# ABSTRACT

Title of Dissertation: CORTICAL REPRESENTATION OF SPEECH IN COMPLEX AUDITORY ENVIRONMENTS AND APPLICATIONS

Venkata Naga Krishna Chaitanya Puvvada
Doctor of Philosophy, 2017

Dissertation directed by: Professor Jonathan Z. Simon
Department of Electrical and Computer Engineering

Being able to attend and recognize speech or a particular sound in complex listening environments is a feat performed by humans effortlessly. The underlying neural mechanisms, however, remain unclear and cannot yet be emulated by artificial systems. Understanding the internal (cortical) representation of external acoustic world is a key step in deciphering the mechanisms of human auditory processing. Further, understanding neural representation of sound finds numerous applications in clinical research for psychiatric disorders with auditory processing deficits such as schizophrenia.

In the first part of this dissertation, cortical activity from normal hearing human subjects is recorded, non-invasively, using magnetoencephalography in two different real-life listening scenarios. First, when natural speech is distorted by reverberation as well as stationary additive noise. Second, when the attended speech is degraded by the presence of multiple additional talkers in the background, simulating a cocktail party. Using natural speech affected by reverberation and noise, it was demonstrated that the auditory cortex maintains both distorted as well as distortion-free representations of speech. Additionally, we show that, while the neural representation of speech remained robust to additive noise in absence of reverberation, noise had detrimental effect in presence of reverberation, suggesting differential mechanisms of speech processing for additive and reverberation distortions. In the cocktail party paradigm, we demonstrated that primary like areas represent the external auditory world in terms of acoustics, whereas higher-order areas maintained an object based representation. Further, it was demonstrated that background speech streams were represented as an unsegregated auditory object. The results suggest that object based representation of auditory scene emerge in higher-order auditory cortices.

In the second part of this dissertation, using electroencephalographic recordings from normal human subjects and patients suffering from schizophrenia, it was demonstrated, for the first time, that delta band steady state responses are more affected in schizophrenia patients compared with healthy individuals, contrary to the prevailing dominance of gamma band studies in literature. Further, the results from this study suggest that the inadequate ability to sustain neural responses in this low frequency range may play a vital role in auditory perceptual and cognitive deficit mechanisms in schizophrenia.

Overall this dissertation furthers current understanding of cortical representation of speech in complex listening environments and how auditory

representation of sounds is affected in psychiatric disorders involving aberrant auditory processing.

# CORTICAL REPRESENTATION OF SPEECH IN COMPLEX AUDITORY ENVIRONMENTS AND APPLICATIONS

by

Venkata Naga Krishna Chaitanya Puvvada

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
2017

Advisory Committee:
Professor Jonathan Z. Simon, Chair
Professor Carol Espy-Wilson
Assistant Professor Behtash Babadi
Assistant Professor Samira Anderson
Professor Sandra Gordon-Salant

# Dedication

*To my mother Guramma*

# Acknowledgements

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 my advisor, Jonathan, for his guidance and unwavering support through out the duration of this dissertation. I would like to thank you for introducing to me to the wonderful world of auditory neuroscience and providing me with numerous opportunities to explore my scientific interests and joining me along the path.

To Dr. Elliot Hong, Thank you for your insights and help with the data analysis of Schizophrenia study in this dissertation.

To Dr. Carol Espy-Wilson, Dr. Behtash Babadi, Dr. Samira Anderson and Dr. Sandra Gordon-Salant. Thank you for serving on this dissertation committee and Dr. Shihab Shamma for serving on the research proposal committee.

To Christian Brodbeck, Sahar Akram, Marisel Villfane-Delgado, Francisco Cervantes Constantino. Thank you for your insightful discussions over the years and most importantly, thank you for your friendship.

To all my friends Praneeth, Varun, RaviTeja, Anup, Sofia, Ishwar, Kapil, Deepthi, Navaneeth, Pradeep, Shashi, Bharat, Ranchu, Jonathan. Thank you for sharing very many wonderful moments with me. Life would not have been nearly as colorful without your presence.

Lastly, I would like to thank my parents for a lifetime of continuous support, motivation and wisdom and for believing in every dream I had and giving me the greatest reason of all to succeed. To my wonderful brother and sister, Anand and Sowjanya for being a great source of support, happiness and joy in my life. To you, I will be forever grateful.

# Contents

# List of Figures

# 1 Introduction

Sound is one of the dominant forms of how we perceive the external world around us, alongside of vision. Speech, a class of sounds, is the vital form of human communication, on its way to become the most important form of human-machine interaction. The ability to attend to and perceive speech in complex listening conditions such as noisy reverberant environments or in presence of additional talkers are mathematically ill-posed problems, yet routinely solved by human brain robustly and with little effort. Such robustness, however, is lost with hearing impairment (Festen and Plomp, 1990; Marrone et al., 2008b, a) and cannot yet be achieved by artificial speech recognition systems (Lippmann, 1997; Cooke et al., 2010; Davis and Scharenborg, 2016), despite recent success of recurrent neural networks in speech recognition (Yu and Deng, 2014). Identifying the neural representations of speech in complex listening conditions is the first key step in understanding the mechanism of respective auditory processing in humans. This is not only of great interest in auditory neuroscience, but also has potential applications in artificial speech recognition systems as well as design of better hearing aids for the impaired. Apart from basic scientific and technological advancement, understanding cortical processing of sounds finds numerous applications in medical community. Deficiencies in cortical processing of sounds can be indicative of certain mental disorders (Iliadou and Iakovides, 2003). For example, auditory hallucinations, which are hallmark feature of schizophrenia, are thought to be cause by aberrant cortical processing and perception of sounds (McLachlan et al., 2013) and hence understanding neural representation of sounds can help in designing new diagnostic measures for psychiatric/mental disorders.

Recognition of speech by humans relies on its spectro-temporal modulations (i.e., variations of energy over time and frequency scales (Chi et al., 1999; Chi et al., 2005)), with slow temporal modulations (<10 Hz) reflecting the syllabic and phrasal structure of speech (Greenberg et al., 2003; Poeppel et al., 2008; Chait et al., 2015) and fast temporal modulations (>100 Hz) indicating the pitch and carrier information (Chi et al., 2005). Non-invasive neuro-imaging techniques such as magnetoecephalography (MEG) (Hamalainen et al., 1993) and electroencephalography (EEG) (Niedermeyer and Lopes da Silva, 2005) are sensitive to activity in human cortex and have millisecond time resolution, enough to resolve the neural activity phase locked to the slow temporal modulations in speech (Ding and Simon, 2009; Wang et al., 2012). Utilizing the neural representations of these slow temporal modulations, this dissertation investigates how complex auditory scenes are encoded, from the mixture of the entire acoustic scene, to its separate individual sources in different areas of auditory cortex, with a special emphasis on speech. We focus on two such complex auditory scenes, speech in the presence of reverberation as well as noise and speech in the presence of multiple background talkers. The investigation focuses on neural mechanisms that employ temporal encoding, allowing us to exploit the strong temporal nature of speech. The slow temporal and spectral modulations in speech and other natural sounds are most critical for their perception (Shannon et al., 1995; Sheft, 2008; Elliott and Theunissen, 2009). Hence, we also investigate the deficiencies in neural representation of sounds at slow rates using auditory steady state response (ASSR) paradigm in Schizophrenia patients, in an effort to discover new objective measures for otherwise subjectively diagnosed disorder. This dissertation consists of three studies and the rest of this chapter presents its organization.

## 1.1 Outline

A review of human auditory processing, neuro-imaging techniques (MEG & EEG) and analysis methods used in this dissertation are provided in Chapter 2.

The first study (Chapter 3) addresses the neural representation of speech in reverberant as well as noisy conditions. Using natural speech distorted by reverberation and spectrally matched additive noise at varying degrees of severity, it is demonstrated that auditory cortex maintains both distorted as well as distortion free representations of speech. Further, these neural representations remained robust to additive noise in absence of reverberation but demonstrated a detrimental effect in presence of reverberation suggesting differential encoding mechanisms for additive and convolutive (reverberation) distortions of speech.

The second study (Chapter 4) addresses the neural representation of an auditory scene, in a multi-source scenario. Using cortical tracking of (continuous) speech, in a multi-talker auditory scene, it is demonstrated that the early neural responses, which primarily originate from core auditory regions, represent the foreground (attended) and background (unattended) speech streams with no significant difference, whereas the late neural responses, which typically originate from higher order auditory regions, represent foreground with significantly higher fidelity than background. Further, it is shown that while early neural responses represent auditory scene in terms of acoustics, the late responses maintain an object-based representation. Additionally, it is shown that even though there is more than one unattended speech stream, the neural representation of background remains unsegregated. *This study has been accepted for publication by the Journal of Neuroscience.*

The third study (Chapter 5) addresses low frequency auditory synchronization deficiencies in schizophrenic patients compared with normal subjects and first-degree relatives of patients. Using electroencephalography recordings in an ASSR paradigm, it is demonstrated that patients exhibit a greater reduction of ASSR power in delta band (2.5 Hz) compared with traditionally investigated gamma band (40 Hz) responses. Critically, reduced delta band ASSR in patients was associated both with more severe longitudinally experienced auditory deficits and also poorer verbal working memory, supporting the use of low frequency ASSR to study the etiology of schizophrenia. *This study has been published in Schizophrenia Bulletin (Puvvada et al., 2017).*

Finally, chapter 6 provides a summary of findings of the current work, along with possible future directions to the studies presented in this dissertation.

# 2 Background:

## 2.1 Human auditory system

### 2.1.1 Auditory pathway

The path through the nervous system of how sound is processed in humans, known as the *auditory pathway*, starts at the ear and proceeds through the cortex. Auditory system can mainly be decomposed into two parts, peripheral and central auditory systems. Here we provide a brief overview of major organs involved and their functional role in forming perception from sound.

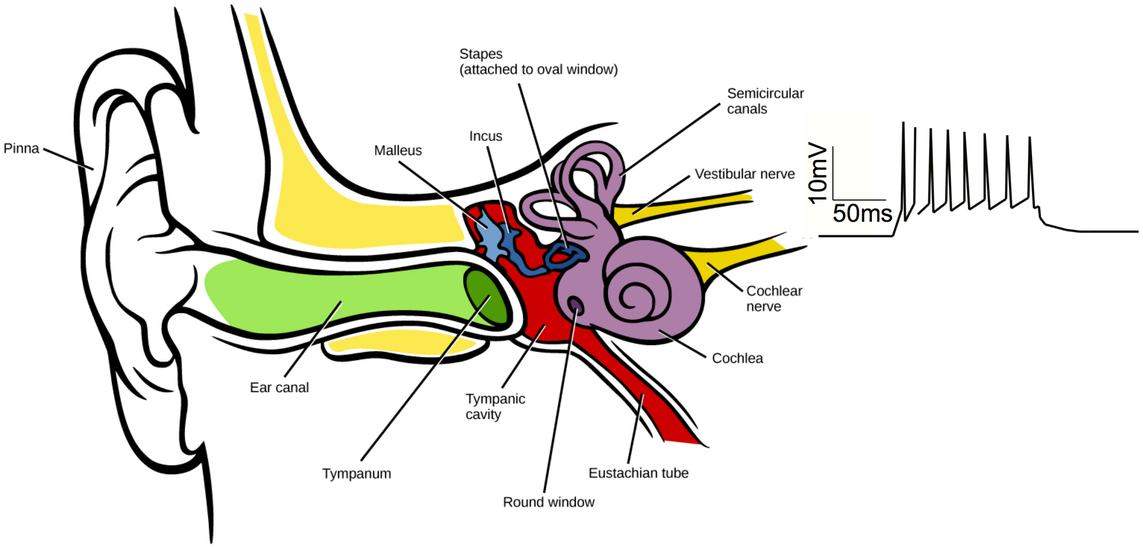

Figure 2.1: A schematic view of peripheral auditory system

Figure 2.1 shows a schematic of peripheral auditory system, consisting mainly of Pinna, Ear canal, Tympanic membrane (Ear drum), Ossicles (Malleus, Incus and Stapes), Cochlea and Auditory/cochlear nerve. The function of the Pinna is to perform spectral transformation through spatial filtering, amplifying the incoming sound in the spectral range of human speech. Also, spatial filtering by Pinna embeds a spectral notch, an important cue in vertical sound localization (Hebrank and Wright, 1974). Further amplification of incoming sound, in speech related spectral range, is performed by Ear canal and Ear drum before delivering the information to Ossicles in the form of mechanical vibrations, through Tympanic membrane, which in turn transfers the acoustic information to fluid-movements in Cochlea. Sound is transferred into (electrical) neural activity in Cochlea. Basilar membrane, a tonotopically arranged structural element that separates two liquid-filled tubes and runs along the Cochlea acts as a filter bank (Figure 2.2, adapted from Ding (2012)). Stretched horizontally, Basilar membrane is about 35 mm long, with its base tuned to high frequencies and apex tuned to low frequencies. The (inner) hair cells on the basilar membrane transform the fluid vibrations into electrical neural activity, which in turn is picked up by Auditory nerve. (Yang et al., 1992; Chi et al., 2005; Zilany et al., 2014) provide computational models of peripheral auditory system.

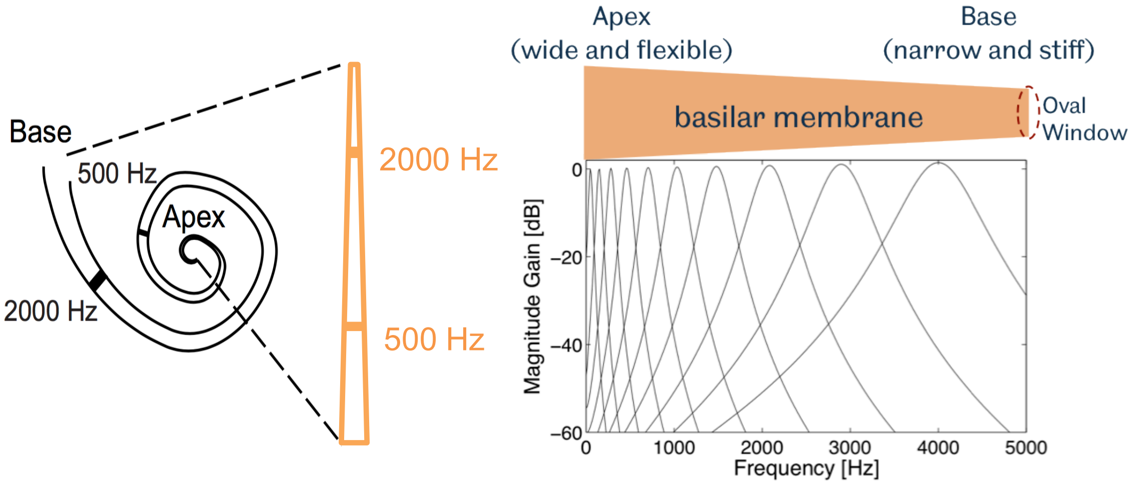

Figure 2.2: A schematic of basilar membrane and filter bank

In the auditory nerve, each nerve fiber is tuned to a specific frequency and particular range of loudness. The neural representations of sound in auditory nerve are then processed by a series of nuclei (neural networks in central auditory system) located in brainstem and thalamus before reaching cortex. The first station auditory nerve innervates is Cochlear Nucleus (CN). The auditory information is further processed by Superior Olivary Complex (SOC), Inferior Colliculus, and Thalamus, before reaching Auditory cortex (AC). While SOC is known to the place of ITD (inter-aural time difference) and ILD (inter-aural level difference) computation, IC acts as a major integration center, receiving (parallel) ascending inputs from auditory nuclei in lower part of brain stem, descending inputs from auditory cortex and inputs from contralateral IC. Overall, these sub-cortical nuclei refine the temporal synchronization of neural responses from auditory nerve (Joris et al., 1994), extract information related to pitch, directionality, integrate inputs from both ears and possibly from somatosensory system (Pickles, 2012) before relaying the acoustic information to cortex. As in cochlea, majority of these sub-cortical nuclei exhibit frequency tuning and tonotopy (Pickles, 2012). Also, it is to be noted that temporal precision of neurons decreases gradually from auditory nerve to cortex (Giraud et al., 2000). Neural phase locking can be observed up to 1 kHz in auditory nerve, up to approximately 200 Hz in thalamus and generally below 40 Hz in the cortex.

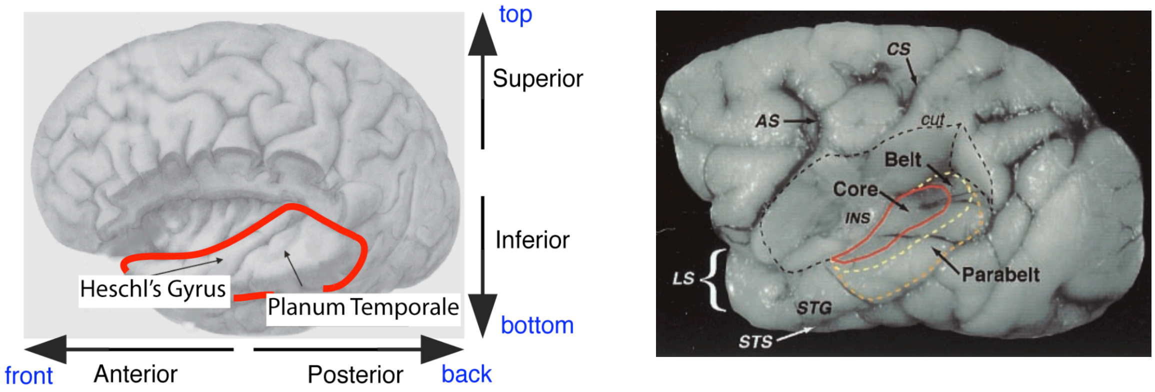

Figure 2.3: Anatomy of primate auditory cortex (adapted from Kaas and Hackett (2000)).

In humans, the auditory cortex is located in the superior temporal lobe (Figure 2.3). It can be divided into three regions namely, core, belt and parabelt regions (Hackett et al., 1998; Kaas and Hackett, 2000). The core region, containing primary auditory cortex (PAC/A1) and other primary-like areas, is centered on Heschl's gyrus (HG) (Da Costa et al., 2011). It is surrounded by the belt and parabelt regions, and by additional higher order auditory regions (Hackett, 2008). The core auditory cortex receives direct input from the thalamus and contains tonotopically organized neurons responding to wide variety of sounds, including speech. A1 encodes spectrotemporal modulations (Depireux et al., 2001), among other acoustic properties, and its responses are affected consistently but modestly by selective attention (Fritz et al., 2003; Fritz et al., 2005). The core region is the origin of MEG M50 response (Yvert et al., 2001), a positive deflection in MEG response around 50 ms after a sound onset, known to be task independent (Chait et al., 2004). Posterior to Heschl's gyrus lies Planum Temporale (PT), containing both parabelt and additional higher order auditory regions (Griffiths and Warren, 2002; Sweet et al., 2005). The PT is supposed to play an important role in high level auditory processing, for instance, identification and segregation of an auditory source in a mixture of acoustic sources (Griffiths and Warren, 2002). PT (Lutkenhoner and Steinstrater, 1998) and/or lateral part of HG (Herdman et al., 2003) are known to be the sources of MEG M100 response, a strong negative response occurring 100 ms post sound onset, strongly modulated by attention (Chait et al., 2004).

### 2.1.2 A Neuron

Neuron is the basic functional unit of central nervous system. It is separated from extracellular medium by a cell membrane and typically consists of three main parts: dendrites, soma (cell body) and an axon (Figure 2.4), which help in receiving, integrating and transmitting the information respectively. The soma is typically tens of microns in diameter (Dayan and Abbott, 2001) and the axon has a typical thickness of few microns (less than 20)  in diameter. A neuron connects to other neuron through axon via synaptic cleft. While the information is transmitted electrically in axon, it is transmitted chemically to the next neuron through neurotransmitters.

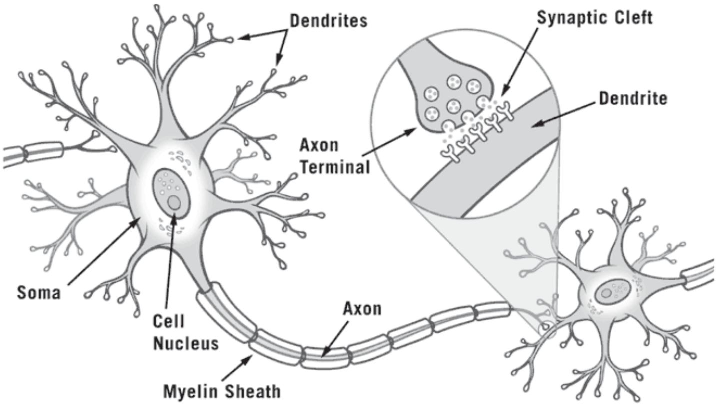

Figure 2.4:  Structure of a neuron

The difference in electrical potential between the interior and exterior (extracellular medium) of a cell is called Membrane potential and the typical resting state membrane potential for a neuron is about -70 mV (exterior is more positive than interior). The dendrites of a neuron receive input from axons (usually from multiple other neurons) via synapses. Action potentials received by a dendrite causes voltage change in dendrite, called post-synaptic potential (PSP). An excitatory post-synaptic potential (EPSP) causes depolarization (reduction of resting state potential) and an inhibitory post-synaptic potential (IPSP) causes hyperpolarization (further increase of resting state potential). When the net voltage change due to spatial and temporal integration of PSPs received by dendrites of a neuron, from multiple pre-synaptic neurons, reaches a threshold, an action potential is fired by the neuron. After firing an action potential, the neuron is ready to fire again after a brief refractory period. Thus the output of neuron is a series of action potentials, which are brief (1-2 ms in duration) voltage changes that propagate along the axon. The number of action potentials generated per second is called *firing rate,* which can be as high as few hundred hertz. The post-synaptic potential also generates current in the dendrites flowing towards soma (Figure 2.5A), called *dendritic current,* typically on the order of fA ($10^{-15}$ Ampere) (Hämäläinen et al., 1993). This dendritic current is the source of electromagnetic fields measured outside the brain. In the following, we provide an introduction to non-invasive neuroimaging and M/EEG sources and measurement.

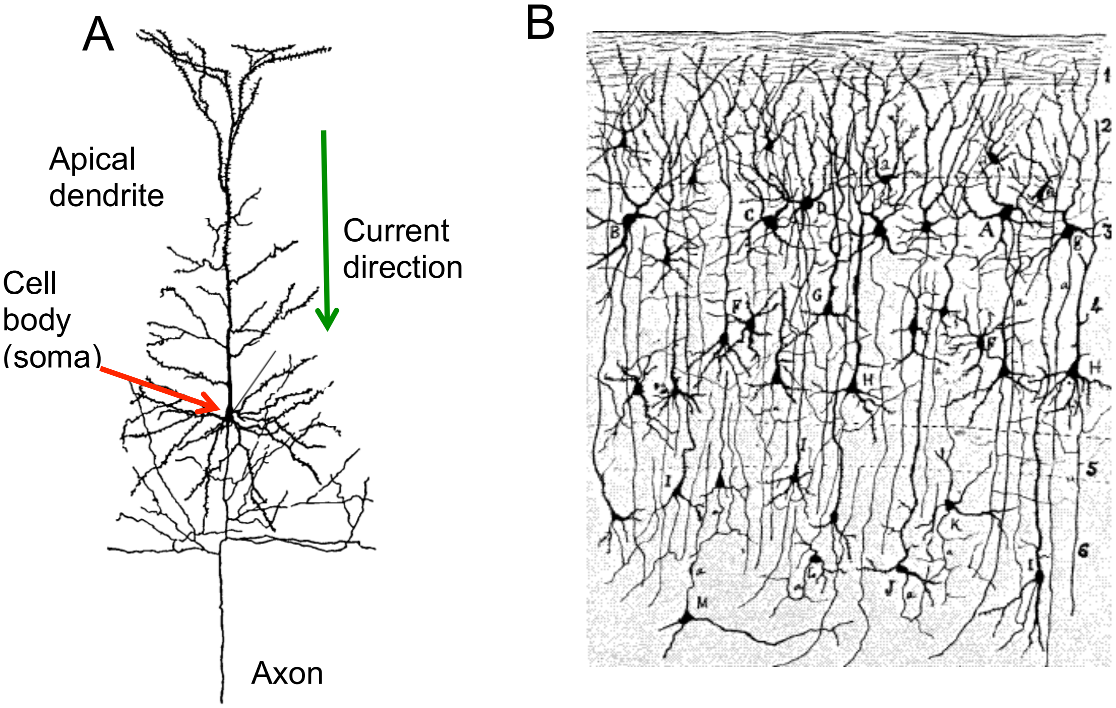

Figure 2.5: (A) A cortical pyramidal neuron. (B) Network of neurons in cortex (adapted from Dayan and Abbott (2001))

## 2.2 Noninvasive neuro-imaging

With at least 10 billion neurons and 10,000 times more connections among them, the human brain is the most complex structure known to exist and the most important organ for humans. It is the seat of information processing, sensing, memory, conscious and unconscious thought, action planning and so on. Investigation of the human brain is great intellectual interest, not only because of the scientific curiosity but also has many practical applications. For example, deep neural networks (LeCun et al., 1998; Bengio, 2009), the current gold standard in classification and recognition systems are inspired from neural networks in the brain. Neuromorphic computation aimed at consuming ultra-low power for a wide array of real-time application is inspired from brain architecture

(Arthur et al., 2012). Also, understanding of the brain has widespread applications in treating psychiatric disorders. Although a great deal about anatomy and physiology of the brain is known, the question of how the brain stores, retrieves and processes information is still largely unknown. Most of our current understanding of the brain comes from studying animal models. Although quite useful, animal models can never replace studies on humans when studying aspects such as speech, which are human specific. While human brain activity may be recorded by inserting electrodes (invasive recording) during surgery on patients suffering from drug-resistant epilepsy or tumors, such studies are restrictive in terms of recording area and are extremely time consuming. Invasive recording can never be performed on healthy humans, due to ethical reasons and hence non-invasive recording techniques provide an excellent alternative for studying human brain. Non-invasive recording techniques can be mainly categorized into two types. The first class of methods directly measure electrical activity associated with neuronal firing such as EEG and MEG. The second class of methods such as Positron emission tomography (PET), functional magnetic resonance imaging (fMRI) measure neuronal activity indirectly, based on the principle that increased neural activity is supported by increased local blood flow and metabolic activity. The dynamics of blood flow (hemodynamic activity) are much slower than dynamics of neural activity and hence PET and fMRI have a time resolution of less than 1 Hz, whereas EEG and MEG can be recorded as fast as 1 ms and can resolve phase locked neural responses to slow temporal modulations of speech (1-16 Hz). In the following we discuss the physiological sources of EEG and MEG and their instrumentation.

### 2.2.1 Physiological sources of electromagnetic fields in brain

Synchronous post-synaptic potentials (PSP) (both excitatory and inhibitory) in pyramidal neurons in cortex are the main source of electric and magnetic activity measured by EEG and MEG respectively. To explain this, consider the cartoon presented in Figure 2.6 (adapted from Niedermeyer and Lopes da Silva (2005)). At rest, membrane potential (potential difference between intracellular and extracellular space) of a neuron is about -70 mV and is thus maintained due to opposing electrical and chemical gradients. The inside of neuron is more negative than outside of neuron. This is represented by negative charge inside and positive charge outside of the neuron at rest. Consider an action potential arriving at the excitatory synapse. The synapse releases neurotransmitters, which opens Na+ ion channels causing the positively charged ions to flow into the cell (step 1). This inward current flow changes the local membrane potential and creates an electrical gradient, along the length of dendride, inside as well as outside of the cell (step 2). This electrical gradient sets up intracellular (primary) currents and extracellular (secondary or volume) currents (step 3).  MEG is sensitive to magnetic fields caused by primary current (mainly) and secondary currents, whereas EEG is sensitive to potential setup by secondary/volume currents.

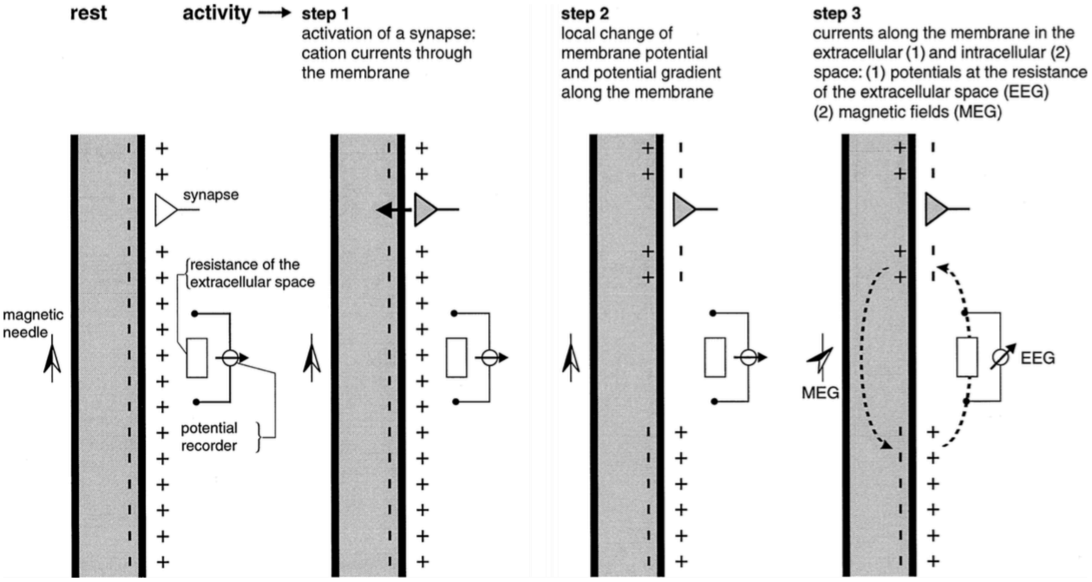

Figure 2.6: Generation of dendritic current in pyramidal neurons (adapted from Niedermeyer and Lopes da Silva (2005)).

