ABSTRACT

Title of Document: TEMPORAL CODING OF SPEECH IN HUMAN AUDITORY CORTEX

Nai Ding, Doctor of Philosophy, 2012

Directed By: Professor, Jonathan Z. Simon, Department of Electrical and Computer Engineering

Human listeners can reliably recognize speech in complex listening environments. The underlying neural mechanisms, however, remain unclear and cannot yet be emulated by any artificial system. In this dissertation, we study how speech is represented in the human auditory cortex and how the neural representation contributes to reliable speech recognition. Cortical activity from normal hearing human subjects is noninvasively recorded using magnetoencephalography, during natural speech listening. It is first demonstrated that neural activity from auditory cortex is precisely synchronized to the slow temporal modulations of speech, when the speech signal is presented in a quiet listening environment. How this neural representation is affected by acoustic interference is then investigated. Acoustic interference degrades speech perception via two mechanisms, informational masking and energetic masking, which are addressed respectively by using a competing speech stream and a stationary noise as the interfering sound. When two speech streams are

presented simultaneously, cortical activity is predominantly synchronized to the speech stream the listener attends to, even if the unattended, competing speech stream is 8 dB more intense. When speech is presented together with spectrally matched stationary noise, cortical activity remains precisely synchronized to the temporal modulations of speech until the noise is 9 dB more intense. Critically, the accuracy of neural synchronization to speech predicts how well individual listeners can understand speech in noise.

Further analysis reveals that two neural sources contribute to speech synchronized cortical activity, one with a shorter response latency of about 50 ms and the other with a longer response latency of about 100 ms. The longer-latency component, but not the shorter-latency component, shows selectivity to the attended speech and invariance to background noise, indicating a transition from encoding the acoustic scene to encoding the behaviorally important auditory object, in auditory cortex. Taken together, we have demonstrated that during natural speech comprehension, neural activity in the human auditory cortex is precisely synchronized to the slow temporal modulations of speech. This neural synchronization is robust to acoustic interference, whether speech or noise, and therefore provides a strong candidate for the neural basis of acoustic background invariant speech recognition.

TEMPORAL CODING OF SPEECH IN HUMAN AUDITORY CORTEX

By

Nai Ding

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
2012

Advisory Committee:
Professor Jonathan Z. Simon, Chair
Professor Shihab A. Shamma
Professor Min Wu
Research Scientist Didier A. Depireux
Professor Rochelle S. Newman

Dedication

To my parents, Ding Xuefeng and Shen Min.

Acknowledgements

First, I would like to thank my advisor Dr. Jonathan Simon for introducing me to auditory neuroscience and providing guidance in the past 5 years. I'm also grateful to Drs. David Poeppel, Monita Chatterjee, and Shihab Shamma for their support and guidance. I would like to thank Drs. Min Wu, Didier Depireux, and Rochelle Newman for serving on the thesis committee, and Dr. Carol Espy-Wilson for serving on the research proposal committee.

I would like to thank Dr. Stephen David for insights into data analysis, Dr. Alain de Cheveigné for insights into MEG processing, and Dr. Mary Howard for support and important feedback on my work.

I also appreciate many inspiring discussions with Xing Tian, Juanjuan Xiang, Pinbo Yin, and Greg Cogan. I want to thank Elana Zion-Golumbic, Yue Zhang, Verónica Figueroa, Ed Smith, and Kai Sun Li for collaborations, Jiachen Zhuo, Viral Tejani, Francisco Constantino, and Marisel Villafañe for comments on the dissertation, Jeff Walker and Max Ehrmann for help with experiments. I also acknowledge all current and past members of the Simon lab, Shamma lab, Poeppel lab, Chatterjee lab and the UMD MEG lab for support and interesting discussions!

Table of Contents

Chapter 1

Introduction

Speech is a dominant form of human communication, and speech communication is remarkably robust to acoustic interference. Such robustness, however, is lost for hearing-impaired listeners (Festen and Plomp, 1990) and cannot yet be emulated by automatic speech recognition systems (Cooke et al., 2010; Lippmann, 1997). Therefore, identifying how speech is represented in the normal-hearing human auditory system and how this neural representation leads to noise-robust speech perception is not only of great interest to neuroscience but also has potential applications in the design of hearing aid devices and noise-robust automatic speech recognition systems.

The recognition of speech relies on the spectro-temporal modulations of speech, i.e. how the energy of speech varies over time and frequency (Chi et al., 1999). In this dissertation, we focus on the neural representation of the slow temporal modulations of speech (< 10 Hz), which reflect the syllabic and phrasal structure of speech (Greenberg et al., 2003; Poeppel et al., 2008). In quiet listening environments, these slow modulations lead to high speech intelligibility, even if accompanied by only very coarse spectral information (Elliott and Theunissen, 2009; Shannon et al., 1995). In complex listening environments, they provide cues for grouping relevant acoustic features into a coherent speech stream (Shamma et al., 2011; Sheft, 2007). Here, the neural representation of slow temporal modulations is investigated using magnetoencephalography (MEG) (Hämäläinen et al., 1993), a noninvasive neural recording tool. MEG is sensitive to neural activity in human auditory cortex (Lütkenhöner and Mosher, 2006) and has

millisecond level time resolution, high enough to resolve neural activity phase locked to these slow temporal modulations (Ding and Simon, 2009; Wang et al., 2012). A review of human auditory processing and MEG is provided Chapter 2.

This dissertation consists of three studies. The first study (Chapter 3) addresses how the temporal modulations of speech are encoded in the human auditory cortex. Instead of excessively repeating a few syllables or short sentences, as done in traditional electrophysiological studies, discourse-level spoken narratives are employed to examine the neural encoding of natural, continuous speech and adopt a systems-theoretic approach to characterize the neural code. It is demonstrated that the slow temporal modulations of speech are encoded in the human auditory cortex by precisely phase-locked neural activity. Furthermore, when two speech signals from the same speaker are simultaneously presented to different ears (dichotic listening), the response to the speech being attended to is substantially stronger than the response to the unattended speech, demonstrating top-down attentional modulation of the neural representation of slow temporal modulations. *This study has been published in the Journal of Neurophysiology (Ding & Simon, 2012).*

The second study (Chapter 4) addresses the neural processing underlying how listeners selectively attend to one of two concurrent speech streams that are mixed into a single acoustic channel, which removes the binaural cues present in the previous study. This study demonstrates that longer-latency (~100 ms) cortical activity is selectively synchronized to the temporal modulations of the attended speech stream, even though the two competing speech streams have strong acoustic overlap. Critically, this neural representation is insensitive to the intensity ratio between the two competing speech streams, at least in the range where the attended speech remains intelligible (intelligibility

> 50%). These results suggest that concurrent speech streams are neurally segregated and encoded differentially in the human auditory cortex, based on their perceptual important rather than physical intensity. *This study has been accepted for publication by Proceedings of the National Academy of Sciences.*

The third study (Chapter 5) addresses the neural encoding of speech embedded in stationary background noise. When processing concurrent streams of speech, the brain benefits from taking clean "glimpses" of the target speech stream when the interfering stream is instantaneously weak (Cooke, 2006). Stationary noise, however, eliminates such clean glimpses and therefore is more detrimental to speech intelligibility (Festen and Plomp, 1990). Neural synchronization to the slow temporal modulations of speech, however, is found to be robust to the background noise until it is 9 dB stronger than speech. Long-term temporal integration (> 100 ms) and neural adaptation to sound intensity are demonstrated to be crucial for the stable neural representation. Critically, the precision of the neural encoding of slow temporal modulations predicts how well a listener can understand speech in noise.

Taken together, this series of studies demonstrate that, during natural speech comprehension, the temporal modulations of speech are encoded precisely by phase-locked activity in the human auditory cortex (Chapter 3), even in the presence of acoustic interference, whether speech or noise (Chapter 3-5). The acoustic degradations caused by speech and noise represent respectively informational masking and energetic masking, two fundamental aspects of the interactions between speech and background (Brungart, 2001; Durlach et al., 2003). Therefore, it is reasonable to infer that in any auditory scene that allows a listener to successfully attend to a speech stream, neural activity in the

listener's auditory cortex is precisely synchronized to the attended speech stream. This noise-robust neural representation of the slow temporal modulations of speech provides a plausible neural basis for noise-robust speech recognition.

Chapter 2

Background

2.1 Overview: auditory neural computations

The auditory system processes sounds through neural computations. Some of these computations are known. For example, the auditory system breaks up sounds into narrow frequency bands and applies a nonlinear compression to the amplitude of the sounds in each frequency band (Hudspeth, 2008). These two kinds of neural computations are the fundamentals of our basis of sound perception (Moore, 2003) and have parallels in signal processing, i.e. the wavelet transform (Mallat, 1999) and the static logarithmic nonlinearity. Furthermore, the auditory system is rapidly adapted to the mean and intensity of the stimulus, which provides a plausible neural basis for intensity independent auditory perception (Robinson and McAlpine, 2009; Zilany et al., 2009). Most of these well-characterized neural computations occur before the neural representations of sounds reach the part of the brain known as the cortex (Fig. 2.1). Little is known, however, about neural computations occurring inside the cortex, which are critical to sound segregation and speech coding (Hickok and Poeppel, 2007; Nelken, 2008).

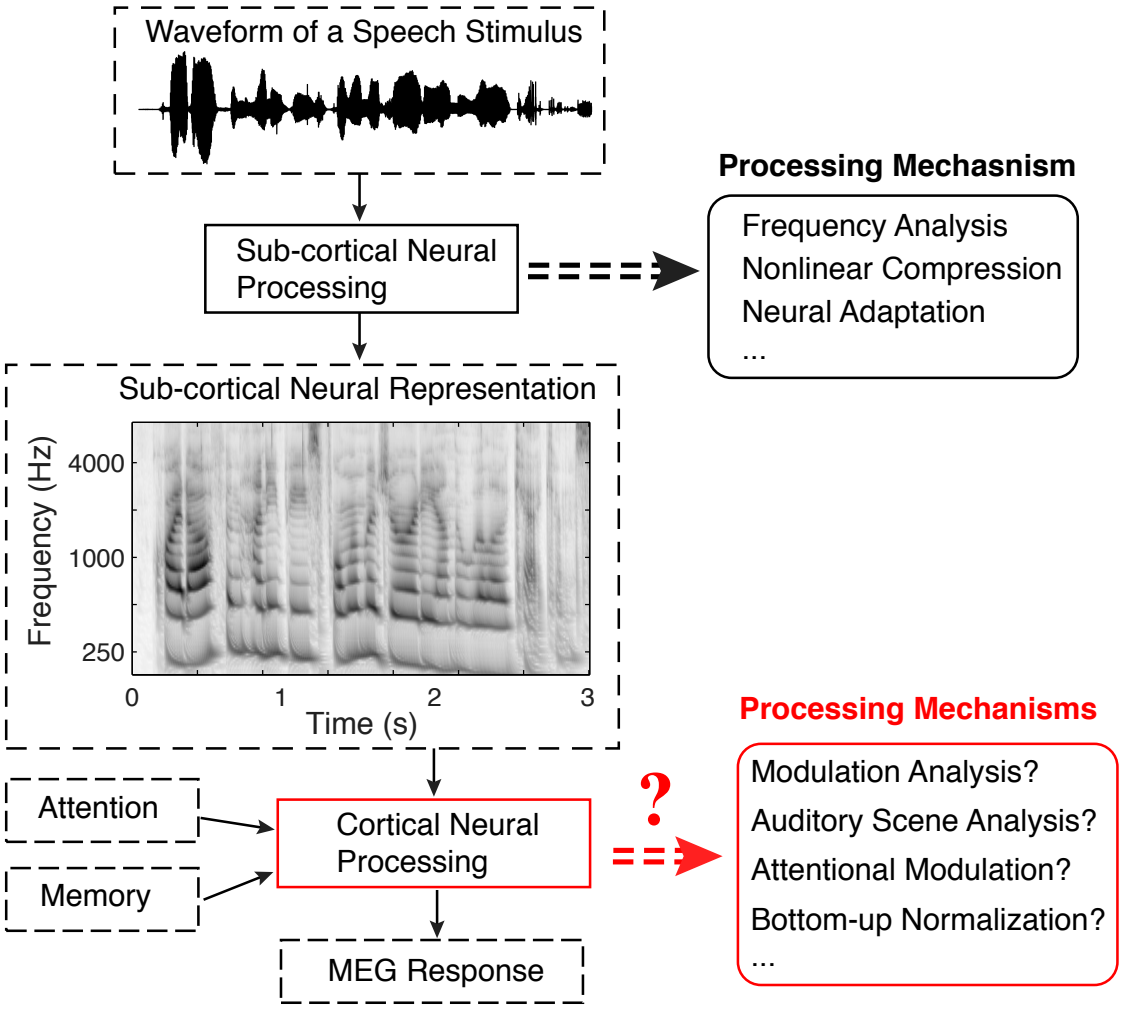

Figure 2.1. Auditory neural computations. The sub-cortical neural processing provides a spectro-temporal representation of the acoustic input (Yang et al., 1992). The cortical neural processing is based on the sub-cortical representation of sounds but is also modulated by cognitive functions such as attention and memory.

One neural computation known to occur in the animal primary auditory cortex is the selectivity to spectro-temporal modulations (Chi et al., 2005; Depireux et al., 2001).

From the perspective of signal processing, the spectral modulation analysis is an analysis of the power spectrum, similar to cepstral analysis and linear predictive analysis (Gold and Morgan, 2000). The temporal modulation analysis selectively processes sound features, e.g. the power spectrum, varying at different rates and is essentially similar to RASTA in speech processing (Gold and Morgan, 2000). The temporal modulation analysis can be viewed as filtering sound features in time, i.e., convolving the sound features with a temporal window. Therefore, temporal modulation analysis is also frequently discussed as a temporal integration process with certain time windows (Poeppel, 2003). Furthermore, cortical processing has been hypothesized to decompose a complex auditory scene into auditory objects, each being the sound generated from a single physical source (Griffiths and Warren, 2004; Nelken and Bar-Yosef, 2008; Shamma et al., 2011), although the neural evidence for this is still lacking.

2.2 Human auditory system

2.2.1 Neurons and neural systems

The fundamental unit of the nervous system is the neuron. A neuron is separated from its outside environment by a cell membrane. A typical neuron contains three parts: the cell body, the axon, and the dendrites (Fig. 2.2A). The cell body is roughly tens of microns in diameter (Dayan and Abbott, 2001, Chapter 1). The dendrites and axons are *processes* (i.e. extensions) a neuron uses to connect with other neurons.

The signal output of a neuron is a series of *action potentials* (or *spikes*), which are brief (1-2 ms in duration) voltage changes that propagate along the axon. An action

potential is generated when the voltage difference across the membrane of a neuron reaches a threshold. The time when a neuron generates an action potential is called the *firing time*. The number of action potentials generated per second is called the *firing rate*. The firing rate can be as high as a few hundred Hertz.

The signal inputs to a neuron are received from the dendrites. A dendrite connects to axons (usually from other neurons) via structures called synapses. Action potentials received by a dendrite cause a voltage change in the dendrite. This voltage change, called a post-synaptic potential, can last tens of milliseconds long. If a dendrite receives several action potentials within a short time period, the post-synaptic potential accumulates. When the accumulated voltage change reaches a threshold, an action potential will be generated by the neuron. The post-synaptic potential also leads to a current in the dendrite flowing towards the cell body (Fig. 2.2B). The *dendritic current* is typically on the order of fA (10^{-15} Ampere) (Hämäläinen et al., 1993).

The activity of a single neuron can be recorded either extracellularly or intracellularly. An extracellular recording measures the electrical activity of a neuron through an electrode placed outside but close to the neuron. It primarily records the action potentials. An intracellular recording measures the voltage across the cell membrane and therefore reflects both action potentials and dendritic activities. A recording from a single neuron is usually called *single unit recording*.

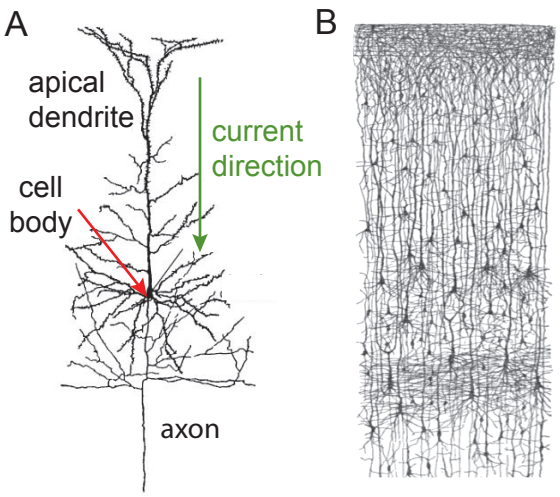

Figure 2.2. (A) A cortical pyramidal neuron. It contains a cell body, an axon, and many dendrites. The green arrow indicates the direction of dendritic current. (B) The drawing of a slice of stained infant cortex by S. Ramon y Cajal. Neurons are densely distributed in the cortex. (Fig. 2.2A is adapted from Fig. 1.1 of Dayan and Abbott (2001) and the Fig. 2.2B is adapted from http://en.wikipedia.org/wiki/File:Cajal_cortex_drawings.png)

Neurons interconnect with one another and form neural networks (Fig. 2.2B). In each mm^2 of the cerebral cortex, there are about 10^5 neurons (Hämäläinen et al., 1993). A common type of neuron in the cortex is the pyramidal neuron. Some dendrites of pyramidal neurons, called the apical dendrites, are roughly perpendicular to the surface of the cortex. The apical dendrites of neighboring neurons are approximately parallel, and therefore the current in the apical dendrites in a local neural network flow in very similar directions. When the currents along the dendrites of many neurons are synchronized in time, they may sum up to be a current source that generates a magnetic field strong enough to measure outside the brain (extracranially). Measurement of this magnetic field

is called magnetoencephalography (MEG). Similarly, synchronized neural activity can also generate an electric potential measurable extracranially, which is called electroencephalography (EEG). MEG/EEG can be measured noninvasively. Nonetheless, since MEG and EEG measure neural activity synchronized over millions of neurons, they have limited spatial resolution (millimeter to centimeter level). This spatial resolution issue is severely aggravated by the fact that reconstructing the spatial distribution of current sources based on its magnetic/electric field is an ill-posed problem (Baillet et al., 2001).

The activity of neural networks can also be indirectly measured by positron emission tomography (PET) and functional Magnetic Resonance Imaging (fMRI) (Logothetis, 2003; Raichle, 1983). PET and fMRI image the dynamics of blood flow inside the brain. Since neural activity has a high metabolic cost, it changes the flow and oxygen level of the blood in local brain areas. The dynamics of blood flow, however, are much slower than the dynamics of neural activity, and PET and fMRI have a time resolution lower than 1 Hz. Therefore, PET and fMRI cannot resolve the neural response phase locked to the slow temporal modulations of speech (1-16 Hz).

2.2.2 Human auditory system

Sounds are transformed from mechanical vibrations to electrical neural activity in the *cochlea*. The neural representations of sounds are then processed by a series of neural networks from the brainstem to the cortex.

In the cochlea, the basilar membrane acts as a filter bank (Fig. 2.3). It is about 35 mm long, with its basal part tuned to high frequencies and its apical part tuned to low frequencies. The frequency tuning of the basilar membrane changes about 1/3-1/4 octave per millimeter (Greenwood, 1990). Besides frequency filtering, another important function of the cochlea is to compress the dynamic range of sound input in a nonlinear fashion (Hudspeth, 2008; Moore, 2003). The inner hair cells on the basilar membrane transform vibrational signals into electrical neuronal activity. The auditory nerves then transmit neural activity of inner hair cells to the central nervous system.

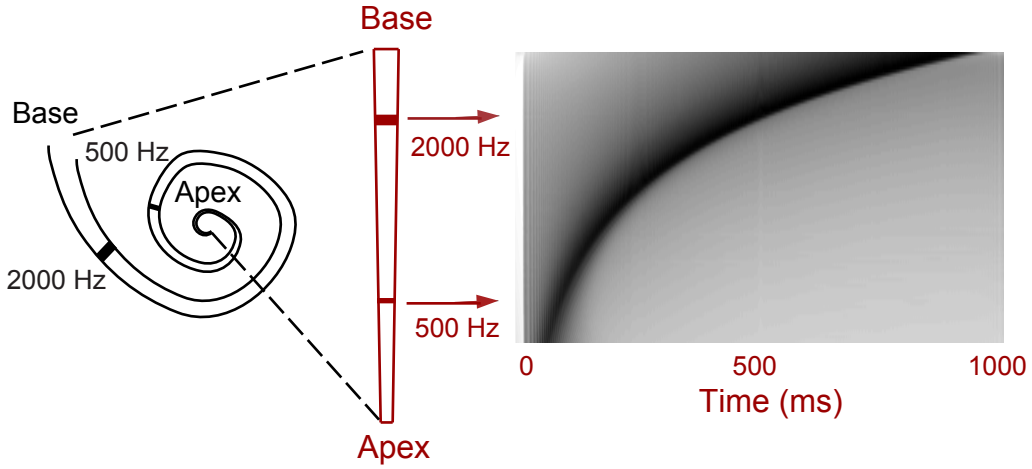

Figure 2.3. A schematic illustration of the function of the basilar membrane in the cochlea. The basilar membrane is a spiral structure in the cochlea (Left). It acts as a filter bank in logarithmic frequency spacing. The response of the basilar membrane to a chirp signal whose frequency linearly increases is simulated and plotted as a function of time (Right), where dark colors mean high activation. The basilar membrane response is simulated based on the model proposed by Yang, Wang, and Shamma (1992). The nonlinearity of the basilar membrane is not simulated.

The neural outputs of the auditory nerves are processed by several nuclei (compact networks of neurons) in the brainstem and thalamus before reaching the cortex. These sub-cortical nuclei refine the temporal synchronization of neural responses (Joris et al., 1994), integrate the inputs from two ears (Chapter 6, Pickles 1988), and may also refine the frequency tuning of neurons through lateral inhibition (Yang et al., 1992). As in the cochlea, a large number of neurons in these sub-cortical nuclei are tuned in frequency and the frequency tuning of these neurons is spatially ordered (Chapter 6, Pickles 1988). The temporal precision of neurons decreases gradually from the cochlea to cortex (Giraud et al., 2000; Joris et al., 1994). Neural phase locking to sound is seen above 1 kHz in auditory nerves, up to ~200 Hz in the thalamus, and generally below 40 Hz in the cortex.

The human auditory cortex is located in the superior part of the temporal lobe (Fig. 2.4). It can be divided into core, belt, and parabelt (association) regions (Hackett et al., 1998; Kaas and Hackett, 2000). The core auditory cortex, including the primary auditory cortex, is located in medial part of the transverse temporal gyrus (Heschl's gyrus). Animal studies show that neurons in primary auditory cortex are generally tuned in frequency (Chapter 7, Pickles 1988). Similarly, human studies demonstrate very fine frequency tuning in some neurons in the Heschl's gyrus (Bitterman et al., 2008). Single unit recording from monkeys and fMRI data from humans suggest a functional dissociation between core and belt auditory cortices (Rauschecker, 1998; Wessinger et al., 2001). The core auditory cortex receives direct input from the thalamus and is most sensitive to pure tones. The belt region receives input from the core auditory cortex and is

more sensitive to narrow band stimuli than pure tones. Neural activity also shows better phase locking to temporal modulations in the core auditory cortex than in some belt regions (Nourski et al., 2009). Since intracranial recording from human subjects are rare while extracranial recordings have limited spatial resolution, the functional division of different auditory regions of human auditory cortex is still far from clear.

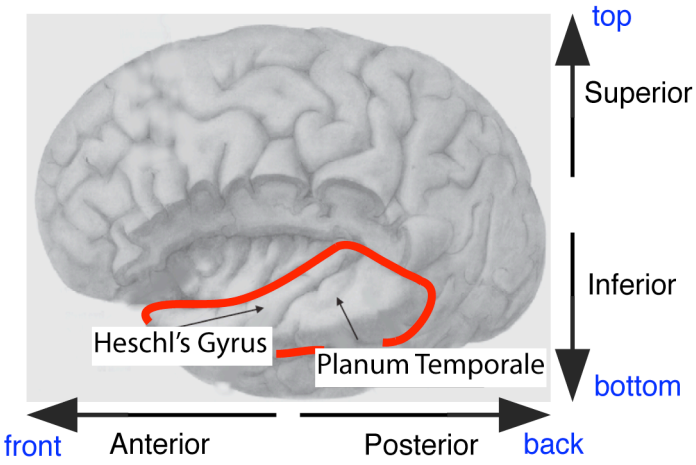

Figure 2.4. Anatomy of human auditory cortex. The human auditory cortex is located in the superior temporal gyrus (circled in red). The primary auditory cortex is in the Heschl's gyrus. The planum temporale is an important area in the association auditory cortex. In this figure, a part of the cortex is removed in order to visualize the auditory cortex (adapted from Tervaniemi and Hugdahl, 2003).

In the association auditory cortex, an important area is the planum temporale (PT), which is posterior to the Heschl's gyrus. The PT is supposed to play an important role in high level auditory processing, e.g. auditory scene analysis (Griffiths and Warren, 2002). The PT (Lütkenhöner and Steinsträter, 1998) and/or the lateral part of Heschl's

gyrus (Herdman et al., 2003) are localized as the sources of the M100, a strong MEG response occurring 100-ms after a sound onset.

2.3 Neural processing in human auditory cortex

2.3.1 Cortical processing of speech

The temporal information of speech, as well as the neural processing of speech, occurs on multiple time scales (Poeppel, 2003; Rosen, 1992; Shamma, 2006). The fastest time scale (> 300 Hz) contains the spectral information of speech, and its variability may be called the spectral modulation of speech (Chi et al., 2005). It is the carrier signal for slower temporal modulations. In auditory cortex, the spectral modulations are represented by the spatial activation pattern of neurons (Chang et al., 2010; Mesgarani et al., 2008). Slower time scales, roughly between 70-300 Hz, are often called temporal periodicity cues (Elhilali et al., 2004; Rosen, 1992) (Fig. 2.5) and are related to pitch perception (Elliott and Theunissen, 2009; Zeng et al., 2005). The neural representation of temporal periodicity is still under debate (de Cheveigné, 2005; Lyon and Shamma, 1996). The slowest time scale of speech, below 16 Hz, is often called the slow temporal modulations of speech (Fig. 2.5) (Chi et al., 2005; Rosen, 1992) and reflects the syllabic and phrasal structure of speech (Greenberg et al., 2003). The slow modulations that are consistent over spectral regions constitute the *temporal envelope* of speech. From the auditory periphery to the auditory cortex, the slow temporal modulations are represented by phase-locked neural activity (Brugge et al., 2009; Liegeois-Chauvel et al., 2004; Nourski et al., 2009).

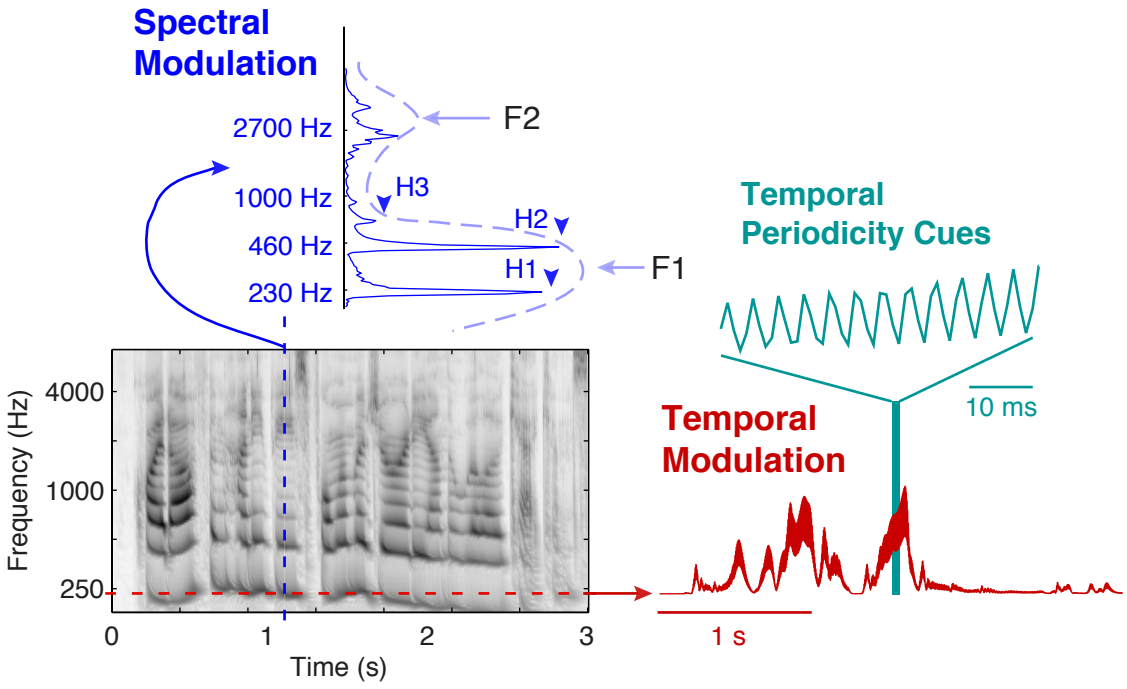

Figure 2.5. Spectro-temporal information of speech. The grayscale graph on the lower left side is the simulated neural representation of speech in the brainstem. This spectro-temporal neural representation is similar to a spectrogram but in log frequency scale. A spectral cross section of the neural representation is shown to visualize the spectral modulations (in blue). The first two formants (F1 and F2) and the first three harmonics (H1, H2 and H3) are marked on the spectral profile. A temporal cross section of the neural representation is shown in red, to visualize the temporal modulations. The temporal modulations in speech are dominated by slow power fluctuations on the order of a couple of Hertz. A short segment of the temporal cross section is zoomed in to visualize the periodicity cue (in cyan), reflects the pitch of the speech.

