sequences. Institute for Perceptual Research.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the edinburgh inventory. Neuropsychologia 9, 97–113.

Oram Cardy, J.E., Flagg, E.J., Roberts, W., Roberts, T.P., 2008. Auditory evoked fields predict language ability and impairment in children. International Journal of Psychophysiology 68, 170–175.

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2014. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. Cerebral Cortex , bht355.

Otazu, G.H., Tai, L.H., Yang, Y., Zador, A.M., 2009. Engaging in an auditory task suppresses responses in auditory cortex. Nature neuroscience 12, 646–654.

Palmer, A.R., Winter, I., Stabler, S., 1996. Responses to simple and complex sounds in the cochlear nucleus of the guinea pig. Advances in speech, hearing and language processing. Ed. WA Ainsworth, JAI Press .

Paltoglou, A.E., Sumner, C.J., Hall, D.A., 2009. Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention. Hearing research 257, 106–118.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS biology 10, e1001251.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., Roberts, T., 2000. Auditory cortex accesses phonological categories: an meg mismatch study. Cognitive Neuroscience, Journal of 12, 1038–1055.

Phillips, D., Hall, S., Hollett, J., 1989. Repetition rate and signal level effects on neuronal responses to brief tone pulses in cat auditory cortex. The Journal of the Acoustical Society of America 85, 2537–2549.

Pinto, D.J., Hartings, J.A., Brumberg, J.C., Simons, D.J., 2003. Cortical damping: analysis of thalamocortical response transformations in rodent barrel cortex. Cerebral Cortex 13, 33–44.

Poeppel, D., Yellin, E., Phillips, C., Roberts, T., Rowley, H., Wexler, K., Marantz, A., 1996. Task-induced asymmetry of the auditory evoked m100 neuromagnetic field elicited by speech sounds. Cognitive Brain Research 4, 231–242.

Polley, D.B., Steinberg, E.E., Merzenich, M.M., 2006. Perceptual learning directs auditory cortical map reorganization through top-down influences. The journal of neuroscience 26, 4970–4982.

Potter, D., Summerfelt, A., Gold, J., Buchanan, R.W., 2006. Review of clinical correlates of p50 sensory gating abnormalities in patients with schizophrenia. Schizophrenia bulletin 32, 692–700.

Pressnitzer, D., Hupé, J.M., 2006. Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. Current Biology 16, 1351–1357.

Pulvermüller, F., Shtyrov, Y., 2009. Spatiotemporal signatures of large-scale synfire chains for speech processing as revealed by meg. Cerebral cortex 19, 79–88.

Ramachandran, V.S., Anstis, S.M., 1985. Perceptual organization in multistable apparent motion. Perception 14, 135–143.

Rasch, R.A., 1978. The perception of simultaneous notes such as in polyphonic music. Acta Acustica united with Acustica 40, 21–33.

Recanzone, G.a., Schreiner, C., Merzenich, M.M., 1993. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. The Journal of Neuroscience 13, 87–103.

Repp, B.H., 1997. Spectral envelope and context effects in the tritone paradox. PERCEPTION-LONDON- 26, 645–666.

Rhodes, G., Jeffery, L., Watson, T.L., Clifford, C.W., Nakayama, K., 2003. Fitting the mind to the world face adaptation and attractiveness aftereffects. Psychological Science 14, 558–566.

Roberts, B., Bregman, A.S., 1991. Effects of the pattern of spectral spacing on the perceptual fusion of harmonics. The Journal of the Acoustical Society of America 90, 3050–3060.

Roberts, T.P., Ferrari, P., Stufflebeam, S.M., Poeppel, D., 2000. Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception. Journal of Clinical Neurophysiology 17, 114–129.

Rupp, A., Uppenkamp, S., Gutschalk, A., Beucker, R., Patterson, R.D., Dau, T., Scherg, M., 2002. The representation of peripheral neural activity in the middle-latency evoked field of primary auditory cortex in humans. Hearing research 174, 19–31.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., Simola, J., 1991. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neuroscience letters 127, 141–145.

Samuel, A.G., Pitt, M.A., 2003. Lexical activation (and other factors) can mediate compensation for coarticulation. Journal of Memory and Language 48, 416–434.

Schnupp, J.W., Honey, C., Willmore, B.D., 2013. Neural correlates of auditory object perception, in: Neural Correlates of Auditory Cognition. Springer, pp. 115–149.

Scholl, B., Gao, X., Wehr, M., 2010. Nonoverlapping sets of synapses drive on responses and off responses in auditory cortex. Neuron 65, 412–421.

Schreiner, C.E., Raggio, M.W., 1996. Neuronal responses in cat primary auditory cortex to electrical cochlear stimulation. ii. repetition rate coding. Journal of neurophysiology 75, 1283–1300.

Schwartz, J.L., Grimault, N., Hupé, J.M., Moore, B.C., Pressnitzer, D., 2012. Multistability in perception: binding sensory modalities, an overview. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 896–905.

Shamma, S.A., 1985. Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. The Journal of the Acoustical Society of America 78, 1622–1632.

Shamma, S.A., Elhilali, M., Micheyl, C., 2011a. Temporal coherence and attention in auditory scene analysis. Trends in neurosciences 34, 114–123.

Shamma, S.A., Elhilali, M., Micheyl, C., 2011b. Temporal coherence and attention in auditory scene analysis. Trends in neurosciences 34, 114–123.

Shamma, S.A., Micheyl, C., 2010. Behind the scenes of auditory perception. Current opinion in neurobiology 20, 361–366.

Shao, Y., Wang, D., 2008. Robust speaker identification using auditory features and computational auditory scene analysis, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE. pp. 1589–1592.