The dipole moment generated due to primary current flow in a pyramidal neuron due to a single PSP is about 20 fA m (Hamalainen et al., 1993) and is directed along the length of neuron as shown in the Figure 2.5A. Usually, a dipole moment on the order of 10 nA m is required to explain the measured magnetic field strengths outside the head. Pyramidal neurons are a common type of neurons in cortex and some dendrites of these pyramidal neurons, called apical dendrites, are roughly perpendicular to the surface of the cortex. Typically, neurons are never present in isolation but interconnected with other neurons, forming networks (Figure 2.5B). In each mm$^2$ of cortex there are roughly 10$^5$ neurons (Hämäläinen et al., 1993). This high density of neurons combined with the fact that apical dendrites of neighboring neurons are approximately parallel implies that the dendritic currents in a local neural network flows in a very similar direction. Thus, when dendritic current in many neurons are synchronized in time (e.g., due to the presence of a stimulus) they add up to a current source that generates a magnetic field strong enough to be measured outside the brain (extracranially). Measurement of this magnetic field is called magnetoencephalography (MEG) and the electrical potential generated due to the corresponding volume currents can be measured extracranially using electroencephalography (EEG). The advantage of MEG/EEG is that they can be measured noninvasively. But, due their requirement of synchronous neural activity over tens of thousands of neurons they have limited spatial resolution (mm to cm).

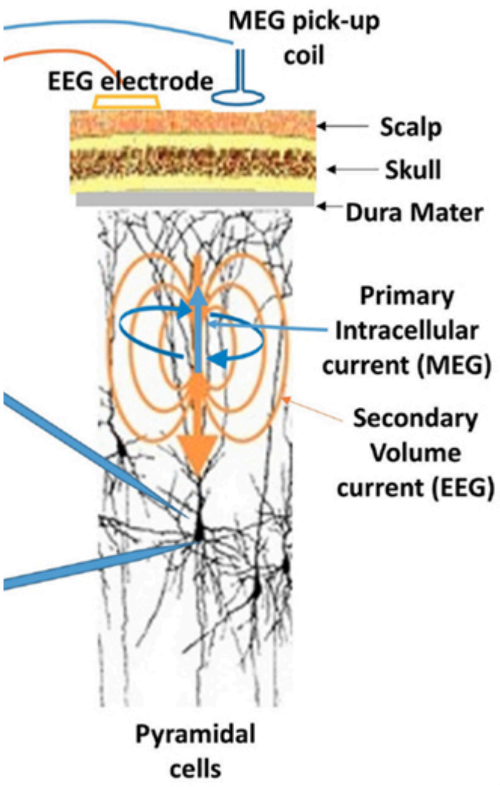
MEG pick-up coil
EEG electrode
Scalp
Skull
Dura Mater
Primary Intracellular current (MEG)
Secondary Volume current (EEG)
Pyramidal cells

Figure 2.7: Primary source of EEG and MEG

Although, the source of both MEG and EEG are post-synaptic potentials, there are two main differences between them worth mentioning, which arise due to the nature of (vector) magnetic and (scalar) electric fields. As mentioned previously, pyramidal cells are oriented perpendicular to cortical surface, which implies that a dipole in gyri is radial to the skull and a dipole in sulci is tangential to the skull, as shown in Figure 2.8 (left). A radial dipole produces no external magnetic field (outside the skull) where as a tangential dipole produces measurable external magnetic field (Figure 2.8, right). In contrast, both radial and tangential dipoles produce measurable surface potentials. Therefore MEG does not see sources in gyri, which are radial, but sees the dipoles in sulci clearly. Another consequence of this dipole orientation (radial vs. tangential) is that the deeper the sources more radial they are (e.g. a dipole at the center of a sphere is always radial). Thus, deeper cortical sources produce more suppressed external magnetic fields. In contrast, potentials measured using EEG due to deeper sources are not as suppressed as magnetic fields. Thus, in general, dipoles located in gyri and dipoles located deeply are more suppressed in MEG compared to EEG. Another important difference between MEG and EEG is due to smearing of potential by the resistivity of skull and various brain tissues whereas the magnetic field is not affected by the same. This implies that MEG source localization is more accurate than EEG source localization. EEG source localization is further made difficult because of the requirement of knowledge of conductivities of various brain tissues. Thus, in general EEG source location accuracy varies about +/- 9 mm whereas MEG can localize with an accuracy of +/- 4 or 5 mm, provided the neural source dipole is tangential.

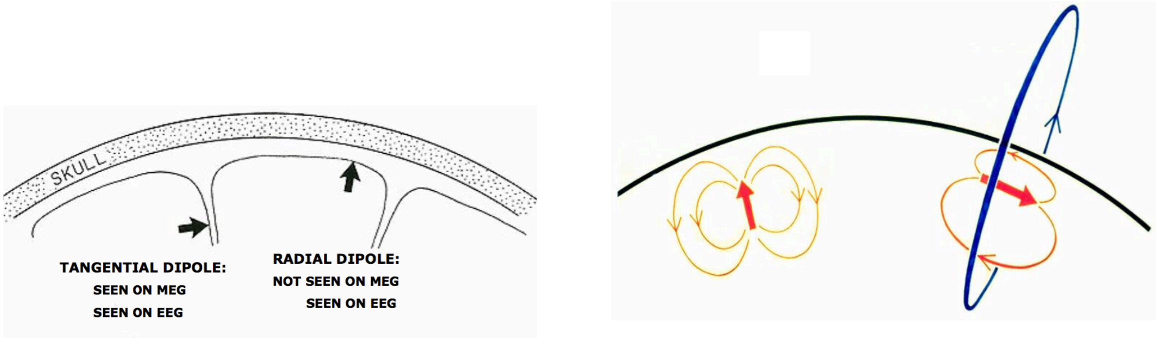

Figure 2.8: Tangential and radial dipoles seen by MEG and EEG

### 2.2.2 Magnetoencephalography

Magnetic field generated by the brain are in the range of 10-100 fT, which is approximately seven orders smaller in magnitude than the earth's magnetic field. Hence special data recording techniques are required to measure magnetic field generated by cortex. A Magnetically shielded room is built around the MEG system, which greatly reduces the interference from external electromagnetic sources, and highly sensitive SQUID (superconducting quantum interference device) sensors are used to measure cortical magnetic fields. The magnetic field is picked up by the *pick-up coils* or *detector coils* and is converted to voltage by SQUID sensors. Typically the pickup coils can be plain magnetometers or gradiometers (Figure 2.9). The simplest of all is the magnetometer, which picks up the magnetic field normal ($B_z$) to its plane. The axial gradiometer measures the difference of $B_z$ in the axial direction z and the planer gradiometer measures the same in tangential direction y.

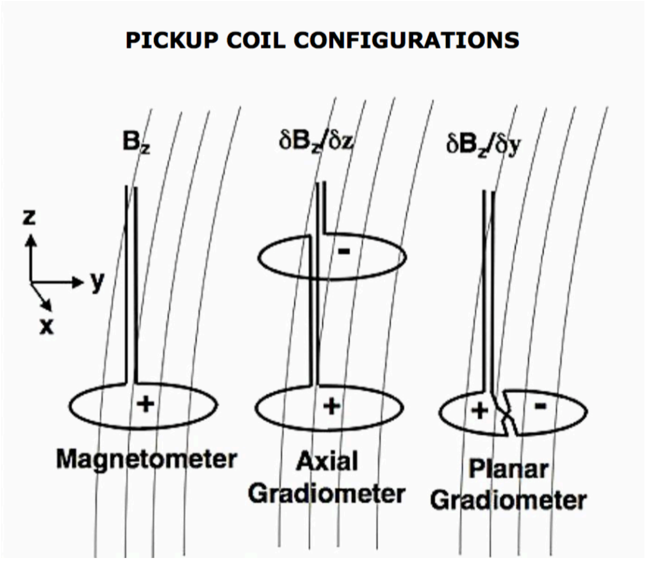

Figure 2.9: Pickup coil configurations in MEG

Modern MEG systems consist of typically 100-300 SQUID sensors distributed over whole head, offering simultaneous recordings (Figure 2.10A). To maintain the super conducting nature of SQUID sensors, they have to be cooled in liquid helium. The MEG studies presented in this dissertation are performed in a magnetically shielded room (MSR) (Yokogawa Electric Corporation) using a 160-channel whole head system (Kanazawa Institute of Technology, Kanazawa, Japan). The pick-up coils are arranged in a uniform array on a helmet-shaped surface (Figure 2.10, adapted from Ding (2012)) with approximately 2.5 cm between the centers of two adjacent 15.5 mm diameter pick-up coils. The pick-up coils are configured as first-order axial gradiometers with 5 cm baseline. Three of the 160 sensors are plain magnetometers located away from the head, and are used to measure ambient magnetic field. These sensors are used as reference channels in de-noising the MEG recordings (de Cheveigne and Simon, 2007). Despite the use of highly sensitive SQUID sensors, the recordings are corrupted by noise due to relatively small nature of magnetic fields produced by sensory or cognitive events compared to variety of noise sources and artifacts (both biological and non-biological). Typical non-biological sources of noise include earth's magnetic field, noise generated by electrical equipment, low frequency fields produced by moving metallic objects such as fans, trains and elevators etc. While most of the noise due to external sources is blocked by MSR, the remaining can be eliminated using reference sensors and de-noising algorithms. Biological sources of noise include magnetic fields generated due to heartbeat, muscle movements, eye-blinks or eye movements, and interference from spontaneous brain activity unrelated to the sensory or cognitive task at hand. Typically spectral and/or spatial filtering (de Cheveigne and Simon, 2008; de Cheveigne and Parra, 2014), component analysis techniques such as Principal component analysis (PCA) and Independent component analysis (ICA) can be used to reduce biological noise. Apart from these, the thermal noise in sensors also contributes to the reduced signal-to-noise ratio (SNR) of the recordings. Apart from de-noising techniques, averaging over multiple trials is one of most common ways of improving SNR of recordings.

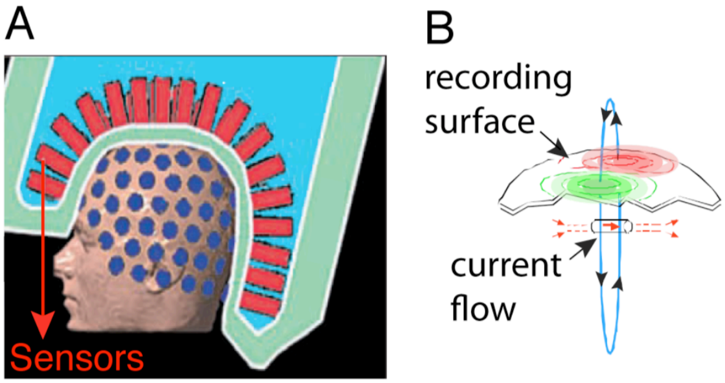

Figure 2.10: An MEG system

**Source estimation in MEG**

The extracranial magnetic field strengths measured in MEG (or electrical potentials in EEG) are superposition of individual fields generated due to multiple (current) sources in the cortex. More than often, we are interested in the location as well as the time course of these cortical sources. Historically, the problem of estimating the sources corresponding to EEG/MEG observations can be viewed as 'localization' or 'imaging'. Localization refers to decomposing observed data into respective contributions of a small number (typically less than ten) of elementary current source models, e.g. point-like equivalent current dipoles, whereas imaging involves tessellating cortical space, either in volume or surface, into a number of patches (~ $10^4$) and estimating the strength of current source, modeled as a dipole with fixed location and possibly variable orientation, in each patch. Thus, localization produces point sources whereas imaging produces a distributed estimate of the cortical activity. While both approaches are mathematically ill-posed inverse problems i.e. given measurements from a number of sensors (~ $10^2$) and an observation model, the goal is to estimate the cortical source activity responsible for the observed measurements, the problem is further complicated in imaging due to its high-dimensionality. In the following, we provide a brief description of Minimum Norm Estimate (MNE) (Hamalainen and Ilmoniemi, 1994), an imaging method, employed in this dissertation, which is also most widely used in general.

Let $N$ denote the number of sensors and $t = 1,...,T$ be the discrete time stamps. Let $y_{i,t}$ be the measurement from *i*th sensor at time *t* and $\mathbf{y}_t := [y_{1,t}, y_{2,t}, \cdots y_{N,t}]'$ be measurements from all sensors at time *t*. Finally let $\mathbf{Y} := [\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y}_T]$ be $N \times T$ dimensional measurement matrix. Let *M* be the total number of dipoles distributed across the cortex and $x_{i,t}$ be the amplitude of *i*th dipole at time *t*. Let $\mathbf{x}_t := [x_{1,t}, x_{2,t}, \ldots x_{M,t}]'$ denote the vector of dipole amplitudes at time *t* and $\mathbf{X} := [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x}_T]$ be $M \times T$ dimensional matrix characterizing source space over time [0, *T*]. With this notation, for a fixed configuration of diploes, the measurement matrix $\mathbf{Y}$ can be related to the source activity matrix $\mathbf{X}$ as follows:

$$\mathbf{Y} = \mathbf{GX} + \mathbf{V} \tag{2.1}$$

where $\mathbf{G}_{\mathrm{N \times M}}$ is the *lead field* matrix relating the source amplitudes to sensor measurements and $\mathbf{V}_{\mathrm{N \times T}}$ is the measurement noise matrix, assumed to be zero mean Gaussian distributed with spatial covariance matrix $\mathbf{C}_{\mathrm{N \times N}}$. Empirically, noise covariance matrix can be computed from empty room or pre-trial sensor measurement data. Further, due to the ill-posed nature of the problem, a regularizer in the form of spatial prior covariance matrix $\mathbf{R} / \lambda^{\mathbf{2}}$ of dimensions $M \times M$ is imposed on $\mathbf{X}$, with $\lambda$ being a scaling factor to control the $\mathbf{R}$-weighted $l_2$-norm of the estimate. With these definitions, the minimum norm estimate is defined as:

$$\widehat{\mathbf{X}} := \underset{\mathbf{X}}{\arg\min} \sum_{t=1}^{T} \left\{ \| \mathbf{y}_t - \mathbf{G}\mathbf{x}_t \|_{\mathbf{C}^{-1}}^{2} + \lambda^2 \| \mathbf{x_t} \|_{\mathbf{R}^{-1}}^{\mathbf{2}} \right\} \tag{2.2}$$

The minimization is separable in time and yields a closed form solution as:

$$\widehat{\mathbf{X}} = \mathbf{RG}'(\mathbf{GRG}' + \lambda^2 \mathbf{C})^{-\mathbf{1}} \mathbf{Y} \tag{2.3}$$

**Auditory studies using MEG**

The auditory cortices are located in the lateral sulcus, which produces tangential dipoles, making MEG ideal for recording the magnetic fields generated by auditory cortices. Apart from this, the setup time for MEG is short making it an ideal candidate for short turnaround time of experiments.

One of the earliest and extensively studied (transient) MEG responses is to onset/offset of a sound (Naatanen and Picton, 1987). The major components of the transient response following an onset (or a click) are defined as 'P1m-N1m-P2m' complex following EEG notation, or M50-M100-M150 complex based on the latency of responses from stimulus onset (Poeppel et al., 1996; Chait et al., 2004). P1m/M50 and P2/M150 are the positive deflections in response occurring at around 50 ms and 150 ms post-stimulus respectively, whereas N1m/M100 is a negative deflection occurring approximately 100 ms post-stimulus. An example of such MEG response is shown in Figure 2.11 and the corresponding topographic layout for M100 response. The M100 response is the most reliable component of the onset/offset response and its latency and amplitude is modulated by various stimulus properties such as loudness, frequency composition and SNR (Naatanen and Picton, 1987; Poeppel et al., 1996; Biermann and Heil, 2000), as well as attention (Ding and Simon, 2012a). While exact neural source of the M100 response is unknown, on a coarse level, M100 is known to originate from superior part of temporal gyrus, with in auditory cortex. Unlike M100, M50 part of transient response appears to be independent of task (Chait et al., 2004).

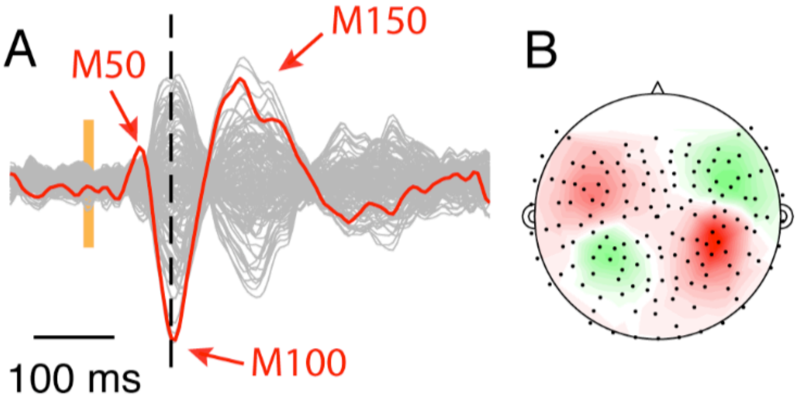

Figure 2.11: M100 response in MEG (adapted from Ding (2012)).

Apart from the responses to onset/offset of sounds, MEG has been used in studies to characterize responses to a wide variety of stimulus, ranging from tone streams (Akram et al., 2014) to amplitude modulated (AM) and frequency modulated (FM) sounds (Ding and Simon, 2009) to parts of speech sounds such as vowels and syllables (Alho et al., 1998; Luo et al., 2005; Tavabi et al., 2007) to more complex speech sounds including words, sentences (Luo and Poeppel, 2007) and stories (Ding and Simon, 2012b) and music (Maess et al., 2001).

### 2.2.3 Electroencephalography

As mentioned previously, EEG measures the voltage fluctuations (approximately in the range of micro-volts) extra-cranially, generated due to cortical current sources and can see both tangential and radial dipoles. EEG typically measures the difference between two electrodes, an active and a reference electrode (Figure 2.12, note that reference electrode is not the ground).

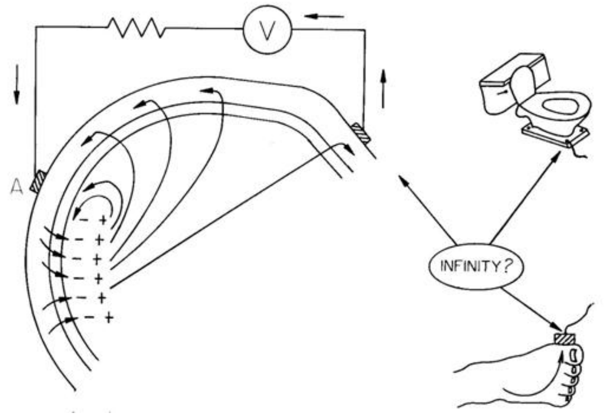

Figure 2.12: Electroencephalography measurement system

Active electrodes are placed over the head, where the activity of interest is to be recorded. Ideally the site of reference electrode should be electrically neutral, with respect to neural activity. However, in reality it is difficult to locate an electrically neutral site and hence the choice of reference electrode site is typically mastoids (conical protrusions of the skull located just behind the ears), which are supposed to pick-up least amount of brain activity of interest. However, the assumption that the mastoids do not pick up signals from regions of interest is unconfirmed (Srinivasan et al., 1998). Another popular referencing method is to use to average of all recorded electrodes as a reference signal. While researchers showed that average referencing may lead to misinterpretations when it is computed using small number of electrodes (Desmedt et al., 1990). However, average re-referencing is known to approximate reference-free recordings when used with high-density recordings (>32) (Bertrand et al., 1985). Since the voltage fluctuations are very small, EEG uses a differential amplifier to amplify the signals. The differential amplifier amplifies the difference between the recording-to-ground voltage and reference-to-ground voltage. So, any electrical activity recorded at the ground site (such as 60 Hz line noise) is cancelled out and anything recorded at the reference site contributes to the recorded signal. Conductive gel is applied between scalp and electrode to ensure good electrical contact and to keep the impedance low. For the study presented in this dissertation, data was recorded continuously from a 64 electrode Quick-Cap (Neuromedical Supplies, TX) with sintered Ag/Ag chloride electrodes and a Neuroscan SynAmp2.

**Auditory studies using EEG**

Transient responses (responses generated to a single event) in EEG are typically referred to as *Event related potentials* (ERPs). *Auditory evoked potentials* (AEPs) are a class of ERPs generated using sound as stimulus. Due to its ability to see both deep and radial sources, AEPs can be recorded from different structures in brain i.e., auditory cortex, brain stem, inferior colliculus and auditory nerve, albeit with progressively increasing difficulty. Although brain stem responses have been recorded using MEG (Parkkonen et al., 2009) they are less common and usually require much larger number of trials. Similar to MEG, the major transient response (from cortex) in EEG is referred to as P1-N1-P2 complex. P1 and P2 are positive deflections occurring around 50 ms and 200 ms respectively, whereas N1 is a negative deflection occurring around 100 ms. Figure 2.13 (Michelini et al., 1982) shows three major groups of AEPs obtained from EEG recordings using sound click as stimulus. *Short latency responses* (SLRs) are the responses from auditory brain stem and occur within 10 ms of stimulus onset and are characterized by 5 main peaks. *Middle latency responses* (MLRs) occur between 10 ms and 80 ms from stimulus onset and originate mainly from upper brain stem and auditory cortex. *Late latency responses* (LLRs) occur after 80 ms from the stimulus onset and originate from higher order auditory cortices.

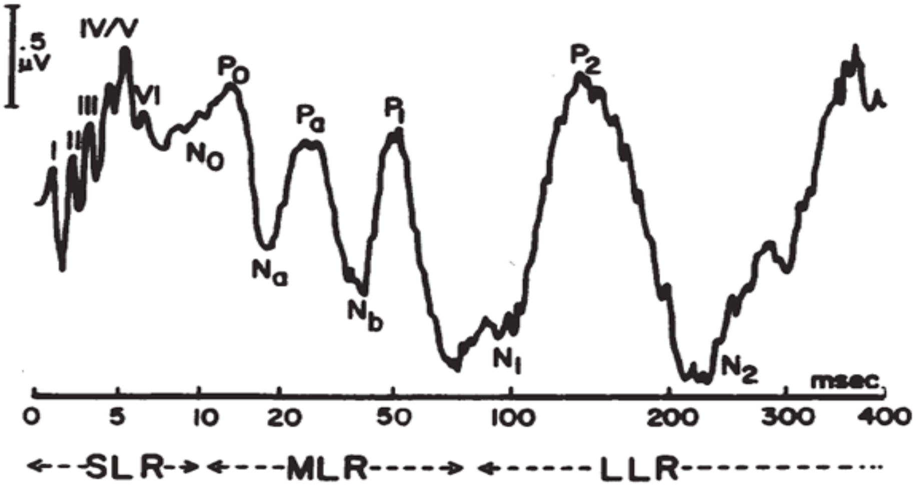

Figure 2.13: Transient auditory evoked potentials elicited by single acoustic click recorded using EEG (adapted from Michelini et al. (1982))

Similar to MEG, EEG has been used in various auditory studies involving odd ball detection, auditory streaming using tones (Snyder et al., 2009), phonemes and speech (Power et al., 2012; O'Sullivan et al., 2015), neural representation of speech (Crosse et al., 2015; Di Liberto et al., 2015), effect of neural oscillations and entrainment on perceptual tasks (Henry and Obleser, 2012, 2013), effects of ageing on hearing (Anderson et al., 2012), as well as in studies identifying biological markers for psychiatric disorders such as schizophrenia (Boutros et al., 2008; Hasey and Kiang, 2013) and so on.

## 2.3 Neural processing in human auditory cortex

Due to their millisecond time resolution, EEG and MEG are able to record phase locked neural responses to temporal modulations in speech and rhythmic sounds. In the following, we briefly discuss the temporal structure of speech and review techniques used to analyzed MEG and EEG data involving speech as stimulus.

### 2.3.1 Temporal modulations in speech

The information in speech is conveyed through its spectro-temporal modulations. The temporal information in speech and similarly the corresponding neural processing occurs on multiple time scales (Rosen, 1992; Poeppel, 2003; Shamma, 2006). These are depicted in a spectrogram in Figure 2.14 (adapted form Chi et al. (2005)) where the waveform from a particular auditory channel at 750 Hz is shown in three different magnifications in the right side panels in Figure 2.14. The top panel shows the slowest modulations, approximately 4 to 6 bursts per second, indicating the syllabic rhythm (Greenberg et al., 2003) in that particular frequency band of the utterance and is affected by dynamics of vocal tract, movement of the formants and onset and offset of consonants which are directly responsible for the speech intelligibility. These modulations are referred to as *slow temporal modulations* in this dissertation and are generally in the range of 1-8 Hz. The slow modulations that are consistent over auditory channels constitutes *temporal envelope* of speech. The slow temporal modulations are coded through phase locked neural activity from periphery to cortex. The middle panel shows intermediate modulation rates (about 200 Hz here, but 70 – 300 Hz in general) corresponding to pitch perception and timbre of the sound. The bottom panel shows the responses at fastest temporal scale (akin to carrier frequency) that carry the energy of stimulus in that particular auditory channel. In this dissertation, we extract the slow temporal modulations using the model proposed by Chi et al. (2005), which simulates central auditory processing using a multi-scale filter bank, with each filter tuned to a spectral and temporal modulations of sound at a particular frequency, denoted as spectral scale (Ω, cycles/octave) and modulation rate (ω, Hz), at a central frequency (CF) respectively.

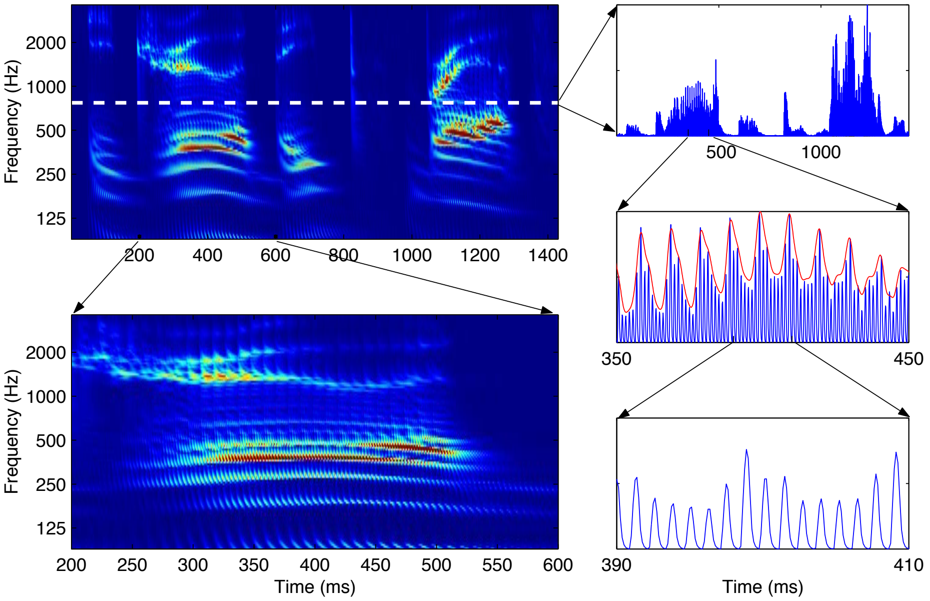

Figure 2.14: Temporal modulations in speech (adapted from (Chi et al., 2005))

## 2.4 Analysis methods

### 2.4.1 Modeling the neural processing of temporal modulations

In this section, we discuss modeling of the neural representations of slow temporal modulations in speech. As mentioned before, temporal modulations are encoded through phase locked neural activity in cortex. We use a linear time-invariant (LTI) system to model auditory cortex. Even though it is known that auditory system is non-linear, it is modeled as a LTI system due to the ease of analysis.

An LTI system is characterized by its impulse response. Given an input (any representation of an auditory stimulus) and output (MEG /EEG recording), the output is modeled as a convolution between input and impulse response. When the input is white noise, impulse response can simply be calculated as the cross correlation between output and input. Lalor et al. (2009) demonstrated this method to estimate the impulse response reflecting cortical processing of temporal modulations. Due to the non-linear nature of auditory system, the impulse response estimated using LTI model is input (stimulus) dependent (LTI model of a non-linear system can be viewed as it linear approximation at a specific operating point, akin to drawing a tangent at a point to an arbitrary shaped input-output curve and hence is input dependent). Hence, several classes of sounds, such as white noise, random chords, natural sounds and speech have been used to model auditory system (deCharms et al., 1998; Theunissen et al., 2001; David et al., 2007; Bitterman et al., 2008; David et al., 2009; Calabrese et al., 2011). Natural sounds, unlike white noise, are correlated over time and hence their auto-correlation has to be taken into account when used as input. In the following, we show how to evaluate the impulse response of LTI model when natural sounds are used as input.

Let $x(t)$ be the envelope (slow temporal modulations) of the input stimulus and $y(t)$ be the neural response, both represented as discrete time signals with $t = 1,...,\tilde{T}$ . The relation between them when modeled as LTI system is given by

$$y(t) = \sum_{\tau=-\infty}^{\infty} x(t-\tau)h(\tau) + \epsilon(t) \qquad (2.4)$$

where $h(t)$ is the impulse response, referred to as *temporal response function* (TRF) in this dissertation and $\epsilon(t)$ is the residual neural response unexplained by the LTI model. Since brain is a causal system (no output before input) and any stimuli generates finite duration neural response, above relation between $x(t)$ and $y(t)$ can be modified as

$$y(t) = \sum_{\tau=0}^{T} x(t-\tau)h(\tau) + \epsilon(t) \tag{2.5}$$

which can be expressed in matrix form as $y(t) = \mathbf{h}^{\mathrm{T}}\mathbf{x}(t) + \epsilon(t)$ , with $\mathbf{h} = [h(0), h(1), ..., h(T)]^{\mathrm{T}}$ and $\mathbf{x}(t) = [x(t), x(t-1), ..., x(t-T)]^{\mathrm{T}}$ . Assuming that input and output are wide-sense stationary,

$$\begin{aligned} &\mathbb{E}[y(t)\mathbf{x}(t)^{\mathrm{T}}] = \mathbf{h}^{\mathrm{T}}\mathbb{E}[\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}] + \mathbb{E}[\epsilon(t)\mathbf{x}(t)^{\mathrm{T}}] \\ &\mathbf{h} = [\mathbb{E}[\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}]]^{-1}\mathbb{E}[y(t)\mathbf{x}(t)] \end{aligned} \tag{2.6}$$

where $\mathbb{E}[.]$ denotes expectation over time. $\mathbb{E}[y(t)\mathbf{x}(t)^{\mathrm{T}}]$ is the cross correlation between input and output and $\mathbb{E}[\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}]$ is the auto-correlation of input. The error and input are uncorrelated at all lags of input and hence the term $\mathbb{E}[\epsilon(t)\mathbf{x}(t)^{\mathrm{T}}]$ is equal to zero. This method of estimating the impulse response is sometimes known as *normalized reverse correlation* (Theunissen et al., 2001). When the input is white noise the autocorrelation matrix is identity and hence the impulse response is just given by the cross correlation between input and output. However, when input is auto-correlated, such as speech, inverting the autocorrelation matrix may be ill-posed. In such cases, dimension reduction techniques such as principal component analysis (PCA) or regularization techniques (Calabrese et al., 2011) can be used.