2.3.2 Cortical processing of temporal modulations

Neural processing of temporal modulations has been mostly studied using amplitude modulated (AM) sounds and frequency modulated (FM) sounds, which are among the simplest sounds that can capture some important dynamic features of speech and other natural sounds. Numerous studies have investigated the neural mechanisms underlying AM and FM processing (see, e.g., Joris et al., 2004 for a review). Slow AM/FM (< 20 Hz) has been most intensively studied since they drive cortical neurons most effectively (Eggermont, 2002; Liang et al., 2002). These slow modulations are encoded by sustained phase-locked neural activity in the cortex, as shown using single unit recording, EEG, and MEG (Alaerts et al., 2009; Ding and Simon, 2009; Eggermont, 2002; Picton et al., 1987; Rees et al., 1986; Wang et al., 2012). The neural response to a 5-Hz amplitude modulated broadband noise is illustrated in Fig. 2.6. When the stimulus modulation frequency is higher than ~40 Hz, phase-locking to the periodic features of an AM or FM is greatly weakened. It has been suggested that slow modulations are represented by a temporal code, while fast modulations are represented by a firing rate code (Wang et al., 2003).

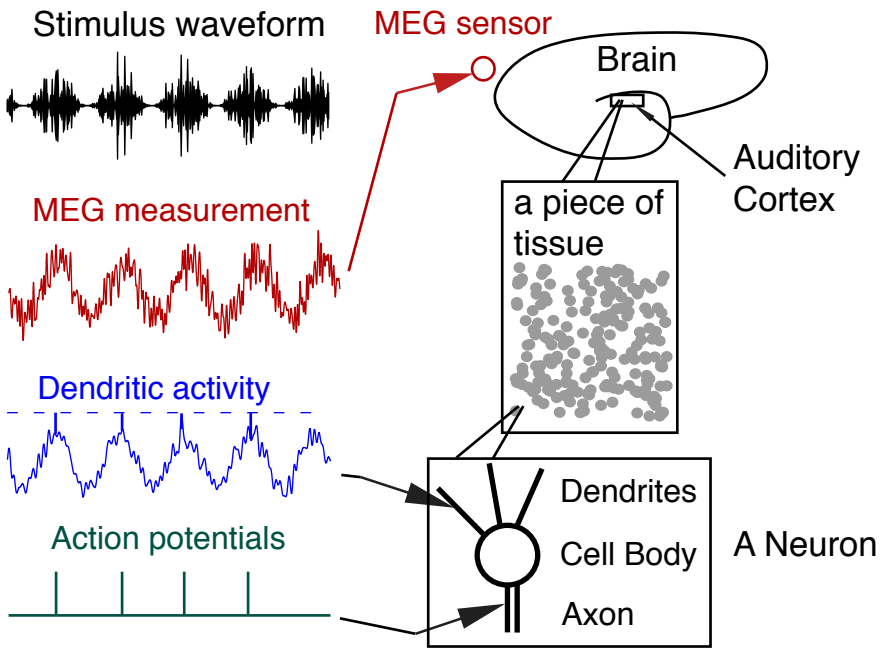

Figure 2.6. A schematic illustration of the neural phase locking to a 5 Hz AM. An idealized neuron generates an action potential at a certain phase in every cycle of the AM. Dendritic activity and the MEG signal follow the AM as a slow oscillation at 5 Hz. The carrier signal of the stimulus, in this case a broadband noise, does not significantly affect the neural coding of temporal modulations.

2.3.3 Modeling the neural processing of temporal modulations

As discussed in the previous sections, in auditory cortex, the spectral modulations of sound are represented spatially (by different neurons) while the temporal modulations of sound are represented temporally (by the response waveforms of neurons). In this section, we discuss theoretical models of the neural processing of temporal modulations.

A linear time-invariant (LTI) system is characterized by its impulse response. The simplest way to characterize the impulse response of a system is to play a white noise and calculate the cross correlation between the system output and stimulus white noise. This method has been applied to estimate the impulse response reflecting cortical processing of temporal modulations (Lalor et al., 2009). (Since the temporal modulations are modulations of a carrier signal, their frequency cannot be very high. Therefore, the white noise in modulation domain has to be band limited, e.g. below 30 Hz.) The auditory system, however, is certainly nonlinear. Therefore, the impulse response is usually stimulus dependent. Spectro-temporally structured sounds, such as random chords and natural sounds, have also been applied to model the auditory system (Bitterman et al., 2008; David et al., 2009; David et al., 2007; deCharms et al., 1998; Theunissen et al., 2001). Natural sounds, unlike white noise, are correlated over time. Therefore, the autocorrelation in natural sounds has to be taken into consideration when estimating the impulse response.

Let us denote the temporal modulation of the stimulus as $x(t)$ and the neural response as $y(t)$. Both $x(t)$ and $y(t)$ are discrete time signals and their relation, when described by an LTI system, is

$$y(t) = \sum_{\tau=-\infty}^{\infty} x(t-\tau)h(\tau) + \varepsilon(t), \tag{2.3.1}$$

where $h(t)$ is the impulse response of the system (called the *temporal* response function (TRF) in this dissertation), and $\varepsilon(t)$ is the neural response cannot be explained by the stimulus using the LTI model. $\varepsilon(t)$ is uncorrelated with $x(t-\tau)$, for arbitrary time delay τ.

For a neural system, the impulse response is causal and of finite duration, i.e. $h(t) = 0$ for $t < 0$ and $t > T$. With this constraint, the relation between $x(t)$ and $y(t)$ is more conveniently expressed as

$$\begin{aligned} y(t) &= \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}(t) + \varepsilon(t), \\ \boldsymbol{h} &= \left(h(0), h(1), \cdots, h(T)\right) \\ \boldsymbol{x}(t) &= \left(x(t), x(t-1), \cdots, x(t-T)\right) \end{aligned}, \tag{2.3.2}$$

By assuming the stimulus and response to be stationary, we have

$$\begin{aligned} \mathrm{E}\left(y(t)\boldsymbol{x}(t)^{\mathrm{T}}\right) &= \boldsymbol{h}^{\mathrm{T}}\mathrm{E}\left(\boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{T}}\right) + \mathrm{E}\left(\varepsilon(t)\boldsymbol{x}(t)^{\mathrm{T}}\right) \\ \boldsymbol{h} &= \mathrm{E}^{-1}\left(\boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{T}}\right)\mathrm{E}\left(y(t)\boldsymbol{x}(t)\right) \end{aligned}, \tag{2.3.3}$$

where $\mathrm{E}(\cdot)$ denotes expectation over time. This solution is commonly known in the neuroscience literature as the normalized reverse correlation (Theunissen et al., 2001). This solution involves inverting the autocorrelation matrix of the stimulus envelope. When the stimulus envelope is white noise, the autocorrelation matrix is an identity matrix and therefore can be ignored. When the stimulus envelope has strong autocorrelation, however, inverting the autocorrelation may be an ill-posed problem, especially when the recording time is limited. To get a robust normalized reverse correlation, a principal component analysis (PCA) based dimension reduction or Tikhonov regularization is usually employed.

Another way to solve the impulse response of neural systems is via boosting (David et al., 2007). The boosting algorithm assumes the impulse response to be sparse in time. It starts with a null impulse response and iteratively updates it to minimize the

prediction error of the model. In each iteration, $\boldsymbol{h}$ is changed by $\Delta\boldsymbol{h}$. Each $\Delta\boldsymbol{h}$ contains only one nonzero element and is optimized to minimize the prediction error:

$$\begin{aligned} \Delta\boldsymbol{h} &= \underset{\Delta\boldsymbol{h}}{\operatorname{argmin}}\, \mathrm{E}\left(\left(y(t) - \left(\boldsymbol{h} + \Delta\boldsymbol{h} \right)^{\mathrm{T}} \boldsymbol{x}(t) \right)^2 \right) \\ \left\| \Delta\boldsymbol{h} \right\|_0 &= 1 \text{ and } \left\| \Delta\boldsymbol{h} \right\|_1 = \delta \end{aligned} \quad , \tag{2.3.4}$$

where $\|.\|_0$ and $\|.\|_1$ are the L_0 and L_1 norm respectively. Since the neural processing of temporal modulations is intrinsically nonlinear, its LTI model is stimulus dependent (Bitterman et al., 2008; David et al., 2009; Theunissen et al., 2001). Each LTI model can be viewed as a linear approximation of the nonlinear system under a certain stimulation condition.

2.4 Noninvasive neural recording using MEG

2.4.1 MEG system

MEG records the magnetic field generated by neural currents in the cortex. MEG is most sensitive to dendritic currents spatially synchronized over a large scale (on the order of millimeter) (Baillet et al., 2001; Hämäläinen et al., 1993). A whole-head MEG system contains an array of sensors laid around the head (Fig. 2.7A). The sensors contain superconducting quantum interference devices (SQUIDs) and are sensitive to magnetic fields at femtoTesla (10^{-15} Tesla) level. A magnetically shielded room is built around the MEG system to reduce the impact of environmental magnetic fields. The MEG system employed by all studies in this dissertation is the University of Maryland and the Kanazawa Institute of Technology joint (UMD-KIT) MEG system, located at University of Maryland, College Park. The UMD-KIT MEG system has 157 sensors, which are

gradiometers. A gradiometer contains a pair of parallel coils and is only sensitive to the local magnetic field in a certain direction (roughly normal to the head surface). Three magnetometers are also built into the system. They are far away from the head and measure the environmental magnetic field. Recordings from the magnetometers and their time-shifted versions are used as regressors to clean up the neural signal recorded by the gradiometers (de Cheveigné and Simon, 2007).

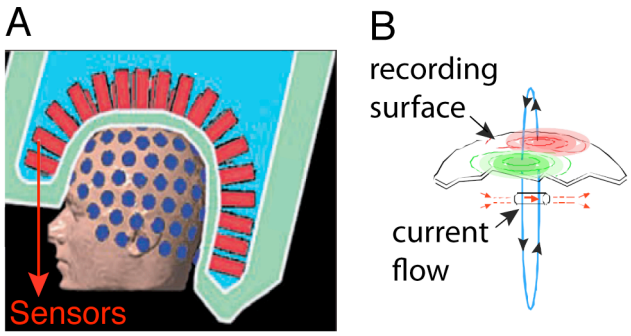

Figure 2.7. The MEG system. A, the MEG sensors are placed on around the head of a subject. B, MEG measures the magnetic field generated by electrical currents inside the brain. The magnetic field going out of the scalp is rendered in red while the magnetic field going inside is rendered in green. (Fig. 2.7A is adapted from Baillet et al. 2001)

2.4.2 Source-space MEG analysis

The relation between a current source in the brain and extracranially measured magnetic field is decided by the electromagnetic conductivity property of the head. A complete characterization of the conductivity property of the head is complicated and usually unrealistic. One way to make the problem tractable is to assume the head to be a

sphere. Although simple, the single sphere model works well and is widely used (Baillet et al., 2001). More sophisticated models include multiple-sphere models and element boundary models (Mosher et al., 1999), which are more computationally heavy and require a precisely digitized head. For the spherical head model, which is employed in this dissertation, there is a closed form relation between the cortical current distribution and the MEG recording.

Single Dipole Model

Suppose a single current dipole is located at $\boldsymbol{r}_q$ and has dipole moment $\boldsymbol{q}$. The magnitude of $\boldsymbol{q}$ is the strength of the current while the orientation of $\boldsymbol{q}$ indicates the direction of the current (in 3-D). When all the MEG sensors are radially oriented towards the center of the spherical head, the magnetic field picked up by a sensor located at $\boldsymbol{r}$ is described as (Baillet et al., 2001),

$$B\left(\boldsymbol{r},\boldsymbol{r}_q,\boldsymbol{q}\right)=\frac{\mu_0}{4\pi}\frac{\boldsymbol{r}\times\boldsymbol{r}_q}{\boldsymbol{r}\left\|\boldsymbol{r}-\boldsymbol{r}_q\right\|_2^3}\cdot\boldsymbol{q}\,, \tag{2.4.1}$$

where $\|.\|_2$ is the L_2 norm. From the equation, it can be seen that the magnetic field $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q})$ is linear with respect to the moment of the dipole, $\boldsymbol{q}$, but nonlinear with respect to the location of the dipole $\boldsymbol{r}_q$.

In most MEG systems, like the UMD-KIT system, the sensors are not perfectly radially oriented. In this case, the magnetic field $B(\boldsymbol{r})$ is still linear with respect to the dipole moment and nonlinear with respect to the dipole location. The magnetic field is expressed as follows (Mosher et al., 1999).

$$B\left(\boldsymbol{r},\boldsymbol{r}_q,\boldsymbol{q}\right)=\frac{\mu_0}{4\pi F^2\left(\boldsymbol{r},\boldsymbol{r}_q\right)}\left(F\left(\boldsymbol{r},\boldsymbol{r}_q\right)\boldsymbol{q}\times\boldsymbol{r}_q-\left(\boldsymbol{q}\times\boldsymbol{r}_q\right)\cdot\boldsymbol{r}\nabla F\left(\boldsymbol{r},\boldsymbol{r}_q\right)\right) \tag{2.4.2}$$

$$F\left(\boldsymbol{r},\boldsymbol{r}_q\right)=d\left(rd+r^2-\left(\boldsymbol{r}\cdot\boldsymbol{r}_q\right)\right)$$

$$\nabla F\left(\boldsymbol{r},\boldsymbol{r}_q\right)=\left(\frac{d^2}{r}+\frac{\left(\boldsymbol{d}\cdot\boldsymbol{r}\right)}{d}+2d+2r\right)\boldsymbol{r}-\left(d+2r+\frac{\left(\boldsymbol{d}\cdot\boldsymbol{r}\right)}{d}\right)\boldsymbol{r}_q$$

$$\boldsymbol{d}=\boldsymbol{r}-\boldsymbol{r}_q, d=\left\|\boldsymbol{d}\right\|_2, \text{and } r=\left\|\boldsymbol{r}\right\|_2$$

When the dipole location $\boldsymbol{r}_q$ and MEG sensor location $\boldsymbol{r}$ are fixed, the magnetic field $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q})$ is solely a linear function of dipole moment $\boldsymbol{q}$. It simply scales when the magnitude of dipole moment changes but behaves in a slightly more complicated way when the dipole moment rotates. The (3-D) dipole moment $\boldsymbol{q}$ can be decomposed into three orthogonal components, $\boldsymbol{q} = q_x\boldsymbol{q}_x + q_y\boldsymbol{q}_y + q_z\boldsymbol{q}_z$, where $\boldsymbol{q}_x$, $\boldsymbol{q}_y$, and $\boldsymbol{q}_z$ are unit length vectors pointing to three orthogonal directions and q_x, q_y and q_z are dipole strength in corresponding directions. Suppose $\boldsymbol{q}_x$ and $\boldsymbol{q}_y$ are tangential to the spherical head surface while $\boldsymbol{q}_z$ is normal to the surface. Inside a spherical conductor, dipoles oriented radially do not generate any magnetic field that is measurable outside the conductor (Baillet et al., 2001). Therefore, component $\boldsymbol{q}_z$ is ignored when calculating $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q})$ and, consequently, $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}) = B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}_x)q_x + B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}_y)q_y$. When $\boldsymbol{q}$ rotates, it is still a weighted sum of $\boldsymbol{q}_x$, $\boldsymbol{q}_y$ and $\boldsymbol{q}_z$ and therefore the new magnetic field is still a weighted sum of $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}_x)$ and $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}_y)$. Since $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}_x)$ and $B(\boldsymbol{r}, \boldsymbol{r}_q, \boldsymbol{q}_y)$ are generated from unit magnitude dipoles, they can be calculated independent of MEG measurements.

For each MEG system, sensor positions, $\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{Ns}$, are fixed. The output of the MEG sensor array can be denoted as $\boldsymbol{B}(\boldsymbol{r}_q, \boldsymbol{q}) = [B(\boldsymbol{r}_1, \boldsymbol{r}_q, \boldsymbol{q}), B(\boldsymbol{r}_2, \boldsymbol{r}_q, \boldsymbol{q}), ..., B(\boldsymbol{r}_{Ns}, \boldsymbol{r}_q, \boldsymbol{q})]$. As a result of linearity,

$$\boldsymbol{B}\left(\boldsymbol{r}_q,\boldsymbol{q}\right)=\left[\boldsymbol{B}\left(\boldsymbol{r}_q,\boldsymbol{q}_x\right)\boldsymbol{B}\left(\boldsymbol{r}_q,\boldsymbol{q}_y\right)\right]\begin{bmatrix}q_x\\q_y\end{bmatrix} \tag{2.4.3}$$

where

$$\boldsymbol{B}\left(\boldsymbol{r}_q,\boldsymbol{q}_x\right)=\left[B\left(\boldsymbol{r}_1,\boldsymbol{r}_q,\boldsymbol{q}_x\right)\;B\left(\boldsymbol{r}_2,\boldsymbol{r}_q,\boldsymbol{q}_x\right),\cdots,B\left(\boldsymbol{r}_{Ns},\boldsymbol{r}_q,\boldsymbol{q}_x\right)\right]^{\mathrm{T}}$$
$$\boldsymbol{B}\left(\boldsymbol{r}_q,\boldsymbol{q}_y\right)=\left[B\left(\boldsymbol{r}_1,\boldsymbol{r}_q,\boldsymbol{q}_y\right)\;B\left(\boldsymbol{r}_2,\boldsymbol{r}_q,\boldsymbol{q}_y\right),\cdots,B\left(\boldsymbol{r}_{Ns},\boldsymbol{r}_q,\boldsymbol{q}_y\right)\right]^{\mathrm{T}}.$$

The magnetic field generated by a single current dipole is illustrated in Fig. 2.7B.

Multiple Dipole Model

There is usually more than one measurable neural source inside the brain. Therefore, the measured MEG response is a linear combination of the magnetic fields generated from multiple dipoles. In a matrix form, the response is expressed as,

$$\boldsymbol{B}_{\mathrm{MEG}}=\sum_{j=1}^{J}\boldsymbol{B}\left(\boldsymbol{r}_{qj},\boldsymbol{q}_j\right)=B\boldsymbol{q} \tag{2.4.4}$$

where

$$B=\begin{bmatrix}\boldsymbol{B}\left(\boldsymbol{r}_{q1},\boldsymbol{q}_{x1}\right) & \boldsymbol{B}\left(\boldsymbol{r}_{q1},\boldsymbol{q}_{y1}\right) & \boldsymbol{B}\left(\boldsymbol{r}_{q2},\boldsymbol{q}_{x2}\right) & \boldsymbol{B}\left(\boldsymbol{r}_{q2},\boldsymbol{q}_{y2}\right) & \cdots & \boldsymbol{B}\left(\boldsymbol{r}_{qJ},\boldsymbol{q}_{xJ}\right) & \boldsymbol{B}\left(\boldsymbol{r}_{qJ},\boldsymbol{q}_{yJ}\right)\end{bmatrix}$$
$$q=\begin{bmatrix}q_{x1} & q_{y1} & q_{x2} & q_{y2} & \cdots & q_{xJ} & q_{yJ}\end{bmatrix}^T$$

In the expression, j is the index of individual dipoles and J is the total number of dipoles. The matrix B is commonly called the lead field matrix. It is independent of the strength of dipoles and is decided by the property of the head model and MEG sensor configuration. Eq. 2.4.4 describes how an electrical activity pattern is transformed into a magnetic field, which is known as the forward problem (Baillet et al., 2001).

Sometimes, it is convenient to digitize the cortex into voxels and treat the neural activity from each voxel as a dipole. If a voxel is not activated, it is represented by a dipole of zero amplitude. Using this voxelized model, one can calculate the lead field matrix purely based on the head model, independent of the properties of neural sources. This way, the lead field matrix B can be viewed as an overcomplete basis to represent MEG signals, and $\boldsymbol{q}$ is a set of coefficients in the overcomplete basis set.

Neural Source Estimation

In experiments, the scalp magnetic field $\boldsymbol{B}_{\mathrm{MEG}}$ is measured while the source activity pattern $\boldsymbol{q}$ is unknown. The number of neural sources J is also unknown. Estimating $\boldsymbol{q}$ based on $\boldsymbol{B}_{\mathrm{MEG}}$ is known as the inverse problem (Baillet et al., 2001), which is not only hard to solve but also has no unique solution. In solving the inverse problem, we map the MEG response to a neural source activation pattern in the brain. Hence, MEG analysis methods based on solving the inverse problem are also called source space analysis methods.

The most classic solution to the inverse problem is current-equivalent dipole fitting (Baillet et al., 2001). This method assumes that only a few neural sources are active, i.e. J small. Dipole fitting is an iterative process. It first tries to explain the measured magnetic field by a single dipole. Estimating the dipole position is a nonlinear problem and can be solved using various nonlinear optimization approaches (Uutela et al., 1998). Estimation of the dipole moment is linear and can be solved by, e.g., least squares methods (Baillet et al., 2001; Mosher et al., 2003). After the best dipole fit is determined, its magnetic field is removed from the measurement and another dipole is

fitted. If the number of neural sources J is known, J iterations are needed. Otherwise, the iteration stops when the measured magnetic field is satisfactorily explained by the fitted dipoles. The goodness of dipole fitting is evaluated as the correlation between the measured magnetic field and the fitted magnetic field,

$$C_{\text{fit}} = \frac{\boldsymbol{B}_{\text{MEG}}^{\text{T}} \boldsymbol{B}_{\text{fit}}}{\left\| \boldsymbol{B}_{\text{MEG}} \right\|_2 \left\| \boldsymbol{B}_{\text{fit}} \right\|_2}, \tag{2.4.5}$$

where $\boldsymbol{B}_{\text{fit}}$ is the magnetic field generated by the dipoles.

Another popular solution to the inverse problem is the minimum norm estimation (MNE) (Hämäläinen and Ilmoniemi, 1994). MNE starts with calculating the lead field for a voxelized brain. Therefore, both $\boldsymbol{B}_{\text{MEG}}$ and B in Eq. 2.4.4 are available. Even in this case, the source activity $\boldsymbol{q}$ still cannot be uniquely solved since Eq. 2.4.4 is highly underdetermined. The MNE method does not only require Eq. 2.4.4 to hold but also minimizes the L_2 norm of $\boldsymbol{q}$. The most basic solution of MNE methods is $\boldsymbol{q} = B^{+}\boldsymbol{B}_{\text{MEG}}$, where B^{+} is the pseudoinverse of B.

A third popular solution is the beamforming approach (Van Veen et al., 1997). Electromagnetic signal travels at light speed and the MEG sensors are closely spaced. Hence, all MEG sensors receive a signal at virtually the same time. As a result, an MEG beamformer cannot rely on the propagation time, but only the amplitude of the signal received by different sensors. The LCMV beamformer and the SAM beamformer are the most popular beamformers (Robinson and Vrba, 1999; Van Veen et al., 1997). They are both closely related to generalized least squares dipole fitting (Mosher et al., 2003).

2.4.3 Data driven MEG analysis

Component Analysis

Any solution to the inverse problem relies on the lead field matrix $\boldsymbol{B}$, which is based on a head model and is therefore unavoidably imprecise. As mentioned in Section 2.4.3, solving the inverse problem is the same as representing the MEG measurement using a basis set determined by the head model. Another way to analyze MEG data is to use a data driven approach to find a set of basis that is statistically optimal to represent the MEG signals. Popular data driven approaches include principal component analysis (PCA), independent component analysis (ICA), common spatial patterns (CSP), and denoising source separation (DSS) (Bell and Sejnowski, 1995; Blankertz et al., 2008; Särelä and Valpola, 2004). For the analysis of MEG responses evoked by auditory stimuli, DSS appears to be an especially effective method (de Cheveigné and Simon, 2008; Wang et al., 2012) and can be used as a preprocessing method for dipole fitting (Ding and Simon, 2009). Each linear component of the MEG measurement is a linear combination of the measurements from different sensors:

$$x(t) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{B}_{\mathrm{MEG}}(t) \tag{2.4.6}$$

The MEG measurement contains neural activity of interest and all other kinds of interferences, e.g. environmental noises and background neural activity. Typically, neural activity of interest is assumed to be independent of any other activity. Symbolically, this decomposition of MEG activity is represented as $\boldsymbol{B}_{\mathrm{MEG}}(t) = \boldsymbol{B}_{\mathrm{s}}(t) + \boldsymbol{B}_{\mathrm{n}}(t)$, where $\boldsymbol{B}_{\mathrm{s}}$ is the neural activity of interest and $\boldsymbol{B}_{\mathrm{n}}$ contains interfering activity. $\boldsymbol{B}_{\mathrm{s}}(t)$ and $\boldsymbol{B}_{\mathrm{n}}(t)$ are assumed to be independent from each other, and therefore $R_{\mathrm{MEG}} = R_{\mathrm{s}} + R_{\mathrm{n}}$, where R_{MEG}, R_{s} and R_{n} are the autocorrelation matrices of $\boldsymbol{B}_{\mathrm{MEG}}(t)$, $\boldsymbol{B}_{\mathrm{s}}(t)$ and $\boldsymbol{B}_{\mathrm{n}}(t)$.

DSS is equivalent to the generalized eigenvalue decomposition of R_s and R_n (Fukunaga, 1972). It is the solution of the following optimization problem

$$\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^T R_s \boldsymbol{a}}{\boldsymbol{a}^T R_n \boldsymbol{a}} . \tag{2.4.7}$$

$\boldsymbol{a}^T R_s \boldsymbol{a}$ is the power of neural activity of interest in a DSS component, and $\boldsymbol{a}^T R_n \boldsymbol{a}$ is the power of other activity in that component. Therefore, DSS maximizes the signal to noise ratio in each component. R_s is usually approximated by the covariance matrix of the evoked response (Section 2.5.1). R_n can be replaced by R_{MEG} based on the property of generalized eigendecomposition (Fukunaga, 1972).

After a DSS component is derived, it is subtracted from the raw data $\boldsymbol{B}_{MEG}(t)$ and another DSS component can be derived in the same way. Each DSS component corresponds to a generalized eigenvector of R_{MEG} and R_s. If we denote the whole set of DSS filters as $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_k]$, then the whole set of DSS components is $\boldsymbol{d}(t) = A\boldsymbol{B}_{MEG}(t)$, where $\boldsymbol{d}(t) = [d_1(t), d_2(t), ..., d_k(t)]^T$. Let $U = A^{-1}$ and $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_k]$, then

$$\boldsymbol{B}_{MEG}(t) = d_1(t)\boldsymbol{u}_1 + d_2(t)\boldsymbol{u}_2 + ... + d_k(t)\boldsymbol{u}_k. \tag{2.4.8}$$

Therefore, DSS is essentially re-representing the measurement using a new set of bases U. The neurophysiological meaning of Eq. 2.4.8 is that the MEG measurement consists of k uncorrelated magnetic field patterns, presumably generated by k current sources.

Combining Component Analysis with Source-space Analysis

Source space analysis methods convert the scalp magnetic field into a neural current distribution over the cortex and therefore have clear physiological meanings.

Nonetheless, source space analysis relies on a conductivity model of the head, which is unavoidably imprecise, and needs to solve the inverse problem, which is not robust. Data driven methods are more flexible and computationally efficient but their results are sometimes hard to interpret neurophysiologically.