Shepard, R.N., 1964. Circularity in judgments of relative pitch. The Journal of the Acoustical Society of America 36, 2346–2353.

Shinn-Cunningham, B.G., Best, V., 2008. Selective attention in normal and impaired hearing. Trends in Amplification .

Shumway, R.H., Stoffer, D.S., 1982. An approach to time series smoothing and forecasting using the em algorithm. Journal of Time Series Analysis 3, 253–264.

Simon, J.Z., Wang, Y., 2005. Fully complex magnetoencephalography. Journal of neuroscience methods 149, 64–73.

Smith, A.C., Brown, E., 2003. Estimating a state-space model from point process observations. Neural Computation 15, 965–991.

Smith, A.C., Frank, L.M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A.M., Suzuki, W.A., Brown, E.N., 2004. Dynamic analysis of learning in behavioral experiments. The Journal of Neuroscience 24, 447–461.

Snyder, J.S., Alain, C., 2007. Toward a neurophysiological theory of auditory stream segregation. Psychological bulletin 133, 780.

Snyder, J.S., Alain, C., Picton, T.W., 2006. Effects of attention on neuroelectric correlates of auditory stream segregation. Journal of cognitive neuroscience 18, 1–13.

Somers, D.C., Dale, A.M., Seiffert, A.E., Tootell, R.B., 1999. Functional mri reveals spatially specific attentional modulation in human primary visual cortex. Proceedings of the National Academy of Sciences 96, 1663–1668.

Srinivasan, R., Russell, D.P., Edelman, G.M., Tononi, G., 1999. Increased synchronization of neuromagnetic responses during conscious perception. The Journal of Neuroscience 19, 5435–5448.

Steinschneider, M., Nourski, K.V., Fishman, Y.I., 2013. Representation of speech in human auditory cortex: is it special? Hearing research 305, 57–73.

Stephens, J.D., Holt, L.L., 2003. Preceding phonetic context affects perception of nonspeech (l). The Journal of the Acoustical Society of America 114, 3036–3039.

Suppes, P., Han, B., 2000. Brain-wave representation of words by superposition of a few sine waves. Proceedings of the National Academy of Sciences 97, 8738–8743.

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

Sussman, E.S., 2007. A new view on the mmn and attention debate. Journal of Psychophysiology 21, 164–175.

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

Taaseh, N., Yaron, A., Nelken, I., 2011. Stimulus-specific adaptation and deviance detection in the rat auditory cortex. PLoS One 6, e23369.

Tanner, M., . Tools for statistical inference, 1996.

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

Theunissen, F.E., Sen, K., Doupe, A.J., 2000. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. The Journal of Neuroscience 20, 2315–2331.

Thoma, R.J., Hanlon, F.M., Moses, S.N., Ricker, D., Huang, M., Edgar, C., Irwin, J., Torres, F., Weisend, M.P., Adler, L.E., et al., 2005. M50 sensory gating predicts negative symptoms in schizophrenia. Schizophrenia research 73, 311–318.

Tiitinen, H., May, P., Näätänen, R., 1997. The transient 40-hz response, mismatch negativity, and attentional processes in humans. Progress in Neuro-Psychopharmacology and Biological Psychiatry 21, 751–771.

Ulanovsky, N., Las, L., Farkas, D., Nelken, I., 2004. Multiple time scales of adaptation in auditory cortex neurons. The Journal of Neuroscience 24, 10440–10453.

Ulanovsky, N., Las, L., Nelken, I., 2003. Processing of low-probability sounds by cortical neurons. Nature neuroscience 6, 391–398.

Vliegen, J., Moore, B.C., Oxenham, A.J., 1999. The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. The Journal of the Acoustical Society of America 106, 938–945.

Wang, D., Brown, G.J., 2006. Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE Press.

Wang, Y., Han, K., Wang, D., 2013. Exploring monaural features for classification-based speech segregation. Audio, Speech, and Language Processing, IEEE Transactions on 21, 270–279.

Warren, R.M., Gregory, R.L., 1958. An auditory analogue of the visual reversible figure. The American journal of psychology .

Webster, M.A., Kaping, D., Mizokami, Y., Duhamel, P., 2004. Adaptation to natural facial categories. Nature 428, 557–561.

Wehr, M., Zador, A.M., 2005. Synaptic mechanisms of forward suppression in rat auditory cortex. Neuron 47, 437–445.

Winkler, I., Czigler, I., Sussman, E., Horváth, J., Balázs, L., 2005. Preattentive binding of auditory and visual stimulus features. Journal of Cognitive Neuroscience 17, 320–339.

Winkler, I., Kushnerenko, E., Horváth, J., Čeponienė, R., Fellman, V., Huotilainen, M., Näätänen, R., Sussman, E., 2003. Newborn infants can organize the auditory world. Proceedings of the National Academy of Sciences 100, 11812–11815.

Wissig, S.C., Kohn, A., 2012. The influence of surround suppression on adaptation effects in primary visual cortex. Journal of neurophysiology 107, 3370–3384.

Xiang, J., Simon, J., Elhilali, M., 2010. Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. The Journal of Neuroscience 30, 12084–12093.

Xu, Y., Chun, M.M., 2009. Selecting and perceiving multiple visual objects. Trends in cognitive sciences 13, 167–174.

Yost, W.A., 1994. Fundamentals of hearing: An introduction . Academic Press.

Zhang, Q.f., Wen, Y., Zhang, D., She, L., Wu, J.y., Dan, Y., Poo, M.m., 2012. Priming with real motion biases visual cortical response to bistable apparent motion. Proceedings of the National Academy of Sciences 109, 20691–20696.

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a cocktail party. Neuron 77, 980–991.