Another method of estimating impulse response of auditory system is through *Boosting* (David et al., 2007). Boosting assumes a sparse prior for the impulse response in time. Starting with null (all zeros) impulse response, the algorithm updates the impulse response iteratively and the updates are incremental in nature. This continues until the correlation between real neural response and the model's response, referred to as *predictive power,* can no longer be improved. In each iteration of the algorithm, $\mathbf{h}$ is either incremented or decremented by $\Delta\mathbf{h}$ . Each $\Delta\mathbf{h}$ contains only one non-zero element and is optimized to minimize the expected value of squared prediction error:

$$\begin{aligned} &\Delta\mathbf{h} = \underset{\Delta\mathbf{h}}{\operatorname{argmin}}\, \mathbb{E}[(y(t) - \widehat{y}(t))^2] \\ &\text{where } \widehat{y}(t) = (\mathbf{h} + \Delta\mathbf{h})^{\mathrm{T}}\mathbf{x}(t) \\ &\text{s.t. } \|\Delta\mathbf{h}\|_0 = 1 \text{ and } \|\Delta\mathbf{h}\|_1 = \delta \end{aligned} \tag{2.7}$$

where $\|.\|_0$ and $\|.\|_1$ are zero and one norms respectively.

### 2.4.2 Stimulus reconstruction from neural responses

In the above, we discussed what is known as *forward problem*, wherein the stimulus is used to predict the neural response. While we discussed the forward problem only in one (time) dimension, the same can be extended to two (spectro-temporal) (Mesgarani and Chang, 2012; Pasley et al., 2012) and well as multi-dimensions (spectro-temporal, phonemic, MFCC etc) (Di Liberto et al., 2015). In this section, we discuss an analogous *reverse problem*, also known as *stimulus reconstruction,* which tries to decode/reconstruct the stimulus from the neural response. While the forward problem tells which acoustic features are transformed into neural responses the reverse problem tells the information contained in the neural responses. Figure 2.15 shows the relation between forward and reverse problems. Mathematically the reverse problem can be formulated as

$$x(t) = \sum_{\tau=0}^{T} y(t + T - \tau)d(\tau) + \epsilon(t) \tag{2.8}$$

where $x(t)$ and $y(t)$ are stimulus (envelope) and neural response (MEG/EEG) respectively and $d(t)$ is the *decoder*. The decoder can be estimated in a similar fashion to impulse response using boosting algorithm.

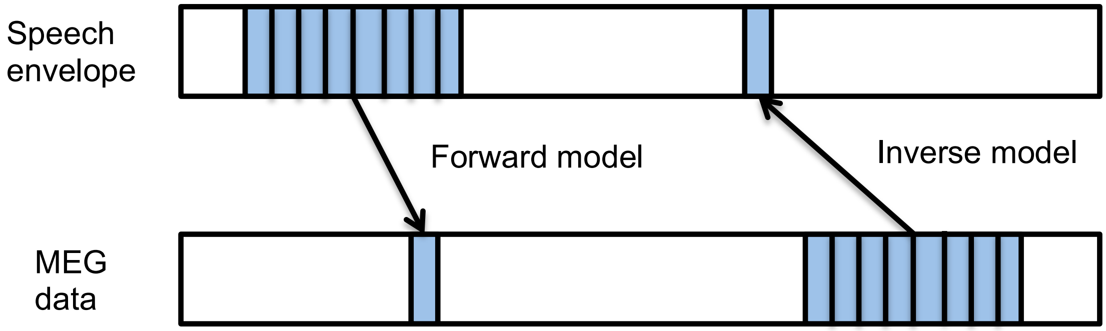

Figure 2.15: Relation between forward and Inverse models

### 2.4.3 Auditory steady state response analysis

Steady state responses (SSRs) are an important class of neural responses recorded using EEG. As mentioned previously, ERPs are the responses generated due a single event, where as SSRs are generated due to repeated and periodic presentation of certain class of stimuli. More precisely, Auditory Steady State Responses (ASSRs) are electrophysiological responses entrained to frequency and phase of a periodic stimulus (Brenner et al., 2009b). Typically, SSRs are generated by synchronous activity of large population of neurons to a temporally modulated stimulus such as amplitude modulated (AM) tones, frequency modulated (FM) tones and click trains (Picton et al., 2003;

Brenner et al., 2009b). Transient ERPs are typically analyzed in time/time-frequency domain and the usual measures include peak latency, amplitude, topography (e.g. Mismatched Negativity (MMN) (Naatanen et al., 2007; Light and Naatanen, 2013)), whereas SSRs are analyzed in frequency domain and typical measures include mean power, Phase locking factor (PLF), Inter-trial coherence (ITC), change in mean power from base line, etc. Figure 2.16 (adapted from Galambos et al. (1981)) shows a pictogram of 40-Hz ASSR generation due the repeated presentation of sound clicks every 25 ms. The waveforms in the upper-half of the figure shows ERPs generated due to each click, 25 ms apart and the bottom-half of figure shows the 40-Hz ASSR generated to overlap-addition of ERPs. While the figure shows the generation of 40 Hz ASSR, SSRs can be elicited using stimuli at wide variety of rates.

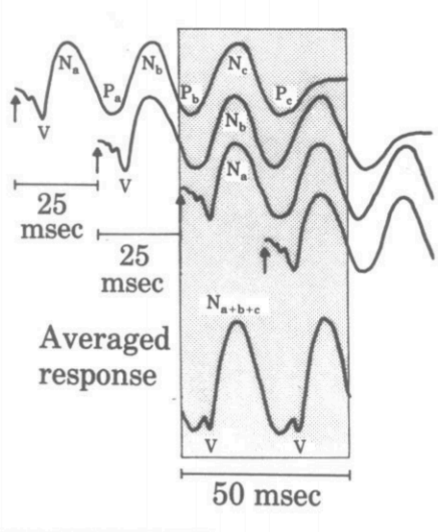

Figure 2.16: Example of 40-Hz ASSR generation (adapted from Galambos et al. (1981))

# 3 Cortical Representation of Noisy Reverberant Speech

## 3.1 Introduction

Speech communication in real-world scenarios, such as in a room or other enclosed space, differs from communication in an isolated environment, since the sound entering the ear is a linear superposition of direct (clean, distortion-free) component and multiple reflections from the surroundings. This general acoustic phenomenon, known as reverberation, is ubiquitous in daily listening environments. The reflections travel a longer path, with correspondingly attenuated amplitudes, before summing linearly with the direct component, thus distorting the clean sound from the original source. Depending on the number of reflections and their attenuation factors (a function itself of the surrounding reflecting surfaces and the paths travelled), the distortion of clean sound can vary from mild (e.g., in large open spaces) to severe (e.g., in a cave, cathedral or a dense forest). The reverberant signal received by the ear can be modeled as $y(t) = s(t)*h(t)$, where $s(t)$ is the clean sound from the source and $h(\mathrm{t})$ is the impulse response of a linear filter representing the delay and attenuation information of reflections (Figure 3.1). On the other hand, knowing only the reverberant signal $y(t)$, to infer the original sound $s(t)$ without knowledge of $h(t)$ is mathematically ill-posed problem, though human listeners are nonetheless able to perform this routinely, with some effort (Sato et al., 2007; Sarampalis et al., 2009; Yang and Bradley, 2009). Comprehension of speech in such a reverberant environment is further complicated by the presence of other sound sources whether stationary (e.g., the sound of an air-conditioner) or non-stationary (e.g., other talkers). The neural mechanisms by which reverberation is accommodated, and the representations employed by the auditory system in that process, in such adverse listening conditions remains unclear.

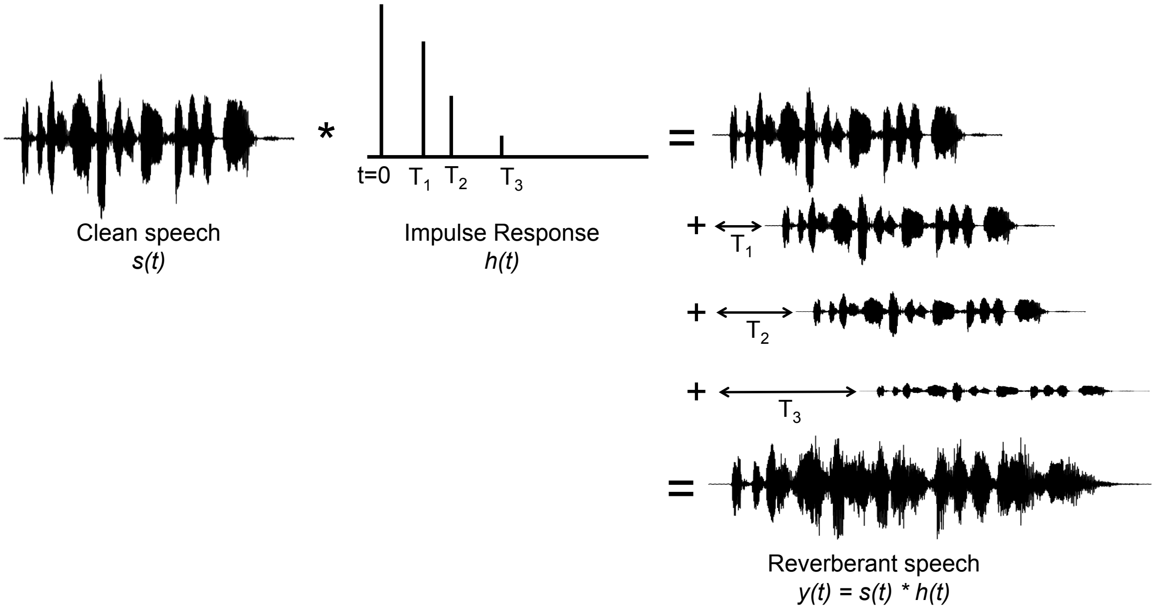

Figure 3.1: Phenomenon of reverberation. A reverberant signal reaching the ear is the sum of the original clean speech and its copies, appropriately time-shifted and scaled. This can be described mathematically as convolution between the clean speech $s(t)$ and the reverberation impulse response $h(t)$ (illustrated here with a schematic impulse response; after Traer and McDermott (2016))

The information in speech is conveyed through its temporal modulations, which can be decomposed into a slow envelope that modulates the fast temporal fine structure (TFS) (Rosen, 1992; Shamma and Lorenzi, 2013). The slower envelope (<10 Hz) corresponds to prosodic, phonemic, syllabic and word rates, whereas the TFS, the fast-varying component of speech, represents pitch, formant structure, timbre, etc. While envelope cues alone may be sufficient for partial speech comprehension in distortion free listening conditions, TFS is also important for speech comprehension, and especially so in the presence of distortions and competing backgrounds (Drullman et al., 1994a, b; Drullman, 1995; Smith et al., 2002; Moore, 2008; Ding et al., 2013; Moon and Hong, 2014; Kong et al., 2015; Rimmele et al., 2015; Swaminathan et al., 2016). While additive noise degrades the speech signal by reducing the intensity contrast, i.e., the depth of modulations, it does not affect the temporal sharpness of the speech signal. In contrast, reverberation, due to its convolutive nature, causes temporal smearing of both the envelope (example shown in Figure 3.2A, top) and TFS (see review by Assmann and Summerfield (2004)). TFS smearing results in spectral blurring (Figure 3.2A, bottom), which can affect the quality of the formant structure, timbre, and even pitch, and envelope smearing affects timing cues in the speech signal such as phoneme and syllable onset and offset. Physiological studies, both in animal models (Moore et al., 2013; Rabinowitz et al., 2013; Mesgarani et al., 2014b) and humans (Ding and Simon, 2013) have demonstrated the robustness of cortical representation of speech in the presence of stationary noise, in spite of degraded representation at the periphery of the auditory system (Delgutte, 1980). Studies of the auditory brainstem (Sayles and Winter, 2008; Sayles et al., 2014; Fujihira et al., 2017) and midbrain (Devore and Delgutte, 2010; Kuwada et al., 2014; Slama and Delgutte, 2015) have shown that peripheral and subcortical neural coding of the temporal envelope can be substantially degraded in a reverberant environment. However, the effects of distortion due to reverberation, as well as the interaction of reverberation and additive noise (if any), on the cortical coding of speech, are less understood.

Using Magnetoencephalography (MEG) recordings of human subjects listening to continuous speech, and linear system methods of neural response prediction (encoding) and stimulus reconstruction (decoding) (Ding and Simon, 2012b; Pasley et al., 2012; Di Liberto et al., 2015), we investigated the effect of noise and reverberation on cortical representation of continuous speech. Mesgarani et al. (2014b) examined the neural responses from single-unit recordings in ferrets, listening to reverberant speech (in absence of additive noise), and found that the corresponding clean speech spectrogram was better reconstructed than reverberant speech spectrogram. Further, Fuglsang et al. (2017), using electroencephalography (EEG) recordings of human subjects listening to speech in reverberation, showed that the clean speech envelope was better reconstructed than the reverberant speech envelope in case of severe reverberant conditions. In contrast to these studies, here, we (1) systematically examined the effects of noise and reverberation on neural encoding of speech by varying the severity of both reverberation and noise, and (2) examined the cortical representation of speech in noisy reverberant environment from both encoding and decoding perspectives, allowing insights into reverberation processing strategies across auditory cortex.

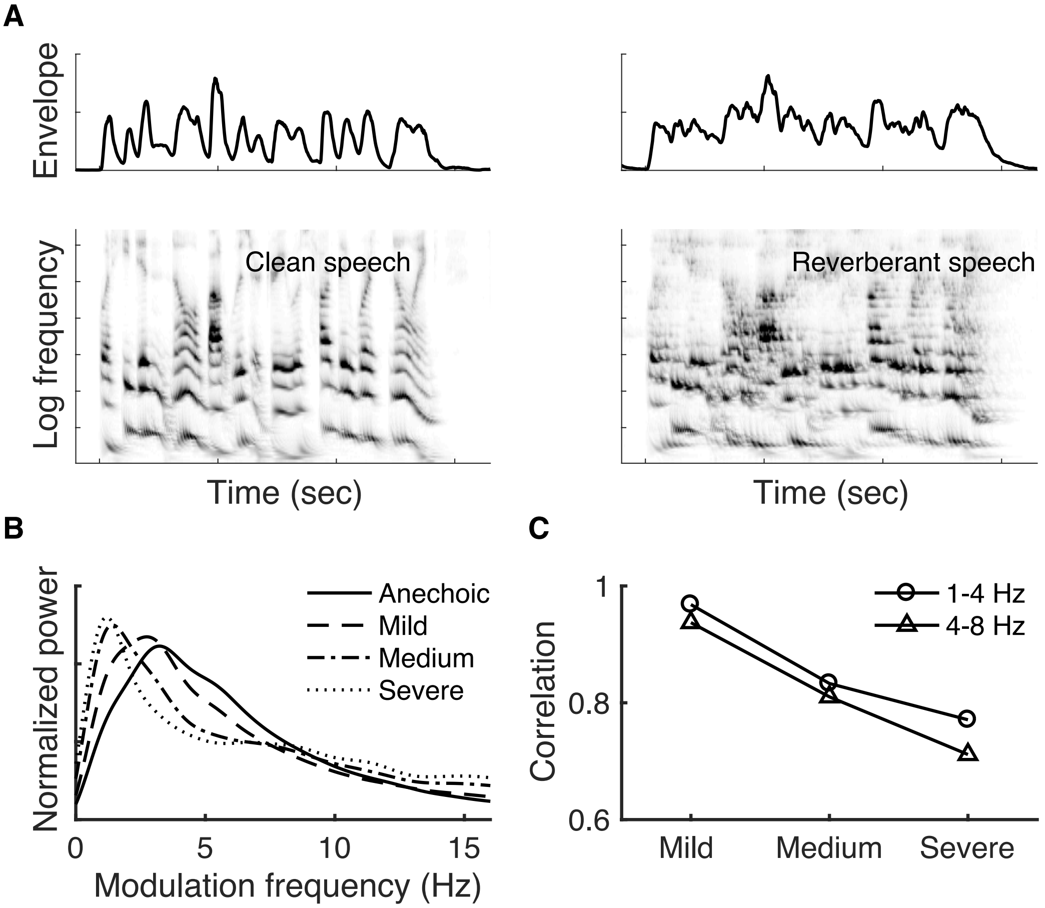

Figure 3.2: Effects of reverberation. **A.** Reverberation smears the temporal envelope (top right) of Clean speech (top left) as multiple reflections superimpose on the direct component from source. Reverberation also distorts the spectral structure of speech as shown by the auditory spectrogram (bottom) of speech without (left) and with (right) reverberation. **B.** The peak of the modulation spectrum occurs around 4 – 5 Hz in clean speech and shifts downward (left) with increasing severity of reverberation. **C.** Correlation coefficients comparing the bandpassed envelopes of reverberant speech, at different levels of severity, with the corresponding clean speech. The distortion effect of reverberation is higher in the 4 – 8 Hz band (corresponding

to neural theta activity) than 1- 4 Hz band (corresponding to neural delta activity).

## 3.2 Materials & Methods

***Subjects and Experimental Design*** Thirteen normal-hearing, young adults participated in the experiment. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland Institutional Review Board and written informed consent was obtained from each subject before the experiment. Subjects listened to 60 s duration speech segments under a full factorial design of three noise and four reverberation levels, totaling twelve stimulus conditions. The three noise levels were no-noise, +6 dB and +3 dB signal-to-noise ratio (SNR). The four reverberation levels are referred to, with increasing severity, as anechoic (clean), mild, medium and severe reverberation with Reverberation Time to 20 dB ($RT_{20}$: time elapsed before the reflections decay by 20 dB with respect to the direct component in terms of energy) values of 0 ms, 150 ms, 300 ms and 450 ms, respectively. The choice of $RT_{20}$ to characterize reverberation instead of the more standard $RT_{60}$ (time elapsed before the reflections decay by 60 dB with respect to the direct component) arises from the usage of listening to reverberant continuous speech: when speech reflections from an earlier time act as a masker for speech at the present time, a target-to-masker ratio (TMR) of 20 dB is perceptually more relevant than a TMR of 60 dB (which is instead more relevant for detection of reverberation in silence). In practice, any given $RT_{20}$ value is approximately one third of the corresponding $RT_{60}$ value. Reverberant speech was generated by convolving a (base) clean speech segment with a Room Impulse Response (RIR) with the desired severity of reverberation. RIRs were generated using the image-source method

(Allen and Berkley, 1979) as implemented by Lehmann and Johansson (2010), by simulating listening conditions in a room of dimensions 7 x 5 x 3 m (length, width, height), with source and listener positioned at (4.5, 2.5, 1.7) m and (3, 2.5, 1.7) m, respectively. Different levels of reverberation were obtained by varying absorption coefficients of walls, floor and roof of the simulated room. Noisy reverberant speech was generated by adding spectrally matched noise to the reverberant speech, at the desired SNR; spectrally matched noise was generated by randomizing the phase of the reverberant speech signal and scaling it appropriately to achieve the required SNR. Mathematically, the stimulus $S(t)$ is constructed as,

$$S(t) = R(t) + N(t) \tag{3.1}$$

where $R(t), N(t)$ are reverberant speech component of the stimulus and spectrally matched noise respectively. Further, $R(t)$ is constructed as,

$$R(t) = C(t) * RIR(t) \tag{3.2}$$

where $C(t), RIR(t)$ are (base) clean speech and RIR, respectively. All twelve (base) speech segments, used to generate twelve stimulus conditions, were taken from a public domain narration of Grimms' Fairy Tales by Jacob & Wilhelm Grimm (https://librivox.org/fairy-tales-by-the-brothers-grimm/), spoken by the same narrator. Periods of silence longer than 300 ms were replaced by a shorter gap whose duration was chosen randomly between 200 ms and 300 ms. When reverberation was added, the amplitude was rescaled so that all exemplars were of approximately equal perceptual loudness. No further scaling was performed when noise was added. Each of the twelve stimulus conditions was presented three times (trials) in succession, with the base speech segment used to generate a particular stimulus condition as well as presentation order of conditions randomized across subjects. To ensure the listeners' attention, a target-word was set before each trial and the subjects were asked to count the number of occurrences of the target-word in the stimulus being played. Additionally, at the end of each trial, subjects answered a different 2-alternative-forced-choice comprehension question. Subjects were required to close their eyes while listening.

***Data recording and pre-processing*** MEG recordings were conducted 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 of 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/$\sqrt{}$Hz or better in the white noise region. Subjects lay horizontally in a dimly lit magnetically shielded room (Yokogawa Electric Corporation). Responses were recorded with a sampling rate of 1 kHz with an online 200-Hz low-pass filter and 60 Hz notch filter. Three reference magnetic sensors and three vibrational sensors were used to measure the environmental magnetic field and vibrations. The reference sensor recordings were utilized to reduce environmental noise from the MEG recordings using the Time-Shift PCA method (de Cheveigne and Simon, 2007). Eye-blinks and heart beat artifacts were removed using Independent Component Analysis (ICA). For analysis in the sensor domain, MEG sensor recordings were decomposed into virtual sensors/components using denoising source separation (DSS) (Särelä and Valpola, 2005b; de Cheveigne and Simon, 2008; de Cheveigne and Parra, 2014), a blind source separation method that enhances neural activity consistent over trials. Specifically, DSS decomposes the multichannel MEG recording into temporally uncorrelated components, where each component is determined by maximizing its trial-to-trial reliability, measured by the correlation between the responses to the same stimulus in different trials. To reduce the computational complexity, sensor domain analysis was performed using DSS components. Additionally, for analysis in the source domain, each subject's head shape was digitized (Polhemus 3SPACE FASTRAK) and the subject's head was localized with respect to the MEG sensors using five marker coils attached to the head. The 'fsaverage' brain provided by FreeSurfer (Fischl, 2012) was fit to each subject's head shape using rotation, translation and uniform scaling. MEG data, after de-noising with time-shift PCA and ICA, were localized to active regions in the cortex using distributed minimum norm estimate (MNE) (Hamalainen and Ilmoniemi, 1994) as implemented in MNE software (Gramfort et al., 2013; Gramfort et al., 2014). The source model comprised of 10242 regularly spaced virtual source dipoles in each hemisphere with orientations perpendicular to the cortical surface. The sensor noise covariance was estimated from the empty room recording data. Due to the auditory nature of the study, further analysis was restricted to the responses estimated at the sources located in the transverse, superior, middle temporal gyri and banks of the superior temporal sulcus (Desikan et al., 2006). Both speech envelope and neural response (either a DSS component in sensor space, or the estimated activity at one source domain location) were band pass filtered between 1 – 8 Hz (delta and theta bands), which correspond to the slow temporal modulations in speech (Ding and Simon, 2012b, a), for further analysis.

***Encoding of stimulus to neural responses*** Encoding models provide a quantitative description of how information in a stimulus is represented in neural responses. Analyzing data from the perspective of encoding (predicting neural responses using the stimulus or some representation of the stimulus) allows investigators to identify, as well as quantify, how features/aspects of the stimulus are represented in the corresponding neural responses (Naselaris et al., 2011). Here, to identify the neural representation of speech distorted by noise and reverberation, three encoding models were compared (namely the Clean, Reverb and Mixed models as described below). Encoding analysis was performed by fitting a linear regression model between the stimulus representation under a particular model (whether Clean, Reverb or Mixed) and the corresponding low frequency (1- 8 Hz) neural responses. This approach has been used previously to describe the temporal relation between a speech stimulus and the corresponding neural response as measured by MEG (Ding and Simon, 2012b), EEG (Di Liberto et al., 2015), or ECoG (Mesgarani and Chang, 2012). The resulting models are commonly referred to as Temporal Response Functions (TRFs) and are mathematically represented as

$$r(t) = \sum_{\tau} s(t-\tau)TRF(\tau) + \varepsilon(t) \qquad (3.3)$$

where $t = 0,1, \ldots, T$ are discretized time instances, $r(t)$ is the neural response (of any individual sensor or DSS component, or the time-course of activity at a source location), $s(t)$ is the choice of stimulus representation in the encoding model under consideration (referred to as 'predictor' here), $TRF(t)$ is the TRF itself, and $\epsilon(t)$ is residual response waveform not explained by the TRF model (Ding and Simon, 2012b). The TRF is estimated using boosting with 10-fold cross-validation (David et al., 2007). Success of the linear model, referred to as 'prediction accuracy', is evaluated by how well it predicts neural responses, as measured by the proportion of the variance explained: the square of the Pearson correlation coefficient between neural response $r(t)$ and the TRF model prediction (right hand side of Eq. (3.3) excluding the error term). The three encoding models compared were: (1) the Clean model, where the stimulus is represented by the broadband envelope of the corresponding clean (base) speech, i.e. the envelope of $C(t)$ of Eq. (3.2); (2) the Reverb model, where the stimulus is represented by the broadband envelope of the reverberant speech component of the stimulus, i.e. the envelope of $R(t)$ of Eq. (3.1); and (3) the Mixed model – a model that allows both Clean and Reverb representations to contribute, i.e., simultaneously using envelopes from both the Clean and Reverb models as predictors. The Clean model tests the hypothesis that despite the distorted acoustic input to the ear, the cortex recovers and maintains neural representations for the underlying distortion free clean speech. The Reverb model tests the hypothesis that acoustic distortions due to reverberation present at the ear are also represented neurally in the cortex. Finally, the Mixed model allows the co-existence of neural representations for both clean and reverberant versions of speech. Such a dual representation is possible due to the hierarchical organization of the auditory cortex, which maintains increasingly complex and distortion robust representations of stimulus (Atencio et al., 2009; Okada et al., 2010; Sharpee et al., 2011). In all the encoding models, the broadband envelope was extracted by averaging the auditory spectrogram of the corresponding speech signal along the spectral dimension (Chi et al., 2005).

In case of the Mixed model, the linear model presented in (1) is modified as

$$r(t) = \sum_{\tau} s_c(t-\tau)TRF_c(\tau) + s_r(t-\tau)TRF_r(\tau) + \varepsilon(t) \qquad (3.4)$$

where $s_c(t)$ is the envelope of clean speech and $s_r(t)$ is the envelope of reverberant component of stimulus and $TRF_c(t), TRF_r(t)$ are the corresponding TRFs. Due to the presence of two predictors, the Mixed model has twice the number of degrees of freedom than the Clean and Reverb models. To ensure that the increased accuracy (if any) of the Mixed model compared to the other two is not merely due to increased degrees of freedom, a non-informative speech envelope was added as an additional predictor in both Clean and Reverb models, thus balancing the number of free parameters across models. For example, in the Clean model, the non-informative speech envelope is obtained by replacing the first half of reverb envelope with its second half and vice versa.

***Decoding speech from neural responses*** While the TRF/encoding analysis described in the previous section predicts neural response from stimulus, decoding analysis reconstructs stimulus envelope using neural responses. Thus, decoding analysis complements the TRF analysis (Mesgarani et al., 2009). Mathematically the envelope reconstruction/decoding operation can be formulated as

$$E(t) = \sum_{k=1}^{N} \sum_{\tau=\tau_b}^{\tau_e} M_k(t+\tau) D_k(\tau) + \epsilon(t) \tag{3.5}$$

where $E(t)$ is the reconstructed envelope, $M_k(t)$ is the MEG recording (neural response) from sensor/component $k$, and $D_k(t)$ is the linear decoder for sensor/component *k*. The times $\tau_b$ and $\tau_e$ denote the beginning and end times of the integration window, 0 and 500 ms respectively here. The decoder is estimated using boosting, analogously to the TRF estimation in the previous section. As decoding analysis integrates information over all data (whether from all sensor or from all source points) recorded in the time window under consideration, we restrict our decoding analysis to sensor space.

***Statistics*** Due to the presence of multiple stimulus conditions (a total of 12 in the full factorial design with three noise and four reverberation levels), the following statistical approach was used to compare between different encoding or decoding models. Considering the example of comparison between Mixed and Reverb models, the difference between both model prediction accuracies was calculated for each subject and condition and a repeated measures Analysis of Variance (ANOVA) is performed on the model differences with noise and reverb as factors (Greenhouse-Geisser corrected when required). Significant effects were followed up with appropriate pairwise t-tests. If a significant interaction effect was observed, a t-test was performed at each stimulus condition, to compare the mean difference of models with zero, correcting for multiple comparisons using False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). In absence of a significant interaction effect, data was pooled according to the main effects, if present, before comparing the average model differences against zero. Here also, FDR was used for multiple comparisons correction.  For example, in the case of significant main effect for the reverberation factor but not noise, data was pooled across noise levels and a t-test was performed at each level of reverb. When comparing two models, either in encoding or decoding analysis, through their differences, anechoic (reverberation free) stimuli were excluded as all models coincide there and so differences would be identically equal to zero for all subjects, with zero variance.