One way to get robust and physiologically meaningful decomposition of the MEG measurement is to combine data driven methods and source space methods sequentially. For example, in the analysis of the three experiments included in this dissertation, DSS is first applied to the MEG recording as a dimensional reduction method. The DSS components containing reliable neural activity are analyzed, and the results, which are in the DSS space, are then converted back to the sensor space, in which a source analysis is applied.

This combined method is more computationally efficient than source space analysis methods, which exhaustively analyzes each voxel of the brain. The combined method is also better than pure data driven methods in that it has a clear physiological interpretation.

2.5 MEG responses to auditory stimuli

Evoked MEG Response

MEG experiments usually repeat the same stimulus several times. The response averaged over trials is called the evoked MEG response. Responses having a random phase in each trial are attenuated by the average. Therefore, the evoked MEG response primarily reflects neural activity synchronized to the stimulus. Suppose, for example, a

neural network, denoted *A,* encodes a 5-Hz AM sound by purely enhancing its own spontaneous activity, while another neural network, *B*, encodes the same stimulus by generating a 5 Hz response phase locked to the stimulus envelope. In this case, only the response of network *B* will survive the average over trials. Several types of MEG responses have been demonstrated to be phase locked to the stimulus, for example the M100 response and the response following slow temporal modulations (Ding and Simon, 2009; Fuentemilla et al., 2006; Luo and Poeppel, 2007).

Transient MEG/EEG Responses

The transient MEG/EEG response to the onset/offset of a sound has been extensively studied (Näätänen and Picton, 1987). The major components of the transient response are first defined in EEG as the P1-N1-P2 response complex. P1 (respectively, P2) is a positive potential measured at around 50 ms (150-200 ms) post-stimulus and N1 is a negative potential measured at around 100 ms post-stimulus. In MEG, a similar response pattern is observed and is called the P1m-N1m-P2m complex or, based on the response latency, M50-M100-M150 complex (Chait et al., 2004; Poeppel et al., 1996). An example of the MEG transient response is shown in Fig. 2.8.

The N1/M100 response is the most reliable component in the transient response complex. Its latency and amplitude are affected by various stimulus properties, e.g. loudness, frequency composition, onset, and signal to noise ratio (Biermann and Heil, 2000; Kaplan-Neeman et al., 2006; Näätänen and Picton, 1987; Poeppel et al., 1996), and is modulated by attention (Hillyard et al., 1973). Although the N1/M100 response is frequently observed at the onset/transition of a sound, it is not observed after every

onset/transient in a sound (Chait et al., 2007; Chait et al., 2004; Gutschalk et al., 2008; Näätänen and Picton, 1987). The generation of the N1/M100 response may be related to the perceptual saliency of the sound onset/offset.

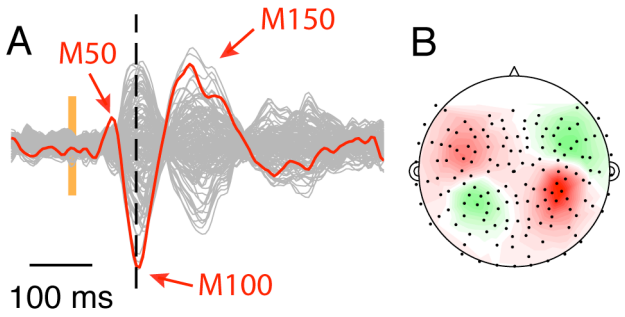

Figure 2.8. The M100. (A) The MEG response to a 20-ms pure tone, with all MEG channels overlaid together. The response from one MEG channel is plotted in red to show the polarities of responses. The sound stimulus is illustrated as a bar in orange. (B) The spatial magnetic field distribution of the MEG response at 100 ms post-stimulus on a flattened head.

The exact neural source of the M100 response is controversial. For example, some researchers localized it to planum temporale (PT) (Lütkenhöner and Steinsträter, 1998) while others localized it to the lateral part of the Heschl's gyrus (Herdman et al., 2003). On a coarse level, however, all the studies agree that the M100 response is from the superior portion of the superior temporal gyrus, within auditory cortex.

Chapter 3

Cortical representation of continuous speech

3.1 Introduction

Spoken language is the dominant form of human communication, and human listeners are superb at tracking and understanding speech even in the presence of interfering speakers (Bronkhorst, 2000; Cherry, 1953). The critical acoustic features of speech are distributed across several distinct spectral and temporal scales. The slow temporal modulations and coarse spectral modulations reflect the rhythm of speech and contain syllabic and phrasal level segmentation information (Greenberg, 1999), and are particularly important for speech intelligibility (Shannon et al., 1995). The neural tracking of slow temporal modulations of speech (e.g. 1–10 Hz) in human auditory cortex can be studied noninvasively using magnetoencephalography (MEG) and electroencephalography (EEG). The low frequency, large-scale, synchronized neural activity recorded by MEG/EEG has been demonstrated to be synchronized by speech stimulus (Luo and Poeppel, 2007) and is phase-locked to the speech envelope, i.e. the slow modulations summed over a broad spectral region (Abrams et al., 2008; Ahissar et al., 2001; Aiken and Picton, 2008; Lalor and Foxe, 2010; Luo and Poeppel, 2007). Temporal locking to features of speech has also been supported by intracranial recordings from human core auditory cortex (Nourski et al., 2009). The temporal features of speech contribute significantly to speech intelligibility, as do key spectral-temporal features in speech such as upward and downward formant transitions. The neural coding of spectro-temporal modulations in natural soundtracks has been studied invasively in human

auditory cortex using intracranial extracellular recordings (Bitterman et al., 2008), where the spectro-temporal tuning of individual neurons was found to be generally complex and sometimes very fine in frequency. At a neural network level, the Blood Oxygen Level Dependent (BOLD) activity measured by functional Magnetic Resonance Imaging (fMRI) also shows complex spectra-temporal tuning and possesses no obvious spatial map (Schönwiesner and Zatorre, 2009). Which spectro-temporal features of speech are encoded in the large-scale synchronized neural activity measurable by MEG and EEG, however, remain unknown and are the focus of the current study.

When investigating the neural coding of speech, there are several key issues that deserve special consideration. One arises from the diversity of speech: language is a productive system permitting the generation of novel sentences. In everyday life, human listeners constantly decode spoken messages they have never heard. In most neurophysiological studies of speech processing, however, small sets of sentences are repeated tens or hundreds of times (though see Lalor and Foxe, 2010). This is primarily due to methodological constraints: neurophysiological recordings, especially noninvasive recordings, are quite variable, and so integrating over trials is necessary to obtain a valid estimate of the neural response. An often-neglected cost of repeated stimuli, however, is that the listener has obtained complete knowledge of the entire stimulus speech after only a few repetitions. Without the demands of speech comprehension, the encoding of this repeated speech might be quite different from the neural coding of novel speech under natural listening conditions. It is pressing, therefore, to develop experimental paradigms that do not require repeating stimuli many times, in order to study how speech is encoded in a more ecologically realistic manner.

Second, speech communication is remarkably robust against interference. When competing speech signals are present, human listeners can actively maintain attention on a particular speech target and comprehend it. The superior temporal gyrus has been identified as a region heavily involved in processing concurrent speech signals (Scott et al., 2009). Recent EEG results have shown that human auditory cortex can selectively amplify the low frequency neural correlates of the speech signal being attended to (Kerlin et al., 2010). This attentional modulation of low frequency neural activity has been suggested as a general mechanism for sensory information selection (Schroeder and Lakatos, 2009). Since speech comprehension is a complex hierarchical process involving multiple brain regions, it is unclear whether the attentional effect seen in the auditory cortex directly modulates feedforward auditory processing or reflects only feedback from language areas, or even motor areas (Hickok and Poeppel, 2007). One approach to test whether feedforward processing is involved in speech segregation is to investigate the latency of the attentional effect. If the attentional modulation of MEG/EEG response has a relatively short latency, e.g. 100 milliseconds, then it is evidence that top-down attention modulates representations that are otherwise dominated by feedforward auditory processing. Otherwise, segregating and selectively processing speech may rely on feedback from non-auditory cortex or complex recursive calculations within auditory cortex.

Additionally, the auditory encoding of speech is lateralized across the two cerebral hemispheres. It has been hypothesized that the right hemisphere is specialized for the encoding the *slow* temporal modulations of speech (Poeppel, 2003). Support for this hypothesis arises from the observation that neural activity in the right hemisphere is

more faithfully synchronized to a speech stimulus than the left, for monaurally and diotically presented speech (Abrams et al., 2008; Luo and Poeppel, 2007). Nevertheless, how this proposed intrinsic lateralization of speech encoding interacts with the asymmetry of the ascending auditory pathway is still unclear.

In this study, we investigate the neurophysiology underlying speech processing in human auditory cortex, using minutes-long spoken narratives as stimuli. To address the robustness of this neural coding of speech under more complex listening conditions, the listeners were presented with two simultaneous (and thus competing) spoken narratives, each presented in a separate ear, as a classical, well-controlled illustration of the cocktail party effect (Cherry, 1953). This design affords us both the opportunity to investigate the spectro-temporal coding of speech under top-down attentional modulation, and the opportunity to separate the intrinsic hemispheric lateralization of speech encoding with the interaction between the left and right auditory pathways. Moreover, previous studies have only demonstrated that speech is encoded in MEG/EEG activity with sufficient fidelity to discriminate among 2 or 3 sentences (Kerlin et al., 2010; Luo and Poeppel, 2007). With a long duration, discourse-level stimulus, we can test the limit of this fidelity by quantifying the maximum number of speech stimuli that can be discriminated based on MEG responses.

Inspired by research on single-unit neurophysiology (deCharms et al., 1998; Depireux et al., 2001), the analysis of MEG activity was performed using the *spectro-temporal response function* (STRF), which can reveal neural coding mechanisms by analyzing the relationship between ongoing neural activity and the corresponding continuous stimuli. The properties of network-level cortical activity, which plays an

important role in auditory processing (Panzeri et al., 2010; Schroeder and Lakatos, 2009), were characterized in terms of features of the STRF, such as the spectro-temporal separability (Depireux et al., 2001; Schönwiesner and Zatorre, 2009), predictive power (David et al., 2009), binaural composition (Qiu et al., 2003), attentional modulation (Fritz et al., 2003), and hemispheric lateralization, in parallel with what has been done in single neuron neurophysiology and fMRI. The quantification of these fundamental neurophysiological features establishes the neural strategy used to encode the spectro-temporal features of speech in mass neural activity, conveying information complimentary to that obtained by single unit neurophysiology and fMRI.

3.2 Methods

3.2.1 Subjects, stimuli and procedures

Subjects

Ten normal hearing, right-handed, young adults (between 19 and 25 years old) participated in the experiment, six female. One additional subject participated in the experiment but was excluded from analysis due to excessive head movement (> 2 cm) during the experiment. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland institutional review board. Written informed consent form was obtained from each subject before the experiment.

Stimuli

Our stimulus consisted of 2 segments from a public domain narration of the short story *The Legend of Sleepy Hollow* by Washington Irving (http://librivox.org/the-legend-of-sleepy-hollow-by-washington-irving/), read by a male speaker. The 2 segments were extracted from different sections of the story and each of the two-minute duration segments was further divided into 2 one-minute long stimuli. The speech signal was low pass filtered below 4 kHz. Periods of silence longer than 300 ms were shortened to 300 ms and white noise, 20 dB weaker than the speech, was added to the signal to mask any possible subtle discontinuities caused by the removal of silent periods. All stimuli were presented at a comfortable loudness level of around 65 dB. The two stimulus segments were sinusoidally amplitude modulated at 95% modulation depth, at 37 and 45 Hz respectively. As determined by Miller and Licklider (1950), gating a speech signal on and off at a high rate (near 40 Hz) does not significantly affect the intelligibility of speech. Such a gating, however, enabled the analysis of auditory steady state response (aSSR), commonly localized to core auditory cortex (Herdman et al., 2003), and therefore allowed us to monitor the activity in the earliest stage of cortical auditory processing. The association between stimulus segment and modulation rate was counterbalanced over subjects.

Procedures

The dichotic listening condition was conducted first. The two audio book excerpts were presented dichotically (separately in each ear) to the subjects using a tube phone plugged into the ear canal. The subjects were instructed to focus on one of the ears until

the stimulus ended. Then the same stimulus was played again but the subjects were instructed to switch focus to the other ear. This process was repeated three times for the same set of stimuli, resulting in 3 identical experimental blocks. All subjects described the dichotic listening task as moderately difficult, and all but one subject reported paying more, or a similar amount of, attention during the second and third presentations of a stimulus, compared with the attention they paid to the first presentation. Which stimulus was played first and which ear was attended to first were counterbalanced over subjects (Table 3.1). After the dichotic listening condition, the monaural speech condition was presented. In this condition, each audio book excerpt was presented monaurally, on the same side as in the dichotic condition. Each stimulus was repeated four times. The subjects kept their eyes closed during the whole experiment and had a break every minute. During the break, they were asked a question related to the comprehension of the passage they just heard. On average, the subjects answered 90% of the questions correctly. The performance of the subjects was not significantly different over the 3 repetition of the stimulus (1-way repeated-measures ANOVA). Additionally, before the main experiment, a pre-experiment was performed. 100 repetitions of a 500-Hz tone pip were presented to each subject to measure the M100 response.

	Left Ear		Right Ear	
	Listening Material	AM rate (Hz)	Listening Material	AM rate (Hz)
Order A	***Segment 1***	***45***	Segment 2	37
Order B	Segment 1	37	***Segment 2***	***45***
Order C	Segment 2	45	***Segment 1***	***37***

Table 3.1. The different stimulus sets used. The stimulus attended to first in the cocktail-party-like condition is in bold italic. Each listening material segment is 2 minutes in duration and is presented to the same ear for both the single speech condition and the cocktail-party-like condition.

3.2.2 Data recording and analysis

Data Recording and Processing

The neuromagnetic signal was recorded using a 157-channel whole-head MEG system (KIT, Kanazawa, Japan), in a magnetically shielded room, with 1 kHz sampling rate. A 200-Hz low-pass filter and a notch filter at 60 Hz were applied on-line. Three reference channels were used to measure and cancel the environmental magnetic field (de Cheveigne and Simon, 2007). Five electromagnetic coils were used to measure each subject's head position inside the MEG machine. The head position was measured twice, once before and once after the experiment, to quantify the head movement during the experiment.

MEG Processing and Neural Source Localization

Recorded MEG signals contain not only responses directly driven by the stimulus, but also stimulus-irrelevant background neural activity. The response component reliably tracking stimulus features is consistent over trials but the stimulus-irrelevant neural activity is not. Based on this property, we decomposed the MEG recording using Denoising Source Separation (DSS) (de Cheveigne and Simon, 2008), a blind source separation method that extracts neural activity consistent over trials. Specifically, DSS decomposes the multi-channel 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. We found that only the first DSS component contains a significant amount of stimulus information (see *Results*), and so analysis was restricted to this component. The spatial magnetic field distribution pattern of this first DSS component was utilized to localize the source of neural responses. In all subjects, the magnetic field corresponding to the first DSS component showed a stereotypical bilateral dipolar pattern, and was therefore well modeled by a single equivalent-current dipole (ECD) in each hemisphere. A spherical head model was derived for each subject using MEG Laboratory software program v.2.001M (Yokogawa Electric, Eagle Technology, Kanazawa Institute of Technology). Position of the ECD was estimated using a global optimization approach (Uutela et al., 1998). The ECD position in each hemisphere was first determined using 54 MEG channels over the corresponding hemisphere. The positions of bilateral ECDs were then refined based on all 157 channels.

After the position of an ECD was determined, the time course of the dipole moment strength was reconstructed using the generalized least squares method (Mosher et al., 2003). In the reconstructed source activity, the polarity of M100 response was defined as negative (to be consistent with the traditional conventions of MEG/EEG research). The temporal activity reconstructed for the neural sources in the left and right hemispheres was employed for further analysis.

STRF Estimation

We modeled the cortical auditory processing using the STRF, which describes the input-output relation between a sub-cortical auditory representation and the cortical MEG response. The sub-cortical auditory representation of the sounds is a function of frequency and time and is denoted as $S_L(f, t)$ or $S_R(f, t)$, for the stimulus in the left or the right ear respectively. The MEG response is a function of time and is denoted as $r(t)$. The linear STRF model can be formulated as

$$r(t) = \sum_f \sum_\tau STRF_L(f,\tau)\, S_L(f, t-\tau) + \sum_f \sum_\tau STRF_R(f,\tau)\, S_R(f, t-\tau) + \varepsilon(t),$$

where $STRF_L(f,t)$ and $STRF_R(f,t)$ are the STRFs associated with the left and right side stimuli and $\varepsilon(t)$ is the residual response waveform not explained by the STRF model. In the monaural stimulus condition, only the relevant stimulus ear is modeled. The sub-cortical auditory representation is simulated using the model proposed by Yang, Wang and Shamma (1992). This auditory model contains 100 frequency channels between 200 Hz and 4 kHz, similar to a spectrogram in log frequency scale. For STRF estimation, the 100 frequency channels are downsampled to five.

The STRF was estimated using boosting with ten-fold cross validation (David et al., 2007). The estimation procedure is described below.

1. *Initialize the STRF.*

   $STRF_0(f,t) = 0$, for all f and t.

2. *Iteratively optimize STRF.*

   The n^{th} iteration is based on the results of the n-1^{th} iteration:

   $$r_{n-1}(t) = \sum_f \sum_\tau STRF_{n-1}(f,\tau)\, S(f, t-\tau) + \varepsilon_{n-1}(t)$$

   In the n^{th} iteration,

   $$r_n(t) = \sum_f \sum_\tau STRF_n(f,\tau)\, S(f, t-\tau) + \varepsilon_n(t), \text{ where}$$

   $$STRF_n(f,\tau) = STRF_{n-1}(f,\tau) + \Delta STRF(f,\tau)$$

   $$\Delta STRF(f,\tau) = \begin{cases} \delta, \text{ if } f = f_0, t = t_0 \\ 0 \end{cases}$$

   The prediction error in the n^{th} iteration is $\varepsilon_n(t) = \varepsilon_{n-1}(t) - \delta S(f_0, t_0)$.

   ΔSTRF is selected to minimize the prediction error, i.e.

   $$\Delta STRF(f,\tau) = \underset{f_0,t_0}{\operatorname{argmin}} \sum_t \varepsilon_n^2(t) = \underset{f_0,t_0}{\operatorname{argmin}} \sum_t \left(\varepsilon_{n-1}(t) - \delta S(f_0,t_0)\right)^2$$

3. *Terminate the iteration when the prediction error of the model drops based on cross validation.*

During STRF estimation, each one-minute long MEG response was divided into 10 segments. Nine segments were used to iteratively optimize the STRF while the

remaining segment was used to evaluate how well the STRF predicts neural responses by its "predictive power": the correlation between MEG measurement and STRF model prediction. Iteration terminated when the predictive power of the STRF decreased for the test segment (e.g. started to demonstrate artifacts of overfitting). The ten-fold cross validation resulted in ten estimates of the STRF, whose average was taken as the final result.

STRF Analysis

The spectral and temporal profiles of the STRF are extracted using singular value decomposition (SVD), $STRF(f,t) = \sum_k \lambda_k SRF_k(f) TRF_k(t)$, $\lambda_1 > \lambda_2 > \cdots$. In SVD, the sign of the singular vectors are arbitrary, but we then further require that the spectral singular vectors be overall positive, i.e. $\sum_f SRF_k(f) > 0$. We refer to the first spectral singular vector, i.e. $SRF_1(f)$, as the *normalized spectral sensitivity function*, and the product of the first temporal singular vector and its singular value, i.e. $\lambda_1 TRF_1(t)$, as the *temporal response function*. Here, the temporal response function reflects the cortical response evoked by a unit broadband power increase of the stimulus. The spectral sensitivity function and temporal response function consider only the first spectral and temporal singular vectors, and therefore only account for some fraction of the total variance of the STRF. This fraction, $\lambda_1^2 / \sum_k \lambda_k^2$, is called the separability of STRF (Depireux et al., 2001). If the separability of STRF is high (near 1), the STRF is well represented as the outer product of the normalized spectral sensitivity function and the

temporal response function, and the spectral and temporal properties of STRF can be discussed separately without any loss of information.

The temporal features of STRF, e.g. the $M100_{STRF}$, were extracted from the temporal response function, since the STRF proved to be highly separable. The $M100_{STRF}$ was determined as the strongest negative peak in the temporal response function between 70 ms and 250 ms. In analysis of the $M100_{STRF}$, the MEG responses to each one-minute long stimulus were averaged within each attentional state unless the experimental block number was employed as an analysis factor.

Decoding Speech Information from Neural Responses

The STRF model addresses how spectro-temporal features of speech are encoded in cortical neural activity. To test how faithful the neural code is, we employ a decoder to reconstruct the speech features from MEG measurements. Since STRF analyses show that only coarse spectro-temporal modulations of speech are encoded in the MEG activity (see *Results*), we concentrate on decoding the envelope of speech in a broad frequency band between 400 Hz and 2 kHz (calculated by summing the auditory channels in this range). The linear decoder is formulated as $\hat{s}(t) = \sum_{\tau} r(t+\tau) D(\tau) + \varepsilon(t)$, where $\hat{s}(t)$, $r(t)$, and $D(t)$ are the decoded speech envelope, the MEG source activity, and the decoder respectively. This decoding analysis naturally complements the STRF analysis (Mesgarani et al., 2009), and the decoder is estimated using boosting in the same way that the STRF is estimated. The time lag between neural activity and stimulus, τ, is assumed to be between 0 ms and 500 ms.

To evaluate the performance of the decoder, we calculate the correlation coefficient between the decoded envelope and the envelope of the actual stimulus, and compare it with the correlations between the decoded envelope and the envelopes of other speech signals. We define the decoding of a neural response as being successfully decoded if the decoded envelope is more correlated with the envelope of the actual stimulus than other non-stimulus envelopes. Using this criterion, when decoding the responses to the 4 one-minute duration spoken narratives, a response is correctly decoded if the reconstructed envelope is more correlated with the actual stimulus than the other 3 stimuli. In this particular case, the decoding task is not very demanding since only 2 bits of information are needed to discriminate four stimuli, while having access to the entire one-minute duration. In order to test the maximum amount of information decodable from the MEG response, we increase the difficulty of the decoding task by splitting the stimulus and the speech envelope decoded from the neural response into multiple segments and determining the relationship between stimulus and response on a segment basis. For example, if the segment duration is 2 seconds, each one-minute long stimulus/response results in 30 segments. To perfectly identify the 30 stimulus segments, one needs at least $\log(30) \approx 5$ bits of information in the 2 second long response, resulting in an information rate of 2.5 bit/s (all uses of the log function are with base 2 implied, as is customary in information theoretic analysis). It is worth noting that the information rate here describes how faithful the decoded envelope resembles the actual envelope, rather the linguistic information carried in speech.

Information theory is employed to characterize how much information can be extracted from the neural encoding of speech. The minimal amount of information

needed to discriminate N patterns is $\log(N)$ bits. When the mutual information between the stimulus and response, I(s,r), is less than $\log(N)$ bits, it is not possible to perfectly decode N equally probable stimulus patterns based on the response. The decoding accuracy is limited by Fano's inequality (Cover and Thomas, 1991).

$$H(P_e) + P_e\log(N-1) > \log(N) - I(s,r),$$

where P_e is percent of correct decoding and $H(P_e) = P_e\log(P_e) + (1-P_e)\log(1-P_e)$. From the inequality, we also have an estimate of the lower bound of the mutual information between stimulus and response: $I(s,r) > \log(N) - H(P_e) - P_e\log(N-1)$. This inequality holds for any N stimulus patterns, even if the stimulus patterns and the decoding algorithm are optimized. For simplicity, we assume the mutual information I(s,r) increases linearly with the duration of stimulus/response and therefore express the result as the mutual information rate, mutual information divided by the stimulus duration.

To avoid overfitting while decoding, we divided the 2 one-minute long stimuli in each ear into two equal size groups. We used one group to train the decoder and the other group to evaluate decoding accuracy. The two groups were then switched. The decoding results, i.e. the correlation between decoded stimuli and real stimuli, were averaged over the two groups.

Significance Tests

The statistical significance of the STRF was estimated by comparing the actual STRF results with the null distribution of the STRF parameters. To estimate the null distribution, we derived pseudo-STRFs based on each spoken narrative and mismatched neural responses. To generate a mismatched response, under each listening condition

(monaural/attended/unattended), we concatenated all the responses to the four spoken narratives and randomly selected a one-minute duration neural recording from the concatenated response. A thousand such mismatched responses were generated and resulted in 1000 pseudo-STRFs in each listening condition.

The predictive power of the actual STRF was viewed as significant if it was greater than any of the predictive powers of the 1000 pseudo-STRFs ($P < 0.001$). Similarly, the $M100_{STRF}$ in actual STRF was viewed as significant if it was stronger than any of the peaks in the pseudo-STRF in the same time window ($P < 0.001$). The amplitude of the $M100_{STRF}$ was further analyzed using repeated-measures ANOVA with Greenhouse-Geisser corrections, using the CLEAVE statistical analysis tool (http://www.ebire.org/hcnlab).

Auditory Steady State Response Analysis

Sinusoidal amplitude modulation of a stimulus would be expected to evoke an aSSR at the modulation rate. In the aSSR analysis, responses to the same stimulus were averaged and converted into the frequency domain using the Discrete Fourier Transform (DFT), with 0.017 Hz resolution (based on the one-minute duration recording). Two stimulus modulation rates, 37 and 45 Hz, were employed in the experiment. In the monaural speech condition each stimulus was only modulated at one rate, and therefore measurements at the other modulation rate were used to evaluate the background neural noise level at that frequency. The significance of the response at a modulation rate was determined by comparing the response magnitude in the presence of the stimulus

modulation and the response magnitude in the absence of the stimulus modulation (permutation test with paired data).

3.3 Results

Representation of Speech in the Low Frequency Neural Response

In the monaural listening condition, two minutes of a single spoken narrative were presented to each ear. We employ the STRF to model how the spectro-temporal modulations of speech are encoded in the MEG activity filtered into different frequency bands. Fig. 3.1 shows the predictive power of STRF, the correlation between the STRF model prediction and the MEG measurement, for every 2-Hz wide frequency band between 1 Hz and 59 Hz. The predictive power is above chance level only in the low frequency region (1 - 8 Hz), which is further analyzed in the following.

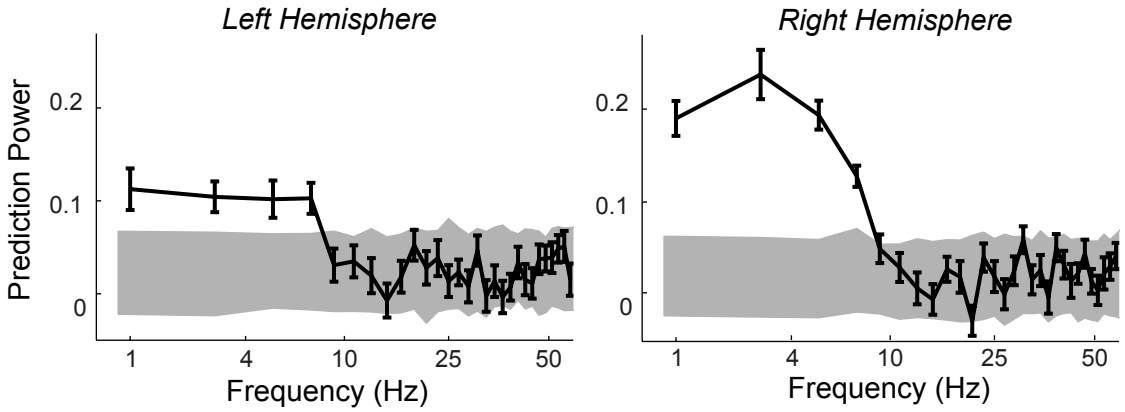

Figure 3.1. The predictive power of STRF model for each 2 Hz wide frequency band between 1 Hz and 57 Hz. The grand averaged predictive power is shown as the black line, with error bars representing one standard error on each side. The shaded gray area covers from 5 to 95 percentile of chance level predictive

power, estimated by bootstrap. The predictive power of STRF of MEG speech response is significantly higher than chance level below 7 Hz.