In the case source domain analysis, nonparametric permutation tests (Nichols and Holmes, 2002; Maris and Oostenveld, 2007), based on the threshold-free cluster-enhancement algorithm (TFCE; Smith and Nichols (2009)), were used to control for multiple comparisons when testing for the significance of a result at a large number of source locations. The precise implementation details are available in the Eelbrain source code (Brodbeck, 2017), but a brief summary follows. First, a test statistic (a t-value in case of t-test or an F-statistic in case of ANOVA) was computed for each source location based on the quantity of interest (here, the difference in prediction accuracies between two models) across subjects. The resulting test statistic map was then processed with TFCE, an image processing algorithm that enhances larger contiguous areas with large values compared to isolated spikes, based on the assumption that meaningful differences have a larger spatial extent than noise. To determine the null distribution for the resulting TFCE values, the procedure was repeated in 10,000 permutations of the data, with condition labels flipped for a randomly selected set of subjects in each permutation. The test statistic computation and TFCE were repeated in each permutation, and the largest value from the cluster-enhanced map is stored as an entry in the null distribution. Thus, a nonparametric distribution for the largest expected TFCE value under the null hypothesis was computed. Any value in the original TFCE map that exceeds the 95th percentile of the distribution is thus significant at the 5% level. Thus, TFCE provides strong control over family-wise type-I error (Nichols and Holmes, 2002).

## 3.3 Results

To examine the neural representation of speech distorted by additive noise and reverberation, three possible encoding models were compared (Clean, Reverb and Mixed models; see Methods for detailed description), using neural responses from the first DSS (most dominant auditory) component (Ding and Simon, 2012b). The performance of each model as measured by prediction accuracy (squared correlation coefficient between actual and predicted response) was computed for each model under each stimulus condition. In particular, if the brain maintains a distortion-free representation of speech in addition to the original distorted acoustic representation of speech, the Mixed model should have higher prediction accuracy than both the Reverb and Clean models, across all stimulus conditions. First, to compare the Mixed and Reverb models, repeated measures two-way ANOVA was performed on the difference of prediction accuracies between Mixed and Reverb models (Figure 3.3A) with noise and reverb as within subject factors (anechoic level in reverb factor was excluded as both models coincide when there is no reverberation). The main effect of reverb was not significant ($F(2, 24) = 3.307$, $p = 0.054$), neither was the effect of noise ($F(2, 24) = 0.436$, $p = 0.652$) or interaction ($F(4, 48) = 0.112$, $p = 0.978$). A post-hoc test, after pooling data across noise and reverb levels, showed that model difference was significantly greater than zero ($t(116) = 6.912$, $p < 0.001$). This suggests that the Mixed model predicts the neural responses better than the Reverb model across all stimulus conditions with reverberation. Similar comparison between Mixed and Clean models (Figure 3.3B) showed that model difference are significant in both noise ($F(2, 24) = 14.380$, $p < 0.001$) and reverb ($F(2, 24) = 13.546$, $p < 0.001$) with significant interaction ($F(4, 48) = 4.774$, $p = 0.003$). Post-hoc tests at each stimulus condition showed that the model difference is significantly greater than zero at all conditions (FDR with $q = 0.01$). This suggests that the Mixed model predicts neural responses better than Clean model. Taken together, these results suggest that when listening to speech in noisy reverberant conditions the auditory cortex maintains representations for both reverberant (distorted) and the corresponding clean (distortion free) versions of the stimulus.

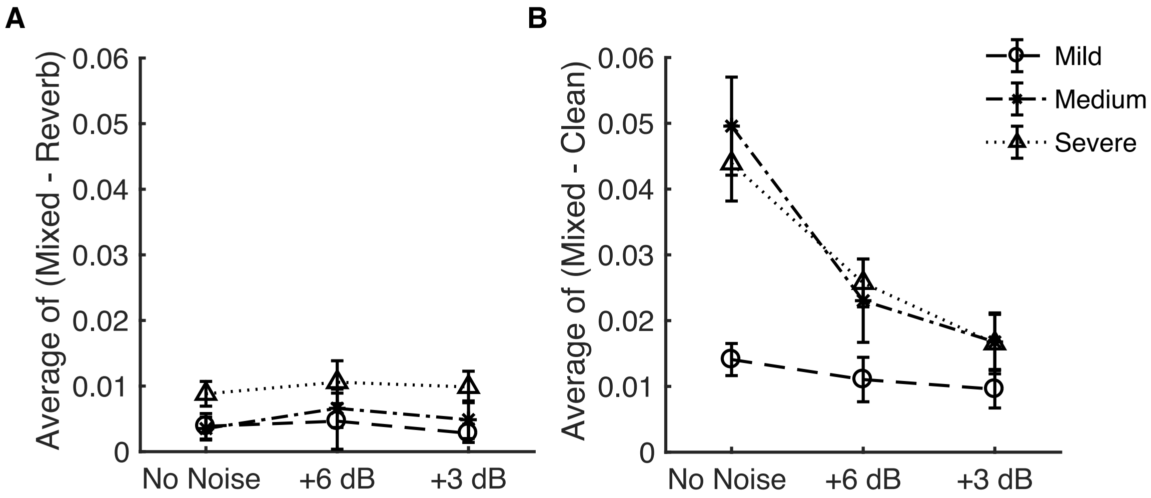

Figure 3.3: Comparing accuracy of encoding models. Difference between prediction accuracies of Mixed and Reverb models (A) as well as Mixed and Clean models (B) are both significantly greater than zero (FDR at q = 0.05 and FDR at q =0.01 respectively). This illustrates that the Mixed model predicts neural responses significantly better than either the Reverb or Clean model for all stimulus conditions with reverberation.

To identify the cortical regions contributing to the increased prediction accuracy of the Mixed model compared with the Reverb model, encoding analysis was performed in the neural source domain (predicting neural activity at each source location). The difference between the prediction accuracies of the two models was computed at each source location for all stimulus conditions. Variation of model difference with respect to reverberation level was modeled, separately for each noise level, as the slope of a line fit between model difference and reverberation level, thus obtaining three data points (one value of slope per noise level) per source location. As ANOVA, correcting for multiple comparisons using TFCE, showed no significance with respect to noise (p >= 0.482), data was pooled by averaging the slope across three noise conditions, resulting in one value of slope per source location.  Any value of slope significantly different from zero indicates significant model difference. A t-test performed at each source location, correcting for multiple comparisons, showed that Heschl's gyrus and middle-to-posterior superior temporal gyrus areas contribute to the increased performance of the Mixed model over the Reverb model (Figure 3.4).

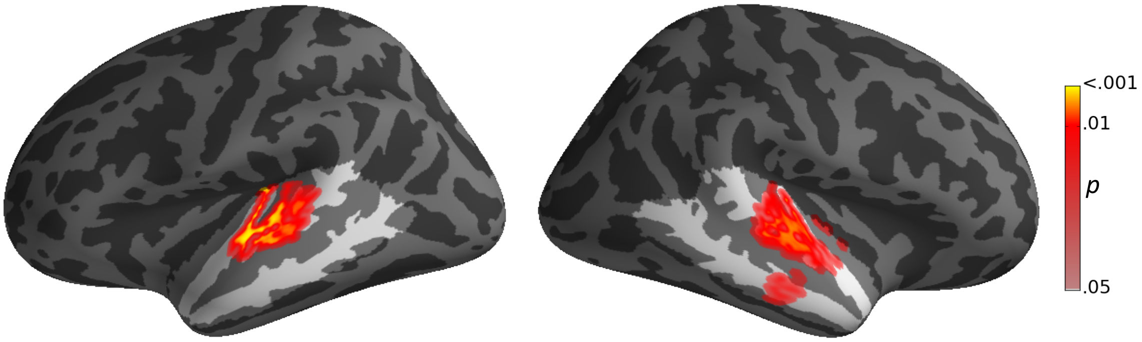

Figure 3.4: Anatomical regions contributing to increased performance of the Mixed model over the Reverb model ($p < 0.05$, corrected), rendered on the inflated brain surface model. These regions are better explained as containing areas with representations of both reverberant (distorted) and the corresponding clean (distortion free) versions of the stimulus, than as containing only representations of the reverberant (distorted) version. Areas that are not included in the analysis are shaded with a dark overlay.

To examine the fidelity of neural encoding of speech under different levels of noise and reverberation, prediction accuracies of the Mixed model (which best explained the neural response among the three encoding models compared) under different stimulus conditions were compared (Figure 3.5). A repeated measures ANOVA was used to assess the effect of noise and reverb on the prediction accuracy of the Mixed model. ANOVA showed a significant interaction between noise and reverb factors (F(2.761, 33.133) = 7.042, p = 0.001). Hence, post-hoc analysis was performed at each reverb level to see the effect of noise. The variation of prediction accuracy with respect to noise, as measured by the slope of the line fit between noise levels and prediction accuracies, at each reverb level, were calculated per subject. A t-test at each reverb level, corrected for multiple comparisons at q = 0.05 FDR, showed that the slope is significantly less than zero for mild (mean = -0.039, t(12) = -2.649, p = 0.021), medium (mean = -0.054, t(12) = -3.285, p = 0.007) and severe (mean = -0.047, t(12) = -3.410, p = 0.005) reverberation, whereas the anechoic condition showed no significant variation with respect to noise (mean = 0.0054, t(12) = 0.793, p = 0.443). This suggests that noise differentially affects the accuracy of neural encoding for conditions with and without reverberation: In the absence of any reverberation, noise did not show a significant effect on the accuracy of neural encoding, whereas its effect was adverse in presence of all reverberation levels tested.

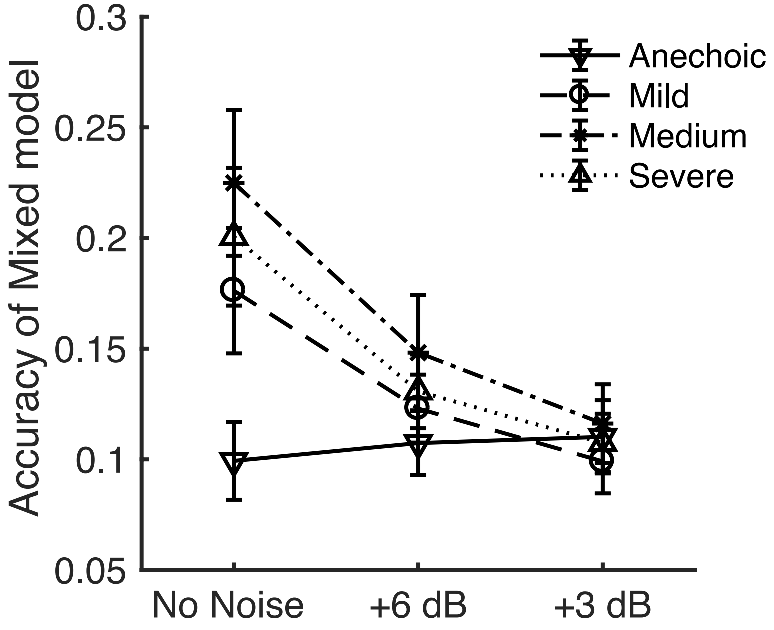

Figure 3.5: Effect of noise and reverberation on accuracy of neural encoding. In the absence of reverberation ("Anechoic"), noise did not show any significant effect on the accuracy of neural encoding. In contrast, encoding accuracy was reduced significantly with increase in noise, in the presence of *any* reverberation.

While the results presented so far provide an encoding perspective of speech in noisy and reverberant listening conditions, the putative role of delta and theta band neural responses in representing different aspects of speech (Ding and Simon, 2014; Kösem and Van Wassenhove, 2017) was examined in the following. The results from encoding models suggest that the auditory cortex maintains representations for both reverberant and clean versions of speech in reverberant environments. To assess the relative contributions of delta and theta band neural responses to the reverberant and clean representations, decoding analysis was employed. Here, both the reverberant and the respective clean versions of the stimulus envelope were reconstructed using delta and theta band neural responses separately, in order to compare which version of the envelope is more faithfully represented by delta and theta neural response. Figure 3.6 shows the difference between reconstruction accuracies of the reverberant and clean envelopes using only delta or only theta band neural responses. A repeated measures ANOVA on model differences (Reverb - Clean), in the delta band, showed a significant effect of noise (F(2, 24) = 7.005, p = 0.004), reverb (F(2, 24) = 8.564, p = 0.002) as well as significant interaction (F(4, 48) = 3.019, p = 0.027). Post-hoc t-tests showed that model difference is significantly greater than zero in all stimulus conditions (multiple comparisons corrected via FDR at q = 0.05). Similar analysis using theta band neural responses showed that model differences are not significantly affected by noise (F(1.395, 16.743) = 0.265, p = 0.691) or reverb (F(2, 24) = 0.904, p = 0.418) with no significant interaction effect (F(2.622, 31.463) = 2.034, p = 0.104). Further, post-hoc analysis showed that the model difference at any stimulus condition was not significantly different from zero (correcting for multiple comparisons using FDR). These results suggest that the delta band responses dominantly maintain reverberant representation, whereas theta band contains nearly equal contributions from both cleaned and reverberant representations. Delta band neural responses maintain a better representation of reverb speech than clean, while theta band shows no such distinction.

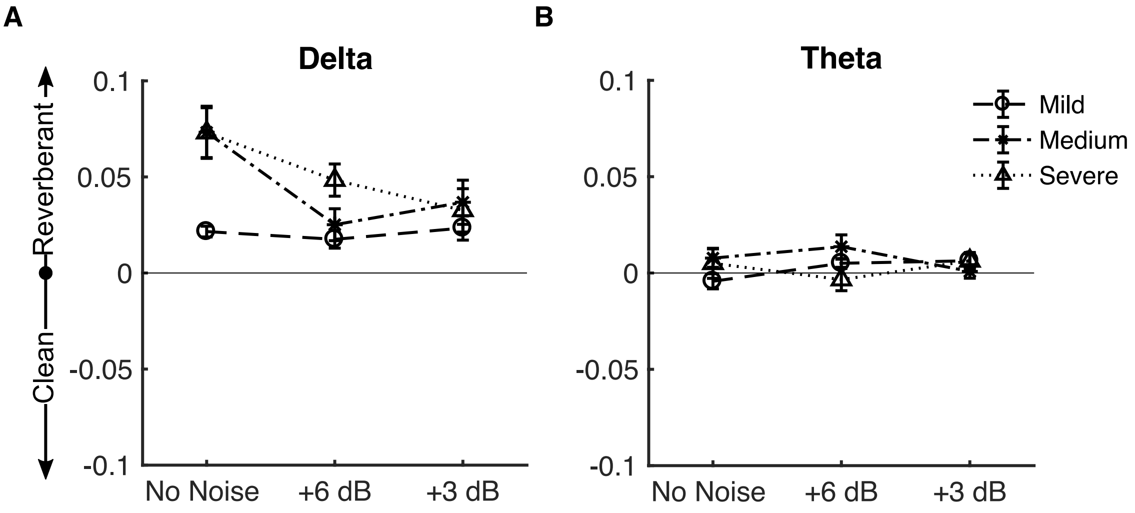

Figure 3.6: Comparing stimulus reconstruction accuracies for reverberant and corresponding clean speech. Results above the midline favor the Reverb model; below the midline favor the Clean model. **A.** Using only delta band (1 – 4 Hz) neural responses, the stimulus reconstruction of reverberant speech is significantly better than the corresponding clean speech (FDR with q = 0.05). **B.** Reconstruction using only theta band (4 – 8 Hz) neural responses did not show significant differences (FDR with q = 0.05) between reconstruction accuracies of the reverberant and respective clean stimulus.

## 3.4 Discussion

Using MEG to record the cortical activity of subjects listening to noisy, reverberant speech, and linear methods of neural response prediction and stimulus reconstruction, we observed that (1) the cortex maintains both distorted as well as the corresponding distortion free representations of distorted speech (2) noise differentially affects the accuracy of neural encoding in absence and presence of reverberation (3) theta band neural responses are a more likely candidate than delta band neural responses to hold the distortion free representation of the (distorted) acoustic stimulus.

That the Mixed model has better encoding accuracy compared to both the Reverb and Clean models (Figure 3.3) suggests that both distorted (reverberant) and distortion free (cleaned) versions of the speech are represented in auditory cortex. Such a dual representation is feasible given the hierarchical nature of auditory processing in cortex (Okada et al., 2010), where progressively distortion free (Moore et al., 2013; Rabinowitz et al., 2013) and categorical representations of speech emerge (Chang et al., 2010; Di Liberto et al., 2015). Historically, echo suppression, in simple stimuli such as lead-lag pairs referred to as the precedence effect, is often explained using inhibition triggered by the leading sound (Litovsky et al., 1999; Xia and Shinn-Cunningham, 2011). Mesgarani et al. (2014b) suggest a similar mechanistic model based on feed-forward synaptic depression and feed-back gain normalization to reduce the distortion due to reverberation. Traer and McDermott (2016) suggest that the problem of speech comprehension in reverberant conditions is solved by the auditory system as a cocktail party problem due to its ill-posed nature. They suggest that the brain uses prior information, accumulated through experience, to separate the clean speech from distorted reverberant speech input to the ear and identify it as an auditory object, separate from the environment in which it was produced. Both of these approaches (Mesgarani et al. (2014b) and Traer and McDermott (2016)) argue for simultaneous cortical representations of cleaned speech and the original reverberant speech, as shown in the current study. Significant difference between the prediction accuracies of the Mixed and Reverb models, reflecting the contribution of the distortion free part of the Mixed model, was confined to Heschl’s gyrus and middle to posterior superior temporal gyrus (Figure 3.4). Similar anatomical areas have been implicated as the substrate of categorical (phonemic) representation of speech in the cortex (Mesgarani et al., 2014a), suggesting that the clean contribution of the Mixed model could be related to the computation of distortion invariant categorical representation of speech.

In the absence of reverberation, the accuracy of neural encoding of speech is not significantly affected by noise (Figure 3.5). Such robustness to stationary noise has been previously demonstrated (Ding and Simon, 2013) and is thought to be the result of neural adaptation to statistics (such as mean and variance) of sound intensity (Dean et al., 2005; Dean et al., 2008; Robinson and McAlpine, 2009). However, in the case of reverberant environments, our results show that the neural encoding of speech is strongly and detrimentally affected by the addition of stationary noise (Figure 3.5). A similar detrimental effect of stationary noise has been previously observed using vocoded speech (Ding et al., 2013), highlighting the importance of TFS integrity for accurate neural encoding of speech in noisy background in contrast to the quiet listening conditions, wherein envelope cues are thought to be sufficient. Further, this suggests that the envelope entrainment to speech observed in MEG and EEG studies is a function of TFS along with the envelope.

On the other hand, in the absence of noise the encoding accuracy of reverberant speech (even under mild reverberation) is significantly higher compared with anechoic condition (Figure 3.5). The low-pass nature of the cortical response modulation transfer function (Simon and Ding, 2010), combined with the downward shift of modulation spectrum with increasing reverberation (Figure 3.2B), could explain the increase in accuracy of neural encoding with reverberation in the absence of noise. However the effect of listening effort due to reverberation cannot be discounted here either. Thus, the observed increase in encoding accuracy with increase in reverberation, in the absence of noise, could be due to combined effect of change in modulation spectrum and listener's effort. Another distinct possibility could be due to the fact that reverberant listening conditions, even mild, are pervasive in daily life, whereas anechoic listening conditions are rarely experienced. Thus, it is possible that ecologically irrelevant anechoic speech is not encoded as accurate as speech in ecologically relevant listening conditions.

Along with successful comprehension of speech in typical reverberant environments, a listener can also perceive and make subjective judgments regarding the reverberant environment, suggesting that such information is readily accessible to the auditory system. The observation that a reverberant envelope is better reconstructed than the corresponding cleaned envelope using only delta band neural responses (Figure 3.6A) suggest that the delta band is a candidate to convey the perception of reverberation. Similar reconstruction results using theta band neural responses (Figure 3.6B) showed no preference for either reverberant or clean envelope. Despite the increased stimulus contrast (reduced correlation) between the reverberant and clean envelopes in the theta band compared to delta band (Figure 3.2C), the shift away from the reverberation-dominated decoding in delta to the more balanced representation in theta provides limited evidence for reverberation removal occurring dominantly in theta band neural responses. These observations are consistent with the hypothesized roles of slow varying delta band and fast varying theta band neural responses to encode information related to the perceived non-speech specific acoustic rhythm and speech specific modulations necessary for intelligibility respectively (Ding and Simon, 2014). As such, it is beneficial for the auditory system to reduce the distortion in the theta band more than the delta band (Figure 3.6). In contrast to the decoding results presented here, using a combination of both delta and theta band neural responses, Fuglsang et al. (2017) showed that clean speech envelope is better reconstructed than reverberant speech envelope in case of severe reverberation. This difference may be due to the lack of binaural cues in the current study, which are known to enhance speech perception in reverberant and noisy environments (Nabelek and Robinson, 1982). Also, using single unit recording from the primary auditory cortex of ferrets, Mesgarani et al. (2014b) showed that clean speech is better reconstructed while listening in reverberant conditions. This difference with the decoding results presented here could be due to the availability of spike/high-gamma (> 40 Hz) neuronal responses in single unit recordings, in contrast to the current study which examined only slow temporal modulations.

In summary, the results suggest that while listening to speech distorted by additive noise and reverberation, the auditory cortex maintains representations for both distorted and the corresponding cleaned (distortion free) speech, possibly in different cortical areas. The additive noise differentially affects the accuracy of neural encoding in presence and absence of reverberation. Finally, theta band neural responses are a candidate for containing distortion free representations of speech in reverberant environments, while the delta band neural responses may convey the non-speech-specific information regarding the reverberant listening environment.

# 4 Cortical Representation of Speech in a Multi-talker Auditory Scene

## 4.1 Introduction

Individual sounds originating from multiple sources in a complex auditory scene mix linearly and irreversibly before they enter the ear, yet are perceived as distinct objects by the listener (Cherry, 1953; Bregman, 1994; McDermott, 2009). The separation, or rather individual re-creation, of such linearly mixed original sound sources is a mathematically ill-posed question, yet the brain nevertheless routinely performs this task with ease. The neural mechanisms by which this perceptual 'un-mixing' of sounds occur, the collective cortical representations of the auditory scene and its constituents, and the role of attention in both, are key problems in contemporary auditory neuroscience.

It is known that auditory processing in primate cortex is hierarchical (Davis and Johnsrude, 2003; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Okada et al., 2010; Peelle et al., 2010; Overath et al., 2015) with subcortical areas projecting onto the core areas of auditory cortex, and from there, on to belt, parabelt and additional auditory areas (Kaas and Hackett, 2000). Sound entering the ear reaches different anatomical/functional areas of auditory cortex with different latencies (Recanzone et al., 2000; Nourski et al., 2014). Due to this serial component of auditory processing, the hierarchy of processing can be described by both anatomy and latency, of which the latter may be exploited using the high temporal fidelity of non-invasive magnetoencephalography (MEG) neural recordings.

In selective listening experiments using natural speech and MEG, the two major neural responses known to track the speech envelope are the $\text{M50}_{\text{TRF}}$ and $\text{M100}_{\text{TRF}}$, with respective latencies of 30 – 80 ms and 80 – 150 ms, of which the dominant neural sources are, respectively, Heschl's gyrus (HG) and Planum temporale (PT) (Steinschneider et al., 2011; Ding and Simon, 2012a). Posteromedial HG is the site of core auditory cortex; PT contains both belt and parabelt auditory areas (here collectively referred to as higher-order areas) (Griffiths and Warren, 2002; Sweet et al., 2005). Hence the earlier neural responses are dominated by core auditory cortex, and the later are dominated by higher-order areas. To better understand the neural mechanisms of auditory scene analysis, it is essential to understand how the cortical representations of a complex auditory scene change from the core to the higher order auditory areas.

One topic of interest is whether the brain maintains distinct neural representations for each unattended source (in addition to the representation of the attended source), or if all unattended sources are represented collectively as a single monolithic background object. A common paradigm used to investigate the neural mechanisms underlying auditory scene analysis employs a pair of speech streams, of which one is attended, which then leaves the other speech stream remaining as the background (Kerlin et al., 2010; Ding and Simon, 2012a; Mesgarani and Chang, 2012; Power et al., 2012; Zion Golumbic et al., 2013b; O'Sullivan et al., 2015). This results in a limitation, which cannot address the question of distinct vs. collective neural representations for unattended sources. This touches on the long-standing debate of whether auditory object segregation is pre-attentive or it is actively influenced by attention (Carlyon, 2004; Sussman et al., 2005; Shinn-Cunningham, 2008; Shamma et al., 2011). Evidence for segregated neural representations of background streams would support the former, whereas a lack of segregated background objects would support the latter.

To address these issues, we use MEG to investigate a variety of potential cortical representations of the elements of a multi-talker auditory scene. We test two major hypotheses: that the dominant representation in core auditory cortex is of the physical acoustics, not of separated auditory objects; and that once object-based representations emerge in higher order auditory areas, the unattended contributions to the auditory scene are represented collectively as a single background object. The methodological approach employs the linear systems methods of stimulus prediction and MEG response reconstruction (Lalor et al., 2009; Mesgarani et al., 2009; Ding and Simon, 2012a; Mesgarani and Chang, 2012; Pasley et al., 2012; Di Liberto et al., 2015).

## 4.2 Materials & Methods

***Subjects & Experimental Design*** Nine normal-hearing, young adults (6 Female) participated in the experiment. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland Institutional Review Board. Subjects listened to a mixture of three speech segments spoken by, respectively, a male adult, female adult and a child speaker. The three speech segments were mixed into a single audio channel with equal perceptual loudness. All three speech segments were taken from public domain narration of Grimms' Fairy Tales by Jacob & Wilhelm Grimm (https://librivox.org/fairy-tales-by-the-brothers-grimm/). Periods of silence longer than 300 ms were replaced by a shorter gap whose duration was chosen randomly between 200 ms and 300 ms. The audio signal was low-pass filtered with cut-off at 4 kHz. In first of three conditions, the subjects were asked to attend to the child speaker, while ignoring the other two (i.e., child speaker as target, with male and female adult speakers as background). In condition two, during which the same mixture was played as in condition one, the subjects were instead asked to attend to the male adult speaker (with female adult and child speakers as background). Similarly, in condition three, the target was switched to the female adult speaker. Each condition was repeated three times successively, producing three trials per condition. The presentation order of the three conditions was counterbalanced across subjects. Each trial was of 220 s duration, divided into two 110 s sections, to reduce listener fatigue. To help participants attend to the correct speaker, the first 30 s of each section was replaced by the clean recording of the target speaker alone, followed by a 5 s upward linear ramp of the background speakers. Recordings of this first 35 s of each segment were not included in any analysis. To further encourage the subjects to attend to the correct speaker, a target-word was set before each trial and the subjects were asked to count the number of occurrences of the target-word in the speech of the attended speaker. Additionally, after each condition, the subject was asked to recount a short summary of the attended narrative. The subjects were required to close their eyes while listening. Before the main experiment, 100 repetitions of a 500-Hz tone pip were presented to each subject to elicit the M100 response, a reliable auditory response occurring ~100 ms after the onset of a tone pip. This data was used check whether any potential subjects gave abnormal auditory responses, but no subjects were excluded based on this criterion.

***Data recording and pre-processing*** MEG recordings were conducted 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 of 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/$\sqrt{}$Hz or better in the white noise region. Subjects lay horizontally in a dimly lit magnetically shielded room (Yokogawa Electric Corporation). Responses were recorded with a sampling rate of 1 kHz with an online 200-Hz low-pass filter and 60 Hz notch filter. Three reference magnetic sensors and three vibrational sensors were used to measure the environmental magnetic field and vibrations. The reference sensor recordings were utilized to reduce environmental noise from the MEG recordings using the Time-Shift PCA method (de Cheveigne and Simon, 2007). Additionally, MEG recordings were decomposed into virtual sensors/ components using denoising source separation (DSS) (Särelä and Valpola, 2005b; de Cheveigne and Simon, 2008; de Cheveigne and Parra, 2014), a blind source separation method that enhances neural activity consistent over trials. Specifically, DSS decomposes the multichannel MEG recording into temporally uncorrelated components, where each component is determined by maximizing its trial-to-trial reliability, measured by the correlation between the responses to the same stimulus in different trials. To reduce the computational complexity, for all further analysis the 157 MEG sensors were reduced, using DSS, to 4 components in each hemisphere. Also, both stimulus envelope and MEG responses were band pass filtered between 1 – 8 Hz (delta and theta bands), which correspond to the slow temporal modulations in speech (Ding and Simon, 2012b, a).

***Neural Model Terminology and Notation*** As specified in the stimulus description, in each condition the subject attends to one among the three speech streams. Neural models of speech stream processing can be compared by contrasting the predicted envelope reconstructions of the different models. The envelope of attended speech stream is referred to as the 'foreground' and the envelope of each of the two unattended speech streams is referred to as the 'individual background'. In contrast, the envelope of the entire unattended part of the stimulus, comprising *both* unattended speech streams, is referred to as the 'combined background'. The envelope of entire acoustic stimulus or auditory scene, comprising of all the three speech streams is referred to as the 'acoustic scene'. Thus, if $S_a, S_b, S_c$ are three speech stimuli, $Env(S_a + S_b + S_c)$ is the acoustic scene. In contrast, the sum of envelopes of three speech streams, $Env(S_a) + Env(S_b) + Env(S_c)$, is referred to as the 'sum of streams', and the two are not mathematically equal: even though both are functions of the same stimuli, they differ due to the non-linear nature of a signal envelope (the linear correlation between the acoustic scene and the sum of streams is typically ~0.75). Combination (unsegregated) envelopes, whether of the entire acoustic scene or the background only, can be used to test neural models that do not perform stream segregation. Sums of individual stream envelopes, whether of all streams or just the background streams, can be used to test neural models that process the (segregated) streams in parallel, given that neurally generated magnetic fields add in linear superposition.

Neural responses with latencies less than ~85 ms (typically originating from core auditory areas) are referred to here as 'early neural responses' and responses with latencies more than ~85 ms (typically from higher-order auditory areas) (Ahveninen et al., 2011; Okamoto et al., 2011; Steinschneider et al., 2011) are referred to as 'late neural responses'.

The next two sections describe models of the neural encoding of stimuli into responses, followed by models of the decoding of stimuli from neural responses. Encoding models are presented here first because of their ease of description over decoding models, but in Results the decoding analysis is presented first, since it is the decoding results that inform the new model of encoding.

***Temporal Response Function*** In an auditory scene with a single talker, the relation between MEG neural response and the presented speech stimuli can be modeled using a linear temporal response function (TRF) as

$$r(t) = \sum_{\tau} s(t-\tau)TRF(\tau) + \varepsilon(t) \tag{4.1}$$

where $t = 0,1,\dots,T$ is time, $r(t)$ is the response from any individual sensor or DSS component, $s(t)$ is the stimulus envelope in decibels, $TRF(t)$ is the TRF itself, and $\epsilon(t)$ is residual response waveform not explained by the TRF model (Ding and Simon, 2012b). The envelope is extracted by averaging the auditory spectrogram, (Chi et al., 2005) along the spectral dimension. The TRF is estimated using boosting with 10-fold cross-validation (David et al., 2007). In case of single speech stimuli, the TRF is typically characterized by a positive peak between 30 ms and 80 ms and a negative peak between 90 ms and 130 ms, referred to as $\mathrm{M50_{TRF}}$ and $\mathrm{M100_{TRF}}$ respectively (Ding and Simon, 2012a) (positivity/negativity of the magnetic field is by convention defined to agree with the corresponding electroencephalography[EEG] peaks). Success/accuracy of the linear model is evaluated by how well it predicts neural responses, as measured by the proportion of the variance explained: the square of the Pearson correlation coefficient between the MEG measurement and the TRF model prediction.

In the case of more than one speaker, the MEG neural response, $r(t)$ can be modeled as the sum of the responses to the individual acoustic sources (Ding and Simon, 2012a; Zion Golumbic et al., 2013b), referred to here as the 'Summation model'. For example, with three speech streams, the neural response would be modeled as

$$r(t) = \sum_{\tau=0}^{\tau=\tau_2} S_a(t-\tau)TRF_a(\tau) + \sum_{\tau=0}^{\tau=\tau_2} S_b(t-\tau)TRF_b(\tau) + \sum_{\tau=0}^{\tau=\tau_2} S_c(t-\tau)TRF_c(\tau) + \epsilon(t) \tag{4.2}$$

where $S_a(t)$, $S_b(t)$ and $S_c(t)$ are the envelopes of the three speech streams, and $TRF_a(t)$, $TRF_b(t)$ and $TRF_c(t)$ are the TRFs corresponding to each stream. $\tau_2$ represents the length of TRF. All TRFs in the Summation model are estimated simultaneously.

In addition to the existing summation model, we propose a new encoding-model referred to as the 'Early-late model', which allows one to incorporate the hypothesis that the early neural responses typically represent the entire acoustic scene, but that the later neural responses differentially represent the separated foreground and background.

$$r(t) = \sum_{\tau=0}^{\tau=\tau_1} S_A(t-\tau)TRF_A(\tau) + \sum_{\tau=\tau_1}^{\tau=\tau_2} S_F(t-\tau)TRF_F(\tau) + \sum_{\tau=\tau_1}^{\tau=\tau_2} S_B(t-\tau)TRF_B(\tau) + \epsilon(t) \tag{4.3}$$

where $S_A(t)$ is the (entire) acoustic scene, $S_F(t)$ is the envelope of attended (foreground) speech stream, and $S_B(t)$ is the combined background (i.e., envelope of everything other than attended speech stream in the auditory scene), and $TRF_A(t), TRF_F(t),$ and $TRF_B(t)$ are the corresponding TRFs. $\tau_1, \tau_2$ represent the boundary values of the integration windows for early and late neural responses respectively, with $0 < \tau_1 < \tau_2$.

The explanatory power of different models, such as the Summation and Early-late models, can be ranked by comparing the accuracy of their response predictions (illustrated in Figure 4.1A). The models differ in terms of number of free parameters, with the Early-late model having fewer parameters than the Summation model. Hence, any improved performance observed in the proposed Early-late model over the Summation model is correspondingly more likely due to a better model fit, since it has less freedom to fit the data (though the converse would not hold).

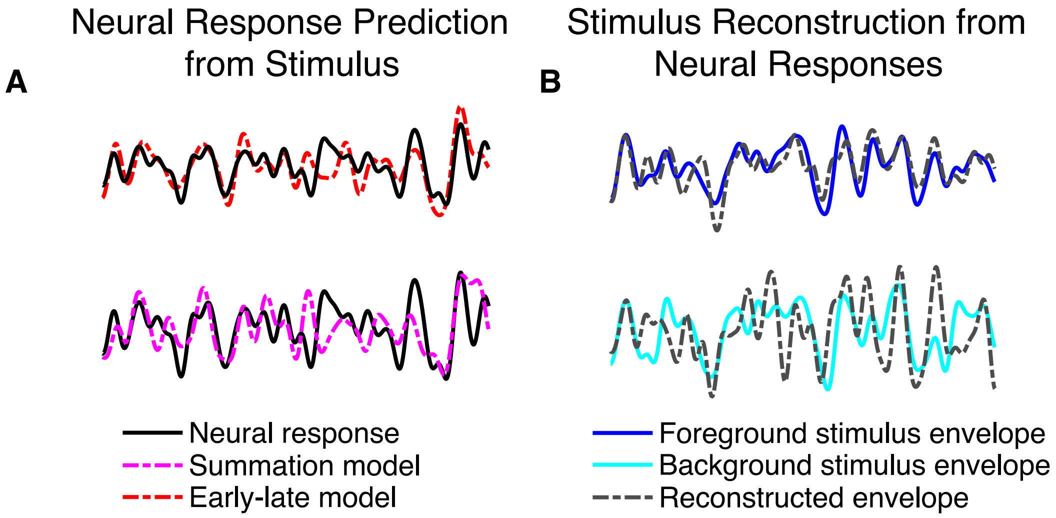

Figure 4.1: Illustrations of outcomes comparing competing encoding- and decoding-based neural representations of the auditory scene and its constituents. All examples are grand averages across subjects (3 seconds duration). **A.** Comparing competing models of *encoding* to neural responses. In both the top and bottom examples, an experimentally measured MEG response (black) is compared to the neural response predictions made by competing proposed models. In the top example, the neural response prediction (red) is from the Early-late model; in the bottom example, the neural response prediction (magenta) is from the Summation model. The

proposed Early-late model prediction shows higher correlation with the actual MEG neural response than Summation model. **B.** Comparing competing models of *decoding* to stimulus speech envelopes. In both the top and bottom examples, an acoustic speech stimulus envelope (blue/cyan) is compared to the model reconstruction of the respective envelope (gray). In the top example, the envelope reconstruction (blue) is of the foreground stimulus, based on late time responses; in the bottom example, the envelope reconstruction (cyan) is of the background stimulus, also based on late time responses. The foreground reconstruction shows higher correlation with the actual foreground envelope, compared to the background reconstruction with the actual background envelope.

***Decoding speech from neural responses*** While the TRF/encoding analysis described in the previous section predicts neural response from the stimulus, decoding analysis reconstructs the stimulus based on the neural response. Thus, decoding analysis complements the TRF analysis (Mesgarani et al., 2009). Mathematically the envelope reconstruction/decoding operation can be formulated as

$$E(t) = \sum_{k=1}^{N} \sum_{\tau=\tau_b}^{\tau_e} M_k(t+\tau) D_k(\tau) + \epsilon(t) \tag{4.4}$$

where $E(t)$ is the reconstructed envelope, $M_k(t)$ is the MEG recording (neural response) from sensor/component $k$, and $D_k(t)$ is the linear decoder for sensor/component $k$. The times $\tau_b$ and $\tau_e$ denote the beginning and end times of the integration window. By appropriately choosing the values of $\tau_b$ and $\tau_e$, envelope reconstructions using neural responses from any desired time window can be compared. The decoder is estimated using boosting analogously to the TRF estimation in the previous section. In the single talker case the envelope is of that talker's speech. In a multi-talker case, the envelope to be reconstructed might be the envelope of the speech of attended talker, or one of the background talkers, or of a mixture of any two or all three talkers, depending on the model under consideration. Chance-level reconstruction (i.e., the noise floor) from a particular neural response is estimated by reconstructing an unrelated stimulus envelope from that neural response. Figure 4.2 illustrates the distinction between reconstruction of stimulus envelope from early and late responses. The stimulus envelope at time point *t* can be reconstructed using neural responses from the dashed (early response) window or dotted (late response) window. (While it is true that the late responses to the stimulus at time point $t - \Delta t$ overlap with early responses to the stimulus at time point *t*, the decoder used to reconstruct the stimulus at time point *t* from early responses is only minimally affected by late responses to the stimulus at time point $t - \Delta t$ when the decoder is estimated by averaging over a long enough duration, e.g., tens of seconds). The cut-off time between early and late responses, $\tau_{boundary}$, was chosen to minimize the overlap between the M50$_{\mathrm{TRF}}$ and M100$_{\mathrm{TRF}}$ peaks, on a per subject basis, with a median value of 85 ms (range 70-100 ms in 5 ms increments); repeating the analysis using the single value of 85 ms for all subjects did not qualitatively change any conclusions. When decoding from early responses only, the time window of integration is from $\tau_b = 0$ to $\tau_e = \tau_{boundary}$. When decoding from late neural responses only, the time window of integration is from $\tau_b = \tau_{boundary}$ to $\tau_e = 500$ ms.

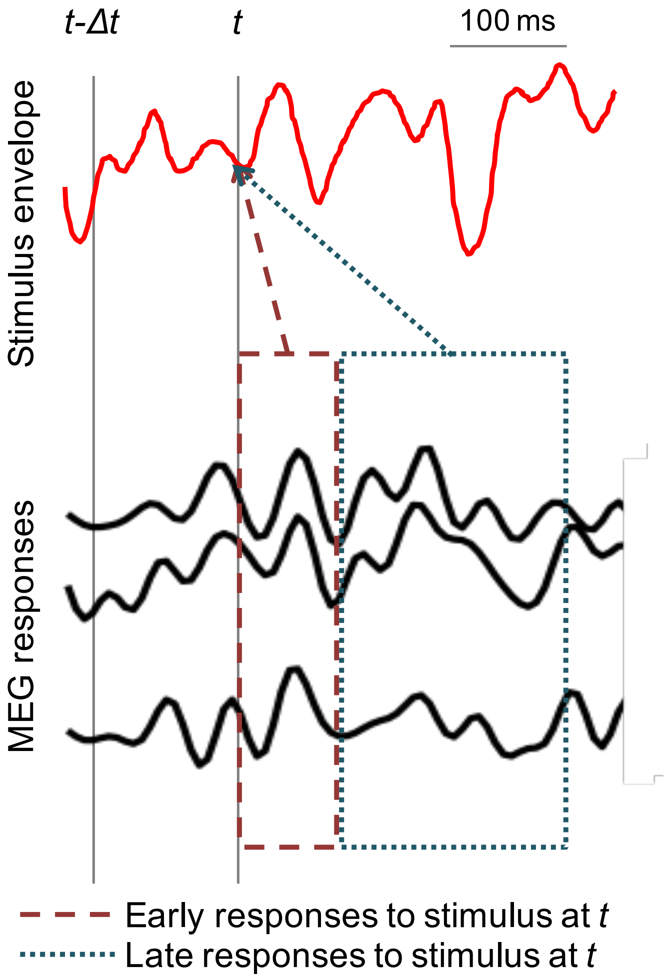
t-Δt
t
100 ms
Stimulus envelope
MEG responses
Early responses to stimulus at t
Late responses to stimulus at t

Figure 4.2: Early vs. late MEG neural responses to a continuous speech stimulus. A sample stimulus envelope and time-locked multi-channel MEG recordings are shown in red and black respectively. The two grey vertical lines indicate two arbitrary time points at *t* - *Δt* and *t*. The dashed and dotted boxes represent the early and late MEG neural responses to stimulus at time point *t* respectively. The reconstruction of the stimulus envelope at time *t* can be based on either early or late neural responses, and the separate reconstructions can be compared against each other.

The robustness of different representations, such as of Foreground vs. Background, can be compared by examining the accuracy of their respective stimulus envelope reconstructions (illustrated in Figure 4.1, right).

***Statistics*** All statistical comparisons reported here are two-tailed permutation tests with $N$=1,000,000 random permutations (within subject). Due to the value of N selected, the smallest accurate $p$ value that can be reported is $2\times 1/N$ (= $2\times 10^{-6}$; the factor of 2 arises from the two-tailed test) and any $p$ value smaller than $2/N$ is reported as $p < 2\times 10^{-6}$. The statistical comparison between foreground and individual backgrounds requires special mention, since each listening condition has one foreground but two individual backgrounds. From the perspective of both behavior and task, both the individual backgrounds are interchangeable. Hence, when comparing reconstruction accuracy of foreground vs. individual background the average reconstruction accuracy of the two individual backgrounds is used. Finally, Bayes factor analysis is used, when appropriate, to evaluate evidence in favor of null hypothesis, since conventional hypothesis testing is not suitable for such purposes. Briefly, Bayes factor analysis calculates the *posterior odds* i.e., the ratio of $P(H_0|observations)$ to $P(H_1|observations)$, where $H_0$ and $H_1$ are the null and alternate hypotheses respectively.

$$\frac{P(H_0|observations)}{P(H_1|observations)} = \frac{P(observations|H_0)}{P(observations|H_1)} \times \frac{P(H_0)}{P(H_1)} \tag{4.5}$$

$$= BF_{01} \times \frac{P(H_0)}{P(H_1)} \tag{4.6}$$

The ratio of $P(observations|H_0)$ and $P(observations|H_1)$ is denoted as the Bayes factor, $BF_{01}$. Then, under the assumption of equal priors ($P(H_0) = P(H_1)$), the posterior odds reduce to $BF_{01}$. A $BF_{01}$ value of 10 indicates that the data is ten times more likely to occur under the null hypothesis than the alternate hypothesis; conversely, a $BF_{01}$ value of 0.1 indicates that the data is 10 times more likely to occur under the alternate hypothesis than the null hypothesis. Conventionally, a $BF_{01}$ value between 3 and 10 is considered as moderate evidence in favor of the null hypothesis, and a value between 10 and 30 is considered strong evidence; conversely, a $BF_{01}$ value between 1/3 & 1/10 (respectively 1/10 & 1/30) is considered moderate (respectively strong) evidence for the alternate hypothesis (for more details we refer the reader to Rouder et al. (2009)).

## 4.3 Results

***Stimulus reconstruction from early neural responses*** To investigate the neural representations of the attended vs. unattended speech streams associated with early auditory areas, i.e., from core auditory cortex, (Nourski et al., 2014), the temporal envelope of attended (foreground) and unattended speech streams (individual backgrounds) were reconstructed using decoders optimized individually for each speech stream. All reconstructions performed significantly better than chance level (foreground vs. noise, $p < 2\times10^{-6}$; individual background vs. noise, $p < 2\times10^{-6}$), indicating that all three speech streams are represented in early auditory cortex. Figure 4.3A shows reconstruction accuracy for foreground vs. individual backgrounds. A permutation test shows no significant difference between foreground and individual background ($p$ = 0.21), indicating that there is no evidence of significant neural bias for the attended speech stream over the ignored speech stream, in early neural responses. In fact, Bayes Factor analysis ($BF_{01}$ = 4.2) indicates moderate support in favor of the null hypothesis (Rouder et al., 2009), that early neural responses do not distinguish significantly between attended and ignored speech streams.

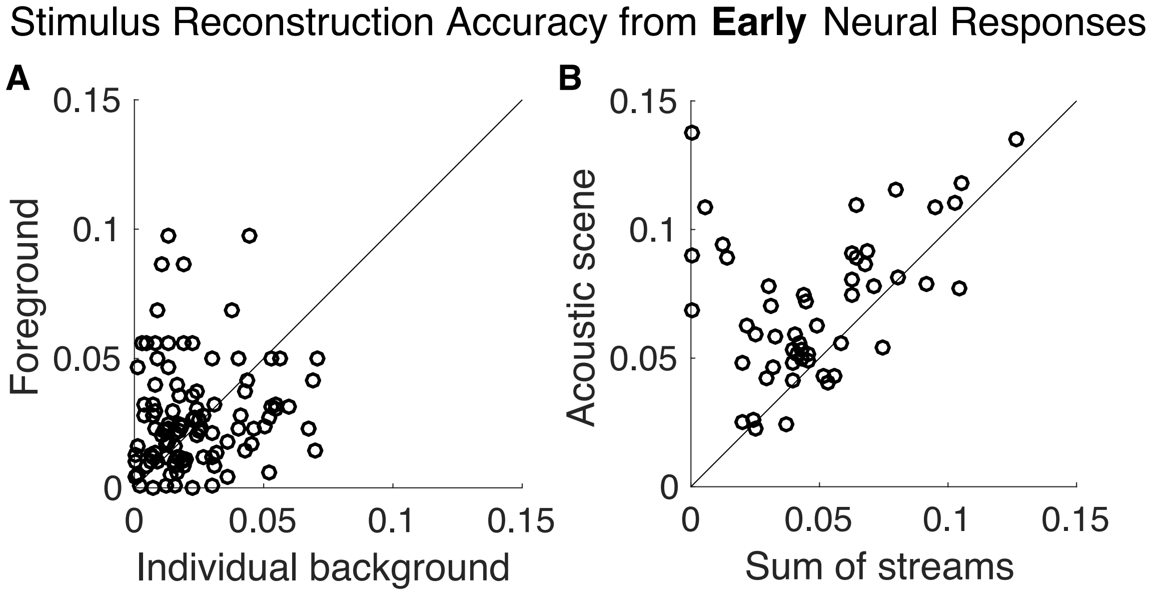

Figure 4.3: Stimulus envelope reconstruction accuracy using *early* neural responses. **A.** Scatter plot of reconstruction accuracy of the foreground vs. individual background envelopes. No significant difference was observed (*p* = 0.21), and therefore no preferential representation of the foreground speech over the individual background streams is revealed in early neural responses. Each data point corresponds to a distinct background and condition partition per subject (with two backgrounds sharing a common foreground). **B.** Scatter plot of reconstruction accuracy of the envelope of the entire acoustic scene vs. that of the sum of the envelopes of all three individual speech streams. The acoustic scene is reconstructed more accurately (visually, most of data points

fall above the diagonal) as a whole than as the sum of individual components in early neural responses ($p < 2 \times 10^{-6}$). Each data point corresponds to a distinct condition partition per subject. In both plots, reconstruction accuracy is measured by proportion of the variance explained: the square of the Pearson correlation coefficient between the actual and predicted envelopes.

To test the hypothesis that early auditory areas represent the auditory scene in terms of acoustics, rather than as individual auditory objects, we reconstructed the acoustic scene (the envelope of the sum of all three speech streams) and compared it against the reconstruction of the sum of streams (sum of reconstruction envelopes of each of the three individual speech streams). Separate decoders optimized individually were used to reconstruct the acoustic scene and the sum of streams. As can be seen in Figure 4.3B, the result shows that the acoustic scene is better reconstructed than the sum of streams ($p < 2\times10^{-6}$). This indicates that early auditory cortex is better described as processing the entire acoustic scene rather than processing the separate elements of the scene individually.

***Stimulus reconstruction from late neural responses*** While the preceding results were based on early cortical processing, the following results are based on late auditory cortical processing (responses with latencies more than ~85 ms). Figure 4.4A shows the scatter plot of reconstruction accuracy for the foreground vs. individual background envelopes based on late responses. A paired permutation test shows that reconstruction accuracy for the foreground is significantly higher than the background ($p < 2\times10^{-6}$).

Even though the individual backgrounds are not as reliably reconstructed as foreground, their reconstructions are nonetheless significantly better than chance level ($p < 2\times10^{-6}$).

In order to distinguish among possible neural representations of the background streams, we compared the reconstructability of the envelope of the entire background as a whole, with the reconstructability of the sum of the envelopes of the (two) backgrounds. If the background is represented as a single auditory object (i.e., “the background”), the reconstruction of the envelope of the entire background should be more faithful than the sum of envelopes of individual backgrounds. In contrast, if the background is represented as distinct auditory objects, each distinguished by its own envelope, the reconstruction of the sum of envelopes of the individual backgrounds should be more faithful. Figure 4.4B shows the scatter plot of reconstruction accuracy for the envelope of combined background vs. the sum of the envelopes of the individual background streams. Analysis shows that the envelope of the combined background is significantly better represented than the sum of the individual envelopes of the individual backgrounds ($p = 0.012$). As noted previously, the envelope of the combined background is actually strongly correlated with the sum of the envelopes of the individual backgrounds, meaning that finding a significant difference in their reconstruction accuracy is *a priori* unlikely, providing even more credence to the result.

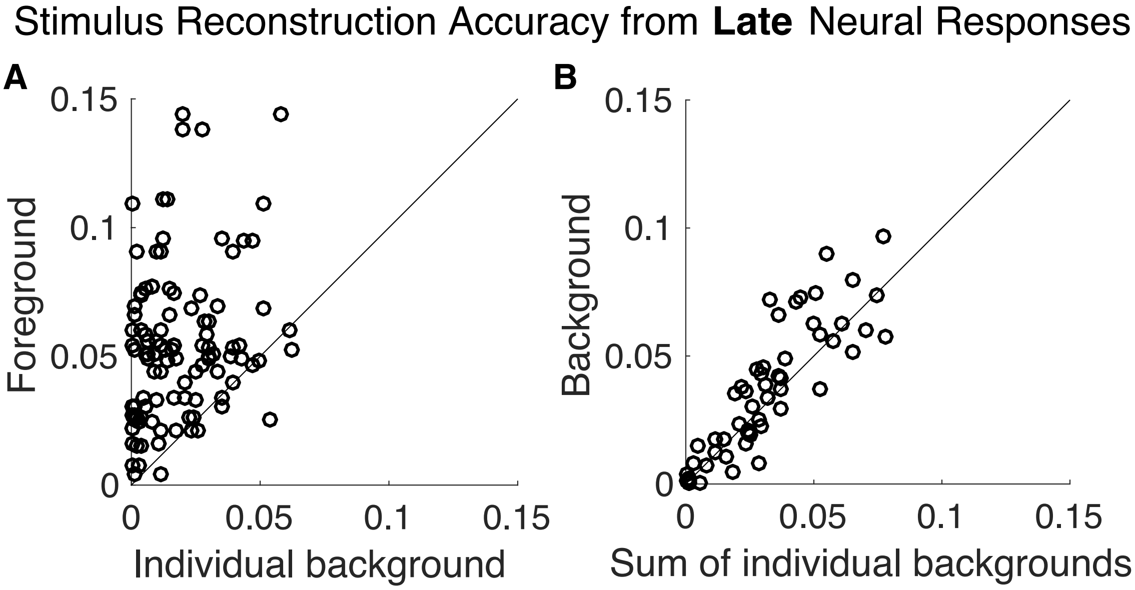

Figure 4.4: Stimulus envelope reconstruction accuracy using *late* neural responses. **A.** Scatter plot of accuracy between foreground vs. individual background envelope reconstructions demonstrates that the foreground is represented with dramatically better fidelity (visually, most of data points fall above the diagonal) than the background speech, in late neural responses ($p < 2 \times 10^{-6}$). Each data point corresponds to a distinct background and condition partition per subject (with two backgrounds sharing a common foreground). **B.** Scatter plot of the reconstruction accuracy of the envelope of the entire background vs. that of the sum of the envelopes of the two individual background speech streams. The background scene is reconstructed more accurately as a monolithic background than as separated individual background streams in late neural responses ($p = 0.012$). Each data point corresponds to a distinct condition partition per subject.

***Encoding analysis*** Results above from envelope reconstruction suggest that while early neural responses represent the auditory scene in terms of the acoustics, the later neural responses represent the auditory scene in terms of a separated foreground and a single background stream. In order to further test this hypothesis, we use TRF-based encoding analysis to directly compare two different models of auditory scene representations. The two models compared are the standard Summation model (based on parallel representations of all speech streams; see Equation 2) and the new Early-late model (based on an early representation of the entire acoustic scene and late representations of separated foreground and background; see Equation 3). Figure 4.5 shows the response prediction accuracies for the two models. A permutation test shows that the accuracy of the Early-late model is considerably higher than that of the Summation model ($p < 2\times10^{-6}$). This indicates that a model in which early/core auditory cortex processes the entire acoustic scene but later/higher-order auditory cortex processes the foreground and background separately has more support than the previously employed model of parallel processing of separate streams throughout auditory cortex.

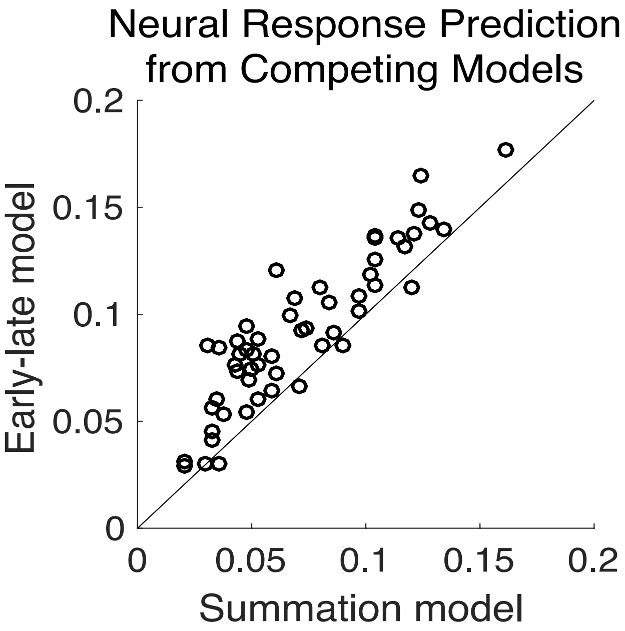

Figure 4.5: MEG response prediction accuracy. Scatter plot of the accuracy of predicted MEG neural response for the proposed Early-late model vs. the standard Summation model. The Early-late model predicts the MEG neural response dramatically better (visually, most of data points fall above the diagonal) than the Summation model ($p < 2 \times 10^{-6}$). The accuracy of predicted MEG neural responses is measured by proportion of the variance explained: the square of the Pearson correlation coefficient between the actual and predicted responses. Each data point corresponds to a distinct condition partition per subject.

## 4.4 Discussion

In this study, we used cortical tracking of continuous speech, in a multi-talker scenario, to investigate the neural representations of an auditory scene. From MEG recordings of subjects selectively attending to one of the three co-located speech streams, we observed that 1) The early neural responses (from sources with short latencies), which originate primarily from core auditory cortex, represent the foreground (attended) and background (ignored) speech streams without any significant difference, whereas the late neural responses (from sources with longer latencies), which originate primarily from higher-order areas of auditory cortex, represent the foreground with significantly higher fidelity than the background; 2) Early neural responses are not only balanced in how they represent the constituent speech streams, but in fact represent the entire acoustic scene holistically, rather than as separately contributing individual perceptual objects; 3) Even though there are two physical speech streams in the background, no neural segregation is observed for the background speech streams.

It is well established that auditory processing in cortex is performed in a hierarchical fashion, in which an auditory stimulus is processed by different anatomical areas at different latencies (Inui et al., 2006; Nourski et al., 2014). Using this idea to inform the neural decoding/encoding analysis allows the effective isolation of neural signals from a particular cortical area, and thereby the ability to track changes in neural representations as the stimulus processing proceeds along the auditory hierarchy. This time-constrained reconstruction/prediction approach may prove especially fruitful in high-time-resolution/low-spatial-resolution imaging techniques such as MEG and EEG. Even though different response components are generated by different neural sources, standard neural source localization algorithms may perform poorly when different sources are strongly correlated in their responses (Lutkenhoner and Mosher, 2007). While the proposed method is not to be viewed as an alternative to source localization methods, it can nonetheless be used to tease apart different components of MEG/EEG response, without explicit source localization.

Even though there is no significant difference between the ability to reconstruct the foreground and background from early neural responses, nonetheless we observe a non-significant tendency towards an enhanced representation of the foreground (foreground > background, $p$ = 0.21). This could be due to task-related plasticity of spectro-temporal receptive fields of neurons in mammalian primary auditory cortex (Fritz et al., 2003), where the receptive fields of neurons are tuned to match the stimulus characteristics of attended sounds. The selective amplification of foreground in late neural responses (from higher-order auditory cortices) but not in early responses (from core auditory cortex) observed here using *decoding* is in agreement with the *encoding* result of Ding and Simon (2012a) where the authors showed that the late $\mathrm{M100_{TRF}}$ component, but not the early $\mathrm{M50_{TRF}}$ component, of TRF is significantly modulated by attention. The increase in fidelity of the foreground as the response latency increases indicates a temporal as well as functional hierarchy in cortical processing of auditory scene, from core to higher-order areas in auditory cortex. Similar preferential representation for the attended speech stream has been demonstrated, albeit with only two speech streams and not differentiating between early and late responses, using delta and theta band neural responses (Ding and Simon, 2012a; Zion Golumbic et al., 2013a; Zion Golumbic et al., 2013b) as well as high-gamma neural responses (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013a), and using monaural (Ding and Simon, 2012a; Mesgarani and Chang, 2012) as well as audio-visual speech (Zion Golumbic et al., 2013a; Zion Golumbic et al., 2013b).

While some researchers suggest selective entrainment (Schroeder and Lakatos, 2009; Ng et al., 2012; Zion Golumbic et al., 2013b; Kayser et al., 2015) as the mechanism for selective tracking of attended speech, others suggest a temporal coherence model (Shamma et al., 2011; Ding and Simon, 2012a). Natural speech is quasi-rhythmic with dominant rates at syllabic, word and prosodic frequencies. The selective entrainment model suggests that attention causes endogenous low frequency neural oscillations to align with the temporal structure of the attended speech stream, thus aligning the high excitability phases of oscillations with events in attended stream. This effectively forms a mask that favors the attended speech. The temporal coherence model suggests that selective tracking of attended speech is achieved in two stages. First, a cortical filtering stage, where feature-selective neurons filter the stimulus, producing a multidimensional representation of auditory scene along different feature axes. This is followed by a second stage, coherence analysis, which combines relevant features streams based on their temporal similarity, giving rise to separate perceptions of attended and ignored streams. In this model, it is hypothesized that attention, acting through in the coherence analysis stage, plays an important role in stream formation. This type of coherence model predicts an unsegregated representation of any (non-attended) background streams.