Neural Representation of Spectro-temporal Features in Speech

The STRF derived from the low frequency MEG response (1 - 8 Hz) is shown in Fig. 3.2A. The STRF can be interpreted in several ways (deCharms et al., 1998; Simon et al., 2007). One is that the STRF at each frequency represents the contribution to the MEG response evoked by a unit power increase of the stimulus in that frequency band. Another, complementary, interpretation is that the STRF, when reversed in time, represents the acoustic features most effective at driving MEG responses. The STRF shows the strongest activation between 400 Hz and 2 kHz, with a peak at ~100 ms post-stimulus. This peak is referred to as the $M100_{STRF}$, in parallel with the M100 evoked by sound onset. This STRF indicates that the MEG response tracks spectro-temporal modulations of speech at latency near 100 ms. From another perspective, the instantaneous MEG response is dominantly driven by spectro-temporal modulations that were present in the stimulus 100 ms ago.

The predictive power of STRF is above chance level (test described in *Methods*, $P < 0.001$) in each hemisphere for each ear, and is significantly higher in the right hemisphere (paired t-test, $t_{19} = 3.3$, $P < 0.004$). In the right hemisphere, the grand averaged predictive power is 0.25 (0.21) for the left (right) side stimulus (significantly higher for the left, paired t-test, $t_9 = 2.4$, $P < 0.05$). In the left hemisphere, the predictive power is similar for stimuli in both ears (0.11 for the left and 0.12 for the right).

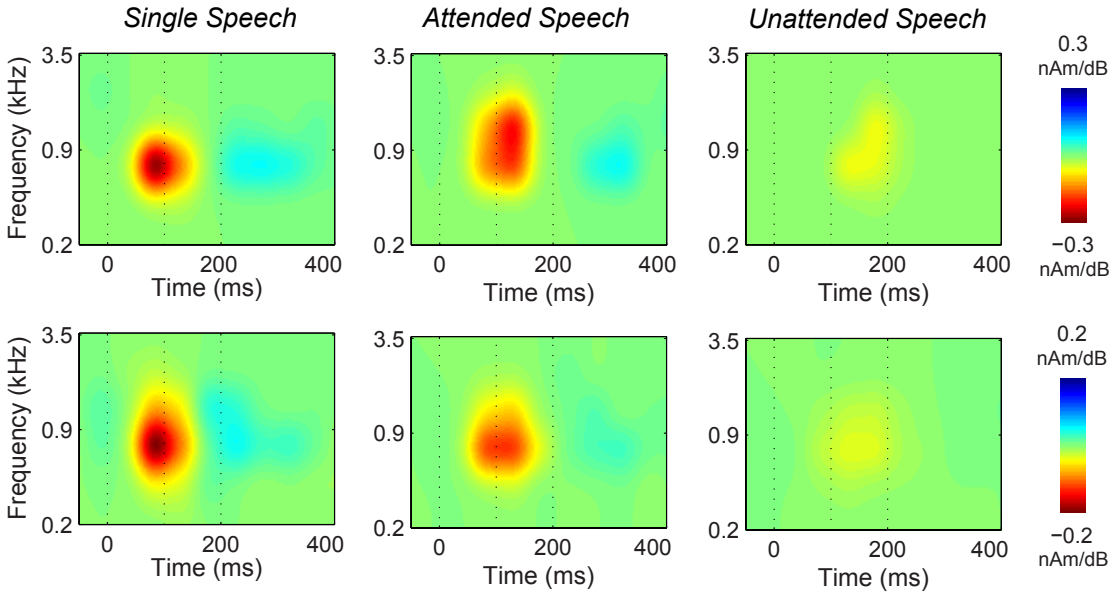

Figure 3.2. STRF derived from the MEG speech response to monaurally presented speech (A) and dichotically presented simultaneous speech signals (B). The most salient feature of the STRF is a negative peak (same polarity as M100/N1) at ~100 ms post-stimulus, sometimes followed by a later peak of opposite polarity. In the dichotic listening condition, the amplitude of the STRF is higher for the attended speech than for the interfering (unattended) speech. All examples are from the right hemisphere for speech presented contralaterally. The STRF is smoothed using a 2-D Gaussian function with standard deviations 5 semitones and 25 ms; representative subject = R1474.

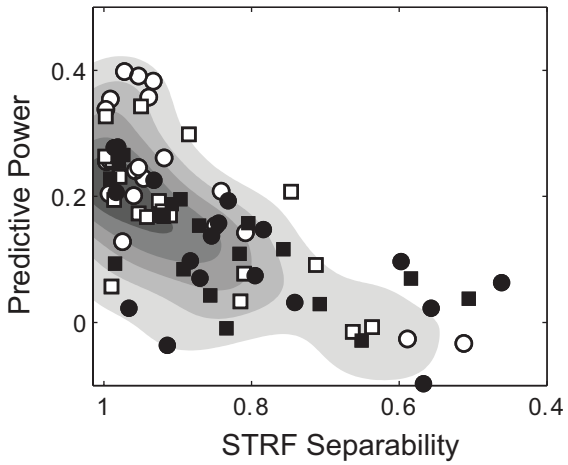

Figure 3.3. Predictive power and separability of the STRF. Each point in the figure is the result from individual subjects in one condition. STRFs with any substantial predictive power are skewed toward high separability. Circles and squares are respectively the results from monaural and binaural listening conditions. Filled and empty symbols are respectively results from left and right hemispheres. The background contour map shows the joint probability distribution density of predictive power and STRF separability. The probability distribution density is obtained by smoothing the 2-D histogram using a Gaussian function (SD = 0.1 in both directions).

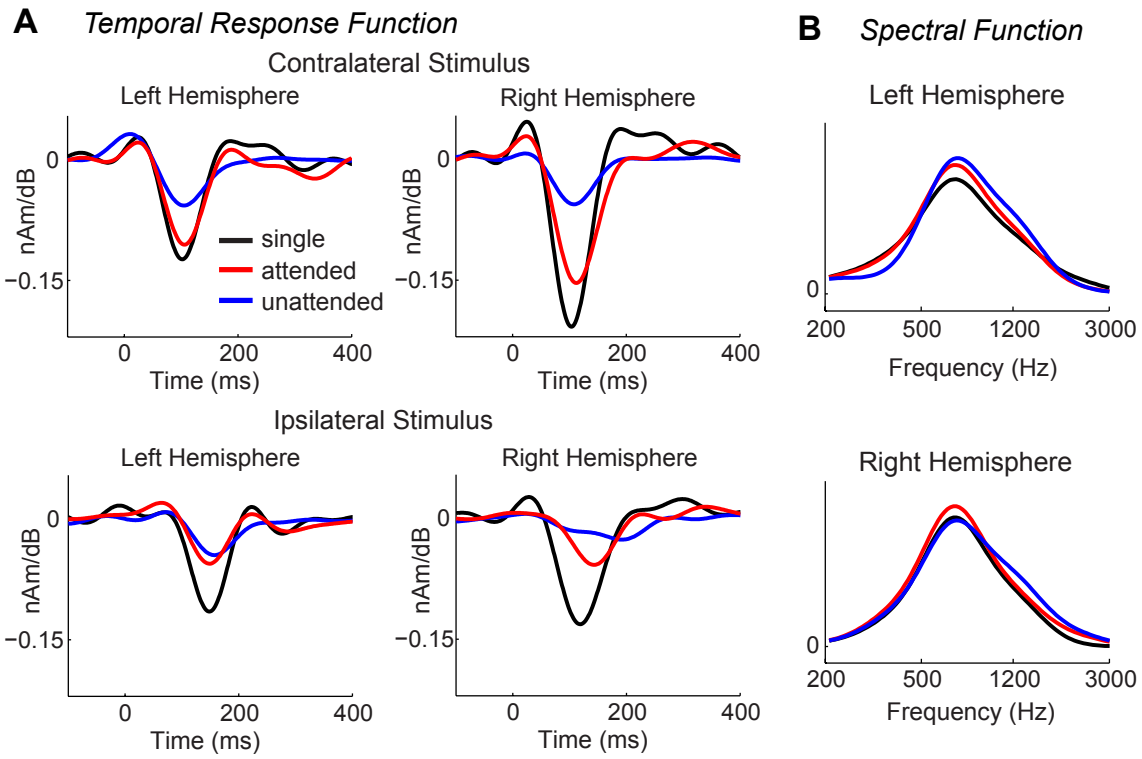

Figure 3.4. Temporal response function and spectral sensitivity functions. (A) Grand average of the temporal response functions to speech stimuli, under three different listening conditions. The amplitude of the temporal response function is higher in the monaural speech condition and is strongly modulated by attention in the dichotic listening condition. (B) The normalized spectral sensitivity function (grand average over subjects) has a peak between 400 and 2000 Hz, in both hemispheres and all listening conditions. Normalized spectral sensitivity functions to contralateral and ipsilateral stimuli are not significantly different and are therefore averaged. The spectral sensitivity function is smoothed using a Gaussian function of 5 semitones standard deviation.

An STRF is called spectro-temporally separable when its temporal and spectral processing are independent of each other (Depireux et al., 2001). The separability of the MEG STRF is very high and is quantitatively illustrated in Fig. 3.3. Furthermore, the STRF separability is positively correlated with the STRF predictive power (Fig. 3.3), indicating that STRFs that predict the MEG response well are generally separable. A separable STRF can be decomposed into the product of a single temporal function (Fig. 3.4A) and a single spectral function (Fig. 3.4B), and therefore the spectral property and temporal property of MEG STRFs are analyzed separately in the following.

The normalized spectral sensitivity function of the STRF shows a broad peak between 400 Hz and 2 kHz (Fig. 3.4B). The spectral sensitivity function significantly changes as a function of frequency (frequency × hemisphere × stimulus side, 3-way repeated-measures ANOVA, $F_{1,359} = 28$, $P < 0.0001$) but is not significantly influenced by stimulus side or by hemisphere.

The $M100_{STRF}$ is well captured in the temporal response function (Fig. 3.4A) and is statistically significant in each hemisphere for each stimulus presentation side (test described in *Methods*, $P < 0.001$). The amplitude and latency of the $M100_{STRF}$ are summarized in Fig. 3.5. The amplitude of this response is larger in the right hemisphere, independent of the stimulus side (hemisphere × stimulus side, 2-way repeated-measures ANOVA, $F_{1,39} = 11.6$, $P < 0.008$), while the latency is shorter for a contralateral stimulus (hemisphere × stimulus side, 2-way repeated-measures ANOVA, $F_{1,39} = 14.6$, $P < 0.005$).

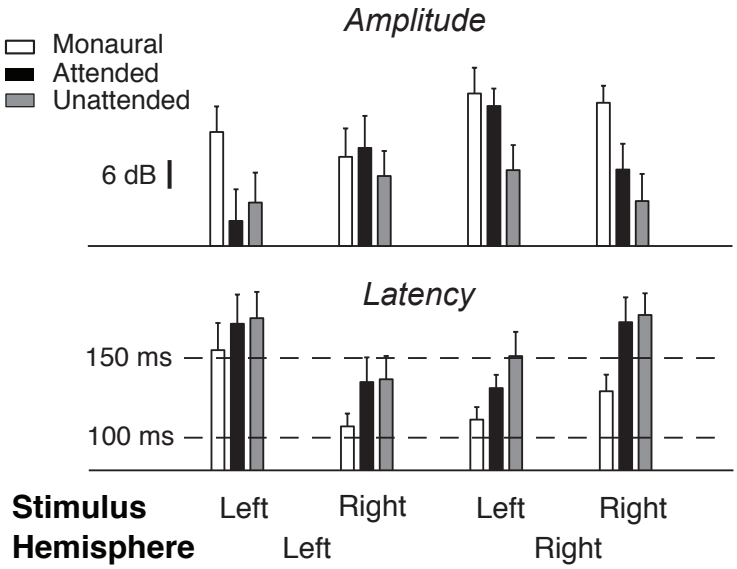

Figure 3.5. Amplitude and latency of the $M100_{STRF}$ (grand average). Error bars represent one standard error. The response amplitude is universally larger and the response latency is universally shorter for monaurally presented speech. In the dichotic condition, the response is stronger for the attended speech than for the unattended speech.

Speech Decoding based on the MEG Response

The STRF analysis above has shown that spectro-temporal modulations of speech are encoded in auditory cortex as a temporal code. The fidelity of this temporal code is further assessed by decoding, i.e. reconstructing, speech features from MEG responses. Since the frequency tuning of STRF is broad, we concentrate on decoding the temporal envelope of speech. In the decoding, we divide the MEG response and corresponding stimulus into multiple segments of equal length and use the decoder (estimated from a non-overlapping data set) to decode the stimulus from each segment. The correlation

between the decoded envelope and real stimulus envelope is shown in Fig. 3.6A, as a grand averaged confusion matrix. This result is based on the right hemisphere's response to a one-minute duration contralateral stimulus, for the case where the stimulus and response are divided into fifty (1.2 s duration) segments. In Fig. 3.6A, the fifty stimulus segments and the fifty envelopes decoded from response segments are indexed sequentially from 1 to 50. If each decoded envelope is attributed to the stimulus whose envelope is most correlated with it, 86 % of the fifty stimulus segments are correctly decoded.

The number and duration of stimulus/response segments have a profound influence on speech decoding performance. Fig. 3.6B shows the speech decoding performance as a function of the number of stimulus segments divided by the duration of each stimulus. Based on Fano's inequality, the speech decoding performance demands that at least 4 bits/s of information in speech is encoded in the right hemisphere MEG response. In the left hemisphere, this value drops to 1 bit/s. This decoding result is based on the confusion matrix averaged over subjects. Analysis of individual subjects confirms that more information is decoded from the right hemisphere than the left (hemisphere × stimulus side 2-way repeated-measures ANOVA, $F_{1,9}$=28.5, $P < 0.0005$) while a similar amount of information is decoded for the left and right side stimuli.

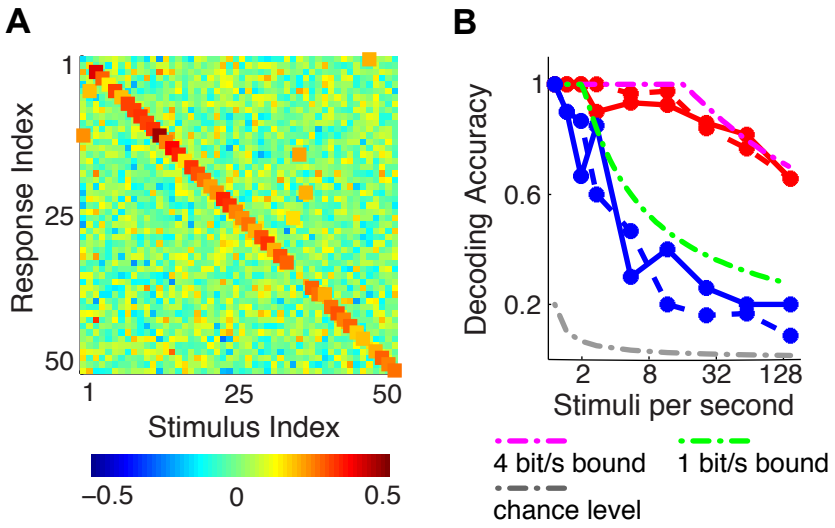

Figure 3.6. Stimulus information in the MEG response. (A) The correlation (color coded) between the stimulus speech envelope and the envelope reconstructed from the right hemisphere MEG response. The stimulus envelope most correlated with each reconstructed envelope is marked by a square. (B) Stimulus decoding accuracy as a function of the number of stimulus segments per second. The blue and the red curves are the results from the left and right hemispheres respectively. Solid and dashed curves are based on the left and right side stimulus. The information decoded from the right (left) hemisphere is roughly 4 bit/s (1 bit/s), and is a conservative estimate of the stimulus information available in the MEG response.

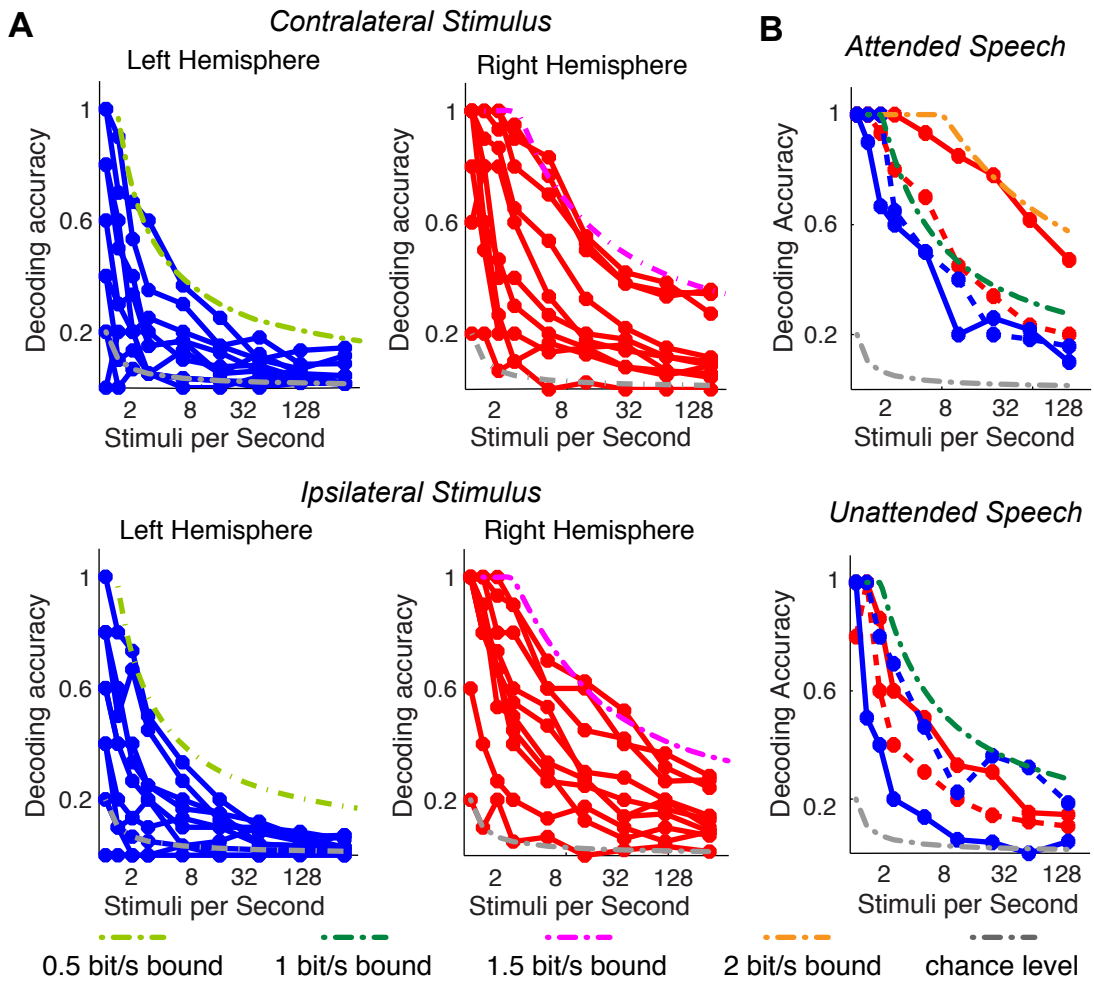

Figure 3.7. Stimulus information in the MEG response. (A) Information decoded from the response to single speech for individual subjects. The decoded information rate is approximately three times higher in the right hemisphere than left, but not significantly influenced by stimulus side. (B) Information decoded from the response to attended speech and unattended speech. The decoding is based on the grand averaged confusion matrix. More information about the stimulus can be decoded when the speech is being attended to than when not.

The decoding result in Fig. 3.6 is based on the confusion matrix averaged over subjects. A significant amount of information can also be decoded from individual subjects (Fig. 3.7A). The information decoded from individual subjects is significantly higher in the right hemisphere than in the left hemisphere (hemisphere × stimulus side two-way repeated-measures ANOVA, $F_{1,9}$=28.5, $P < 0.0005$) but is not significantly influenced by the stimulus side.

Spectro-temporal Representation of Simultaneous Speech Signals

Beyond the monaural listening condition analyzed above, subjects also took part in a dichotic listening experiment. In this condition, on top of the single spoken narrative in one ear, another spoken narrative was presented simultaneously in the opposite ear, resulting in a dichotic listening condition. In each experimental block, the subjects were first instructed to listen to the spoken narrative in one ear, and then, when the stimulus was repeated, to listen to the spoken narrative in the other ear. Therefore, the speech signal in each ear serves both as a target (when attended to) and as an interference signal (when not being attended to). Each experimental block was presented three times. The STRF is determined separately for the stimulus in each ear, under each attentional condition and for each hemisphere.

The STRF shows salient $M100_{STRF}$ for both attended and unattended speech (Fig. 3.2 and Fig. 3.4A), similar to the STRF for monaural speech. The STRFs obtained from this dichotic listening condition remain highly separable (Fig. 3.3). Frequency × hemisphere × attentional state (attended vs. unattended) 3-way repeated-measures

ANOVA shows that the normalized spectral sensitivity function is not influenced by attentional state and is not different between the two hemispheres (Fig. 3.4B).

The $M100_{STRF}$ is statistically significant for both attended and unattended speech (test described in *Methods*, $P < 0.001$). Compared with the $M100_{STRF}$ for monaural stimuli, the $M100_{STRF}$ for dichotic stimuli is weakened (paired t-test. $P << 0.0001$ for both attended response and unattended response) and delayed (paired t-test. $P < 0.002$ for attended response and $P << 0.0001$ for unattended response). A 4-way repeated measures ANOVA (attentional state × hemisphere × stimulus side × experimental block) shows that the latency of this peak in each hemisphere is shorter for the contralateral stimulus ($F_{1,239} = 13.5$, $P < 0.006$).

In the dichotic listening condition, the neural representation of speech remains faithful. The predictive power of the STRF is far above chance level (test described in *Methods*, $P < 0.001$). It is not significantly affected by hemisphere or which ear is attended to individually but is affected by the interaction between the two (2-way repeated-measures ANOVA, $F_{1,39} = 20.0$, $P < 0.002$). The predictive power is higher when attention is paid to the contralateral stimulus (0.17 vs. 0.10), for either hemisphere. A considerable amount of speech information can be decoded from the MEG responses to both the attended and the unattended speech (Fig. 3.7B). The amount of information extracted from individual subjects is analyzed using a 3-way repeated-measures ANOVA (attentional state × hemisphere × stimulus side). More information is decoded when the stimulus is being attended to ($F_{1,79} = 23$, $P < 0.0009$) and in the right hemisphere ($F_{1,79} = 6.5$, $P < 0.03$).

Attentional Modulation during Dichotic Listening

The amplitude of the $M100_{STRF}$ (Fig. 3.5) is substantially modulated by attention. A 4-way repeated-measures ANOVA (with attentional state, hemisphere, stimulus side and experimental block number as factors) reveals that the neural response to attended speech is significantly stronger than the neural response to unattended speech ($F_{1,239}$ = 10.0, $P < 0.02$). There is a significant interaction among attentional state, hemisphere and stimulus side ($F_{1,239} = 9.1$, $P < 0.02$). For the speech stimulus in each ear, the attentional effect is more salient in the contralateral hemisphere (paired t-test, $t_{59} = 3.3$, $P < 0.002$). There is also an interaction between hemisphere and stimulus side ($F_{1,239} = 16.2$, $P < 0.003$). The response to the stimulus on either side is stronger in the contralateral hemisphere. None of the factors interact with experimental block number. Even when only the first experimental block is considered, the attention effect is significant (attentional state × hemisphere × stimulus side, 3-way repeated-measures ANOVA, $F_{1,79}$ = 28.1, $P < 0.0005$, stronger when attended) and the interaction among attentional state, hemisphere and stimulus side is significant ($F_{1,79} = 9.0$, $P < 0.02$, attentional modulations stronger in the contralateral hemisphere).

To investigate the temporal dynamics of the attentional gain effect within a single presentation of the stimulus, each one-minute response was divided into 10 six-second segments and the temporal response function estimated for each segment independently. The attentional gain of the $M100_{STRF}$ was extracted from each temporal response function as the gain difference between attended response and unattended response, in dB. A 3-way repeated-measures ANOVA (hemisphere × stimulus side × segment) on the

attentional gain of the $M100_{STRF}$ reveals no significant interaction between the attention gain and segment number.

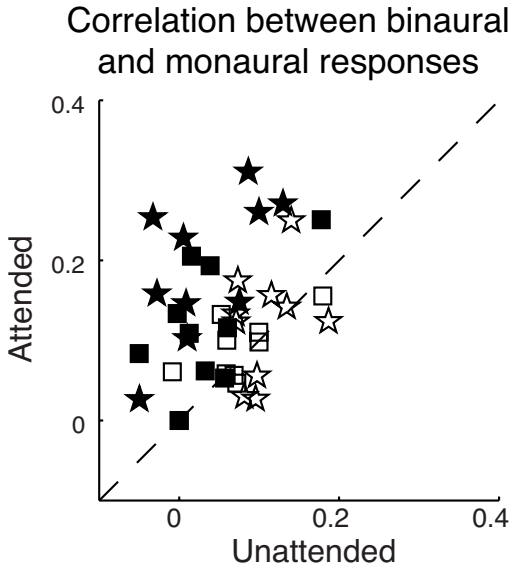

Figure 3.8. Correlation between the MEG response to dichotic speech stimuli and the MEG responses to the two speech components presented monaurally. Each symbol in the figure is the result from one subject. The response in the right (left) hemisphere is plotted as stars (squares). For each hemisphere, if the attended ear in the dichotic condition is the contralateral ear, the result is plotted as a filled symbol but otherwise a hollow symbol. The response to dichotic stimuli is more correlated with the response to the attended speech component, especially in the contralateral hemisphere.

As a result of the attentional gain effect, one might expect the neural response to the speech mixture to be more similar to the neural response to the attended speech than the response to the unattended speech. This hypothesis is confirmed by the analysis of the correlation between the MEG response to the speech mixture and the MEG responses to

individual speech components measured during monaural listening (Fig. 3.8). A 3-way repeated-measures ANOVA, with speech component (attended or unattended), hemisphere, and stimulus side as factors, confirms that the response to the mixture is more correlated with the response to the attended speech component ($F_{1,79} = 36.2$, $P < 0.0002$). The ANOVA analysis also reveals a significant interaction among speech component, hemisphere, and stimulus side ($F_{1,79} = 39.7$, $P < 0.0001$): the response to the mixture is especially dominated by the response to the attended speech in the hemisphere contralateral to the ear the attended speech is presented to.

Neural source localization

In the STRF and decoding analyses, the MEG speech response is decomposed into components using a blind source separation method, DSS (de Cheveigné and Simon, 2008). Only the first DSS component, which has the strongest trial-to-trial reliability, produced any STRF with substantive predictive power (Fig. 3.9). The topography of the spatial magnetic field associated with this first DSS component is quantitatively similar to that of the well-known M100 response. The correlation between them is 96.0% for the grand average magnetic fields (with a 95% confidence interval of 94.6% to 97.0% correlation, estimated by bootstrap sampling). The magnetic field patterns associated with the first DSS component and the M100 are separately modeled by a single ECD in each hemisphere. The correlation between the measured magnetic field and that of the dipole model is 94 % ± 5% and 92% ± 7% (mean ± SD) for the DSS component and the M100 respectively. The ECD locations for the two responses are not distinguishable ($P > 0.1$ in all directions), consistent with their topographical similarity, which implies that both are

centered in association auditory cortex (Lütkenhöner and Steinsträter, 1998; Woldorff et al., 1993).

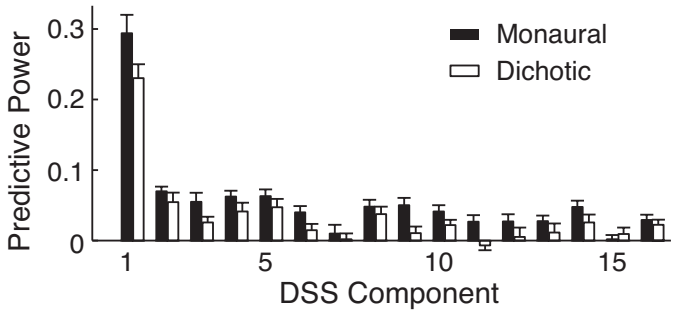

Figure 3.9. Predictive power of the STRF derived from each DSS component. The first DSS component results in significantly higher predictive power than other components and therefore was the only one used to localize the source of the MEG response.

Auditory Steady State Response

The sinusoidal modulation of the speech waveforms near 40 Hz generates a small but observable aSSR. For the monaural speech condition, the aSSR at both modulation rates is statistically significant ($P < 0.05$). In the dichotic listening condition, the attentional modulation of aSSR power is assessed by 2-way repeated-measures ANOVA (attentional state × hemisphere) but no significant effects are seen.

3.4 Discussion

In this study, we have characterized how spectro-temporal features of speech are encoded in spatially synchronized activity in auditory cortex, by quantifying the relationship between ongoing MEG response and continuous speech stimulus. To summarize the major results: (1) the neural activity in auditory cortex precisely encodes the slow temporal modulations of speech (< 8 Hz) in a broad spectral region between 400 and 2000 Hz, which roughly encompasses the first and second formants of speech. (2) The neural coding of slow temporal modulations is stronger and more precise in the right hemisphere, regardless of which ear the speech stimulus is presented to. In the right hemisphere the neural code is faithful enough to discriminate the responses to hundreds of speech stimuli based a few seconds of neural recording. (3) The neural response in each hemisphere is weaker, and has a longer latency, for speech stimulus monaurally presented to the ipsilateral ear, similar to what is observed for the M100 response (Pantev et al., 1986; Rif et al., 1991).

Using a dichotic listening paradigm, we have further demonstrated how competing speech signals are encoded. (1) Auditory cortex precisely tracks the temporal modulations of both incoming speech signals but substantially more strongly for the attended one. (2) The effect of attentional modulation in auditory cortex has latency of only 100 ms, indicating that the segregation of dichotic speech stimuli must still involve feedforward neural processing. (3) The attentional modulation of auditory activity is present even during the subjects' first exposure to a dichotic speech mixture. (4) The attentional gain effect is more salient in the hemisphere contralateral to the attended ear.

(5) The neural response to speech in either ear is weakened (c.f. Fujiki et al. 2002; Penna et al. 2007) and delayed by speech in the other ear.

These results on the spectro-temporal neural encoding of speech provides a clear explanation for stimulus-synchronized neural response observed in previous experiments (Abrams et al., 2008; Ahissar et al., 2001; Aiken and Picton, 2008; Lalor and Foxe, 2010; Luo and Poeppel, 2007; Nourski et al., 2009). The properties and indications of this neural code are discussed in the following.

Attentional Gain Control for Unfamiliar Speech

The attention-modulated sensory gain control shown in this study is largely independent of specific knowledge of the content of the speech, since it is effective even on the first exposure to the speech. As far as we know, this is the first evidence that attentional gain modulation is active with a relative short latency when human listeners strive to comprehend novel speech in the presence of interfering speech. While natural speech has built-in contextual and rhythmic cues, these do not predict the content of the speech by any means. It is known that even without any rhythmic cues the auditory evoked response to an attended stimulus can be enhanced (Hillyard et al., 1973). It is also possible, however, that contextual cues, and especially the rhythm of natural speech, facilitate the neural amplification of speech encoding (Lakatos et al., 2008). Experiments using dichotically presented tone sequences demonstrate that the effect of attention on the M100 (N1) is observed for stimuli with some kinds of rhythm (typically fast) (Ahveninen et al., 2011; Hillyard et al., 1973; Power et al., 2010; Rif et al., 1991; Woldorff et al., 1993), but not others (Hari et al., 1988; Ross et al., 2000). Therefore, it is critical to show

directly whether early auditory response to speech, with its unique complex temporal structure, is modulated by attention.

Equally as important, the attentional gain effect is seen in auditory cortex, directly affecting a neural response component whose latency is only about 100 ms, and which is phase locked to low-level acoustic features of speech. Therefore, the segregation and selective processing of two dichotically presented speech signals almost certainly involve feedforward auditory neural computations. Also, because of the relatively short latency and the neural source location, it is unlikely that this observed speech segregation occurs during or after the semantic processing of speech. It is also worth noting, however, that the early sensory response to the unattended speech is suppressed but not eliminated. This relatively weak auditory response may be further processed, leading to the interaction between dichotically presented speech signals seen behaviorally. Additionally, although the $M100_{STRF}$ is modulated by attention, the aSSR is not. This result is consistent with previous observations that 40-Hz aSSR is not, or only very weakly, modulated by attention or even awareness of sounds (Gutschalk et al., 2008; Lazzouni et al., 2010; Linden et al., 1987). Compared with the $M100_{STRF}$, the aSSR has a shorter latency at about 50 ms (Ross et al., 2000). Moreover, the neural source location of the aSSR is commonly believed to be in core auditory cortex (Herdman et al., 2003), while the neural source location of the $M100_{STRF}$ is centered in association auditory cortex (Lütkenhöner and Steinsträter, 1998). Therefore, although feedforward processing is clearly involved in dichotic speech segregation, it may not occur at the level of core auditory cortex. It is also possible, however, that the lack of statistically significant attentional effects on the aSSR

is due to the weakness of the aSSR; it is known that aSSR is attenuated by slow temporal modulations, such as those present in speech (Ding and Simon, 2009).

Although dichotic speech segregation is reflected in the feedforward early auditory response seen here, it is certainly under the modulation of higher order cortical networks. Further experiments are still necessary to identify the network controlling the attentional gain effects seen in auditory cortex, which may include areas in the frontal and parietal cortex (Hill and Miller, 2010; Shomstein and Yantis, 2006). The attention-control signals by no means need to be phase-locked to acoustic features of the speech stimulus and therefore cannot be extracted using the STRF analysis employed here.

In addition, since the current experiment uses the same speaker and same narrative source for both ears, rather than tones of different frequencies, we have demonstrated that this attentional sensory gain control can be driven entirely by the stimulus ear, not needing, e.g., spectral cues. Of course other monaural cues, such as pitch and rhythm, and binaural cues, such as interaural time difference (ITD) and interaural level difference (ILD), can also be utilized to segregate concurrent sounds (Bronkhorst, 2000). Previous experiments with simple non-speech stimuli have demonstrated that monaural cue based segregation of spectrally non-overlapping sounds is reflected neurally in human auditory cortex (Bidet-Caulet et al., 2007; Elhilali et al., 2009b; Xiang et al., 2010). Future experiments are needed to address whether speech segregation itself, which is a much harder problem, also occurs in human auditory cortex at a short latency.

Hemispheric Lateralization of Speech Coding in Auditory Cortex

Although the neural tracking of spectro-temporal modulations in speech is seen bilaterally, it is strongly lateralized to the right hemisphere, independent of the stimulus ear. This lateralization is demonstrated by the amplitude of the $M100_{STRF}$ (Fig. 3.5) and more critically by the fidelity of neural coding (Fig. 3.6B). This strong right hemisphere dominance effect is surprising, however, since it is not observed in the M100 response to sound onsets or to aSSR to 40-Hz amplitude modulations (Rif et al., 1991; Ross et al., 2005), both of which are instead stronger in the hemisphere contralateral to the ear receiving the stimulus or equally strong in both hemispheres. Furthermore, even for responses to speech, if both the response tracking speech features and other responses are considered, the total response is stronger in the left rather than right hemisphere (Millman et al., 2011). Nor can the rightward lateralization of the neural representation of speech be explained anatomically, since the dominant excitatory input to each hemisphere is from the contralateral ear (Pickles, 1988). Therefore, this result gives further support to the hypothesis that the right hemisphere is intrinsically dominant in processing the slow modulations (<10 Hz) in speech during natural speech comprehension (Poeppel, 2003). This right hemisphere dominance has also been observed in the neural response to speech (Abrams et al., 2008; Kerlin et al., 2010; Luo and Poeppel, 2007), and even in endogenous neural oscillations (Giraud et al., 2007).

On top of this intrinsic right hemisphere dominance, however, during dichotic listening the effect of attentional gain control is even more prominent in the hemisphere contralateral to the attended side. This hemispheric lateralization effect likely arises from the anatomical asymmetry between the left and right afferent pathways to each

hemisphere. When two different sounds are presented to the two ears separately, their neural representations form a competition (elaborated below in *Binaural Interaction*). One result of this competition may be that each hemisphere primarily processes information from the contralateral ear, where most of the excitatory afferent inputs are from (Pickles, 1988). Therefore, the neural processing of each stimulus can be most strongly modulated by the attentional gain change in the contralateral hemisphere.

Neural Coding of Spectro-Temporal Dynamics of Speech Signals

Using STRF analysis, we have demonstrated that slow temporal modulations of speech (particularly of coarse spectral modulations) are precisely encoded in human auditory cortex. Taking advantage of the fine time resolution of MEG, we show that the observed neural responses encode at least 4 bit/s information. This indicates that, using a linear decoder, we can errorlessly discriminate about 16 speech stimuli (4 bits) of one-second duration based on their MEG responses. Similarly, this same information rate allows one to errorlessly discriminate about 256 speech stimuli (8 bits) of two-second duration. The possibility of discriminating MEG/EEG responses to speech has been suggested by earlier studies but only shown based on a small number of several-second duration sentences (Kerlin et al., 2010; Luo and Poeppel, 2007). The MEG response is also robust: an $M100_{STRF}$ is observed even for unattended speech. This contrasts with the observation that the neural representation of sounds in anesthetized avian auditory forebrain is severely degraded by acoustic interference (Narayan et al., 2007) and therefore suggests that the robust neural coding may require top-down attentional modulation.

In speech, temporal modulations below 10 Hz convey syllabic and phrasal level information (Greenberg, 1999). In quiet, these slow modulations, in concert with even a very coarse spectral modulation, accomplish high speech intelligibility (Shannon et al., 1995). When speech is masked by acoustic interference, slow temporal modulations of the interference releases the masking of the target speech (Festen and Plomp, 1990). Faster acoustic fluctuations of speech, e.g. spectral and pitch cues, that contain phonetic and prosodic information, are gated by the slow modulations (Rosen, 1992). Similarly, the neural processing of speech features on short time scales (< 100 ms) may also be modulated by the low frequency neural activity analyzed in this study. The phonetic information of speech has been suggested to be spatially coded over neural populations in auditory cortex (Chang et al., 2010). This spatial code discriminates different syllables most effectively at around 100 ms after the syllable onset, consistent with the latency of the $M100_{STRF}$. Other possible neural signatures of higher level processing of speech are high frequency neural oscillations (40 - 150 Hz), which are also coupled to slow neural oscillations below 10 Hz (Lakatos et al., 2008). Therefore, the slow activity noninvasively measured by MEG probably reflects the timing of such microscopic neural computations of the phonetic level information of speech.

The STRF of the MEG Speech Response

The mathematical linear system bridging the speech stimulus and the neural representation of that speech can be represented graphically by the STRF. The predictive power of the MEG STRF compares well to that obtained from single cortical neurons for speech stimuli (Biermann and Heil, 2000; David et al., 2009; David et al., 2007). The

MEG STRF is highly separable: the temporal processing of the speech stimulus is consistent over the entire frequency range of the STRF and is equally sensitive to upward and downward changes in frequency content. This contrasts with the variety of separability seen in the STRFs of single neurons in primary auditory cortex (Depireux et al., 2001) and the inseparability seen using fMRI (Schönwiesner and Zatorre, 2009). This difference in separability reflects differences between the spectro-temporal tuning of individual neurons, spatially-synchronized activity and non-spatially-synchronized activity. MEG and fMRI recordings reflect the activity of large neural populations. Additionally MEG records only spatially synchronized components of the response (and in this study stimulus-synchronized neural activity), while fMRI measures the indirect hemodynamic response, which is influenced by both synchronized and asynchronous neural activity. Hence, MEG and fMRI demonstrate very different aspects of the population level distribution of the spectro-temporal tuning properties of neurons and are therefore naturally complementary.

In summary, in this study we demonstrate the existence of a neural encoding of speech in human auditory cortex that can be measured extracranially and non-invasively. We have also demonstrated that this neural encoding is based on the acoustic modulations of the spectro-temporal features of speech. The encoding is quite faithful (perhaps even surprisingly so given that the neural signal is measured extracranially), and able to distinguish among hundreds of different stimuli in the course of only a few seconds. Additionally, on the one hand, the encoding strategy is very strongly tied to the physical properties of speech, which would normally imply a bottom-up encoding process. But on the other, the encoding strategy is also strongly modulated by the

attentional state of the listener, demonstrating that top-down processes directly modulate the neural representation of the fundamental acoustic features of speech. Finally, we have also developed a practical experimental paradigm that allows single-trial analysis of the auditory cortical encoding of continuous speech in an ecologically realistic manner.

Chapter 4

Cortical representation of simultaneous speakers

4.1 Introduction

In a complex auditory scene, humans and other animal species can perceptually detect and recognize individual auditory objects or auditory streams, i.e. the sound arising from a single source, even if strongly overlapping acoustically with sounds from other sources. To accomplish this remarkably difficult task, it has been hypothesized that the auditory system first decomposes the complex auditory scene into separate acoustic features, and then binds the features, as appropriate, into auditory objects (Bregman, 1990; Griffiths and Warren, 2002; Shamma et al., 2011; Shinn-Cunningham, 2008). The neural representations of auditory objects, each the collective representation of all the features belonging to the same auditory object, have been hypothesized to emerge in auditory cortex to become fundamental units for high-level cognitive processing (Fishman and Steinschneider, 2010; Nelken, 2008; Snyder et al., 2012). The process of parsing an auditory scene into auditory objects is computationally complex and cannot as yet be emulated by computer algorithms (Wang and Brown, 2006), but it occurs reliably, and often effortlessly, in the human auditory system. For example, in the classic "cocktail party problem", where multiple speakers are talking at the same time (Cherry, 1953), human listeners can selectively attend to a chosen target speaker, even if the competing speakers are acoustically more salient, e.g. louder, or perceptually very similar, e.g. of the same gender (Brungart, 2001).

To demonstrate an object-based neural representation that could subserve the robust perception of an auditory object, several key pieces of evidence are needed. The first is to demonstrate neural activity that exclusively represents a single auditory object (Griffiths and Warren, 2004; Nelken and Bar-Yosef, 2008). In particular, such an object-specific representation must be demonstrated in a range of auditory scenes with reliable perception of that auditory object, and especially in challenging scenarios where the auditory object cannot be easily segregated by any basic acoustic features, such as frequency or binaural cues. For this reason we investigate the existence of object-specific auditory representations by using an auditory scene consisting of a pair of concurrent speech streams mixed into a single acoustic channel. In this scenario, the two speech streams each form a distinct perceptual auditory object, but they overlap strongly in time and frequency, and are not separable using spatial cues. Therefore any neural representation of an auditory object, i.e. in this case, a single stream of speech, would not emerge without complex segregation and grouping processes.

Secondly, the neural processing of an auditory object must also be adaptive and independent (Griffiths and Warren, 2004; Shinn-Cunningham, 2008). In particular, the neural processing of each auditory object should be modulated based on its own behavioral importance and acoustic properties, without being influenced by the properties of other auditory objects or the stimulus as a whole. Building on the well-established phenomena of *feature-based* top-down attentional modulation (Elhilali et al., 2009a; Fritz et al., 2003; Hillyard et al., 1973; Xiang et al., 2010) and feature-based bottom-up neural adaptation to sound intensity (Robinson and McAlpine, 2009), here, we investigate whether such top-down and bottom-up modulations occur separately for individual

auditory objects, i.e., in an *object-based* manner. Specifically, using this speech segregation paradigm, we ask the listeners to attend to one of the two speakers while manipulating separately the intensity of the attended and background speaker. If an observed neural representation is object-based: not only must it be enhanced by top-down attention, but it must also adapt to the intensity change of that speech stream alone, without being affected by the intensity change of the other stream or of the mixture as a whole.

In this study, we investigate whether a robust neural representation of an auditory object can be observed in the brain, and when and where it might emerge. In the experiment, the subjects selectively listened to one of two concurrent spoken narratives mixed into a single acoustic channel, answering comprehension questions about the attended spoken narrative after each one-minute block. The neural recordings were obtained using magnetoencephalography (MEG), which is well suited to measure spatially-coherent neural activity synchronized to speech rhythms, i.e. the slow temporal modulations that define the speech envelope (Abrams et al., 2008; Ahissar et al., 2001; Koskinen et al., 2012; Luo and Poeppel, 2007). Such spatially-coherent phase-locked activity is strongly modulated by attention (Ding and Simon, 2012; Kerlin et al., 2010; Schroeder and Lakatos, 2009) and has been hypothesized to play a critical role in grouping acoustic features into auditory objects (Shamma et al., 2011).

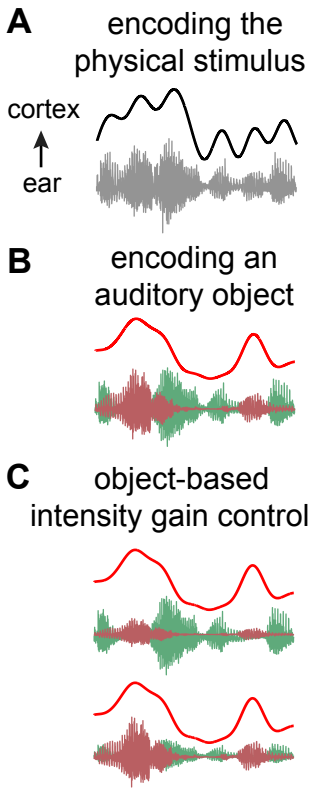

Figure 4.1. Illustration of object-based neural representations. Here the auditory scene is illustrated using a mixture of two concurrent speech streams. (A) If a complex auditory scene is not neurally parsed into separate auditory objects, cortical activity (upper curve) phase locks to the temporal envelope of the physical stimulus, i.e. the acoustic mixture (lower waveform). (B) In contrast, using the identical stimulus (but illustrated here with the unmixed instances of speech in different colors), for a hypothetical neural representation of an individual auditory object, neural activity would instead selectively phase lock to the temporal envelope only of that auditory object. (C) The neural representation of an auditory object should, furthermore, neurally adapt to an intensity change of its own object (left) but should remain insensitive to intensity changes in another auditory object (right).

Specifically, we hypothesize that, in cortical areas with an object-based representation, neural activity should phase lock to the rhythm of a single auditory object, while in cortical areas where object-based representations are not yet formed, or formed only weakly, the neural response should phase lock to the envelope of the entire physical stimulus, i.e. the speech mixture (both examples are illustrated in Fig. 4.1A & B). In other words, whether a neural response is encoding one speech stream, the other speech stream, or the mixture, can be easily distinguished by which sound's rhythm it is synchronized to. Critically, bottom-up neural adaptation to sound intensity is also investigated. Neural adaptation also determines whether a neural representation is object-based based or not, based upon which sound stream (or mixture) the neural representation adapts to. We do this by analyzing the phase-locked neural activity when the intensity of the attended and the background speakers are manipulated separately (Fig. 4.1C). These hypothesized, object-specific, neural representations are investigated, and revealed, using single-trial neural recordings, and a novel neural decoding method that parallels state-of-the-art analysis methods used in fMRI (Kay et al., 2008) and intracranial recording (Pasley et al., 2012).

4.2 Methods

4.2.1 Subject, stimuli and Procedures

Subjects

Twenty normal hearing, right-handed, young adult native speakers of American English (between 18 and 26 years old) participated in the experiment in total. Eleven (5 female) participated in the Equal-Loudness experiment, six (3 female) participated in the Varying-Loudness experiment, and three (2 female) participated in the Same-Gender experiment. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland institutional review board. Written informed consent form was obtained from each subject before the experiment.

Stimuli and Procedures

The stimuli contain three segments from the book *A Child's History of England* by Charles Dickens (http://librivox.org/a-childs-history-of-england-by-charles-dickens/), narrated by three different readers (2 female). All speaker pauses (periods of silence longer than 300 ms) were shortened to 300 ms, and then each chapter was divided into one-minute duration sections. The speech mixtures were constructed by mixing the two chapters digitally in a single channel. All stimuli were low-pass filtered below 4 kHz and delivered identically to both ears using tube phones plugged into the ear canals. The subjects were required to close their eyes when listening. Before each main experiment, 100 repetitions of a 500-Hz tone pip were presented to each subject to elicit the M100 response, which is a reliable auditory response measured 100 ms after the onset of a tone pip and whose neural source is easy to localize within auditory cortex (Lütkenhöner and Steinsträter, 1998). The three main experiments were conducted as follows.

Equal-Loudness Experiment: In this and the Varying-Loudness experiment, two speakers of opposite gender were mixed. The average pitch of the two speakers was

separated by 5.5 semitones (de Cheveigné and Kawahara, 2002). The first 2 sections from each chapter were mixed with equal root mean square values (RMS) for sound amplitude. The subjects were instructed to focus on one speaker until the mix was finished, and then to switch focus to the other speaker while the same mix was played again. The same process was repeated 3 times, resulting in 3 trials with identical stimulus and attentional focus. Each trial contains 2 one-minute duration sections. To help the listeners attend to the correct speaker, the first second of each section was replaced by the clean recording from the target speaker. The speaker attended to first was counterbalanced across subjects. After each section, the subjects were asked to answer a question related to the comprehension of the passage they had just attended to. On average, 69% of the questions were correctly answered (not depending on the number of trials, $P > 0.8$, $F(2,32) = 0.2$, 1-way repeated measures ANOVA). After this part of the experiment, the unmixed stimuli (each speaker alone) were presented to the listeners, 4 times. Comprehension questions were interspersed occasionally to ensure the subjects were awake during the whole experiment.

Varying-Loudness experiment: In this experiment, the intensity of Speaker Two was fixed at roughly 75 dB SPL and Speaker One was presented at either the same intensity, as evaluated by RMS value, or at intensity 5 dB or 8 dB lower. Therefore, the attended speaker had constant intensity while the background speaker was reduced, when Speaker One was attended. In contrast, the background speaker was kept constant while the attended speaker was damped, when Speaker Two was attended. The TMR varied overall from -8 dB to 8 dB. Each TMR condition contained 2 one-minute duration sections, after each of which a question was asked. The experiment was divided into 4

blocks. In each block, the listener focused on one speaker (balanced over subjects), and switched focus after every block. Each block started with 2 sections of clean speech from the target speaker and was followed by sections of speech mixtures with decreasing TMR. The story continued naturally throughout each block. Such an experimental design produces two trials, from alternative blocks, for each stimulus for each attentional condition. Five out of the six subjects in this experiment were asked to subjectively rate what percentage of words was correctly recognized after the first listening to each stimulus.

Same-Gender experiment: The two chapters read by female speakers were mixed digitally with equal intensity and then divided into 6 thirty-second duration sections. The average pitch of the two speakers differs by 3.2 semitones. The subjects were instructed to focus on one speaker throughout the 6 sections and then switch attention to the other speaker when all the sections were played again. This whole process was repeated again, resulting in two trials for each attentional state. To help the subjects to identify which speaker to listen to, the first 5 seconds of each section were replaced by clean speech from the target speaker. The neural recording during the first 5 seconds was therefore not included in any analysis. A comprehension question was asked after each session. Additionally, each listener went through two initial training sessions before attending to each speaker. In the first session, the non-attended speaker was turned on gradually, using a sigmoidal ramp that saturated after 20 seconds. The second session used stimuli having the same intensity as the stimuli used in the experiment. The training sessions were repeated upon the subjects' request, to make sure the subjects were able to identify and focus on the target speaker after the last training session.

4.2.2 Data recording and analysis

Data Recording and Processing

The neuromagnetic signal was recorded using a 157-channel whole-head MEG system (KIT, Kanazawa, Japan), in a magnetically shielded room, with 1 kHz sampling rate. A 200-Hz lowpass filter and a notch filter at 60 Hz were applied online. Three reference magnetic sensors and three vibrational sensors were used to measure the environmental magnetic field and vibrations, and were employed to denoise the MEG signals (de Cheveigné and Simon, 2007). Five electromagnetic coils were used to measure each subject's head position inside the MEG machine. The ongoing neural response (excluding the first second) during each 1-min duration stimulus was filtered between 1 and 8 Hz (Ding and Simon, 2012) and then used for the decoding and STRF analysis.

Speech Decoding

A linear model was employed to decode the temporal envelope of each speaker in the stimulus by linearly integrating the spatial-temporal brain activity. The decoder was optimized using generalized eigen-decomposition so that the decoded envelope was maximally correlated with the speaker to decode and minimally correlated with the other speaker (mathematically formulated in the following paragraphs). All correlations in this study were measured by the absolute value of the Pearson's correlation coefficient. The decoder optimized this way was a discriminative model that reconstructed an envelope similar to one speaker but distinct from the other and was therefore employed to explore the neural code unique to each speaker. A 2-fold cross validation was employed to

evaluate the performance of decoders: half of the data in each experimental condition were used to train the decoder and the other half were used to calculate the correlations between the decoder output and the stimulus envelopes. The decoder was applied to individual trials, and the percent of trials where decoding was successful (decoded envelope being more correlated with the intended speaker) is always reported as the grand average. This decoding approach effectively explores both the spatial and temporal information in MEG and avoids the sometimes ill-posed problem of estimating the neural source locations.

Mathematically, the decoding operation can be formulated as $ENV(t) = \Sigma_k \Sigma_\tau M_k(t+\tau) D_k(\tau)$, where $ENV(t)$ is the decoded envelope, $M_k(t)$ is the MEG measurement from a single sensor k, and $D_k(t)$ is the linear decoder for the same sensor k. In the following, we first discuss the case of a single MEG sensor and therefore drop the index k. In matrix form, the decoding is expressed as $\boldsymbol{v} = \boldsymbol{Md}$, where $\boldsymbol{v} = [ENV(0), ENV(\Delta t), ..., ENV(T_{MAX})]^T$, $\boldsymbol{d} = [D(0), D(\Delta t), ..., D(T_D)]^T$, and the matrix $\boldsymbol{M}$ is $[M(0), M(0 + \Delta t), ... , M(0 + T_D); M(\Delta t), M(\Delta t + \Delta t), ... , M(\Delta t + T_D); ... ; M(T_{MAX}), M(T_{MAX} + \Delta t), ... , M(T_{MAX} + T_D)]$. T_D, the maximal time delay considered by the decoder, is selected to be 500 ms.

Suppose the envelopes of the speech streams of the two speakers are $\boldsymbol{s}_1 = [s_1(0), s_1(\Delta t), ..., s_1(T_{MAX})]$ and $\boldsymbol{s}_2 = [s_2(0), s_2(\Delta t), ..., s_2(T_{MAX})]$ and they are normalized to have the same L_2 norm, i.e. $||\boldsymbol{s}_1|| = ||\boldsymbol{s}_2||$. The envelope was extracted by summing, over frequency, the spectro-temporal representation of the speech (Yang et al., 1992) with its amplitude expressed in logarithmic scale. The correlation between the decoded envelope and the envelopes of the two speech streams are $c_1 = \alpha^{-1} \boldsymbol{s}_1^T \boldsymbol{v} = \alpha^{-1} \boldsymbol{s}_1 \boldsymbol{Md}$ and $c_2 = \alpha^{-1} \boldsymbol{s}_2 \boldsymbol{Md}$

respectively, where $\alpha^{-1} = \|\boldsymbol{s}_1\|.\|\boldsymbol{v}\| = \|\boldsymbol{s}_2\|.\|\boldsymbol{v}\|$. Let us denote $\boldsymbol{r}_1 = \boldsymbol{s}_1\boldsymbol{M}$, and $\boldsymbol{r}_2 = \boldsymbol{s}_2\boldsymbol{M}$, then $(c_1/c_2)^2 = (\boldsymbol{d}\boldsymbol{r}_1\boldsymbol{r}_1^{\mathrm{T}}\boldsymbol{d}^{\mathrm{T}})/(\boldsymbol{d}\boldsymbol{r}_2\boldsymbol{r}_2^{\mathrm{T}}\boldsymbol{d}^{\mathrm{T}})$. Denote $\boldsymbol{R}_1=\boldsymbol{r}_1\boldsymbol{r}_1^{\mathrm{T}}$ and $\boldsymbol{R}_2=\boldsymbol{r}_2\boldsymbol{r}_2^{\mathrm{T}}$. Then it is known that the quantity $(c_1/c_2)^2$ is maximized when $\boldsymbol{d}$ is the generalized eigen-vector of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ with the largest eigen-value (Fukunaga, 1972).

The conclusion from this single MEG sensor case is easily generalized to the case for multiple MEG sensors by concatenating the recording from all the MEG sensors. For example, in the case of 100 MEG sensors, the first row of $\boldsymbol{M}$ becomes [$M_1(0)$, $M_1(0 + \Delta t)$, ... , $M_1(0 + T_{\mathrm{D}})$, $M_2(0)$, $M_2(0 + \Delta t)$, ... , $M_2(0 + T_{\mathrm{D}})$, ..., $M_{100}(0)$, $M_{100}(0 + \Delta t)$, ... , $M_{100}(0 + T_{\mathrm{D}})$] after concatenation. In this study, to reduce the computational complexity, the 157 MEG sensors were compressed into 3 virtual sensors using DSS in each hemisphere (de Cheveigné and Simon, 2008). Therefore, first the 6 virtual sensors were concatenated, then the two covariance matrices, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, were calculated, and finally the decoder was obtained by generalized eigen-decomposition.