The representation of an auditory scene in core auditory cortex, based on the early responses, is here shown to be more spectro-temporal- or acoustic-based than object-based (e.g., Figure 4.3B). This is further supported by the result that the Early-late model predicts MEG neural responses significantly better than Summation model (e.g., Figure 4.5). This is consistent with previous demonstrations that neural activity in core auditory cortex was highly sensitive to acoustic characteristics of speech and primarily reflects spectro-temporal attributes of sound (Nourski et al., 2009; Okada et al., 2010; Ding and Simon, 2013; Steinschneider et al., 2014). In contrast, Nelken and Bar-Yosef (2008) suggest that neural auditory objects may form as early as primary auditory cortex, and Fritz et al. (2003) show that representations of dynamic sounds in primary auditory cortex are influenced by task. As a working principle, it is possible that less complex stimuli are resolved earlier in the hierarchy of auditory pathway (e.g., sounds that can be separated via tonotopy) whereas more complex stimuli (e.g., concurrent speech streams), which need further processing, are resolved only much later in auditory pathway. In addition, it is worth noting that the current study uses co-located speech streams, whereas mechanisms of stream segregation will also be influenced by other auditory cues, including spatial cues, differences in acoustic source statistics (e.g., only speech streams vs. mixed speech and music; strong statistical differences might drive stream segregation in a more bottom-up manner than the top-down attentional effects studied here), perceptual load effects (e.g., tone streams vs. speech streams), as well as visual cues. Any of these additional cues has the potential to alter the timing and neural mechanisms by which auditory scene analysis occurs.

It is widely accepted that an auditory scene is *perceived* in terms of auditory objects (Bregman, 1994; Griffiths and Warren, 2004; Shinn-Cunningham, 2008; Shamma et al., 2011). Ding and Simon (2012b) demonstrated evidence for an object-based cortical representation of an auditory scene, but did not distinguish between early and late neural responses. This, coupled with the result here that early neural responses provide an acoustic, not object-based, representation, strongly suggest that the object-based representation emerges only in the late neural responses/higher-order (belt and parabelt) auditory areas. This is further supported by the observation that acoustic invariance, a property of object-based representation, is observed in higher order areas but not in core auditory cortex (Chang et al., 2010; Okada et al., 2010). When the foreground is represented as an auditory object in late neural responses, the finding that the combined background is better reconstructed than the sum of envelopes of individual backgrounds (Figure 4.4B) suggests that in late neural responses the background is not represented as separated and distinct auditory objects. This result is consistent with that of Sussman et al. (2005), who reported an unsegregated background when subjects attended to one of three tone streams in the auditory scene. This unsegregated background may be a result of an 'analysis-by-synthesis' (Yuille and Kersten, 2006; Poeppel et al., 2008) mechanism, wherein the auditory scene is first decomposed into basic acoustic elements, followed by top-down processes that guide the synthesis of the relevant components into a single stream, which then becomes the object of attention. The remainder of the auditory scene would be the unsegregated background, which itself might have the properties of an auditory object. When attention shifts, new auditory objects are correspondingly formed, with the old ones now contributing to the unstructured background. Shamma et al. (2011) suggest that this top down influence acts through the principle of temporal coherence. Between the two opposing views, that streams are formed pre-attentively and that multiple streams can co-exist simultaneously, or that attention is required to form a stream and only that single stream is ever present as separated perceptual entity, these findings lend support to the latter.

In summary, these results provide evidence that, in a complex auditory scene with multiple overlapping spectral and temporal sources, the core areas of auditory cortex maintains an acoustic representation of the auditory scene with no significant preference to attended over ignored source, and with no separation into distinct sources.

It is only the higher-order auditory areas that provide an object based representation for the foreground, but even there the background remains unsegregated.

# 5 Delta vs. Gamma Auditory Steady State Synchrony in Schizophrenia

## 5.1 Introduction

Many schizophrenia-related symptoms, such as auditory hallucination, speech disorganization, disorganized thoughts, and verbal working memory deficits are in the auditory–verbal domain, suggesting that the schizophrenia disease process impacts the auditory processing pathway. Electrophysiological abnormalities in schizophrenia are consistently reported in patients during auditory paradigms such as auditory mismatch negativity (Javitt et al., 1996; Light and Braff, 2005), steady-state response (Kwon et al., 1999; Hirano et al., 2015), sensory gating (Freedman et al., 1996; Hong et al., 2008), and word, language and speech processing (Ford et al., 2007; Kiang et al., 2008). Auditory–verbal processing deficits in schizophrenia may thus be associated with fundamental electrophysiological deficits in the auditory processing network.

Cortical oscillations are thought to play an important role in cognitive functioning, communication, and integration of information across different regions of the brain (Basar et al., 2001; Ward, 2003; Uhlhaas et al., 2009). In healthy subjects, low frequency oscillations (<10 Hz) regulate speech processing (Ding and Simon, 2012a), where accurate perception of attended speech is associated with more precise delta band (1-4 Hz) neuronal responses (Ding et al., 2013; Zion Golumbic et al., 2013b) than in other bands. These low frequency oscillations, especially in the delta band, appear to serve a stabilization and enhancement function while attending to, and during the processing of, auditory streams (Ding and Simon, 2014). Auditory selective attention is also associated with the entrainment of ongoing neuronal oscillations in the delta band, which modulates neuronal excitability in primary auditory cortex (Lakatos et al., 2013b). We hypothesized that in schizophrenia a reduced ability to generate synchronous delta oscillations in response to auditory stimuli would disrupt auditory processing, leading to auditory-verbal domain problems.

The auditory steady-state response (ASSR) can be used to test the integrity of cortical oscillatory activity (Picton et al., 2003; Uhlhaas and Singer, 2010; O'Donnell et al., 2013). It is a robust activation paradigm to elicit frequency-specific auditory responses. It is generated using stimuli that are repeated (periodic) at a specified frequency, resulting in electroencephalographic neural entrainment at the presentation frequency. We used a 2.5 Hz (mean of 1 to 4 Hz) stimulus train to elicit ASSR in the delta band, and to test the joint hypotheses that 1) schizophrenia is associated with an inability to support delta synchronization, and 2) this impairment is associated with cognitive disturbances and other symptoms in the auditory–verbal domain.

Our study is the first to investigate the delta (1-4 Hz) range ASSR in schizophrenia. Previous studies using attention-based paradigms have shown that delta entrainment is associated with clinical symptoms in schizophrenia (Lakatos et al., 2013a). ASSR has appeal in translational research for studying intrinsic neurobiological abnormalities without the need to rely on explicit behavioral performance. The first study to investigate ASSR in the 20 to 40 Hz range in schizophrenia reported reduced ASSR at 40 Hz (Kwon et al., 1999). Subsequent ASSR studies expanded the range down to 5-10 Hz or up to 80-160 Hz (Hamm et al., 2011;

Tsuchimoto et al., 2011), and have generally confirmed ASSR deficits in schizophrenia (Clementz et al., 2004; Hong et al., 2004; Light et al., 2006; Spencer et al., 2008; Brenner et al., 2009a; Hamm et al., 2011; Tsuchimoto et al., 2011; Kirihara et al., 2012; Hirano et al., 2015) primarily in the 40 to 80 Hz range. The strong interest in 40 Hz has also been justified by finding a reduced 40 Hz ASSR in the first-degree relatives (FDR) of schizophrenia patients (Hong et al., 2004; Rass et al., 2012). FDR typically have about a 10-fold increase in risk for schizophrenia compared with the general population, although risks do not necessitate a transition to psychosis as the rate of schizophrenia is about 10% in FDR (Kendler et al., 1993; Erlenmeyer-Kimling et al., 1997). Given these findings, we investigated delta band ASSR, along with higher frequencies at 5, 10, 20, 40, and 80 Hz, with a focus on assessing the relationship between delta (2.5 Hz) vs. gamma (40 to 80 Hz). This design also allowed an unbiased assessment across a very broad range of frequencies, to determine whether there may be frequency specificity in schizophrenia psychopathology.

## 5.2 Materials & Methods

Participants: The study included 128 schizophrenia spectrum disorder (SSD) patients and 108 healthy controls (HC) (Table 5.1). The Structured Clinical Interview for DSM was used to make Axis I diagnoses. All patients were recruited from outpatient clinics; media advertisements were used for HC. Subjects with medical and neurological illnesses, head injury, and substance dependence or abuse (except nicotine) were excluded. Six patients were not on antipsychotics, 19 on typical, 74 on one atypical, 18 on more than one atypical, and 11 on a combination of atypical and typical antipsychotics. Patients on daily GABAergic hypnotics were excluded. Significant findings in SSD were re-examined in

FDR of patients (n=55) who have no psychosis. 70% of the FDR were from families of the patients in this study. All patient probands of the FDR were interviewed by SCID regardless of whether the patients were in the current study.

Table 5.1: Demographic and clinical information

| | HC<br>(n=108) | SSD<br>(n=128) | FDR<br>(n=55) | HC vs. SSD | | HC vs. FDR | |
|---|---|---|---|---|---|---|---|
| | | | | F or $\chi^2$ | p value | F or $\chi^2$ | p value |
| Age mean (SD) | 37.9<br>(13.8) | 37.8<br>(13.1) | 46.6<br>(13.6) | 0.003 | 0.96 | 15.1 | <0.001** |
| %Male | 65.7 | 67.2 | 31.6 | 0.06 | 0.89 | 17.5 | <0.001** |
| Verbal working memory | 20.4<br>(4.4) | 17.0<br>(5.3) | 18.9<br>(5.1) | 24.8 | <0.001** | 3.4 | 0.067 |
| Auditory perception trait | 3.8<br>(4.8) | 19.4<br>(11.8) | 4.8<br>(7.7) | 140.3 | <0.001** | 0.9 | 0.35 |
| Auditory perception state | 1.1<br>(2.4) | 9.8<br>(12.0) | 3.8<br>(4.8) | 46.4 | <0.001** | 0.5 | 0.46 |
| BPRS | n/a | 40.4<br>(11.2) | n/a | n/a | n/a | n/a | n/a |

BPRS: Brief Psychiatric Rating Scale; HC: healthy controls; SSD: schizophrenia spectrum disorder patients; FDR: first degree relatives of SSD patients

Auditory Clinical and Cognitive Symptoms: We developed the Auditory Perceptual Trait and State Scale (APTS) to measure perceptual abnormalities. The anomalies are rated for "trait", defined as longitudinally experienced symptoms over one's lifetime, and "state", defined as symptoms recently experienced in the past week. The full scale is available at http://www.mdbrain.org/APTS.pdf. The APTS is self-rated. Its test-retest reliability was assessed in 41 participants about 4 months apart, which showed ICC=0.81 for both the trait and state measures, suggesting good reliability. The Brief

Psychiatric Rating Scale (BPRS) was used to rate overall symptoms. The APTS was administered to all participants; the BPRS only to patients. Finally, auditory-verbal working memory was assessed using the digit sequencing task (Hong et al., 2004; Rass et al., 2012).

ASSR Paradigm: ASSRs were recorded in a sound-attenuated chamber while participants listened to click trains, delivered by headphones, at 2.5, 5, 10, 20, 40, and 80 Hz. Seventy-five stimulus trains (trials) each consisting of 15 clicks, with each click at 72 dB and of 1 ms duration were delivered at each stimulus frequency. The duration ranged from 6 s per train for 2.5 Hz, to 0.1875 s per train for 80 Hz. The inter-train interval was 0.7 s. Therefore, the durations for 2.5, 5, 10, 20, 40, and 80 Hz were 8.38, 4.69, 2.82, 1.89, 1.42, and 1.19 minutes, respectively, presented in six separate blocks separated by two minutes. The order of the blocks was randomized. This design allows steady-state neural entrainment for each frequency (Figure 5.2 & Figure 5.3). A hearing screening test excluded apparent hearing impairment. EEG was recorded using a 64 electrode Quick-Cap with sintered Ag/Ag chloride electrodes and a Neuroscan SynAmp$^2$ (Compumedics, Charlotte, NC) at 1000 Hz with a 0.1 to 200 Hz bandpass filter. Impedance was kept below 5 kΩ. Offline, electrodes were average referenced, highpass-filtered at 0.8 Hz, and detrended. Ocular artifacts were removed using the time-shift-PCA algorithm, with ocular channels as references(de Cheveigne and Simon, 2007). The full-duration waveforms from each channel were epoched into 75 individual trials.

Normalized ASSR Power: While typical ASSR analysis uses individual channels (often CZ or FZ), we adapted signal processing techniques (Wang et al., 2012; de Cheveigne and Arzounian, 2015) where individual EEG channels are spatially combined to maximize response reliability using the Denoising Source Separation (DSS) algorithm (Särelä and Valpola, 2005a; de Cheveigne and Simon, 2008; de Cheveigne and Parra, 2014). DSS is a blind source separation technique related to Principal Component Analysis (PCA) and Independent Component Analysis (ICA) but specifically designed for use with data from multi-trial evoked responses or narrowband signals. DSS works by enhancing stimulus-driven activity over stimulus-unrelated activity, with its components ordered according to their reliability(Särelä and Valpola, 2005a; de Cheveigne and Simon, 2008; de Cheveigne and Parra, 2014).

Raw ASSR power was calculated as the magnitude squared of the Fourier transform at the stimulus frequency. The Fourier transform was calculated using concatenated trials rather than averaged trials to increase spectral resolution (Elhilali et al., 2009; Xiang et al., 2010). Background power was calculated as average spectral power over 1 Hz width frequency bands (on either side of the stimulus frequency, after leaving a guard band of 0.5 Hz on either side). Normalized ASSR power was then calculated as the mean over DSS components of the ratio of raw ASSR power and respective background power. This normalization with respect to background power dramatically reduces subject-to-subject variability of frequency response profiles (Elhilali et al., 2009). This combined use of DSS and normalized ASSR power represents the two critical improvements over previous ASSR power extraction methods. The accompanying reduction of noise is particularly critical for low frequency ASSR, which is known to be more susceptible to background low frequency fluctuations (Picton et al., 2003; Wang et al., 2012).

ASSR Phase Locking Value (PLV): PLV (Jervis et al., 1983) has been extensively used in ASSR analysis. Increased variability of neural responses across trials reduces the PLV value towards 0, where as increased reliability increases the value towards 1 (Herrmann et al., 2013). First, intra-electrode PLV was calculated for each stimulus frequency at each electrode according to the following formula

$$PLV = \frac{1}{N}\left|\sum_{k=1}^{N} e^{-i\theta_k}\right|$$

where $N$ is the number of trials and $\theta_k$ denotes the phase at the steady state frequency from the Discrete Fourier Transform. Based on topographic analysis showing that ASSR was strongest at fronto-central locations (Figure 5.1), the PLV of 16 fronto-central electrodes (AF3, AFZ, AF4, F3, F1, FZ, F2, F4, FC3, FC1, FCZ, FC2, FC4, C1, CZ, C2) were averaged and used for the final PLV assessment.

Statistics: Data processing was performed without the knowledge of group and demographic information. Repeated measures ANOVA was performed to compare normalized ASSR power by stimulus frequency (six) and group (SSD vs. HC). The Greenhouse-Geisser correction was applied. Significant effect was followed by post-hoc comparison using Bonferroni correction (p<0.008). If a significant difference was found for SSD vs. HC, we then further tested whether the same frequency was significantly different between FDR vs. HC (no further Bonferroni correction was applied). A similar analysis was followed for the PLV measure. Contributions of ASSR to clinical measures were examined using stepwise linear regression, where at each step the ASSR power at the six frequencies were the predictors and one clinical measure was the dependent variable. Multi-collinearity was examined using variance inflation factor (VIF) (Stevens, 2002). A regression model was considered significant if the overall model was significant at p<0.05 and all predictors had VIF<5. All tests were two-tailed.

## 5.3 Results

ASSR in Schizophrenia Spectrum Disorder Patients

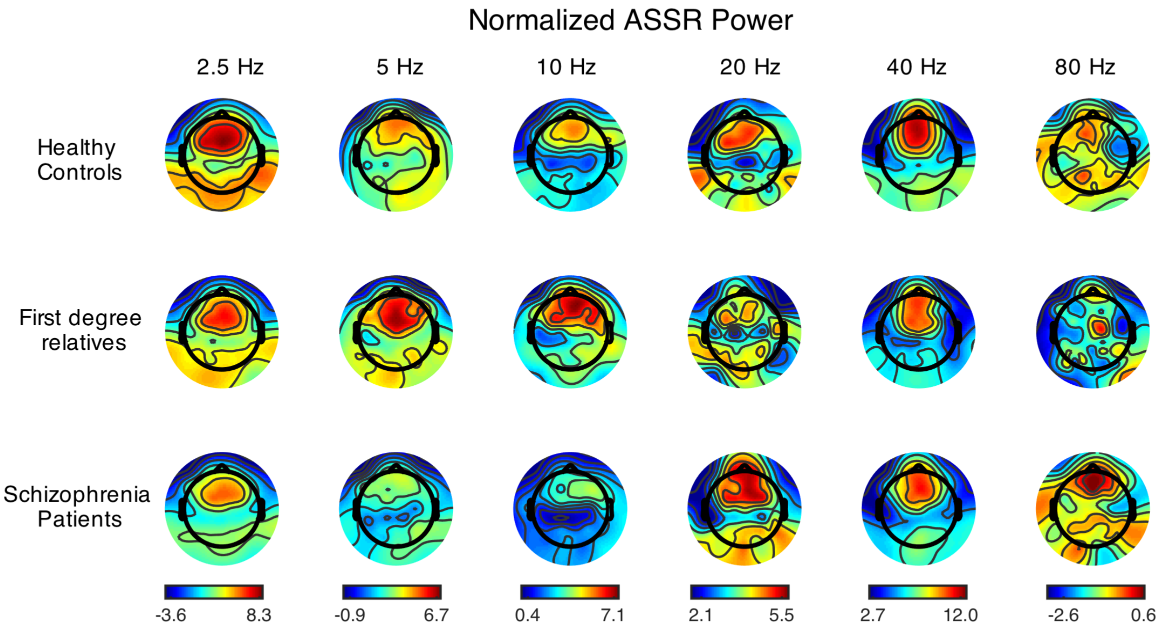

Figure 5.1: Grand averages of topographies of normalized ASSR power for HC, FDR and SSD patients. Scaled based on lowest (blue) to highest power value (red) within each frequency. Refer to Results for statistical group differences.

Figure 5.1 shows that the spatial distribution of normalized ASSR power has a fronto-central accentuation; Figure 5.2 & Figure 5.3 shows grand average time courses of the ASSR responses from electrode FZ and first DSS component respectively. Repeated measures ANOVA on normalized ASSR power extracted by DSS showed significant effects for stimulus frequency (F=390.1, p<0.001), group (SSD vs. HC; F=13.4, p<0.001) and a frequency × group interaction (p=0.039). Post-hoc tests showed that the SSD group had significantly reduced power at 2.5 Hz (F=18.3, p<0.001), 5 Hz (F=10.1, p=0.002), 10 Hz (F=9.9, p=0.002), and 40 Hz (F=8.7, p=0.004), but not 20 Hz (p=0.18) or 80 Hz (p=0.03) after Bonferroni correction (Figure 5.4A). When the analogous analysis was performed on ASSR responses without the use of DSS (e.g., from the single electrode FZ) and without normalization, most findings of significance were lost: only 40 Hz ASSR showed nominally significant reduction in SSD compared with HC (p=0.017), which was then lost after correcting for multiple comparisons**.**

To formally test the frequency × group interaction between 2.5 Hz and 40 Hz, ANOVA was repeated contrasting these frequencies. It showed significant group (p<0.001) and interaction effects (p=0.023), where the interaction was due to a greater reduction of 2.5 Hz ASSR than of 40 Hz ASSR, in patients compared with controls, as seen in Figure 5.4A.

Re-examining these findings using PLV, significant effects were seen for frequency (p<0.001) and group (p<0.001) without interaction (p=0.09). Patients had reduced PLV at 2.5 Hz (F=9.5, p=0.002), 5 Hz (F=8.2, p=0.004), 10 Hz (F=5.9, p=0.016), 40 Hz (F=5.3, p=0.022) and 80 Hz (F=4.0, p=0.045) but not 20 Hz (p=0.50). Findings from 2.5, 5, 10, and 40 Hz replicated power-based analyses and thus no further Bonferroni correction was applied (Figure 5.4B). Therefore, reduced ASSR was found in 2.5, 5, 10, and 40 Hz in both power and phase based analysis.

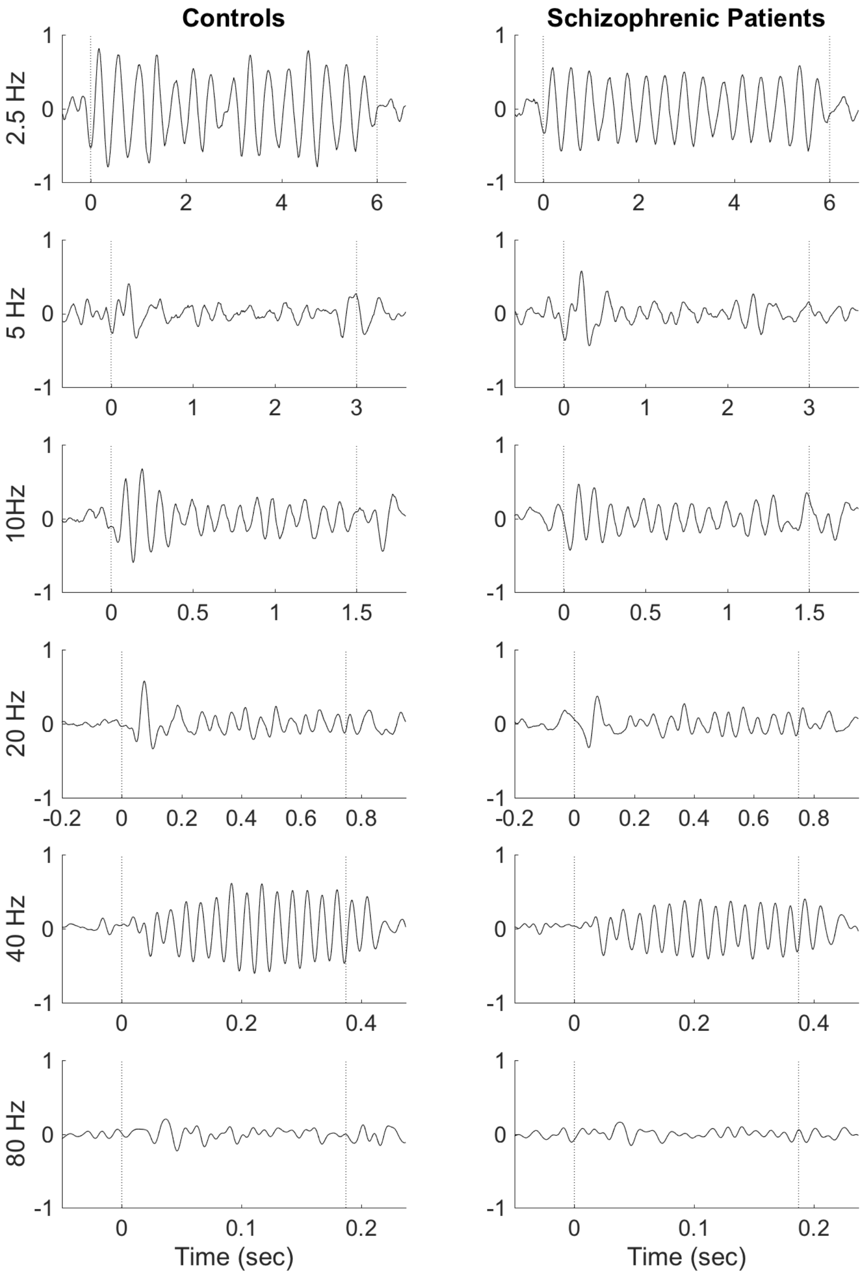

Figure 5.2: Time-domain grand averages from electrode FZ. The vertical axis shows amplitude in μV. The vertical dotted lines indicate begin and end points of a stimulus train. Preferential entrainment in the delta (2.5 Hz) and gamma (40 Hz) bands can be seen in both the controls and schizophrenic patient

group, and 2.5 Hz and 40 Hz stimuli are also associated with larger patient-control differences.

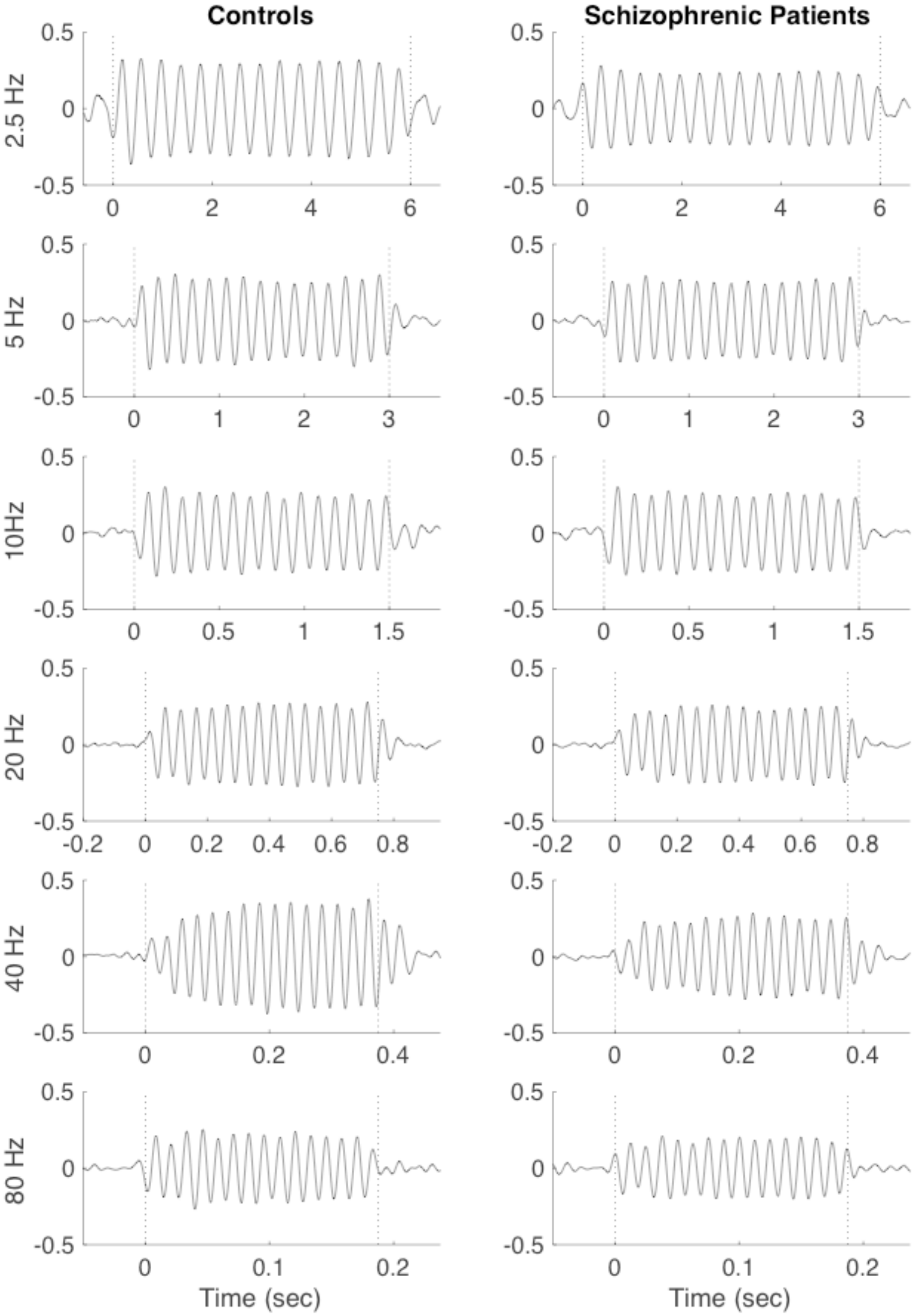

Figure 5.3: Time-domain grand averages using the first DSS component. The vertical dotted lines indicate begin and end points of a stimulus train. DSS is

able extract ASSR responses with higher SNR compared to single electrode responses (Figure 5.2). In controls, 2.5 Hz and 40 Hz stimuli elicit larger ASSR amplitudes than at other frequencies, and 2.5 Hz and 40 Hz stimuli are also associated with larger patient-control differences.