The chance level performance of the decoders was simulated by independently shuffling the order of each one-minute long duration stimulus (independently between the two simultaneous speakers) and the order of all the responses 4096 times. At this chance level, obtained by reconstructing the stimulus based on unmatched responses, the reconstructed envelope is similarly correlated with the speech envelopes of both speakers ($P > 0.8$, paired t-test), and the 95th percentile of the correlation with each speech stream envelope is below 0.01, showing that the decoder does not show bias toward either speaker.

In the Varying-Loudness experiment, the same decoder was employed to decode the stimulus at every TMR. The stimulus and response in every TMR condition were

divided into a training and a testing set. All training sets are then pooled together to train the decoder. After training, the decoder was applied to individual TMR conditions to assess the neural encoding accuracy. Therefore, if the decoding results were consistent over TMR conditions, it would imply that not only is the encoding accuracy unaffected by the intensity change of a speaker but also the underlying spatial-temporal neural code. In Fig. 4.4B, the decoding accuracy for each speech stream is normalized separately. Specifically, the decoding accuracy for one speech stream, the first or the second, is divided by the decoding accuracy of that speech stream when presented individually, c_{s1} or c_{s2}, and then multiplied by the mean accuracy of decoding a speech stream presented individually, i.e. $(c_{s1} + c_{s2})/2$.

STRF

The STRF models how speech features are encoded in the MEG response (Ding and Simon, 2012), in contrast to how the decoders transform MEG activity (backwards) to speech features. A single STRF transforms the spectro-temporal features of speech to a single response waveform. Therefore, due to the multi-channel nature of MEG, a complete forward model is described as a 3-D spatial-STRF model (MEG sensor position × frequency × time). The MEG data was averaged over trials in the STRF analysis, for each stimulus and attentional condition.

The mathematical formulation of the STRF analysis is as follows. The spectro-temporal representations of the speech streams of the two speakers are denoted as $S_A(f, t)$ and $S_B(f, t)$ respectively, and the MEG response is denoted as $r(t,k)$, where k is an index for MEG sensors. The linear STRF model can then be formulated as

$$r(t,k) = \Sigma_f \Sigma_\tau \, STRF_A(f,\tau,k) S_A(f, t-t) + \Sigma_f \Sigma_\tau \, STRF_B(f,\tau,k) S_B(f, t-t) + \varepsilon(t,k),$$

where $STRF_A(f,t,k)$ and $STRF_B(f,t,k)$ are STRFs for the attended and background speech respectively for every MEG sensor, and $\varepsilon(t,k)$ is the residual response waveform not explained by the STRF model. The spectro-temporal representations of the speech of the two speakers were calculated from unmixed speech using an auditory filterbank model (Yang et al., 1992). The amplitude of the stimulus, $S_A(f, t)$ and $S_B(f, t)$, is represented in logarithmic amplitude scale (i.e. in dB) and the mean amplitude is normalized to 0 dB. The mean amplitude of the envelope of each speech stream is normalized since, in a linear model such as the STRF, the mean of the stimulus is represented by the DC component of the neural response, which is not reliably measurable by MEG. The same model is used in all the three experiments, regardless of the actual intensity of either speech stream. Therefore, the amplitude of the STRFs should co-vary with the intensity of either speech stream, unless such the change of stimulus intensity is compensated by the auditory system in an object-based manner.

The STRF model was applied separately to individual sensors. For the sake of computational efficiency, however, the 157-channel MEG dataset was dimensionally reduced to 30 channels when estimating the STRF, using denoising source separation (DSS) (de Cheveigné and Simon, 2008), but then transformed back to the MEG sensor space.

The temporal profile of an STRF is extracted using singular value decomposition (SVD). For the STRF from a MEG sensor or a neural source location, the SVD of STRF is $STRF(f,t) = \Sigma_p \, \lambda_p TRF_p(t) SRF_p(f)$. The temporal profile of the STRF, or the temporal response function, is defined as $\lambda_1 TRF_1(t)$ (Ding and Simon, 2012).

Extraction of the $M50_{STRF}$ and $M100_{STRF}$ magnetic fields

The $M50_{STRF}$ and $M100_{STRF}$ were extracted from two time intervals: 10 - 100 ms and 50 - 200 ms, respectively. The approximate latency of each response peak was determined based on the temporal response function extracted from the spatial-STRF using singular value decompositions (SVD) (Ding and Simon, 2012). The $M100_{STRF}$, also known as the $M100_{STRF}$ response, is known to have the same polarity as the M100 response evoked by a tone pip(Ding and Simon, 2012), while the $M50_{STRF}$ has the opposite polarity. Therefore the $M100_{STRF}$ was determined by the strongest response peak with a magnetic field topology positively correlated with that of the M100, and similarly for the $M50_{STRF}$ but with a negative correlation. The magnetic field pattern extracted for each peak was averaged over speakers and attentional conditions, and then used for neural source localization.

Source Space Analysis

In the neural source analysis, subjects from the Equal-Loudness experiment and the Varying-Loudness experiment were pooled together, and the responses at different TMRs were also averaged. The neural source of each peak in the STRF was modeled by an equivalent-current dipole (ECD) in each hemisphere. A spherical head model was derived for each subject using MEG Laboratory software program v.2.001M (Yokogawa Electric, Eagle Technology, Kanazawa Institute of Technology). The position of the ECD was estimated using a global optimization approach (Uutela et al., 1998). The grand averaged correlation between the fitted ECD magnetic field and the measured magnetic

field is above 95% in both hemispheres and for both $M50_{STRF}$ and $M100_{STRF}$. When comparing the ECD positions of different peaks in STRF, we included only ECDs successfully capturing the measured magnetic field, characterized by a higher than 85% correlation between the ECD magnetic field and the measured magnetic field. No more than 2 out of the 17 subjects were excluded this way, for each STRF peak. After the ECD positions were determined, the moment of the dipole was estimated using the generalized least squares method (Mosher et al., 2003). In the dipole analysis, the sign of the magnetic field of the $M50_{STRF}$ is flipped, in order to make the polarity of its moment consistent with that for the $M100_{STRF}$, and the polarity of the $M100_{STRF}$ is defined as negative, to be consistent with the polarity of the N1 peak of EEG.

In the analysis of the amplitude and latency of the $M50_{STRF}$ and the $M100_{STRF}$, the STRFs are projected to the lead field of the dipole in each hemisphere. Mathematically, if the STRF is $STRF(f,t,k)$ and the lead field is $L(k)$, the projection is $\Sigma_k STRF(f,t,k)L(k)$.

Models of Gain Control

The intensity of the stimulus or an auditory object is normalized in the auditory system by response gain control. The neural phenomena associated with different gain control models are simulated. In the simulation, the envelope of speech is assumed to be faithfully encoded in auditory cortex, and the imperfect decoding of speech envelope is assumed to be due to (stimulus-irrelevant) neural background activity. Therefore, the MEG measurement is modeled as a linear mixture of neural activity phase-locked to each speech stream and stimulus-irrelevant spontaneous activity. To simplify the simulation, but without loss of generality, we further assume that the neural encoding of each stream

instantaneously follows that speech stream, e.g. $r_{A0}(t) = s_A(t)$ and $r_{B0}(t) = s_B(t)$, where $r_{A0}(t)$ and $r_{B0}(t)$ are the raw neural response to the attended and background speech respectively and $s_A(t)$ and $s_B(t)$ are the corresponding speech envelopes.

We model the intensity gain control of neural activity using two models. One model normalizes the MEG activity by the strength (measured by the root mean square (RMS)) of the envelope of the acoustic mixture, i.e. $s_{mix}(t)$; the second model by the strength of each speaker individually. The two models are described as follows.

Global gain control model: $r_A(t) = r_{A0}(t)/\mathrm{RMS}(s_{mix}(t))$

$r_B(t) = r_{B0}(t)/\mathrm{RMS}(s_{mix}(t))$

Object-based gain control model: $r_A(t) = r_{A0}(t)/\mathrm{RMS}(s_A(t))$

$r_B(t) = r_{B0}(t)/\mathrm{RMS}(s_B(t))$

The neural reconstruction of the attended speech, a linear combination of MEG activity, is modeled as $\hat{s}_A(t) = r_A(t) + \lambda_B r_B(t) + \lambda_N n(t)$. In the reconstruction, $r_B(t)$ and $n(t)$ are attenuated but not eliminated due to, e.g., the limited spatial resolution of MEG. The two free parameters λ_B and λ_N are fit based on the Equal-Loudness experiment, i.e. when $s_A(t)$ and $s_B(t)$ have equal intensity: λ_B and λ_N are adjusted so that the simulated decoding results, i.e. the correlation between $r_A(t)$ and $s_A(t)$ and the correlation between $r_A(t)$ and $s_B(t)$, match the experimental observations in the Equal-Loudness experiment (Fig. 4.2B). The model is then used to predict the decoding results in the Varying-Loudness experiment, where the intensity of two speakers are changed separately. The model predictions are generally insensitive to the values of λ_B and λ_N.

4.3 Results

Deciphering the Spatial-Temporal Code for Individual Speakers

In the first experiment, listeners selectively listened to one of two competing speakers of different gender, mixed into a single acoustic channel with equal intensity. To probe object-specific neural representations, we reconstructed the temporal envelope of each of the two simultaneous speech streams by optimally integrating MEG activity over time and space (i.e. sensors). Such a reconstruction of the envelope of each speech stream, not the physical stimulus, can be successful only if the stimulus mixture is neurally segregated ("unmixed") and the speech of the two speakers are represented differentially. We first reconstructed the temporal envelope of the attended speech. Figure 4.2A shows representative segments of the different envelopes reconstructed by this decoder, from listeners hearing the identical speech mixture but attending to different speakers in it. The reconstructed envelope is clearly more correlated with the envelope of the attended speech than with that of the background one, despite the fact that the stimuli were identical in both cases. The higher correlation with the attended speech is statistically significant ($P < 0.001$, paired permutation test, Fig. 4.2B, left) and is seen in 92% of trials (Fig. 4.2C).

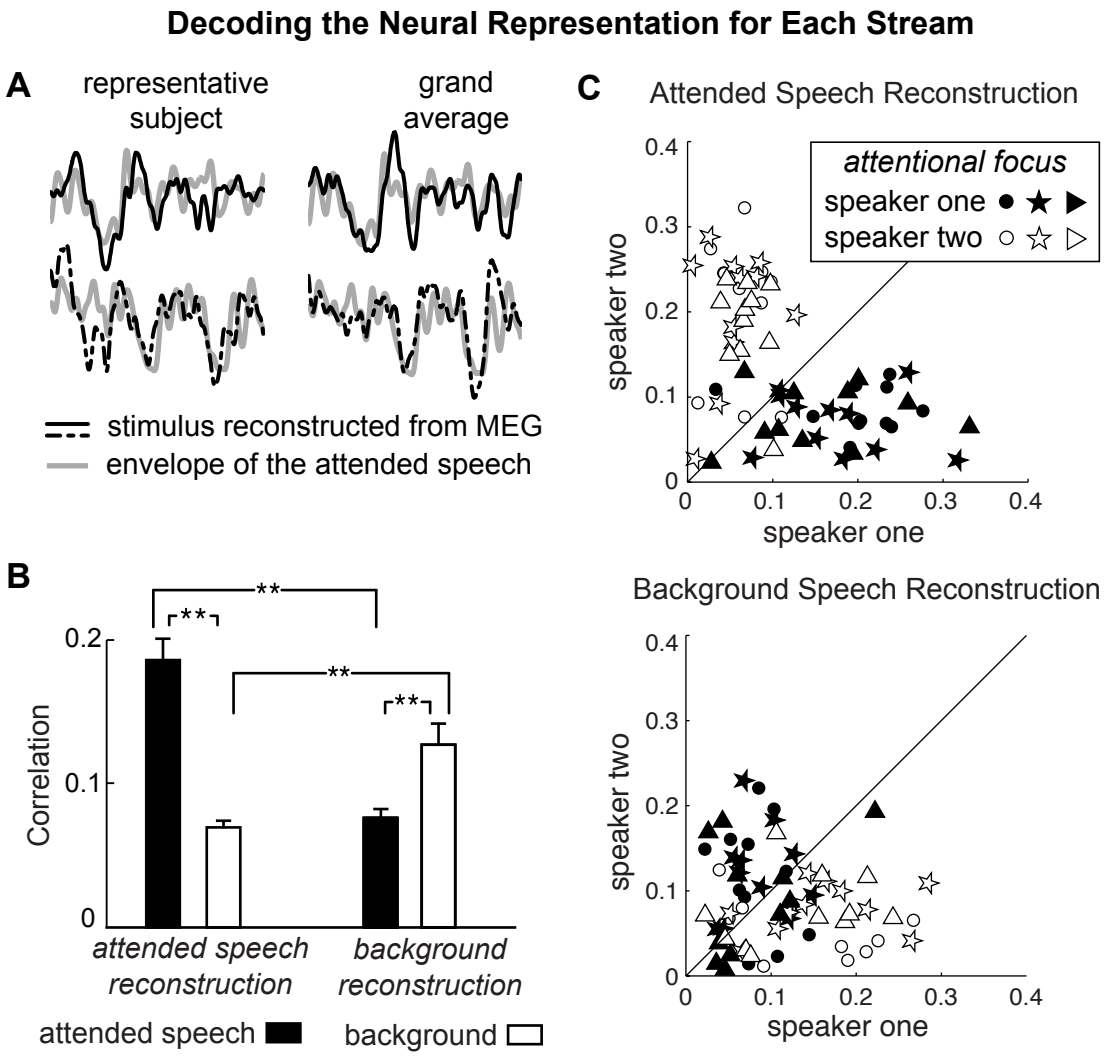

Figure 4.2. Decoding the cortical representation specific to each speech stream. (A) Examples of the envelope reconstructed from neural activity (black), superimposed on the actual envelope of the attended speech (gray). *Different* envelopes (in the upper and lower panels) are decoded from neural responses to *identical* stimuli, depending on whether the listener attends to one or the other speaker in the speech mixture, with each resembling the envelope of the attended speech. Here, the signals, 5 seconds in duration, are averaged over three trials for illustrative purposes, but all results in the study are based on

single trial analysis. **(**B) Two separate decoders reconstruct the envelope of the attended and background speech respectively from their separate spatial-temporal neural responses to the speech mixture. The correlation between the decoded envelope and the actual envelope of each speech stream is shown in the bar graph (averaged over trials and speakers), with each error bar denoting one SEM across subjects (** $P < 0.005$, paired permutation test). The separate envelopes reconstructed by the two decoders selectively resemble that of attended and background speech, demonstrating a separate neural code for each speech stream. (C) Decoding of the Speech Representations from Single Trials. Scatter plots of the correlation coefficients measured from individual trials and individual subjects, between the decoded envelope and the actual envelope. The attentional focus of listeners is denoted by marker color and the separate trials are denoted by marker shapes. Comparing the results of the two decoders, it can be seen that the speech of the attended and background speakers can be decoded separately from the same response, even on a single trial basis.

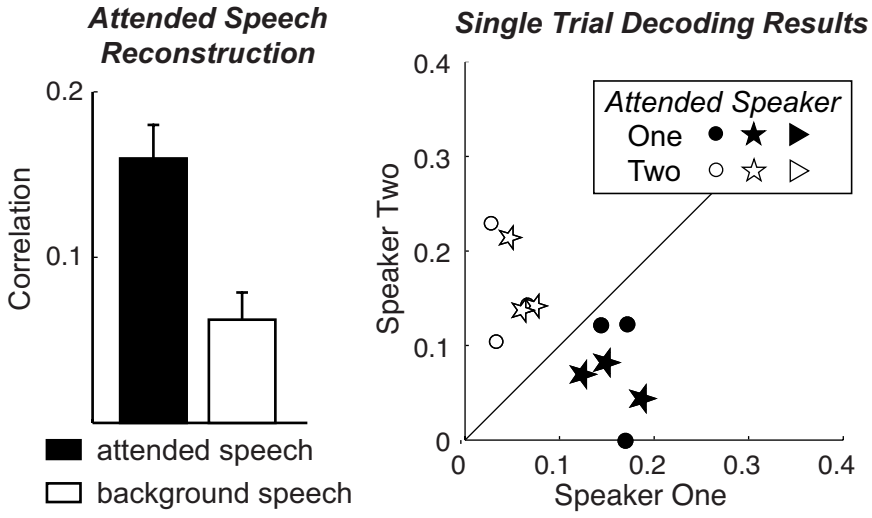

Figure 4.3. Decoding of the speech representations for two competing female speakers. The correlation between the decoded envelope and the actual envelope is shown by the bar graph (averaged over trials and subjects) and the scatter plot (each trial and subject separately). The attended speech can be decoded exclusively from the neural response to the mixture.

We also reconstructed the temporal envelope of the background speech using a second decoder that integrates neural activity spatial-temporally in a different way. The result of this reconstruction is indeed more correlated with the envelope of the background speech rather than the attended speech ($P < 0.005$, paired permutation test, Fig. 4.2B, right). Therefore, by integrating the temporal and spatial neural responses in two distinct ways, the attended and background speech can be successfully decoded separately. On average, the reconstruction for the background speech is more correlated with the background speech in 73% of the trials from individual subjects (Fig. 4.2C; significantly above chance level, $P < 0.002$, binomial test).

In this experiment, the speakers are of opposite gender, but the neural representations of segregated speech streams can be similarly demonstrated even for the more challenging scenario where the two speakers are of the same gender. In the Same-Gender experiment, after a training session, the subjects can successfully recognize and follow the target speaker and answer 74% of the comprehension questions asked during the experiment. From the neural response, the temporal envelope of the attended speaker is decoded (Fig. 4.3) and the decoded envelope is more correlated with the attended speaker than the unattended speaker (paired t-test based on individual trials from individual listeners, $P < 0.01$, for both speakers). Consequently, all these results (Fig. 4.2 and Fig. 4.3) demonstrate that the neural representation in auditory cortex goes beyond encoding just the physically presented stimulus (the speech mixture) and shows selective phase-locking to auditory objects.

Robustness to the Intensity of Either Speaker

When the intensity of either of the two competing speaker changes, up to 10 dB, human listeners can still understand either speaker with more than 50% intelligibility (Brungart, 2001). Intensity gain control may contribute to this robustness in speech perception. Here, we address whether intensity gain control occurs globally for an auditory scene or separately for each auditory object.

A second 'Varying-Loudness' experiment was carried out, where the intensity level of one speech stream, either the attended or the background, is kept constant while the other is reduced (up to 8 dB). Under this manipulation, the intensity ratio between the attended and background speaker, i.e. the target to masker ratio (TMR), ranges between -

8 dB and 8 dB. The listeners correctly answered 71% of the questions asked after each minute of listening, which did not significantly vary with TMR ($P > 0.7$, one-way repeated measures ANOVA), indicating that the listeners understood the story without any obvious difficulty, even when the acoustics of stimulus changed dramatically. The averaged subjective speech intelligibility is 88%, 80%, 68%, 60%, and 48% at 8 dB, 5 dB, 0 dB, -5 dB, and -8 dB TMR respectively, which varies significantly with TMR ($P < 10^{-4}$, $F(4,24) = 12.6$, 1-way repeated measures ANOVA).

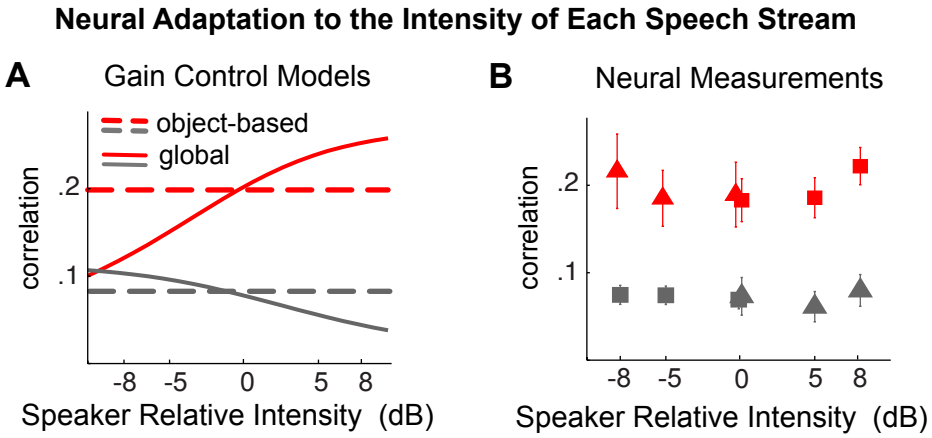

Figure 4.4. Decoding the attended speech over a wide range of relative intensity between speakers. (A) Decoding results simulated using different gain control models. The x-axis shows the intensity of the attended speaker relative to the intensity of the background speaker. Object-based intensity gain control predicts a speaker intensity invariant neural representation while the global gain control mechanism does not. (B) Neural decoding results in the Varying-Loudness experiment. The cortical representation of the target speaker (red symbols) is insensitive to the change in physical dominance of the speech (dashed orange

curve). The acoustic envelope reconstructed from cortical activity is much more correlated with the attended speech (red symbols) than the background speech (gray symbols). Triangles and squares are results from the two speakers respectively.

To distinguish how different intensity gain control mechanisms would affect the neural representation of each speech stream, we simulate possible decoding outcomes (SI Methods). The MEG activity is simulated by the sum of activity precisely phase-locked to each speech stream, and interfering stimulus-irrelevant background activity. The strength of the phase-locked activity is normalized by either the strength of whole stimulus, for a global gain control mechanism, or the strength of the encoding auditory object, for an object-based gain control mechanism. The simulated decoding outcomes under different gain control mechanisms are shown in Fig. 4.4A.

The neural decoding from actual MEG measurements is shown in Fig. 4.4B. For the decoding of the attended speech, the decoded envelope is significantly more correlated with the envelope of the attended speech ($P < 0.004$, $F(1,71) = 25.8$, attentional focus × TMR 2-way repeated measures ANOVA), and this correlation is not affected by TMR. The result is consistent with the object-based gain control model, and not with the global gain control model. Similarly, the neural decoding of the background speech is also affected by the attentional focus ($P < 0.02$, $F(1,71) = 14.65$, higher correlation with the background speech, 2-way ANOVA), without interaction between attention and TMR. Consequently, the neural representation of a speech stream is stable both against the intensity change of that stream and also against the intensity change of

the other stream, consistent with the hypothesized object-specific gain control (c.f. the examples shown in Fig. 4.1C).

Spatial Spectro-temporal Response Function and Neural Sources

The decoding analysis above integrates neural activity, spatial-temporally, to optimally reveal an object-specific neural representation. To characterize the neural code that the decoder extracts information from, we analyze the neural *encoding* process via the spectro-temporal response function (STRF), for each MEG sensor (deCharms et al., 1998; Depireux et al., 2001). The linear STRF and the linear decoder are respectively the forward and backward models describing the same relationship between the stimulus and neural response. Nevertheless, only the forward STRF model can reveal the timing and spatial information of the neural encoding process.

An STRF functionally describes how the spectro-temporal acoustic features of speech are transformed into cortical responses. It deconvolves the neural activity evoked by the continuous envelope of speech. In this STRF model, the encoding of each speech stream is modeled using the auditory spectrogram (Yang et al., 1992) of the "unmixed" speech signal with unit intensity. For any given frequency the horizontal cross-section of the STRF characterizes the time course of the neural response evoked by a unit power increase of the stimulus at that frequency, for one MEG sensor.

The MEG STRF contains two major response components (Fig. 4.5A & B): one with latency near 50 ms, here called the $M50_{STRF}$, and the other with latency near 100 ms, here called the $M100_{STRF}$. This indicates that two major neural response components continuously follow the temporal envelope of speech, with delays of 50 ms and 100 ms

respectively. Since an STRF is derived for each MEG sensor, the neural source locations of the $M50_{STRF}$ and $M100_{STRF}$ can be estimated based on the distribution over all sensors of the strength of each component, i.e. the topography of the magnetic fields at each latency (Fig. 4.5C). An equivalent current dipole (ECD) based neural source analysis reveals that the $M50_{STRF}$ and $M100_{STRF}$ responses arise from different neural sources, with the source of $M100_{STRF}$ being 5.5 ± 1.5 mm and 7.1 ± 2.0 mm more posterior in the left and right hemispheres respectively (Fig. 4.5D, $P < 0.005$ for both hemispheres, paired t-test). The ECD location of the neural source of the $M100_{STRF}$ peak is consistent with that observed for the M100 response to tone pips, localized to the superior temporal gyrus (STG) and roughly in the planum temporale (Lütkenhöner and Steinsträter, 1998).

The amplitudes of the $M50_{STRF}$ and $M100_{STRF}$ are further analyzed in the neural source space, based on the STRF at the ECD location of each component. The amplitude of the $M100_{STRF}$ is much stronger for the attended speech than for the background speech (Fig. 4.5B, $P < 0.007$, $F(1,87) = 11.85$, attentional focus × hemisphere × speaker, 3-way repeated-measures ANOVA). The amplitude of the $M50_{STRF}$ is, in contrast, not significantly modulated by either attention or TMR. The latency of the $M50_{STRF}$ and $M100_{STRF}$ are also modulated by attention ($P < 0.03$, $F(1,87) > 7$ for both peaks, 3-way repeated-measures ANOVA) and are on average 11 and 13 ms shorter, respectively, when attended.

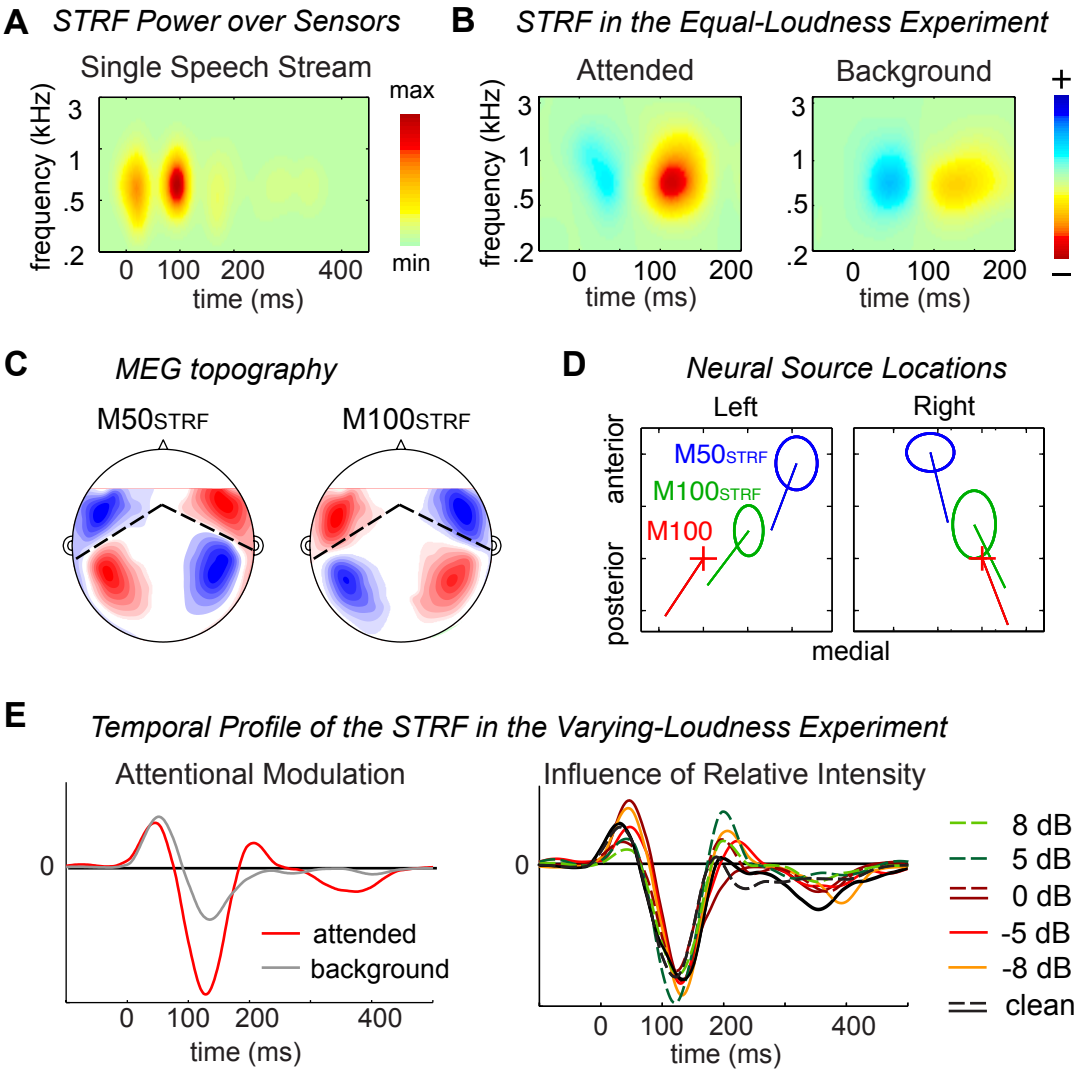

Figure 4.5. Cortical Encoding of the Spectral-temporal Features of Different Speech Streams. (A) The STRF power as function of frequency and time (summed over all sensors and subjects) for unmixed speech. It is dominated by two response components, $M50_{STRF}$ and $M100_{STRF}$, with respective latencies near 50 ms and 100 ms. (B) The STRFs for the attended and background speech, at the neural source location of the $M100_{STRF}$. Attention strongly enhances the response with latency near 100 ms. (C) The MEG topography of the $M50_{STRF}$ and $M100_{STRF}$, averaged over subjects and experiments. (D) The neural source

locations for the $M50_{STRF}$ and $M100_{STRF}$ in each hemisphere, as estimated by dipole fitting. The location of neural source of the $M50_{STRF}$ is anterior and medial to that of the $M100_{STRF}$ and M100. The source location for each subject is aligned based on the source of the M100 response to tone pips, shown by the cross. The span of each ellipse is 2 SEM across subjects. The line from each dipole location illustrates the grand averaged orientation of each dipole. Each tick represents 5 mm. (E) The temporal profile of the STRF in the Varying-Loudness Experiment for the attended speech. The $M100_{STRF}$ (averaged over TMR) is modulated by attention while the $M50_{STRF}$ is not. Neither response peak is affected by the intensity change of the two speakers.

The temporal profile of the STRF in the Varying-Loudness experiment is shown in Fig. 4.5E, which is extracted by applying a singular value decomposition to the STRF. The $M100_{STRF}$ is modulated by attention ($P < 0.03$, $F(1,143) = 9.4$, attentional focus × hemisphere × speaker × TMR, 4-way repeated-measures ANOVA) while the $M50_{STRF}$ is not. Neither response component is affected by TMR (c.f. the examples shown in Fig. 4.1C). The invariance of the $M50_{STRF}$ and $M100_{STRF}$ to the intensity of both the attended and background speech streams provides further evidence for the hypothesized object-specific gain control.

4.4 Discussion

This study investigates whether a multi-source auditory scene, perceptually represented in terms of auditory objects, is neurally represented in terms of auditory objects as well. From subjects selectively listening to one of two spectro-temporally overlapping speech streams, we do observe neural activity selectively synchronized to the speech of a single speaker (as was illustrated in Fig. 4.1B). Furthermore, in an ecologically valid listening setting, this selective representation of an individual speech stream is both modulated by top-down attention, and normalized by the intensity of that sound stream alone (as was illustrated in Fig. 4.1C). In sum, this meets all the criteria of an object-based representation, e.g. those specified by Griffiths and Warren (Griffiths and Warren, 2004): the observed neural representation is selective to the sound from a single physical source, is minimally affected by competing sound sources, and is insensitive to perceptually unimportant acoustic variations of the stimulus, e.g. changes in intensity.

Temporal Coherence, Attention and Object-based Representations

The object-specific representations seen here are precisely synchronized to the temporal envelope of speech. In speech and natural sounds in general, the temporal envelope gates on and off, and therefore synchronizes, various acoustic features, including pitch and formant structures. Therefore, they provide important cues for perceptual auditory grouping (Sheft, 2007) and are critical for robust speech recognition. For example, acoustic cues that can be used to segregate concurrent speech streams are dominantly content-independent voice features of each speaker, e.g. the pitch, which are not sufficient for the recognition of speech. At the same time, it is difficult to extract the

acoustic features necessary for speech recognition, e.g. the spectro-temporal envelope of speech, from a speech mixture, in the absence of strong speech segregation cues (such as pitch) (Stickney et al., 2004). A solution to this dilemma would be to group acoustic features belonging to the same auditory object, both speech segregation and intelligibility-relevant cues, through temporal coherence analysis, and then selectively process the attended auditory object as a whole (Shamma et al., 2011). In other words, the auditory cortex selects the attended speech stream by amplifying neural activity synchronized to the coherent acoustic variations of speech, i.e. the envelope. This idea is highly consistent with the large-scale synchronized and object-specific activity seen in this study.

At the neuronal mechanistic level, it is plausible that the low frequency phase-locked neural activity binds features belonging to the same object by regulating the excitability of neurons, so that a given neural network will be more responsive when processing features from the corresponding auditory object (Schroeder and Lakatos, 2009). Furthermore, such a rhythmic regulation of neuronal excitability may also contribute to the segmentation of continuous speech into perceptual units, e.g. syllables (Luo and Poeppel, 2007).

In the current experiment, the auditory scene consists of only two auditory objects and we demonstrate that the attended object and the background object are represented differentially. For the case of more than two auditory objects in an auditory scene, whether the neural system divides the scene into multiple objects, or only the attended object and the background, must be determined by future experiments.

Hierarchical Processing of Auditory Objects Auditory Cortex

Of the two major neural response components that track the speech envelope, the $M100_{STRF}$ is significantly modulated by attention but the $M50_{STRF}$ is not. These two neural response components track the speech envelope with different latencies and are generated from distinct neural sources. Based on their positions relative to the neural source of the M100 (Lütkenhöner and Steinsträter, 1998), the $M50_{STRF}$ and $M100_{STRF}$ arise roughly from Heschl's gyrus and the planum temporale, respectively. The latency and source location of the two components demonstrate a hierarchy of auditory processing (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009), and the representation of the attended object become dominant from shorter- to longer-latency activity and from core to posterior auditory cortex. Therefore, although auditory object representations may start to emerge as early as primary auditory cortex (Nelken and Bar-Yosef, 2008), the substantial top-down attentional modulation of the large-scale object neural representation measured here only emerges with later processing.

The routing of the neural processing of the attended auditory object into posterior auditory cortex may generally underlie the selection of auditory information when there are competing spectro-temporally complex auditory objects. MEG studies have shown that selectively listening to sound embedded in a complex auditory scene modulates longer latency (~100 – 250 ms) responses in association auditory cortex but not the shorter latency (~50 ms) steady state response in core auditory cortex (Ding and Simon, 2012; Gutschalk et al., 2008; Okamoto et al., 2011). The specific latency (whether ~100 ms (Hillyard et al., 1973; Rif et al., 1991) or longer (Ahveninen et al., 2011; Hari et al., 1988; Ross et al., 2010)) of the attentional modulation in association auditory cortex

shows evidence of variation with the rhythm of the stimulus. The attentional modulation near 100 ms seen here, therefore, might only occur for some dynamic stimuli, e.g. those with a speech-like rhythm. PET studies also indicate that the areas posterior to core auditory cortex are more activated when speech is interfered by temporally modulated noise than stationary noise (Scott et al., 2009; Scott et al., 2004), since modulated noise contains speech-like features and requires additional processes of information selection. Furthermore, a recent fMRI study has also shown that, in a multi-talker environment, the planum temporale is increasingly activated when the number of information sources, i.e. speakers, increases (Smith et al., 2010). Taken together, these results support the idea that posterior auditory cortex plays a major role in the generation of auditory objects (Griffiths and Warren, 2002; Zatorre et al., 2002) and the selection of information based on the listener's interest.

Neural Adaptation to the Intensity of Individual Auditory Object

The recognition of speech relies on the shape of its spectro-temporal modulations and not its mean intensity. This study demonstrates that cortical activity is precisely phase locked to the temporal modulations, but insensitive to the mean intensity of the speech streams, and therefore is effectively encoding the only shape of the modulations. Intensity gain control has been demonstrated in multiple stages of the auditory system (Robinson and McAlpine, 2009; Watkins and Barbour, 2009) and constitutes an auditory example of neural normalization, which has been suggested as a canonical neural computation (Carandini and Heeger, 2012). For example, in the cochlear nucleus,

neurons encode the shape of the spectral modulation of speech, e.g. a vowel, invariantly to its mean intensity (Young, 2008).

Critically different from these previous studies, however, the encoding of temporal modulations seen here is invariant to the intensity of each speech *stream* rather than the overall intensity of the mixture. In the Varying-Loudness experiment, the intensity of one speaker changes while the other is kept constant. Maintaining a stable representation despite the altered speech requires the observed neural adaptation to the sound intensity of the specific speech stream. The stable representation of the constant speaker, in contrast, requires the observed lack of adaptation to the overall intensity of the sound mixture, which co-varies with the intensity of the altered speech. These both contrast with the simpler mechanism of global intensity gain control, which would require the neural representation of both speech streams to be modulated in the same way based on the overall intensity of the acoustic stimulus. Therefore, the data strongly suggest the existence of an *object-specific* intensity gain control, which normalizes the strength of neural activity based on the intensity of individual auditory objects.

In sum, this study demonstrates the key signatures of an object-specific neural representation arising from the analysis of a complex auditory scene. Such object-specific neural representations are phase-locked to the slow rhythms (<10 Hz) of the encoded auditory object, and adapt to the intensity of that object alone. Under the modulation of top-down attention, the auditory response in posterior auditory cortex (latency near 100 ms) dominantly represents the attended speech, even if the competing speech stream is physically more intense. This object-specific auditory representation provides a bridge

between feature-based, precisely phase-locked sensory responses, and interference-resilient cognitive processing and recognition of auditory objects.

Chapter 5

Cortical representation of speech in noise

5.1 Introduction

Normal hearing human listeners are remarkably good at understanding speech in complex listening environments (Brungart, 2001; Festen and Plomp, 1990). The recognition of speech relies on the spectro-temporal modulations of speech (Chi et al., 1999; Elliott and Theunissen, 2009), including the important component of slow temporal modulations (< 16 Hz). These slow temporal modulations, which constitute the envelope of speech (Rosen, 1992), contribute to robust speech recognition in two ways. First, they reflect the syllabic and phrasal rhythm of speech (Greenberg et al., 2003; Poeppel et al., 2008) and, in quiet listening environments, lead to high intelligibility even with only very coarse spectral information (Shannon et al., 1995). Second, in complex auditory scenes, they provide primary cues to group together features belonging to the same sound stream and therefore play a critical role in extracting a target speech stream from the acoustic background (Shamma et al., 2011).

The functional importance of the slow temporal modulations makes it a plausible hypothesis that noise-robust speech recognition relies on stable neural synchronization to the speech envelope. Specifically, it has been proposed that cortical activity synchronized to the speech envelope underlies the parsing of continuous speech into basic processing units, e.g. syllables, and regulates the allocation of neural resources to the processing of each unit (Giraud and Poeppel, 2012; Schroeder and Lakatos, 2009). Furthermore,

selective neural synchronization to a speech stream embedded in a complex auditory scene has been hypothesized as the mechanism to segregate the speech stream from the acoustic background (Shamma et al., 2011). Both the segregation of speech from background and the parsing of speech into perceptual units are prerequisites for robust speech recognition. Therefore, if cortical activity synchronized to the speech envelope is truly involved in these processes, it must reliably occur in the presence of any acoustic background that does not eliminate speech intelligibility. This critical prediction is tested in this study.

The acoustic background interferes with speech in two ways, energetic masking and information masking. *Energetic masking* is caused by the physical, acoustic overlap between the speech and the background. It causes strong degradation to the neural representation of speech in the auditory periphery, though how the degraded peripheral representation is further processed in the central auditory system is not yet well understood. Energetic masking is exemplified by speech masked by spectrally matched stationary noise (Festen and Plomp, 1990), the scenario investigated here. *Informational masking* refers to any speech background interaction that is not caused by acoustic overlap, but instead by the perceptual similarity between the speech and background. It is exemplified by speech masked by another stream of speech (Brungart, 2001), the scenario for which evidence of reliably neural synchronization to speech has very recently been demonstrated (e.g. Chapter 3-4). In this study, we demonstrate how energetic masking affects neural synchronization to the speech envelope, and test the link between this neural synchronization and noise-robust recognition of speech.

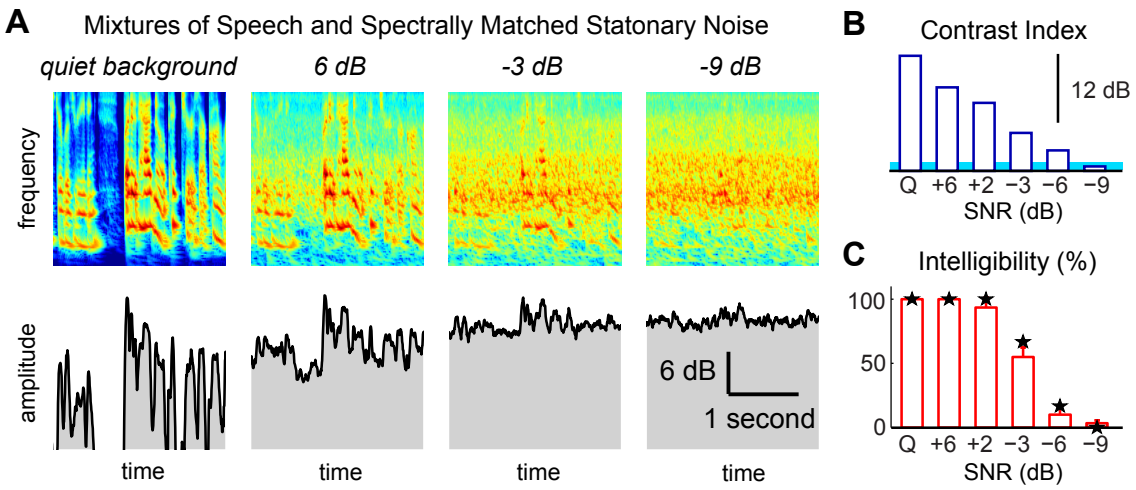

Figure 5.1. Speech embedded in spectrally matched stationary noise. (A) The auditory spectrogram (upper panel) and the temporal envelope (lower panel) of speech embedded in noise, at 4 SNRs. The background noise causes severely degradation to the spectro-temporal features of speech. (B) The contrast index characterizes the spectro-temporal contrast of the stimulus at each SNR. The shaded blue area covers the 5th to 95th percentile of the contrast index calculated for stationary noise alone, and the SNR condition Q indicates a quiet background. The intensity contrast of the stimulus decreases continuously with SNR. In this illustration, though not in the experiment, the same speech segment is used in every SNR condition. (C) Subjectively rated intelligibility of speech (bars), and percent of comprehension questions correctly answered (stars). The intelligibility remains unaffected by SNR until -3 dB SNR.

Spectrally matched stationary noise causes severe acoustic degradation to speech. It reduces the intensity contrast of the speech and distorts its spectro-temporal modulations (Fig. 5.1A & B). Speech intelligibility, however, is robust to such acoustic

degradations until the noise is about 3 dB stronger than speech (Fig. 5.1C). Psychoacoustic studies suggest that this robustness is maintained by insensitivity to the intensity contrast of speech (Stone et al., 2011) and by selectively processing the temporal modulations at those modulation rates less corrupted by noise (Jorgensen and Dau, 2011). Although it is still unclear whether these computational strategies suggested by psychophysical studies are indeed implemented in the human brain, their possible neural underpinnings have been suggested by animal studies: Insensitivity to stimulus contrast can be achieved by neural adaptation to the mean and variance of stimulus intensity (Dean et al., 2005; Nagel and Doupe, 2006; Rabinowitz et al., 2011), and the selective encoding of temporal modulations can result from stimulus-dependent spectro-temporal tuning of neurons (Escabí et al., 2003; Lesica and Grothe, 2008; Woolley et al., 2005; Woolley et al., 2006)

In this study, subjects listened to a spoken narrative, mixed with spectrally matched stationary noise, at different signal-to-noise ratios (SNR). We hypothesize that cortical synchronization to the speech envelope is robust against the acoustic degradations caused by noise, at least when speech intelligibility remains high. In other words, it is hypothesized that the severe acoustic degradations caused by noise are compensated for neurally. To test the hypothesis, we record from subjects using magnetoencephalography (MEG), which can directly measure cortical activity synchronized to the envelope of speech (Ding and Simon, 2012; Luo and Poeppel, 2007). The neural computations underlying the hypothesized stable neural representation are also investigated with particular attention to how the brain overcomes, or compensates for, the loss of stimulus dynamic range and distortions of the stimulus temporal

modulations. Furthermore, we investigate the specific relationship between the cortical encoding of slow temporal modulations and individual subjects' ability to recognize speech in noise.

5.2 Methods

5.2.1 Subject, stimuli and Procedures

Subjects

Eleven normal hearing, right-handed, young adults (7 females, all between 20 and 31 years old) participated in the experiment. One subject was excluded due to the lack of auditory responses to both tones and speech. Subjects were paid for their participation. The experimental procedures were approved by the University of Maryland institutional review board. Written informed consent form was obtained from each subject before the experiment.

Stimuli and Procedure

The stimuli were taken from the beginning of a narration of the story *Alice's Adventures in Wonderland* (http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-4/). The sound recording was low-pass filtered below 4 kHz and divided into twelve 50-second duration sections, after long speaker pauses (> 300 ms) were shortened to 300 ms. A spectrally matched stationary noise was generated based on a 12-order linear predictive model estimated from the speech recording, and mixed into speech with one of six SNRs, i.e. quiet (no noise added in), +6 dB, +2 dB, -3 dB, -6 dB, and -9 dB.

The intensity of speech, measured by RMS, was the same for all sections and the intensity of noise was varied to create different SNRs. All the sections were presented sequentially and then repeated twice. The subjects were asked a comprehension question after each section, and also rated intelligibility of speech (in percentage) during the first presentation of each section.

The SNR decreased or increased every two sections. For the decreasing SNR order (applied to five subjects), no noise was added to the first two sections; noise 6 dB weaker than speech was added to the following two sections, and then the noise level kept increasing over sections. The increasing SNR order, in contrast, started with the lowest SNR, i.e. -9 dB, and finished with a quiet condition. The SNR order affects neither speech intelligibility (SNR × Order, 2-way repeated measures ANOVA) nor the neural reconstruction of speech (SNR × Order × Trial, 3-way repeated measures ANOVA), and therefore was not distinguished in the analysis.

All stimuli were presented identically to both ears and the subjects were required to close their eyes when listening. Before the main experiment, 100 repetitions of a 500-Hz tone pip were presented to elicit the M100 response, which is a reliable auditory response measured 100 ms after the onset of a tone pip and whose neural source is easy to localize within auditory cortex. The neuromagnetic signal was recorded using a 157-channel whole-head MEG system (KIT, Kanazawa, Japan), with 1 kHz sampling rate. A 200-Hz lowpass filter and a notch filter at 60 Hz were applied online and environmental noise was removed offline. Details of the recording procedure were described in Ding & Simon (2012).

Stimulus Characterization

The auditory spectrogram of the stimulus was calculated using a sub-cortical auditory model (Yang et al., 1992) and expressed in linear amplitude scale. The broadband envelope of stimulus was defined the as the sum of the auditory spectrogram over frequency. The spectro-temporal contrast of a stimulus was characterized using a contrast index, the coefficient of variation of the auditory spectrogram, an extension of the fluctuation index (Nelken et al., 1999). The coefficient of variation is the standard deviation of the amplitude of the auditory spectrogram divided by the mean. It is zero for a sound with its energy being constant over time and frequency and grows as the contrast, i.e. depth, of the spectro-temporal modulations increases.

5.2.2 Data recording and analysis

Neural Reconstruction of Stimulus

The temporal envelope of the actual stimulus (the speech-noise mixture) or the speech only (embedded in the stimulus) was reconstructed by linearly integrating MEG activity over time and sensors. The reconstructed speech envelope, $E(t)$, is expressed as $\hat{E}(t) = \Sigma_k \Sigma_{1 \le \tau \le T} M_k(t + \tau) D_k(\tau)$, where $M_k(t)$ is the MEG signal from a sensor k and $D_k(t)$ is the linear decoder for the same sensor. The envelope to reconstruct, $E(t)$, is either the envelope of the actual stimulus or the envelope of the underlying speech. The decoder $D_k(t)$ was optimized using boosting with 10-fold cross-validation (David et al., 2007) to maximize the correlation between $\hat{E}(t)$ and $E(t)$. To reduce computational complexity, the MEG sensors in each hemisphere were compressed into 3 components using denoising

source separation (DSS) (de Cheveigné and Simon, 2008). Both hemispheres were used unless otherwise specified.

The decoder $D_k(t)$ integrates MEG activity over a time period T, which is set to 500 ms when not specified. This assumes that the information of the stimulus at time t is encoded in the neural response in a time window between t and $t + T$. This time window is parametrically adjusted between 50 ms and 1000 ms to investigate which time interval carries more information. During this varying integration window analysis, however, the auto-correlation of the speech envelope must be taken into consideration. For example, the response at time $t - 50$ ms, $M(t - 50)$, should contain no information of the stimulus at a future time t, $E(t)$. Nevertheless, if $M(t - 50)$ encodes information of the stimulus at time $t - 100$ ms, which is heavily correlated with $E(t)$, then, apparently, from $M(t - 50)$ some information about $E(t)$ can be reconstructed, implicitly through $E(t - 100)$. Therefore, in the integration window analysis, we partialed out the auto-correlation of the envelope using an extended model $E(t) = \Sigma_k \Sigma_{1 \le \tau \le T} M_k(t + \tau) D_k(\tau) + \Sigma_{1 \le \tau \le T^*} E(t - \tau) D_A(\tau) + \varepsilon(t)$, where $\varepsilon(t)$ is the unexplained residual. $D_k(t)$ and $D_A(t)$, the decoder and the regressor for speech autocorrelation, are estimated together using boosting (David et al., 2007). T^*, the maximal time range where the autocorrelation of speech is considered, is set to 500 ms. In this case, the reconstructed neural response, $\hat{E}^*(t) = \Sigma_k \Sigma_{1 \le \tau \le T} M_k(t + \tau) D_k(\tau)$, is a reconstruction of the component in speech envelope that cannot be predicted by its own history due to the rhythm of speech, i.e. the innovation information at a given moment.

Amplitude-Intensity Function

To systematically characterize the gain of cortical responses, the relationship between the stimulus and response is further modeled using a linear-nonlinear model: $E(t) = \Gamma(\Sigma_k \Sigma_{1 \le \tau \le T} M_k(t+\tau)\underline{D}_k(\tau)) + \varepsilon(t)$, where $E(t)$ is the envelope of the actual stimulus and the decoder $\underline{D}_k(t)$ is subject to the constraint that $\Sigma_k \Sigma_{1 \le \tau \le T} (\underline{D}_k(\tau))^2 = 1$. Since the decoder $\underline{D}_k(t)$ is normalized, the response gain is only reflected in the amplitude-intensity function Γ. The nonlinear function Γ is obtained by fitting the stimulus envelope $E(t)$ as a function of the linearly reconstructed envelope $\underline{\hat{E}}(t) = \Sigma_k \Sigma_{1 \le \tau \le T} M_k(t+\tau)\underline{D}_k(\tau)$ using the following procedure. For stimulus intensity I_0, the corresponding response amplitude A_0 is estimated by averaging the reconstructed envelope at time moments when the stimulus intensity is close to I_0, i.e. $T_0 = \{t \mid I_0 - \Delta I < E(t) < I_0 + \Delta I\}$. ΔI is one 10^{th} of the range between the 5^{th} and 95^{th} percentiles of $E(t)$. The estimated AIF is smoothed using a Gaussian function with SD as ΔI.

Phase-locking Spectrum

The phase locking of the neural response was investigated in narrow frequency bands (2-Hz wide), by calculating the inter-trial correlation of the neural response. The major component of MEG response was extracted using the first DSS component (de Cheveigné and Simon, 2008) and applied to this analysis. The phase-locking spectrum of the neural response to speech has a low-pass shape. To estimate the low-pass cutoff frequency, the phase-locking spectrum is modeled using a sigmoidal function $1 - 1/\exp(-\alpha(f - f_T))$. The slope parameter α and location parameter f_T are fitted in the least squares sense. In this modeling, since a sigmoidal function is bounded between 0 and 1, the inter-

trial correlation as a function of frequency is normalized so that the maximal value is 1 and the minimal value is 0.

TRF

The TRF deconvolves the continuous neural response evoked by the continuous speech stream, and obtains a response waveform due to a unit power increase of the stimulus (Ding and Simon, 2012). A TRF was estimated based on each MEG sensor, and the MEG data was averaged over trials in the TRF analysis. To estimate the TRF, an spectro-temporal response function (STRF) is first estimated using boosting with 10-fold cross validation (David et al., 2007), using the procedure described in Ding & Simon (2012). The TRF is obtained by summing the STRF over frequency. The $M50_{STRF}$ was determined as the response peak between 0 and 140 ms, which has a magnetic field topography negatively correlated with that of the M100. The $M100_{STRF}$ was determined as the response peak between 80 and 180 ms, which has a magnetic field positively correlated with that of M100 (detailed procedures described in Chapter 4).

Neural Source Analysis

The neural sources of the $M50_{STRF}$ and $M100_{STRF}$ were modeled by an equivalent-current dipole (ECD) in each hemisphere, based on a spherical head model (Ding and Simon, 2012). The $M50_{STRF}$ and $M100_{STRF}$ magnetic fields were well fitted by the ECD model. The median correlation between the fitted ECD magnetic field and the measured magnetic field is above 90% in both hemispheres and for both the $M50_{STRF}$ and $M100_{STRF}$. When comparing the ECD positions of different peaks in TRF, we included

only ECDs successfully capturing the measured magnetic field, characterized by a higher than 80% correlation between the ECD magnetic field and the measured magnetic field. Only one subject was excluded this way. After the ECD positions were determined, the moment of the dipole was estimated using the generalized least squares method (Mosher et al., 2003). For the dipole moment, the polarity of the $M100_{STRF}$ is defined as negative, to be consistent with the polarity of the N1 peak of EEG. The TRF projected to the ECD location was employed to analyze the amplitude and latency of the $M50_{STRF}$ and $M100_{STRF}$ (see Chapter 4).

5.3 Results

Noise Robust Cortical Reconstruction of Speech

The stimulus consists of a narrated story that is divided into six 100-second duration sessions. Each is presented either in quiet (alone) or with spectrally matched stationary noise (SNR ranging from -9 to +6 dB). A contrast index is used to characterize how the background noise reduces the intensity contrast, i.e. the depth of the spectro-temporal modulations, of the stimulus. As shown in Fig. 5.1B, the intensity contrast of the speech-noise mixture decreases monotonically with decreasing SNR, until finally reaching the intensity contrast of stationary noise alone, at -9 dB SNR.

To investigate how the cortical representation of speech is affected by noise, we attempted to reconstruct the temporal envelope of the underlying speech (as opposed to the actual stimulus including noise), from the cortical response to the noisy stimuli (Fig. 5.2A). The accuracy of the reconstruction reflects how precisely cortical activity is

synchronized to the speech envelope, even in the presence of background noise, and is a lower bound to how accurately the bare speech (with the background noise removed) is encoded in cortex.