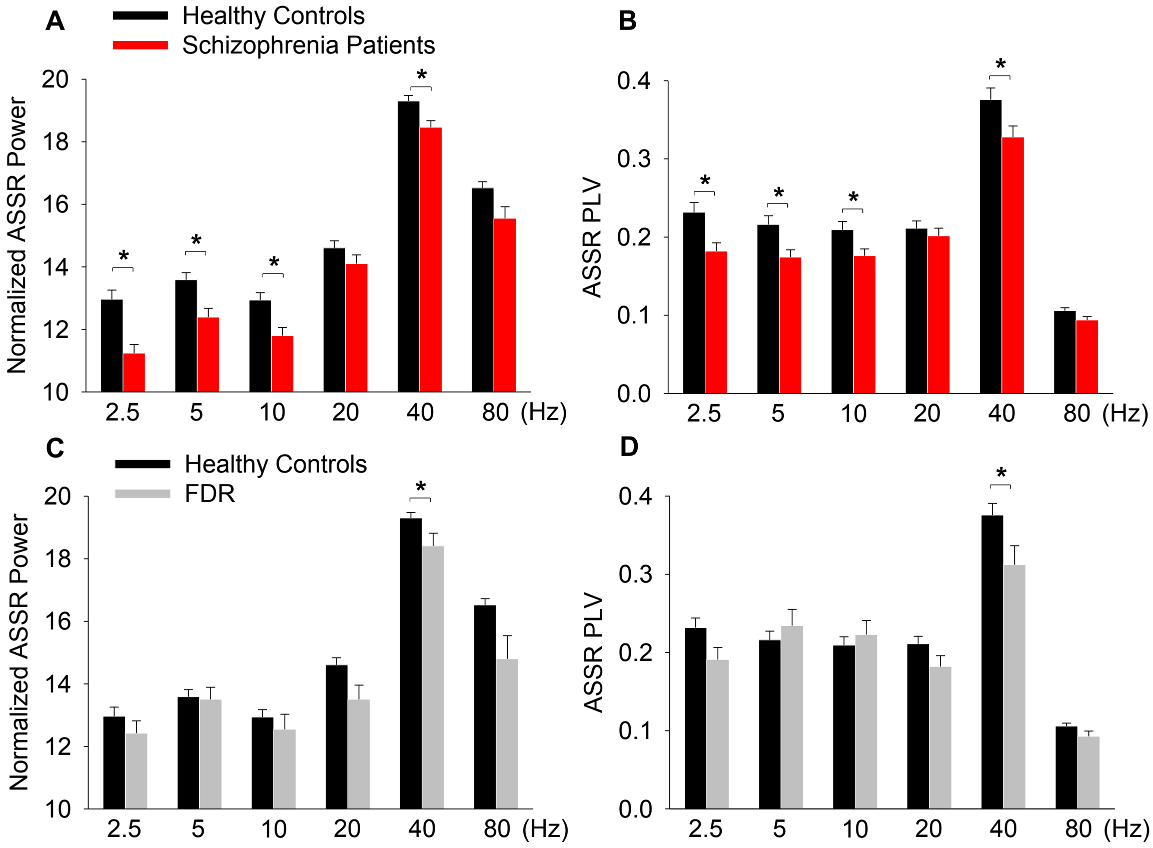

Figure 5.4: Mean and s.e. of normalized power (in dB) and phase locking values (PLV). **A**: Power at 2.5, 5, 10 and 40 Hz are significantly lower for patients than controls. **B**: Replicable findings using PLV. **C** and **D**: FDR showed replicable ASSR reduction compared with controls only at 40 Hz. * Statistically significant. Effect sizes are tabulated in Appendix.

### ASSR in First Degree Relatives

Age and sex were not matched between FDR and HC (Table 5.1). However, neither age ($p=0.44$) nor sex ($p=0.44$) were significant in the repeated measures ANCOVA and so were removed. The results showed significant effects for stimulus frequency ($F=171.0$, $p<0.001$), group ($F=5.1$, $p=0.025$), and frequency × group interaction ($p=0.036$) in FDR vs. HC. Post-hoc tests at the frequencies for which SSD and HC were significantly different (2.5, 5, 10 and 40 Hz) showed that FDR had lower ASSR power than HC at 40 Hz ($F=5.4$, $p=0.022$) but not at 2.5, 5, or 10 Hz ($p=0.28$ to 0.85) (Figure 5.4C). Only the reduction in gamma band ASSR at 40 Hz was considered a replication of findings in patients.

For PLV, age ($p=0.71$) and sex ($p=0.88$) were not significant. There was a significant stimulus frequency effect ($F=84.4$, $p<0.001$) and a frequency × group interaction ($F=3.5$, $p=0.007$). Post hoc tests showed that only the 40 Hz ASSR reduction ($F=5.5$, $p=0.021$) was replicated (Figure 5.4D).

In summary, findings were largely consistent between normalized power and PLV except with PLV generally having smaller effect sizes (Figure 5.4**;** Appendix Table A1 & A2). In subsequent analyses, we opted to use only normalized power based ASSR.

### ASSR and Verbal Working Memory (VWM)

Working memory is impaired in SSD (Barch et al., 2009; Forbes et al., 2009). While the SSD group had lower VWM compared with HC ($p<0.001$, Table 5.1), the FDR did not significantly differ from HC on VWM ($p=0.067$). The regression model was significant in SSD ($F=15.8$, $\Delta R^2=11.8\%$, $p<0.001$; all $VIFs<1.5$) where only 2.5 Hz ASSR significantly contributed to VWM ($t=4.0$, $p<0.001$): patients with lower delta power showed worse VWM ($r=0.36$, $p<0.001$) (Figure 5.5A). The correlation of 2.5 Hz ASSR with VWM was not significant in either the HC or FDR groups.

We calculated the correlation coefficients between VWM and ASSR at each frequency: 2.5 Hz: $r=0.34$, $p<0.001$; 5 Hz: $r=0.30$, $p=0.001$; 10 Hz: $r=0.22$, $p=0.016$; 20 Hz: $r=0.$ 19, $p=0.040$; 40 Hz: $r=0.20$, $p=0.033$; and 80 Hz: $r=0.15$, $p=0.11$. The relationship between ASSR and VWM, quantified through these correlation coefficients, was strongly linked to the stimulus frequency ($r=-0.95$, $p=0.003$) (Figure 5.5C): the correlation between ASSR and VWM significantly decreases with increasing stimulus frequency.

The model was also significant in FDR ($F=9.8$, $\Delta R^2=17.5\%$, $p=0.003$; $VIFs<3.6$) where only the 40 Hz ASSR significantly contributed to VWM ($t=3.13$, $p=0.003$) (Figure 5.5B). The model was not significant in controls (model $p>0.05$).

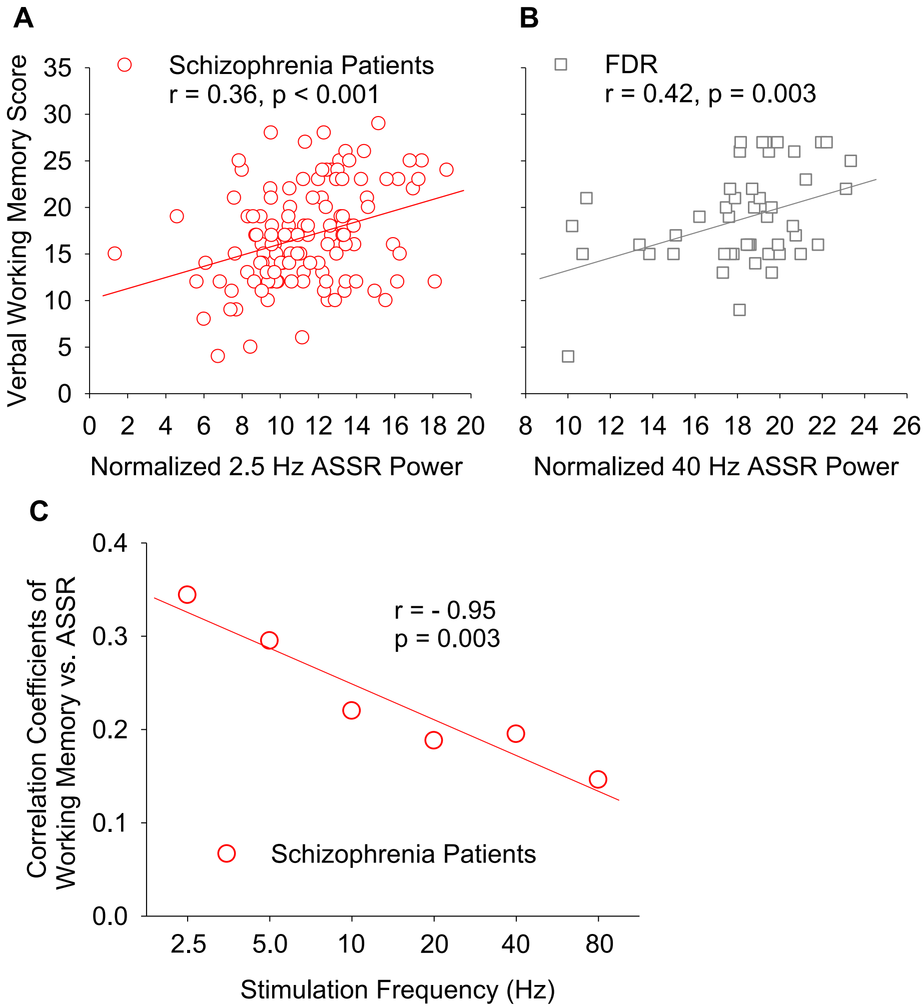

Figure 5.5: Frequency-specific associations between verbal working memory and ASSR. **A**: In SSD, higher 2.5 Hz ASSR was associated with better working memory. **B**: In FDR, 40 Hz ASSR was associated with working memory. **C**: The ASSR-working memory relationships (by their correlation coefficients: y axis) were strongly (negatively) associated with stimulus frequencies.

### ASSR and Auditory Perception Abnormality

A regression model with APTS trait score as the dependent variable and normalized ASSR power as predictors in the SSD group was significant (F=7.7, $\Delta R^2$=13.8%, p=0.001; VIFs<1.5). Only 2.5 Hz (t=-2.8, $\Delta R^2$=6.8%, p=0.007) and 40 Hz (t=3.6, $\Delta R^2$=6.9%, p=0.001) normalized ASSR powers were significant predictors but in opposite directions: reduced 2.5 Hz and increased 40 Hz ASSR were associated with more longitudinally experienced auditory symptoms in SSD patients (Figure 5.6B and Figure 5.6C). The model was not significant for state auditory symptoms (p=0.056) although the trends were the same.

### Medication and Other Clinical Measurements

Chlorpromazine equivalent (CPZ) of antipsychotic dosages was not correlated with ASSR power at any frequency (all r<0.13, all p>0.20). BPRS total or psychosis score was not significantly correlated with ASSR power at any frequency (all r<0.11, all p>0.30).

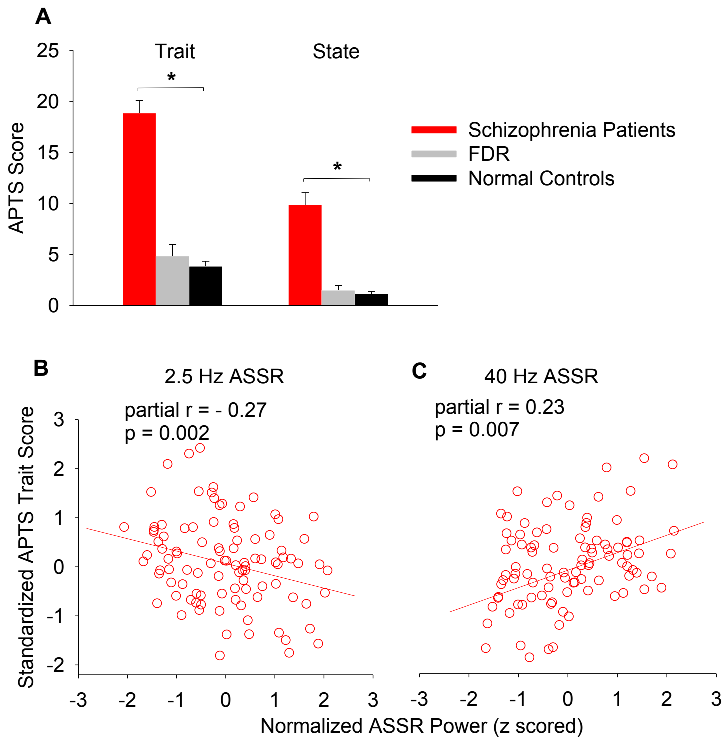

Figure 5.6: Auditory perception as measured by APTS was significantly higher (*) in patients compared with controls as experience in lifetime (trait) or the past seven days (state), but not in FDR compared with controls (**A**). Regression analyses reviewed that 2.5 Hz (**B**) and 40 Hz (**C**) ASSR contributed to auditory perceptual trait score in patients but in opposite directions. Partial r refers to having partialled out the effects of 2.5 Hz for 40 Hz, or vice versa, in the regression analyses.

## 5.4 Discussion

We found that delta and gamma ASSR were both reduced in patients, with delta showing a more pronounced reduction. Critically, reduced delta ASSR was associated both with more severe longitudinally experienced auditory symptom “traits” and also poorer verbal working memory. The observed reduction in gamma ASSR, on the other hand, which was also present in non-psychotic FDR, was found to be more associated with the risk of SSD than with SSD itself.

The finding of reduced 40 Hz ASSR in SSD replicates other studies (Kwon et al., 1999; Light et al., 2006; Spencer et al., 2008; Vierling-Claassen et al., 2008) while the finding of reduced delta ASSR is new. While delta band power may be significantly increased in SSD in resting EEG (Sponheim et al., 1994), delta band oscillations in SSD are also found to be significantly reduced in stimulus or behavior activated paradigms (Ford et al., 2002; Brenner et al., 2003; Ford et al., 2008; Basar-Eroglu et al., 2009; Bates et al., 2009; Doege et al., 2009; Hamm et al., 2011; Donkers et al., 2013). Of particular relevance are findings of reduced delta power and fronto-temporal coherence during talking (Ford et al., 2002) and auditory target detection (Ford et al., 2008) in SSD. Unlike a task-related auditory paradigm depending on performance, the delta ASSR is a passive paradigm and its deficit may indicate difficulty in generating normal delta synchronization to auditory stimuli. Reduced ability to generate delta entrainment might serve as a tangible mechanism for the frequently observed auditory–verbal domain issues in SSD, as neural entrainment in the delta and theta band is critical for normal speech perception (Ding and Simon, 2012a; Ding et al., 2013; Zion Golumbic et al., 2013b).

Auditory hallucinations are experienced by most patients with SSD in their lifetime (Sartorius et al., 1974; Andreasen, 1991). We tested the hypothesis that an ASSR delta deficit contributes to their auditory symptoms, as low frequency temporal modulations (<4 Hz) are more critical in speech perception than faster (22–40 Hz) modulations (Ding and Simon, 2012a; Ding et al., 2013; Zion Golumbic et al., 2013b). This hypothesis was supported. Patients have impaired ability to generate delta ASSR more so than any other frequency tested, and this deficit is associated with more severe longitudinally experienced auditory anomalies. We did not find this correlation with "state" symptoms in APTS or BPRS, perhaps due to fluctuations in state symptoms or to variability in treatment or treatment response.

ASSR based delta entrainment was significantly associated with VWM (Figure 5.5A). The linear relationship between stimulus frequency and the ASSR-VWM correlation coefficients (r=-0.95; Figure 5.5C) further highlights a potentially prominent role of low frequency oscillations in the auditory cognitive system in patients with SSD.

In evaluating the association between ASSR power and VWM, we observed that the most strongly correlated frequency band changed from delta with SSD to gamma with FDR (Figure 5.5A vs. Figure 5.5B). In the auditory cortex, the amplitude of neural oscillations are controlled in a nested fashion, where delta (1-4 Hz) phase modulates theta (4-8Hz) amplitude, and theta phase modulates gamma (30+ Hz) amplitude(Lakatos et al., 2005). This oscillatory hierarchy is thought to control baseline excitability (Lakatos et al., 2005). Under this assumption, we speculate that in individuals without a major deficit in delta generation, as in FDR, gamma band abnormality may yield a more apparent relationship with VWM (Figure 5.5B). However, in individuals with major deficits in delta generation, as in the patients, delta deficits may exert a more fundamental role and thus stronger contribution to VWM (Figure 5.5A & Figure 5.5C).

Figure 5.2 illustrates the 'preferential' entrainment at 2.5 Hz and 40 Hz using absolute power analysis even at a single electrode. The special 40 Hz entrainment in human brains is well known but the 2.5 Hz case is a new observation. Using normalization and DSS analysis, the reduction of 2.5 Hz ASSR for SSD was significant, but not when using simple spectral power; this may explain the lack of earlier observations. DSS and normalization reduce variability separately in delta ASSR (normalization does not improve gamma ASSR), allowing the delta reduction to even surpass the gamma reduction.

Delta ASSR was not significantly different between FDR and controls, suggesting that this deficit does not indicate a genetic vulnerability for SSD. The finding of reduced 40 Hz ASSR in FDR replicated our previous finding (Hong et al., 2004), now in an independent, much larger cohort. Combined with another independent replication (Rass et al., 2012), the data support a 40 Hz ASSR deficit as a genetic biomarker for SSD.

Greater 40 Hz ASSR (within overall reduction) in patients was associated with more auditory symptoms (Figure 5.6C). This matches findings in the visual domain, where higher gamma during gestalt perception was associated with more visual hallucinations (Spencer et al., 2004). Recent animal and human studies are converging to show that glutamatergic receptor antagonists increase gamma neural oscillations (Hong et al., 2010; Sullivan et al., 2015). A leading hypothesis in psychosis generation is excitatory glutamatergic receptor hypofunction, based on observations that glutamatergic receptor antagonism by phencyclidine and ketamine mimics aspects of schizophrenia symptomatology (Kantrowitz and Javitt, 2010; Snyder and Gao, 2013). Therefore, the link between higher gamma power and more visual and auditory symptoms could be through abnormal glutamatergic mechanisms.

Gamma oscillations are generated by inhibitory GABAergic interneurons regulating excitatory glutamatergic pyramidal neurons (Bartos and Elgueta, 2012; Gonzalez-Burgos and Lewis, 2012). Abnormal GABAergic regulation of gamma is thought to underlie working memory deficits in SSD (Maldonado-Aviles et al., 2009; Hines et al., 2013; Kim et al., 2015) and is often hypothesized to be genetic in origin (Straub et al., 2007). Therefore, reduced gamma ASSR in FDR, and correlation with working memory in FDR (Figure 5.5B), appear to support the hypothesis that gamma-working memory deficit confers risk for SSD. The neural mechanisms underlying delta oscillations are less well understood. Studies of sleep and waking state delta oscillations (Neske, 2015) suggest that N-Methyl-D-aspartate (NMDA) receptors play a role in maintaining these slow oscillations, and the NMDA receptor antagonist ketamine reduces slow wave 1-5 Hz oscillations(Hong et al., 2010). Whether reduced delta ASSR reflects an NMDA hypofunction origin of schizophrenia (Coyle, 2012) would require follow-up studies.

Compared to conventional single-channel-based spectral power analysis, this study employed the techniques of DSS, which integrates over channels, and normalization, which takes into account background power; both contribute separately to increase statistical power in the ASSR analysis. Normalization particularly improves ASSR analysis at lower frequencies, due to the 1/f nature (strong rise at low frequencies) of noisy background activity in electrophysiological recordings (Miller et al., 2009; Voytek et al., 2015). DSS enhances amplitude contrast, due to its ability to optimally combine responses across electrodes and so extract ASSR responses with higher fidelity. DSS performs only spatial, not spectral filtering, and hence does not introduce artifacts associating with spectral filtering.

An important limitation is that we did not test for the specificity of the findings. We tested auditory working memory, but not deficits in other cognitive domains. This limits interpretation regarding whether the correlations with clinical features were specific to auditory working memory or more general to other cognitive deficits. Reduced delta frequency ASSR in SSD might also arise from antipsychotic medications, though no correlations were found between delta frequency ASSR and current antipsychotic medication dosage. Finally, the failure to synchronize to delta frequency stimulation could also be related to abnormal baseline delta activity, although the DSS procedure was designed to account for this effect.

In summary, the results from this study support that inadequate ability to sustain neural oscillatory responses in the lower frequency range may play a role in the auditory perceptual and cognitive deficit mechanisms in schizophrenia. The findings from this study support the use of delta range ASSR as part of the effort to build translational animal models to study the etiology of SSD.

# 6 Summary

Most normal-hearing listeners can hold a conversation with a partner in everyday settings with an ease that is unmatched by any artificial recognition system available today. Listeners can reliably extract meaning from a sound source in a cacophony of multiple competing talkers, broadband machine noise, room reflections etc. in busy offices, crowded restaurants, noisy streets and so on. Inspired from Cherry (1953) these complex auditory scenarios are generally referred to as "cocktail party problems" and Bregman (1994) dubbed our ability to solve these problems as "auditory scene analysis". While we hear with ears, it is with the brain that we listen. Neural computations in the brain facilitate our ability to navigate these complex auditory scenarios and reasonably, many neurological illnesses are correlated with deficiencies in cortical auditory processing. In this dissertation, we investigated the neural representations, which form the basis for neural computations, in two different abstractions of cocktail party. Listening to continuous speech in (1) reverberant as well as noisy environments and (2) in the presence of multiple competing background talkers. Further, using auditory steady state response (ASSR) paradigm, we showed that delta band neural responses, which are commensurate with slow temporal rhythms in speech, are better correlated with auditory processing deficits observed in schizophrenia than historically focused gamma band responses.

Distinct from existing studies, which investigated cortical encoding of speech either in anechoic conditions or corrupted by simple additive noise, the study in chapter 3 focused on cortical encoding of speech distorted by noise as well as the reverberation which is ubiquitous in our daily listening conditions. Evidence shown in this work suggests that cortex maintains both distorted and distortion free representation of speech in reverberant listening conditions. Further, the results showed that noise affected cortical encoding of speech only in presence of reverberation, arguing for differential encoding mechanisms for additive and convolutive distortions of speech, with implications for distortion robust speech perception observed in humans. Given the limited spatial resolution of current MEG source localization techniques, especially in presence of correlated sources, the study presented here did not explore the corresponding regions of brain hosting distorted and distorted free representations of reverberant speech. Hence, it would be of great interest to replicate the current study using intercranial recording techniques. Further, it would be of interest to investigate the effect of binaural cues, given their relevance in accurate speech perception in noisy conditions.

Elaborating on the important mechanistic questions that remain open since Cherry's seminal cocktail party work, the study presented in chapter 4 addressed how speech streams in a multi-talker auditory scene are represented, parsed and attended in different hierarchical levels of auditory cortex as measured by how the speech envelope is represented at lower (1-8 Hz) MEG frequencies. The results strongly suggest that early auditory cortex represents auditory scene acoustically and holistically, not as objects and with out any preference for attended speech over unattended. Later auditory cortex represents auditory scene in terms of objects with attended speech represented much more strongly than unattended. Further, the attentional "background," comprising those parts of auditory scene that are not the focus of attention, is shown to be organized as an unsegregated background, rather than separate individual objects. The results imply that the auditory object segregation is influenced by attention, though not measured directly here, rather than pre-attentive – an important debate in the field, with implications for all downstream speech processing in attention, memory and language. An interesting future direction for this work is to study the neural representation of auditory scene with additional streaming cues such as directional sources, differing statistics of auditory sources as well as integration with visual cues.

Given that low frequency neuronal responses can reliably track speech and that schizophrenia patients exhibit frequent auditory-verbal domain deficits, chapter 5 explored the integrity of neural responses to auditory stimuli using ASSR paradigm over broad range of frequencies. Examining the association between auditory steady state responses and cognitive performance and auditory hallucination symptom severity in patients, first-degree relatives and controls, the results, apart from replicating the gamma band deficiencies observed in patients, revealed for the first time that delta band deficiencies are much greater in patients compared with controls. Interestingly, observed reduction in delta ASSR was better correlated with cognitive deficits observed in patients than gamma band, thus extending and shifting the focus of neural response impairments away from what has been a relatively narrow focus on high-frequency gamma responses in the community and highlights the potential importance of slow wave activity in studying the etiology of schizophrenia. Complimenting the passive listening paradigm used in this study, an important extension would be to use an active listening paradigm, such as a cocktail party used in chapter 4, which also takes in to consideration the attention deficits observed in schizophrenia patients.

# Appendix

Mean, standard error, standard deviation and effect size data for ASSR power and PLV measures for controls vs. patients and controls vs. FDR from chapter 5.

**Table A1: Mean, Standard Error, Standard Deviation, and Effect Size Data: Controls vs. Patients**

| | | **Controls** | | | | | **Patients** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Measure** | **Stimulation Frequency** | **Mean** | **s.e** | **s.d.** | **N** | | **Mean** | **s.e.** | **s.d.** | **N** | **Effect Size** |
| Power | 2.5 Hz | 12.96 | 0.29 | 3.02 | 108 | | 11.24 | 0.27 | 3.08 | 128 | 0.57 |
| | 5 Hz | 13.58 | 0.23 | 2.35 | 108 | | 12.39 | 0.28 | 3.20 | 128 | 0.42 |
| | 10 Hz | 12.94 | 0.24 | 2.43 | 108 | | 11.80 | 0.26 | 2.96 | 128 | 0.42 |
| | 20 Hz | 14.60 | 0.23 | 2.36 | 108 | | 14.10 | 0.28 | 3.17 | 128 | 0.18 |
| | 40 Hz | 19.30 | 0.18 | 1.85 | 108 | | 18.46 | 0.21 | 2.41 | 128 | 0.39 |
| | 80 Hz | 16.52 | 0.20 | 2.04 | 108 | | 15.55 | 0.37 | 4.20 | 128 | 0.29 |
| PLV | 2.5 Hz | 0.23 | 0.01 | 0.13 | 108 | | 0.18 | 0.01 | 0.12 | 128 | 0.40 |
| | 5 Hz | 0.22 | 0.01 | 0.12 | 108 | | 0.17 | 0.01 | 0.11 | 128 | 0.38 |
| | 10 Hz | 0.21 | 0.01 | 0.11 | 108 | | 0.18 | 0.01 | 0.10 | 128 | 0.32 |
| | 20 Hz | 0.21 | 0.01 | 0.10 | 108 | | 0.20 | 0.01 | 0.11 | 128 | 0.09 |
| | 40 Hz | 0.38 | 0.01 | 0.15 | 108 | | 0.33 | 0.01 | 0.16 | 128 | 0.30 |
| | 80 Hz | 0.11 | 0.00 | 0.04 | 108 | | 0.09 | 0.00 | 0.05 | 128 | 0.26 |

**Table A2: Mean, Standard Error, Standard Deviation, and Effect Size Data: Controls vs. FDR**

| | | **Controls** | | | | | **FDR** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Power | 2.5 Hz | 12.96 | 0.29 | 3.02 | 108 | | 12.42 | 0.39 | 2.92 | 55 | 0.18 |
| | 5 Hz | 13.58 | 0.23 | 2.35 | 108 | | 13.50 | 0.39 | 2.86 | 55 | 0.03 |
| | 10 Hz | 12.94 | 0.24 | 2.43 | 108 | | 12.54 | 0.49 | 3.60 | 55 | 0.14 |
| | 20 Hz | 14.60 | 0.23 | 2.36 | 108 | | 13.50 | 0.46 | 3.38 | 55 | 0.40 |
| | 40 Hz | 19.30 | 0.18 | 1.85 | 108 | | 18.41 | 0.41 | 3.02 | 55 | 0.38 |
| | 80 Hz | 16.52 | 0.20 | 2.04 | 108 | | 14.80 | 0.74 | 5.49 | 55 | 0.48 |
| PLV | 2.5 Hz | 0.23 | 0.01 | 0.13 | 108 | | 0.19 | 0.02 | 0.12 | 55 | 0.33 |
| | 5 Hz | 0.22 | 0.01 | 0.12 | 108 | | 0.23 | 0.02 | 0.16 | 55 | -0.14 |
| | 10 Hz | 0.21 | 0.01 | 0.11 | 108 | | 0.22 | 0.02 | 0.13 | 55 | -0.11 |
| | 20 Hz | 0.21 | 0.01 | 0.10 | 108 | | 0.18 | 0.01 | 0.10 | 55 | 0.29 |
| | 40 Hz | 0.38 | 0.01 | 0.15 | 108 | | 0.31 | 0.02 | 0.18 | 55 | 0.39 |
| | 80 Hz | 0.11 | 0.00 | 0.04 | 108 | | 0.09 | 0.01 | 0.05 | 55 | 0.30 |

PLV: Phase locking value

FDR: First degree relatives

# Bibliography

Ahveninen J, Hamalainen M, Jaaskelainen IP, Ahlfors SP, Huang S, Lin FH, Raij T, Sams M, Vasios CE, Belliveau JW (2011) Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proc Natl Acad Sci U S A 108:4182-4187.

Akram S, Englitz B, Elhilali M, Simon JZ, Shamma SA (2014) Investigating the neural correlates of a streaming percept in an informational-masking paradigm. PLoS One 9:e114427.

Alho K, Connolly JF, Cheour M, Lehtokoski A, Huotilainen M, Virtanen J, Aulanko R, Ilmoniemi RJ (1998) Hemispheric lateralization in preattentive processing of speech sounds. Neurosci Lett 258:9-12.

Allen JB, Berkley DA (1979) Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America 65:943-950.

Anderson S, Parbery-Clark A, White-Schwoch T, Kraus N (2012) Aging affects neural precision of speech encoding. J Neurosci 32:14156-14164.

Andreasen NC (1991) Schizophrenia: the characteristic symptoms. Schizophrenia Bulletin 17:27-49.

Arthur JV, Merolla PA, Akopyan F, Alvarez R, Cassidy A, Chandra S, Esser SK, Imam N, Risk W, Rubin DBD, Manohar R, Modha DS (2012) Building block of a programmable neuromorphic substrate: A digital neurosynaptic core. In: Neural Networks (IJCNN), The 2012 International Joint Conference on, pp 1-8.

Assmann P, Summerfield Q (2004) The perception of speech under adverse conditions. In: Speech processing in the auditory system, pp 231-308: Springer.

Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. Proc Natl Acad Sci U S A 106:21894-21899.

Barch DM, Berman MG, Engle R, Jones JH, Jonides J, MacDonald A, III, Nee DE, Redick TS, Sponheim SR (2009) CNTRICS final task selection: working memory. Schizophr Bull 35:136-152.

Bartos M, Elgueta C (2012) Functional characteristics of parvalbumin- and cholecystokinin-expressing basket cells. J Physiol 590:669-681.

Basar E, Basar-Eroglu C, Karakas S, Schurmann M (2001) Gamma, alpha, delta, and theta oscillations govern cognitive processes. Int J Psychophysiol 39:241-248.

Basar-Eroglu C, Schmiedt-Fehr C, Mathes B, Zimmermann J, Brand A (2009) Are oscillatory brain responses generally reduced in schizophrenia during long sustained attentional processing? International journal of psychophysiology : official journal of the International Organization of Psychophysiology 71:75-83.

Bates AT, Kiehl KA, Laurens KR, Liddle PF (2009) Low-frequency EEG oscillations associated with information processing in schizophrenia. Schizophr Res 115:222-230.

Bengio Y (2009) Learning deep architectures for AI. Foundations and trends® in Machine Learning 2:1-127.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57:289-300.

Bertrand O, Perrin F, Pernier J (1985) A theoretical justification of the average reference in topographic evoked potential studies. Electroencephalogr Clin Neurophysiol 62:462-464.

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. J Neurophysiol 84:2426-2439.

Bitterman Y, Mukamel R, Malach R, Fried I, Nelken I (2008) Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature 451:197-201.

Boutros NN, Arfken C, Galderisi S, Warrick J, Pratt G, Iacono W (2008) The status of spectral EEG abnormality as a diagnostic test for schizophrenia. Schizophr Res 99:225-237.

Bregman AS (1994) Auditory scene analysis: The perceptual organization of sound: MIT press.

Brenner CA, Sporns O, Lysaker PH, O'Donnell BF (2003) EEG synchronization to modulated auditory tones in schizophrenia, schizoaffective disorder, and schizotypal personality disorder. Am J Psychiatry 160:2238-2240.

Brenner CA, Kieffaber PD, Clementz BA, Johannesen JK, Shekhar A, O'Donnell BF, Hetrick WP (2009a) Event-related potential abnormalities in schizophrenia: a failure to "gate in" salient information? Schizophr Res 113:332-338.

Brenner CA, Krishnan GP, Vohs JL, Ahn WY, Hetrick WP, Morzorati SL, O'Donnell BF (2009b) Steady state responses: electrophysiological assessment of sensory function in schizophrenia. Schizophr Bull 35:1065-1077.

Brodbeck C (2017) https://doi.org/10.5281/zenodo.438193. In.

Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN (2011) A Generalized Linear Model for Estimating Spectrotemporal Receptive Fields from Responses to Natural Sounds. PLoS ONE 6:e16104.

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

Chait M, Simon JZ, Poeppel D (2004) Auditory M50 and M100 responses to broadband noise: functional implications. Neuroreport 15:2455-2458.

Chait M, Greenberg S, Arai T, Simon JZ, Poeppel D (2015) Multi-time resolution analysis of speech: evidence from psychophysics. Front Neurosci 9:214.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428-1432.

Cherry EC (1953) Some Experiments on the Recognition of Speech, with One and with 2 Ears. Journal of the Acoustical Society of America 25:975-979.

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

Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106:2719-2732.

Clementz BA, Keil A, Kissler J (2004) Aberrant brain dynamics in schizophrenia: delayed buildup and prolonged decay of the visual steady-state response. Brain Res Cogn Brain Res 18:121-129.

Cooke M, Hershey JR, Rennie SJ (2010) Monaural speech separation and recognition challenge. Computer Speech & Language 24:1-15.

Coyle JT (2012) NMDA receptor and schizophrenia: a brief history. Schizophr Bull 38:920-926.

Crosse MJ, Butler JS, Lalor EC (2015) Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. J Neurosci 35:14195-14204.

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

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Network 18:191-212.

David SV, Mesgarani N, Fritz JB, Shamma SA (2009) Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J Neurosci 29:3374-3386.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423-3431.

Davis MH, Scharenborg O (2016) Speech perception by humans and machines. Speech Perception and Spoken Word Recognition:181.

Dayan P, Abbott LF (2001) Theoretical neuroscience: Cambridge, MA: MIT Press.

de Cheveigne A, Simon JZ (2007) Denoising based on time-shift PCA. J Neurosci Methods 165:297-305.

de Cheveigne A, Simon JZ (2008) Denoising based on spatial filtering. J Neurosci Methods 171:331-339.

de Cheveigne A, Parra LC (2014) Joint decorrelation, a versatile tool for multichannel data analysis. Neuroimage 98:487-505.

de Cheveigne A, Arzounian D (2015) Scanning for oscillations. J Neural Eng 12:066020.

Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci 8:1684-1689.

Dean I, Robinson BL, Harper NS, McAlpine D (2008) Rapid neural adaptation to sound level statistics. J Neurosci 28:6430-6438.

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

Delgutte B (1980) Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J Acoust Soc Am 68:843-857.

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85:1220-1234.

Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968-980.

Desmedt JE, Chalklin V, Tomberg C (1990) Emulation of somatosensory evoked potential (SEP) components with the 3-shell head model and the problem of 'ghost potential fields' when using an average reference in brain mapping. Electroencephalogr Clin Neurophysiol 77:243-258.

Devore S, Delgutte B (2010) Effects of reverberation on the directional sensitivity of auditory neurons across the tonotopic axis: influences of interaural time and level differences. J Neurosci 30:7826-7837.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr Biol 25:2457-2465.

Ding N (2012) Temporal coding of speech in human auditory cortex. In: Electrical and Computer Engineering: University of Maryland, College park.

Ding N, Simon JZ (2009) Neural representations of complex temporal modulations in the human auditory cortex. J Neurophysiol 102:2731-2743.

Ding N, Simon JZ (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854-11859.

Ding N, Simon JZ (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78-89.

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728-5735.

Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci 8:311.

Ding N, Chatterjee M, Simon JZ (2013) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88C:41-46.

Doege K, Bates AT, White TP, Das D, Boks MP, Liddle PF (2009) Reduced event-related low frequency EEG activity in schizophrenia during an auditory oddball task. Psychophysiology 46:566-577.

Donkers FC, Englander ZA, Tiesinga PH, Cleary KM, Gu H, Belger A (2013) Reduced delta power and synchrony and increased gamma power during the P3 time window in schizophrenia. Schizophr Res 150:266-268.

Drullman R (1995) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97:585-592.

Drullman R, Festen JM, Plomp R (1994a) Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95:2670-2680.

Drullman R, Festen JM, Plomp R (1994b) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95:1053-1064.

Elhilali M, Xiang J, Shamma SA, Simon JZ (2009) Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol 7:e1000129.

Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. PLoS Comput Biol 5:e1000302.

Erlenmeyer-Kimling L, Adamo UH, Rock D, Roberts SA, Bassett AS, Squires-Wheeler E, Cornblatt BA, Endicott J, Pape S, Gottesman, II (1997) The New York High-Risk Project. Prevalence and comorbidity of axis I disorders in offspring of schizophrenic parents at 25-year follow-up. Arch Gen Psychiatry 54:1096-1102.

Festen JM, Plomp R (1990) Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J Acoust Soc Am 88:1725-1736.

Fischl B (2012) FreeSurfer. Neuroimage 62:774-781.

Forbes NF, Carrick LA, McIntosh AM, Lawrie SM (2009) Working memory in schizophrenia: a meta-analysis. Psychol Med 39:889-905.

Ford JM, Krystal JH, Mathalon DH (2007) Neural synchrony in schizophrenia: from networks to new treatments. Schizophr Bull 33:848-852.

Ford JM, Roach BJ, Hoffman RS, Mathalon DH (2008) The dependence of P300 amplitude on gamma synchrony breaks down in schizophrenia. Brain Res 1235:133-142.

Ford JM, Mathalon DH, Whitfield S, Faustman WO, Roth WT (2002) Reduced communication between frontal and temporal lobes during talking in schizophrenia. Biological Psychiatry 51:485-492.

Freedman R, Adler LE, Myles-Worsley M, Nagamoto HT, Miller C, Kisley M, McRae K, Cawthra E, Waldo M (1996) Inhibitory gating of an evoked response to repeated auditory stimuli in schizophrenic and normal subjects. Human recordings, computer simulation, and an animal model. Arch Gen Psychiatry 53:1114-1121.

Fritz J, Elhilali M, Shamma S (2005) Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. Hear Res 206:159-176.

Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6:1216-1223.

Fuglsang SA, Dau T, Hjortkjaer J (2017) Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage.

Fujihira H, Shiraishi K, Remijn GB (2017) Elderly listeners with low intelligibility scores under reverberation show degraded subcortical representation of reverberant speech. Neurosci Lett 637:102-107.

Galambos R, Makeig S, Talmachoff PJ (1981) A 40-Hz auditory potential recorded from the human scalp. Proc Natl Acad Sci U S A 78:2643-2647.

Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R, Kleinschmidt A (2000) Representation of the temporal envelope of sounds in the human brain. J Neurophysiol 84:1588-1598.

Gonzalez-Burgos G, Lewis DA (2012) NMDA receptor hypofunction, parvalbumin-positive neurons, and cortical gamma oscillations in schizophrenia. Schizophr Bull 38:950-957.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hamalainen MS (2014) MNE software for processing MEG and EEG data. Neuroimage 86:446-460.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hamalainen M (2013) MEG and EEG data analysis with MNE-Python. Front Neurosci 7:267.

Greenberg S, Carvey H, Hitchcock L, Chang S (2003) Temporal properties of spontaneous speech—a syllable-centric perspective. Journal of Phonetics 31:465-485.

Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. Trends Neurosci 25:348-353.

Griffiths TD, Warren JD (2004) What is an auditory object? Nature Reviews Neuroscience 5:887-892.

Hackett TA (2008) Anatomical organization of the auditory cortex. J Am Acad Audiol 19:774-779.

Hackett TA, Stepniewska I, Kaas JH (1998) Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. J Comp Neurol 394:475-495.

Hamalainen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV (1993) Magnetoencephalography - Theory, Instrumentation, and Applications to Noninvasive Studies of the Working Human Brain. Rev Mod Phys 65:413-497.

Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV (1993) Magnetoencephalography\char22{}theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 65:413-497.

Hamalainen MS, Ilmoniemi RJ (1994) Interpreting magnetic fields of the brain: minimum norm estimates. Med Biol Eng Comput 32:35-42.

Hamm JP, Gilmore CS, Picchetti NA, Sponheim SR, Clementz BA (2011) Abnormalities of neuronal oscillations and temporal integration to low- and high-frequency auditory stimulation in schizophrenia. Biological Psychiatry 69:989-996.

Hasey GM, Kiang M (2013) A review of recent literature employing electroencephalographic techniques to study the pathophysiology, phenomenology, and treatment response of schizophrenia. Curr Psychiatry Rep 15:388.

Hebrank J, Wright D (1974) Spectral cues used in the localization of sound sources on the median plane. J Acoust Soc Am 56:1829-1834.

Henry MJ, Obleser J (2012) Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc Natl Acad Sci U S A 109:20095-20100.

Henry MJ, Obleser J (2013) Dissociable neural response signatures for slow amplitude and frequency modulation in human auditory cortex. PLoS One 8:e78758.

Herdman AT, Wollbrink A, Chau W, Ishii R, Ross B, Pantev C (2003) Determination of activation areas in the human auditory cortex by means of synthetic aperture magnetometry. Neuroimage 20:995-1005.

Herrmann B, Henry MJ, Grigutsch M, Obleser J (2013) Oscillatory phase dynamics in neural entrainment underpin illusory percepts of time. J Neurosci 33:15799-15809.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393-402.

Hines RM, Hines DJ, Houston CM, Mukherjee J, Haydon PG, Tretter V, Smart TG, Moss SJ (2013) Disrupting the clustering of GABAA receptor alpha2 subunits in the frontal cortex leads to reduced gamma-power and cognitive deficits. Proceedings of the National Academy of Sciences of the United States of America 110:16628-16633.

Hirano Y, Oribe N, Kanba S, Onitsuka T, Nestor PG, Spencer KM (2015) Spontaneous Gamma Activity in Schizophrenia. JAMA psychiatry 72:813-821.

Hong LE, Summerfelt A, Mitchell BD, McMahon RP, Wonodi I, Buchanan RW, Thaker GK (2008) Sensory gating endophenotype based on its neural oscillatory pattern and heritability estimate. Arch Gen Psychiatry 9:1008-1016.

Hong LE, Summerfelt A, Buchanan RW, O'Donnell P, Thaker GK, Weiler MA, Lahti AC (2010) Gamma and delta neural oscillations and association with clinical symptoms under subanesthetic ketamine. Neuropsychopharmacology 35:632-640.

Hong LE, Summerfelt A, McMahon R, Adami H, Francis G, Elliott A, Buchanan RW, Thaker GK (2004) Evoked gamma band synchronization and the liability for schizophrenia. Schizophr Res 70:293-302.

Iliadou V, Iakovides S (2003) Contribution of psychoacoustics and neuroaudiology in revealing correlation of mental disorders with central auditory processing disorders. Ann Gen Hosp Psychiatry 2:5.

Inui K, Okamoto H, Miki K, Gunji A, Kakigi R (2006) Serial and parallel processing in the human auditory cortex: a magnetoencephalographic study. Cereb Cortex 16:18-30.

Javitt DC, Steinschneider M, Schroeder CE, Arezzo JC (1996) Role of cortical N-methyl-D-aspartate receptors in auditory sensory memory and mismatch negativity generation: implications for schizophrenia. ProcNatlAcadSciUSA 93:11962-11967.

Jervis BW, Nichols MJ, Johnson TE, Allen E, Hudson NR (1983) A fundamental investigation of the composition of auditory evoked potentials. IEEE Trans Biomed Eng 30:43-50.

Joris PX, Carney LH, Smith PH, Yin TC (1994) Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency. J Neurophysiol 71:1022-1036.

Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci U S A 97:11793-11799.

Kantrowitz JT, Javitt DC (2010) N-methyl-d-aspartate (NMDA) receptor dysfunction or dysregulation: the final common pathway on the road to schizophrenia? Brain Res Bull 83:108-121.

Kayser C, Wilson C, Safaai H, Sakata S, Panzeri S (2015) Rhythmic auditory cortex activity at multiple timescales shapes stimulus-response gain and background firing. J Neurosci 35:7750-7762.

Kendler KS, McGuire M, Gruenberg AM, O'Hare A, Spellman M, Walsh D (1993) The Roscommon Family Study. I. Methods, diagnosis of probands, and risk of schizophrenia in relatives. Archives of General Psychiatry 50:527-540.

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". J Neurosci 30:620-628.

Kiang M, Kutas M, Light GA, Braff DL (2008) An event-related brain potential study of direct and indirect semantic priming in schizophrenia. Am J Psychiatry 165:74-81.

Kim T, Thankachan S, McKenna JT, McNally JM, Yang C, Choi JH, Chen L, Kocsis B, Deisseroth K, Strecker RE, Basheer R, Brown RE, McCarley RW (2015) Cortically projecting basal forebrain parvalbumin neurons regulate cortical gamma band oscillations. Proceedings of the National Academy of Sciences of the United States of America 112:3535-3540.

Kirihara K, Rissling AJ, Swerdlow NR, Braff DL, Light GA (2012) Hierarchical organization of gamma and theta oscillatory dynamics in schizophrenia. Biol Psychiatry 71:873-880.

Kong YY, Somarowthu A, Ding N (2015) Effects of Spectral Degradation on Attentional Modulation of Cortical Auditory Responses to Continuous Speech. J Assoc Res Otolaryngol 16:783-796.

Kösem A, Van Wassenhove V (2017) Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. Language, Cognition and Neuroscience 32:536-544.

Kuwada S, Bishop B, Kim DO (2014) Azimuth and envelope coding in the inferior colliculus of the unanesthetized rabbit: effect of reverberation and distance. J Neurophysiol 112:1340-1355.

Kwon JS, O'Donnell BF, Wallenstein GV, Greene RW, Hirayasu Y, Nestor PG, Hasselmo ME, Potts GF, Shenton ME, McCarley RW (1999) Gamma frequency-range abnormalities to auditory stimulation in schizophrenia. Arch Gen Psychiatry 56:1001-1005.

Lakatos P, Schroeder CE, Leitman DI, Javitt DC (2013a) Predictive suppression of cortical excitability and its deficit in schizophrenia. J Neurosci 33:11692-11702.

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J Neurophysiol 94:1904-1911.

Lakatos P, Musacchia G, O'Connel MN, Falchier AY, Javitt DC, Schroeder CE (2013b) The spectrotemporal filter mechanism of auditory selective attention. Neuron 77:750-761.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol 102:349-359.

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2278-2324.

Lehmann E, Johansson AM (2010) Diffuse reverberation model for efficient image-source simulation of room impulse responses. Audio, Speech, and Language Processing, IEEE Transactions on 18:1429-1439.

Light GA, Braff DL (2005) Mismatch negativity deficits are associated with poor functioning in schizophrenia patients. Arch GenPsychiatry 62:127-136.

Light GA, Naatanen R (2013) Mismatch negativity is a breakthrough biomarker for understanding and treating psychotic disorders. Proc Natl Acad Sci U S A 110:15175-15176.

Light GA, Hsu JL, Hsieh MH, Meyer-Gomes K, Sprock J, Swerdlow NR, Braff DL (2006) Gamma band oscillations reveal neural network cortical coherence dysfunction in schizophrenia patients. Biological Psychiatry 60:1231-1240.

Lippmann RP (1997) Speech recognition by machines and humans. Speech communication 22:1-15.

Litovsky RY, Colburn HS, Yost WA, Guzman SJ (1999) The precedence effect. J Acoust Soc Am 106:1633-1654.

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

Luo H, Husain FT, Horwitz B, Poeppel D (2005) Discrimination and categorization of speech and non-speech sounds in an MEG delayed-match-to-sample study. Neuroimage 28:59-71.

Lutkenhoner B, Steinstrater O (1998) High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiol Neurootol 3:191-213.

Lutkenhoner B, Mosher JC (2007) Source Analysis of Auditory Evoked Potentials and Fields. In: Auditory evoked potentials : basic principles and clinical application (Burkard RF, Eggermont JJ, Don M, eds), pp xix, 731 p., 716 p. of plates. Philadelphia: Lippincott Williams & Wilkins.

Maess B, Koelsch S, Gunter TC, Friederici AD (2001) Musical syntax is processed in Broca's area: an MEG study. Nat Neurosci 4:540-545.

Maldonado-Aviles JG, Curley AA, Hashimoto T, Morrow AL, Ramsey AJ, O'Donnell P, Volk DW, Lewis DA (2009) Altered markers of tonic inhibition in the dorsolateral prefrontal cortex of subjects with schizophrenia. Am J Psychiatry 166:450-459.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177-190.

Marrone N, Mason CR, Kidd G, Jr. (2008a) Evaluating the benefit of hearing aids in solving the cocktail party problem. Trends Amplif 12:300-315.

Marrone N, Mason CR, Kidd G, Jr. (2008b) The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. J Acoust Soc Am 124:3064-3075.

McDermott JH (2009) The cocktail party problem. Curr Biol 19:R1024-1027.

McLachlan NM, Phillips DS, Rossell SL, Wilson SJ (2013) Auditory processing and hallucinations in schizophrenia. Schizophr Res 150:380-385.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233-236.

Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol 102:3329-3339.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014a) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006-1010.

Mesgarani N, David SV, Fritz JB, Shamma SA (2014b) Mechanisms of noise robust representation of speech in primary auditory cortex. Proc Natl Acad Sci U S A 111:6792-6797.

Michelini S, Arslan E, Prosser S, Pedrielli F (1982) Logarithmic display of auditory evoked potentials. J Biomed Eng 4:62-64.

Miller KJ, Sorensen LB, Ojemann JG, den Nijs M (2009) Power-law scaling in the brain surface electric potential. PLoS Comput Biol 5:e1000609.

Moon IJ, Hong SH (2014) What is temporal fine structure and why is it important? Korean J Audiol 18:1-7.

Moore BC (2008) The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. J Assoc Res Otolaryngol 9:399-406.

Moore RC, Lee T, Theunissen FE (2013) Noise-invariant neurons in the avian auditory cortex: hearing the song in noise. PLoS Comput Biol 9:e1002942.

Naatanen 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.

Naatanen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118:2544-2590.

Nabelek AK, Robinson PK (1982) Monaural and binaural speech perception in reverberation for listeners of various ages. J Acoust Soc Am 71:1242-1248.

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400-410.

Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. Front Neurosci 2:107-113.

Neske GT (2015) The Slow Oscillation in Cortical and Thalamic Networks: Mechanisms and Functions. Front Neural Circuits 9:88.

Ng BS, Schroeder T, Kayser C (2012) A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. J Neurosci 32:12268-12276.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1-25.

Niedermeyer E, Lopes da Silva FH (2005) Electroencephalography : basic principles, clinical applications, and related fields, 5th Edition. Philadelphia: Lippincott Williams & Wilkins.

Nourski KV, Steinschneider M, McMurray B, Kovach CK, Oya H, Kawasaki H, Howard MA, 3rd (2014) Functional organization of human auditory cortex: investigation of response latencies through direct recordings. Neuroimage 101:598-609.

Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Howard MA, 3rd, Brugge JF (2009) Temporal envelope of time-compressed speech represented in the human auditory cortex. J Neurosci 29:15564-15574.

O'Donnell BF, Vohs JL, Krishnan GP, Rass O, Hetrick WP, Morzorati SL (2013) The auditory steady-state response (ASSR): a translational biomarker for schizophrenia. Suppl Clin Neurophysiol 62:101-112.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb Cortex 25:1697-1706.

Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20:2486-2495.

Okamoto H, Stracke H, Bermudez P, Pantev C (2011) Sound processing hierarchy within human auditory cortex. Journal of Cognitive Neuroscience 23:1855-1863.

Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903-911.

Parkkonen L, Fujiki N, Makela JP (2009) Sources of auditory brainstem responses revisited: contribution by magnetoencephalography. Hum Brain Mapp 30:1772-1782.

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10:e1001251.

Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci 4:51.

Pickles JO (2012) An introduction to the physiology of hearing: BRILL.

Picton TW, John MS, Dimitrijevic A, Purcell D (2003) Human auditory steady-state responses. Int J Audiol 42:177-219.

Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. Speech Communication 41:245-255.

Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. Philos Trans R Soc Lond B Biol Sci 363:1071-1086.

Poeppel D, Yellin E, Phillips C, Roberts TP, Rowley HA, Wexler K, Marantz A (1996) Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. Brain Res Cogn Brain Res 4:231-242.

Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. Eur J Neurosci 35:1497-1503.

Puvvada KC, Summerfelt A, Du X, Krishna N, Kochunov P, Rowland LM, Simon JZ, Hong LE (2017) Delta Vs Gamma Auditory Steady State Synchrony in Schizophrenia. Schizophrenia Bulletin:sbx078.

Rabinowitz NC, Willmore BD, King AJ, Schnupp JW (2013) Constructing noise-invariant representations of sound in the auditory pathway. PLoS Biol 11:e1001710.

Rass O, Forsyth JK, Krishnan GP, Hetrick WP, Klaunig MJ, Breier A, O'Donnell BF, Brenner CA (2012) Auditory steady state response in the schizophrenia, first-degree relatives, and schizotypal personality disorder. Schizophr Res 136:143-149.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718-724.

Recanzone GH, Guard DC, Phan ML (2000) Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. J Neurophysiol 83:2315-2331.

Rimmele JM, Zion Golumbic E, Schroger E, Poeppel D (2015) The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. Cortex 68:144-154.

Robinson BL, McAlpine D (2009) Gain control mechanisms in the auditory pathway. Curr Opin Neurobiol 19:402-407.

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 336:367-373.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 16:225-237.

Sarampalis A, Kalluri S, Edwards B, Hafter E (2009) Objective measures of listening effort: effects of background noise and noise reduction. J Speech Lang Hear Res 52:1230-1240.

Särelä J, Valpola H (2005a) Denoising Source Separation. The Journal of Machine Learning Research 6:233-272.

Särelä J, Valpola H (2005b) Denoising source separation. Journal of Machine Learning Research 6:233-272.

Sartorius N, Shapiro R, Jablensky A (1974) The international pilot study of schizophrenia. Schizophrenia Bulletin Winter:21-34.

Sato H, Sato H, Morimoto M, Ota R (2007) Acceptable range of speech level for both young and aged listeners in reverberant and quiet sound fields. J Acoust Soc Am 122:1616.

Sayles M, Winter IM (2008) Reverberation challenges the temporal representation of the pitch of complex sounds. Neuron 58:789-801.

Sayles M, Stasiak A, Winter IM (2014) Reverberation impairs brainstem temporal representations of voiced vowel sounds: challenging "periodicity-tagged" segregation of competing speech in rooms. Front Syst Neurosci 8:248.

Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci 32:9-18.

Shamma S (2006) Analysis of speech dynamics in the auditory system. Dynamics of Speech Production and Perception: Life and Behavioural Sciences:335-342.

Shamma S, Lorenzi C (2013) On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. J Acoust Soc Am 133:2818-2833.

Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34:114-123.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303-304.

Sharpee TO, Atencio CA, Schreiner CE (2011) Hierarchical representations in the auditory cortex. Curr Opin Neurobiol 21:761-767.

Sheft S (2008) Envelope processing and sound-source perception. In: Auditory perception of sound sources, pp 233-280: Springer.

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182-186.

Simon JZ, Ding N (2010) Magnetoencephalography and auditory neural representations. In: 26th Southern Biomedical Engineering Conference: Sbec 2010, pp 45-48.

Slama MC, Delgutte B (2015) Neural coding of sound envelope in reverberant environments. J Neurosci 35:4452-4468.

Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83-98.

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87-90.

Snyder JS, Holder WT, Weintraub DM, Carter OL, Alain C (2009) Effects of prior stimulus and prior perception on neural correlates of auditory stream segregation. Psychophysiology 46:1208-1215.

Snyder MA, Gao WJ (2013) NMDA hypofunction as a convergence point for progression and symptoms of schizophrenia. Front Cell Neurosci 7:31.

Spencer KM, Salisbury DF, Shenton ME, McCarley RW (2008) gamma-Band Auditory Steady-State Responses Are Impaired in First Episode Psychosis. Biological Psychiatry.

Spencer KM, Nestor PG, Perlmutter R, Niznikiewicz MA, Klump MC, Frumin M, Shenton ME, McCarley RW (2004) Neural synchrony indexes disordered perception and cognition in schizophrenia. ProcNatlAcadSciUSA 101:17288-17293.

Sponheim SR, Clementz BA, Iacono WG, Beiser M (1994) Resting EEG in first-episode and chronic schizophrenia. Psychophysiology 31:37-43.

Srinivasan R, Nunez PL, Silberstein RB (1998) Spatial filtering and neocortical dynamics: estimates of EEG coherence. IEEE Trans Biomed Eng 45:814-826.

Steinschneider M, Liégeois-Chauvel C, Brugge JF (2011) Auditory evoked potentials and their utility in the assessment of complex sound processing. In: The auditory cortex, pp 535-559: Springer.

Steinschneider M, Nourski KV, Rhone AE, Kawasaki H, Oya H, Howard MA, 3rd (2014) Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Front Neurosci 8:240.

Stevens JP (2002) Applied multivariate statistics for the social sciences (4th ed.). . Mahwah, NJ.

Straub RE, Lipska BK, Egan MF, Goldberg TE, Callicott JH, Mayhew MB, Vakkalanka RK, Kolachana BS, Kleinman JE, Weinberger DR (2007) Allelic variation in GAD 1 (GAD67) is associated with schizophrenia and influences cortical function and gene expression. MolPsychiatry Online publication

Sullivan EM, Timi P, Hong LE, O'Donnell P (2015) Reverse translation of clinical electrophysiological biomarkers in behaving rodents under acute and chronic NMDA receptor antagonism. Neuropsychopharmacology 40:719-727.

Sussman ES, Bregman AS, Wang WJ, Khan FJ (2005) Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. Cogn Affect Behav Neurosci 5:93-110.

Swaminathan J, Mason CR, Streeter TM, Best V, Roverud E, Kidd G, Jr. (2016) Role of Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening. J Neurosci 36:8250-8257.

Sweet RA, Dorph-Petersen KA, Lewis DA (2005) Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. J Comp Neurol 491:270-289.

Tavabi K, Obleser J, Dobel C, Pantev C (2007) Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. Eur J Neurosci 25:3155-3162.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 12:289-316.

Traer J, McDermott JH (2016) Statistics of natural reverberation enable perceptual separation of sound and space. Proc Natl Acad Sci U S A 113:E7856-E7865.

Tsuchimoto R, Kanba S, Hirano S, Oribe N, Ueno T, Hirano Y, Nakamura I, Oda Y, Miura T, Onitsuka T (2011) Reduced high and low frequency gamma synchronization in patients with chronic schizophrenia. Schizophr Res 133:99-105.

Uhlhaas PJ, Singer W (2010) Abnormal neural oscillations and synchrony in schizophrenia. Nat Rev Neurosci 11:100-113.

Uhlhaas PJ, Pipa G, Lima B, Melloni L, Neuenschwander S, Nikolic D, Singer W (2009) Neural synchrony in cortical networks: history, concept and current status. Front Integr Neurosci 3:17.

Vierling-Claassen D, Siekmeier P, Stufflebeam S, Kopell N (2008) Modeling GABA alterations in schizophrenia: a link between impaired inhibition and altered gamma and beta range auditory entrainment. J Neurophysiol 99:2656-2671.

Voytek B, Kramer MA, Case J, Lepage KQ, Tempesta ZR, Knight RT, Gazzaley A (2015) Age-Related Changes in 1/f Neural Electrophysiological Noise. J Neurosci 35:13257-13265.

Wang Y, Ding N, Ahmar N, Xiang J, Poeppel D, Simon JZ (2012) Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. J Neurophysiol 107:2033-2041.

Ward LM (2003) Synchronous neural oscillations and cognitive processes. Trends Cogn Sci 7:553-559.

Xia J, Shinn-Cunningham B (2011) Isolating mechanisms that influence measures of the precedence effect: theoretical predictions and behavioral tests. J Acoust Soc Am 130:866-882.

Xiang J, Simon J, Elhilali M (2010) Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. J Neurosci 30:12084-12093.

Yang W, Bradley JS (2009) Effects of room acoustics on the intelligibility of speech in classrooms for young children. J Acoust Soc Am 125:922-933.

Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. Information Theory, IEEE Transactions on 38:824-839.

Yu D, Deng L (2014) Automatic Speech Recognition: A Deep Learning Approach: Springer Publishing Company, Incorporated.

Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends in cognitive sciences 10:301-308.

Yvert B, Crouzeix A, Bertrand O, Seither-Preisler A, Pantev C (2001) Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in humans. Cereb Cortex 11:411-423.

Zilany MS, Bruce IC, Carney LH (2014) Updated parameters and expanded simulation options for a model of the auditory periphery. The Journal of the Acoustical Society of America 135:283-286.

Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D (2013a) Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J Neurosci 33:1417-1426.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013b) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77:980-991.