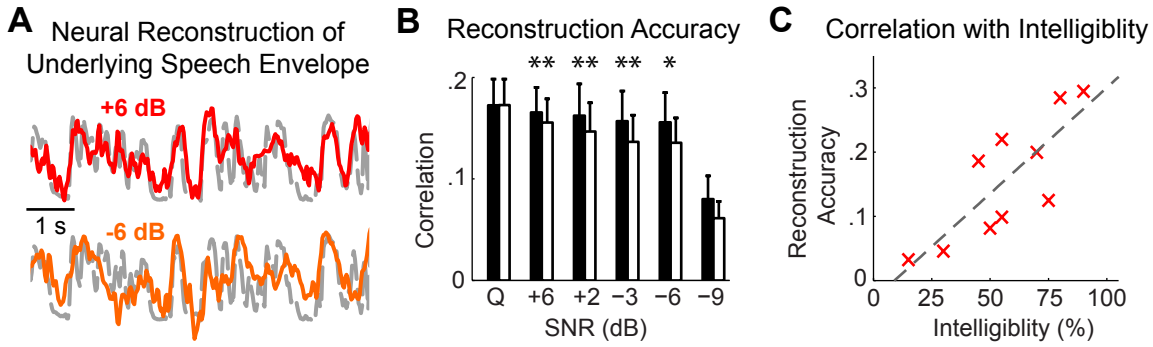

Figure 5.2. Neural Reconstruction of the Temporal Envelope of Speech. (A) The red and orange waveforms are the envelopes reconstructed from the neural responses in two SNR conditions. The dashed gray waveform is the envelope of the underlying speech in each stimulus. The neural construction matches the speech envelope well at both SNRs. The neural reconstructions illustrated are averaged over trials and subjects. (B) Correlation between the single-trial neural reconstruction and the envelope of either the underlying speech (black) or the actual stimulus (white). The correlation is averaged over trials and the error bar is 1 SEM over subjects. The reconstruction of the underlying speech is more accurate than the reconstruction of the actual noisy stimulus (** $P < 0.01$, * $P < 0.05$, paired t-test). (C) Relationship between the neural reconstruction accuracy and speech intelligibility, at -3 dB SNR. Each subject is shown by a red cross. The neural and behavioral results are highly correlated, with the regression line shown by the dashed line.

At the intermediate SNR of -3 dB, the subjectively rated speech score varies broadly over subjects, with a median of 55%. At this SNR, individual speech score is strongly correlated with the accuracy of neural reconstruction (Fig. 5.2C). The correlation coefficient is 0.79 ± 0.15 (Mean ± SEM, the SEM is consistently used in the paper to describe subject variations and is calculated using bootstrap), significantly positive ($P < 0.005$, bootstrap). When the two hemispheres are analyzed separately, the reconstruction in each hemisphere is correlated with speech intelligibility (mean correlation coefficient: 0.81, no significant difference between hemispheres, $P = 0.41$, bootstrap).

At other SNR conditions the speech scores clump near ceiling (median > 90%) or floor (≤ 10%) values (Fig. 5.1C), precluding the computation of analogous correlations. In other words, the transition from an intelligible stimulus to unintelligible stimulus typically occurs near -3 dB SNR. To better characterize this transition SNR for individuals, we identify the SNR for which the speech score drops to 50% (the speech recognition threshold, SRT) by fitting the relationship between individual's speech score and SNR as a sigmoidal function. The SRT is negatively correlated with neural reconstruction accuracy (correlation coefficient -0.67 ± 0.17; significantly negative, $P < 0.005$, bootstrap). This correlation indicates that subjects with more accurate neural synchronization to speech can recognize speech more robustly at lower SNRs.

To investigate whether the neural encoding of the underlying clean speech is a result of the neural encoding of the actual stimulus, we also reconstructed the envelope of the actual noisy stimulus from cortical activity. This decoding, although seemingly more straightforward, is less accurate than the decoding of the underlying speech for SNRs between +6 dB and -6 dB (Fig. 5.2B). Therefore, auditory cortex predominantly

synchronizes to the underlying speech rather than the physically presented sound mixture. The mechanisms underlying this robust neural representation are analyzed in following sections.

Contrast Gain Control

Background noise reduces the intensity contrast of the stimulus but not the accuracy of the neural representation of speech. This indicates that the loss of stimulus contrast is compensated for neurally through contrast gain control. To test this hypothesis explicitly, we analyze the relationship between the instantaneous intensity of the stimulus and the instantaneous amplitude of the neural response (Fig. 5.3A). This relationship, referred to as the amplitude-intensity function (AIF), analogous to a single neuron's spike rate-intensity function, strongly depends on the SNR and the slope of the AIF increases with decreasing SNR until SNR reaches -9 dB. The slope of the AIF reflects how quickly the amplitude of neural activity increases with the intensity of the stimulus. It is steeper for lower SNRs, indicating that the neural response is more sensitive to subtle intensity changes in the stimulus when the overall contrast of the stimulus is low. The slope of the AIF, extracted by a linear regression, increases 16 ± 2 dB (Mean ± SE) as SNR decreases from infinity (quiet) to -6 dB (Fig. 5.3A). To test how much this increase of response gain compensates the noise-induced loss of stimulus contrast, we use a modified AIF to describe the relationship between neural response amplitude and the instantaneous intensity of the underlying speech (not the actual stimulus). This modified AIF is SNR independent until -9 dB SNR (Fig. 5.3B), indicating that the amplitude of cortical activity encodes the intensity of the underlying speech rather than the intensity of the actual

stimulus. In other words, the noise-induced change of stimulus contrast is completely compensated for by response gain control in this SNR range.

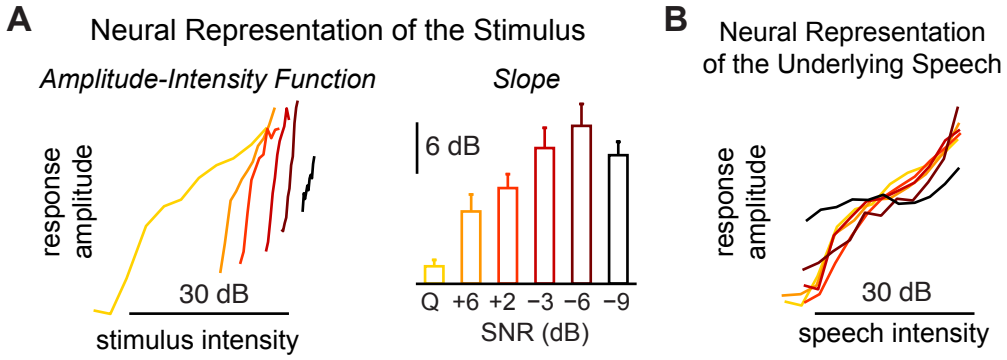

Figure 5.3. Neural Encoding of Stimulus Intensity. (A) The amplitude of neural response is plotted as a function of the instantaneous intensity of stimulus, for each SNR (left, color code the same as the bar graph on the right). The AIF strongly depends on SNR, as is characterized by the slope of the AIF (right). The stimulus dependent AIF indicates contrast gain control (cf. the stimulus contrast index illustrated in Fig. 5.1B). The error bar is 1 SEM over subjects (bootstrap). (B) The amplitude of neural response is plotted as a function of the instantaneous intensity of the underlying speech. This modified AIF shows invariance to SNR, except for -9 dB SNR, indicating noise-invariant neural encoding of speech intensity.

Modulation Sensitivity

Speech and noise each have a distinct modulation spectrum (the power spectrum of the temporal envelope), with the noise possessing more energy at higher modulation rates. Therefore, when noise is introduced, the energy of the stimulus envelope spreads

into higher modulation rates (Fig. 5.4A). Consequently, if cortical activity were simply following the temporal modulations of the stimulus, it would also spread into higher frequencies. This conjecture, however, can be ruled out (Fig. 5.4B). In fact, at the higher frequencies (e.g. near 7 Hz), the most reliable phase locking, measured by inter-trial response correlation, is seen with a quiet acoustic background, and the phase-locking spectrum of the cortical response progressively shifts towards low frequency as more noise is introduced.

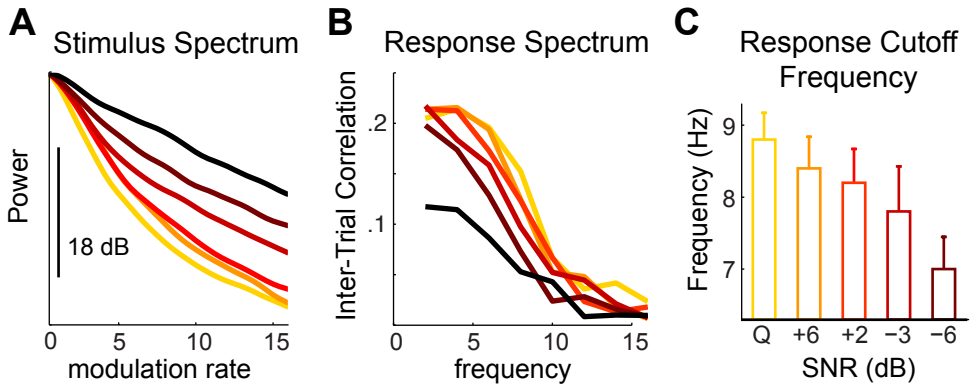

Figure 5.4. Neural Encoding of Temporal Modulations. The color code is the same in all panels and is specified in (C). (A) The power spectrum of the stimulus envelope, at different SNRs. Each spectrum is normalized based on its power density at 0.1 Hz, to emphasize changes in shape. The modulation spectrum of speech in quiet background (yellow) has the sharpest low-pass shape, and background noise increases the proportion of the stimulus power in higher modulation rates. (B) The phase-locking spectrum of the cortical response. It is consistently low-pass in shape but with a cutoff frequency that decreases with poorer SNR. (C) The cutoff frequency of the phase-locking

spectrum (not reliably estimable at -9 dB SNR) decreases with SNR. Error bar is 1 SEM over subjects.

The cutoff frequency of phase-locking spectrum (Fig. 5.4C, estimated by fitting the spectrum as a sigmoidal function) decreases significantly and monotonically from 8.7 ± 0.4 Hz to 7.0 ± 0.5 Hz when the SNR decreases from infinity (quiet background) to -6 dB ($P < 0.005$, bootstrap). Between +6 dB and -6 dB, the cutoff frequency decreases 0.72 ± 0.29 Hz every 6 dB (linear regression). Therefore, as the noise level rises, the auditory system reduces its sensitivity to fast temporal modulations, so that it does not respond to the increasingly stronger fast modulations introduced by the noise.

Temporal Integration

A separate measure of how phase locking of the response depends on SNR and frequency is to analyze the phase locking as a function SNR, at each frequency (Fig. 5.5A). At very low frequencies (e.g. 2 Hz), the response phase locking is not affected by noise until the poorest SNR of -9 dB. At higher frequencies (e.g. 6 and 8 Hz), however, phase locking decreases continuously with SNR. Specifically, the lowest SNR that does not affect neural phase locking is -6 dB, +2 dB, and +6 dB, for neural activity at 2 Hz, 4 Hz, and 6 Hz ($P > 0.5$, one-way repeated-measures ANOVA for the neural phase locking at each frequency, including the conditions between quiet and the lowest SNR; $P < 0.01$ if the SNR range is broadened). The stability of neural phase locking at lower, but not higher, frequencies suggests that the long-term temporal integration is important in maintaining a noise-robust neural representation.

To confirm the role of long-term integration in encoding the speech envelope, we again applied the neural reconstruction analysis, but with varying time integration windows. In the analysis shown in Fig. 5.2, the reconstruction of the stimulus at each time moment is based on the response in a 500-ms time window starting from that moment. When this window size is allowed to vary, the reconstruction results show a strong dependency on the integration time (Fig. 5.5B). At the poorer SNRs, e.g. -3 to -9 dB, the decoding results improve substantially when the window of integration is allowed to increase in size from 100 ms to 200 ms. This demonstrating the importance of long-term (> 100 ms) integration in encoding speech in a strong noise background.

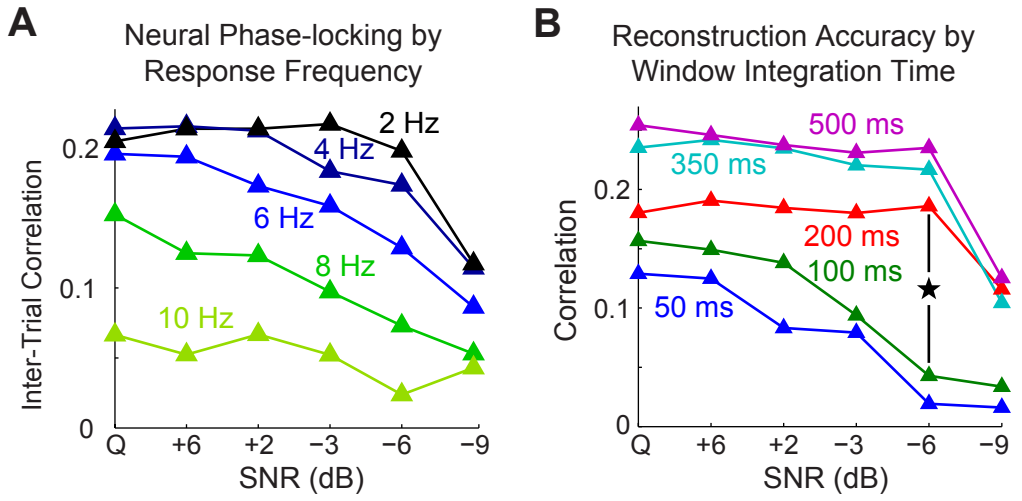

Figure 5.5. Neural Reconstruction with Various Temporal Integration Duration. (A) The phase locking of neural activity as a function of SNR. When SNR decreases from +6 to -6 dB, the neural phase locking at 2 Hz is stable but the neural phase locking at 8 Hz continuously decreases, with intermediate trends of decrease at intermediate frequencies. (B) The ability to reconstruct the speech envelope from the neural response depends on the temporal integration window. Each color-coded curve is the reconstruction accuracy for a different integration

window. The strongest window-dependent change in reconstruction accuracy is observed near -6 dB (marked by a star), where the decoding results improves substantially when the window of integration is allowed to increase in size from 100 ms to 200 ms.

Temporal Response Function

To explicitly characterize how the spectro-temporal features of the stimulus are encoded cortically as a function of time, and by cortical area, for each MEG sensor we estimate the temporal response function (TRF), which characterizes the time course of neural activity evoked by a unit power increase of the stimulus (Ding & Simon, 2012). While the neural reconstruction integrates responses over a specified duration, the TRF describes the neural response at each time moment, i.e. each time lag between the stimulus and the response, through deconvolution. In the TRF analysis, the stimulus amplitude is normalized in each SNR condition by z-score. With the stimulus thus normalized, an SNR-independent TRF amplitude would demonstrate a neural representation independent of the mean and variance (i.e. contrast) of the stimulus intensity.

The instantaneous TRF power, averaged over all MEG sensors, is shown in Fig. 5.6A, upper panel. The TRF is clearly delayed as the noise level increases. The onset latency of TRF (the earliest time point when the TRF amplitude passes the 99^{th} percentile of the pre-stimulus TRF amplitude) is continuously delayed as the noise level rises (Fig. 5.6A, lower panel). This latency elongation is statistically significant, since the relationship between onset latency and SNR, when fitted by a line, has a significantly negative slope ($P < 0.001$, bootstrap). The earliest two components of the TRF, called the

$M50_{STRF}$ and $M100_{STRF}$ are extracted and further analyzed. These two components are generated bilaterally in auditory cortex (Ding & Simon, 2012).

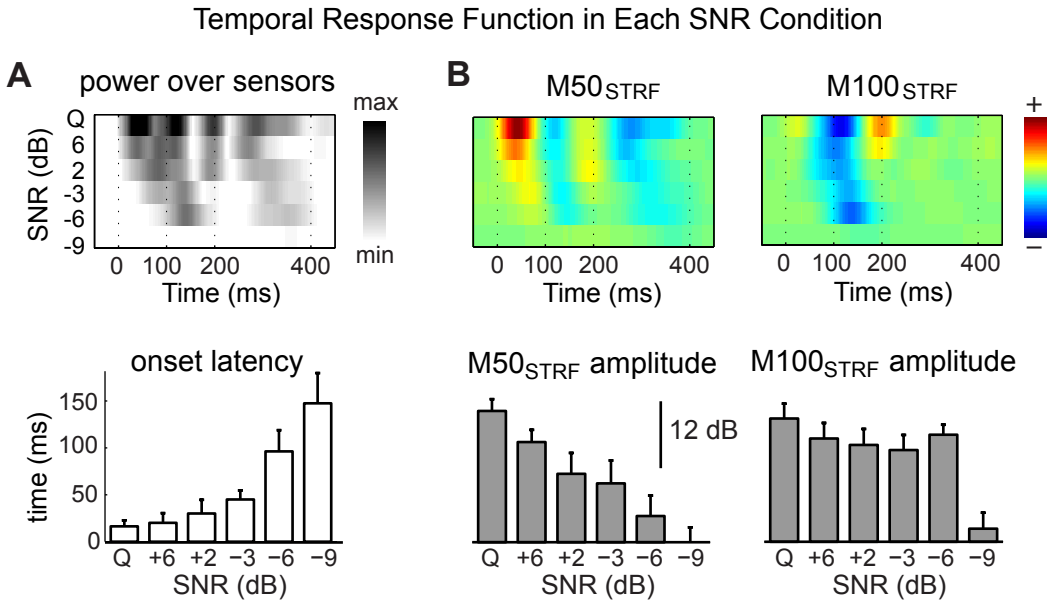

Figure 5.6. SNR dependent temporal response function. (A) The instantaneous TRF power, summed over sensors. The TRFs from all SNR conditions are stacked vertically. The latency at which the TRF amplitude surpasses the noise floor is shown in the lower panel. The TRF onset is significant delayed by noise. (B) The TRFs at the neural sources of the $M50_{STRF}$ and $M100_{STRF}$ (upper panels). The amplitude of the $M50_{STRF}$ decreases when the level of noise increases (c.f. the stimulus contrast index illustrated in Fig. 5.1B), while the amplitude of the $M100_{STRF}$ remains stable until -9 dB SNR.

A bilateral equivalent current dipole (ECD) model shows that the ECD position of the $M50_{STRF}$ is on average 10 (13) mm more anterior than that of the $M100_{STRF}$ in the left

(right) hemisphere (statistically significant in the right hemisphere; paired t-test, $t_8>6$, $P < 0.02$). The TRF at the ECD position of $M50_{STRF}$ and $M100_{STRF}$ are shown in Fig. 5.6B. The TRF is averaged over the two hemispheres since very similar results are seen in each of them. The amplitude of the $M50_{STRF}$ decreases continuously with SNR, while the amplitude of the $M100_{STRF}$ is insensitive to SNR until the SNR decreases to -9 dB. A linear regression analysis shows that, in between -6 dB and 6 dB SNR, the amplitude of the $M50_{STRF}$ decreases 1.0 ± 0.2 dB (significantly negative, $P < 0.001$, bootstrap) while the amplitude of the $M100_{STRF}$ changes 0.0 ± 0.2 dB (N.S.) each 1 dB SNR change. The same regression analysis reveals that the latency of the $M50_{STRF}$ increases with SNR, with a change of 3.0 ± 0.6 ms/dB.

5.4 Discussion

By recording from human subjects listening to continuous speech embedded in noise, this study demonstrates that the temporal modulations of speech are reliably represented in auditory cortex, at least until the noise is more than twice as strong as the speech (-6 dB SNR). Two distinct types of acoustic degradation caused by noise, i.e. the compression of stimulus dynamic range and the severe distortion at fast temporal modulations, are separately compensated for in the auditory system by contrast gain control and a shift in modulation sensitivity. The noise-robust neural representation of slow temporal modulations provides a plausible neural basis for noise-robust recognition

of speech and is directly correlated with individual subjects' ability to recognize speech in noise.

Reliable Neural Encoding of Slow Temporal Modulations of Speech

The slow temporal modulations of the speech reflect the syllabic structure of speech (Greenberg et al., 2003), and, in complex auditory scenes, serve as primary cues to bind acoustic features belonging to the same speech stream (Sheft, 2007). In parallel to the functional importance of the slow temporal modulations, neural activity synchronized to them has been hypothesized as a marker for the formation of a coherent neural representation of an auditory stream (Shamma et al., 2011), and also the neural basis for segmenting continuous speech into basic processing units, e.g. syllables, and allocating neural resources to the processing of each perceptual unit (Giraud and Poeppel, 2012; Schroeder and Lakatos, 2009).

A key prediction for these hypotheses is that neural synchronization to the speech envelope, i.e., spectrally coherent temporal modulations, is robust against any acoustic degradation that does not eliminate speech intelligibility, since the segregation of speech from acoustic background and the parsing of speech into perceptual units are both prerequisites for speech recognition. Consistent with this prediction, we demonstrate that the neural synchronization to slow modulations of speech is indeed resilient to the strong energetic masking of the background noise. Previous studies have demonstrated that the neural synchronization is also resilient to the strong informational masking of a competing speech stream. Taken together, therefore, it is now demonstrated that cortical

encoding of temporal modulations is robust to both energetic and informational masking, at least for those cases where speech remains partly intelligible. This reliable neural encoding of slow temporal modulations is likely a key step in transforming the variable acoustic properties of an auditory scene into a stable perception of a speech stream.

The robust neural encoding of slow temporal modulations is only achievable by complex neural computations, including what can be characterized as contrast gain control and long-term temporal integration, as will be discussed in the following.

Contrast Gain Control in Auditory Cortex

The dynamic range of speech is severely compressed by acoustic degradation such as background noise and reverberation. Therefore, to achieve robust speech recognition, an adaptive neural coding scheme for sound intensity is unavoidable. Indeed, in single unit studies with non-speech stimuli, neural adaptation to the mean and/or variance of sound intensity has been observed and the gain control effect enhances along the ascending auditory pathway (Dean et al., 2005; Robinson and McAlpine, 2009; Wen et al., 2009) (Rabinowitz et al., 2011; Watkins and Barbour, 2009; Zilany et al., 2009).

In this study, a hierarchy of contrast gain control is seen in auditory cortex. The early $M50_{STRF}$ component, localized to an area consistent with core auditory cortex (Chapter 4), is significantly weakened as the dynamic range of the stimulus is compressed by background noise, reflecting incomplete contrast gain control. Similar phenomena have been seen for the MEG auditory steady state response (aSSR) to 40-Hz amplitude modulations, which also has short latency and localizes to core auditory cortex (Ross et al., 2000). The aSSR is substantially weakened by a reduction of the stimulus

modulation depth (Ross et al., 2000) or an increase of the level of background noise, regardless of the subjects' attentional state (Okamoto et al., 2011). These MEG results are also consistent animal studies which demonstrate that neurons in core auditory cortex show contrast gain control but are still sensitive to the modulation depth of the stimulus (Malone et al., 2010; Rabinowitz et al., 2011).

In contrast, almost complete contrast gain control is seen in the long latency $M100_{STRF}$ component, localized to posterior association auditory cortex (Chapter 4). When the subjects actively listen to noise-corrupted speech, the amplitude of the $M100_{STRF}$ remains unaffected for all SNRs higher than -6 dB. Similarly, for subjects engaged in a syllable discrimination task, the EEG N1 response to isolated syllables (latency near 100 ms) is also stable to background noise, at least for positive SNRs (Kaplan-Neeman et al., 2006; Whiting et al., 1998). This robustness, however, is not observed during passive listening and therefore may require attention. The EEG N1 response to isolated syllables (Cunningham et al., 2001) or pure tones (Billings et al., 2009) is significantly weakened by background noise during passive listening. Similarly, the aSSR evoked by slow amplitude modulations (e.g. at 4 Hz), which has latency near 100 ms, also diminishes when the stimulus modulation depth decreases, during passive listening (Rees et al., 1986). In sum, neural adaptation to the dynamic range of stimulus enhances along the ascending auditory pathway, even from the shorter latency (~50 ms) response from core auditory cortex to the longer latency (> 100 ms) response from association auditory cortex.

Encoding of Slow Temporal Modulations and Long-term Integration

Noise-robust neural synchronization to speech is only observed in low frequency (< 4 Hz) neural activity (Fig. 5.5A). The precision of higher frequency neural synchronization (4-8 Hz) decreases continuously as the level of noise increases. This suggests that, in noisy environments, the stress information of speech, reflected by very slow (< 4 Hz) temporal modulations (Greenberg, 1999), is more reliably encoded in cortex than faster linguistic structures such as unstressed syllables and phonemes. This phenomenon may also be related to the intrinsic properties of cortical neural circuits, as delta (1-4 Hz) and theta (4-8 Hz) have been classified as two distinct frequency bands for cortical oscillations. The current results is highly consistent with the hypothesis that delta band cortical activity is a more fundamental rhythm regulating the excitability of neurons, while theta band activity is more closely tied to the physical properties of the sensory stimulus (Schroeder and Lakatos, 2009).

The robust neural synchronization to slow but not fast rhythms of speech reflects a change in the modulation transfer function (MTF), i.e. the cortical sensitivity to temporal modulations at different modulation rates. The cutoff frequency of the MEG measured MTF shifts towards low frequency as the level of noise increases (Fig. 5.4C). Similar plasticity of the modulation transfer function has also been demonstrated in individual neurons (Woolley et al., 2005; Woolley et al., 2006). Neurons in the midbrain also lose sensitivity to fast modulations when encoding animal vocalization in noise (Lesica and Grothe, 2008), and due to their anesthetized condition, this suggests a bottom-up contribution of the plasticity. Top-down attention, however, can also modulate the temporal properties of neurons, e.g. response latency and duration (Fritz et al., 2007).

Therefore, both bottom-up and top-down modulations may contribute to the noise-induced low-frequency shift of the MTF.

A decrease of the MTF cutoff frequency suggests a longer-term temporal integration in the auditory system. The involvement of long-term integration can also be seen from the elongation of neural response latency (Fig. 5.6A) and that the neural reconstruction of speech requires a temporal integration window over 100 ms at low SNRs (Fig. 5.5B). Noise-induced latency elongation has been commonly seen for EEG/MEG responses to sound onsets (Billings et al., 2009; Kaplan-Neeman et al., 2006). Moreover, such as delay in neural response is associated with an elongation in reaction time to discriminate syllables in noise (Whiting et al., 1998). The elongation of neural response latency and behavioral reaction time suggest the detection of sound target in noise requires integrating information over a longer time window in a noisy environment than in quiet condition. This is consistent with the optimal signal detection theory, which states that an accurate decision can only be made when enough information is cumulated, a process that will take longer if the less information can be extracted at each time moment due to noise (Gold and Shadlen, 2007).

Parsing of Continuous Speech and Intelligibility

The very slow temporal modulations of speech are accurately encoded in human auditory cortex until the SNR is as low as -6 dB. The intelligibility of speech, however, starts to decrease at +2 dB SNR. Therefore, the robustness of neural synchronization to speech is more likely to reflect the perception of the syllabic structure of speech rather than, for example, the decoding of lexical information. Parsing continuous speech into

syllables or phrases is a prerequisite for speech intelligibility and is more robust to noise than speech intelligibility. For example, listeners can reliably make use of stress cues to detect word boundaries, even at very low SNRs that allow little intelligibility (Woodfield and Akeroyd, 2010).

Although the grand averaged neural encoding accuracy does not predict speech intelligibility as a function of SNR, individual decoding accuracy does predict how well a subject can recognize speech in noise (Fig. 5.2C). This suggests that, in noise, the recognition of speech is limited by the neural processing in auditory cortex. More precise neural synchronization to speech is likely a marker of auditory system's success in extracting speech information, e.g. syllables, from the noisy stimulus.

In summary, this study demonstrates noise-robust neural synchronization to the slow temporal modulations of speech, even under the difficult condition of energetic masking. This neural synchronization is correlated with speech intelligibility in noise, and acts as a marker of the segregation of speech from the acoustic background.

Chapter 6

Summary and Future Work

6.1 Summary and general discussions

Cortical Restoration of Speech Embedded in a Complex Auditory Scene

Based on the three studies described in Chapter 3-5, it is well demonstrated that large-scale cortical activity measured by MEG is reliably synchronized to the temporal envelope of speech. In other words, the rhythm of auditory cortex, temporally coherent current flux in millions of neurons, is synchronized to the rhythm of speech, the temporally coherent variations of spectro-temporal features. Critically, this synchronization occurs robustly even in the presence of acoustic interference, which affects speech intelligibility through two distinct mechanisms, i.e. informational masking and energetic masking (Brungart, 2001; Durlach et al., 2003; Moore, 2003; Stone et al., 2011), both of which are addressed in this dissertation. Informational masking is caused the perceptual similarity between the competing auditory objects. For example, for the auditory scene studied in Chapters 3 and 4, where the two competing auditory objects are both audible and intelligible speech streams. In this scenario, the difficulty of speech recognition is to correctly *select* the auditory features belonging to the target speech stream. Energetic masking, in contrast, is caused by the physical, acoustic overlapping between auditory objects. For example, for the auditory scene studied in Chapter 5, the two auditory objects, speech and noise, are perceptually very distinct sounds. Nevertheless, the stationary noise causes strong masking effects since its energy strongly

overlaps with the energy of speech and therefore reduces the audibility of speech. Taking together the results from Chapter 3-5, it is clear that cortical synchronization to speech is robust to both informational and energetic masking (Fig. 6.1), and therefore is likely a general phenomenon underlying speech listening in complex auditory scenes. Furthermore, all experiments described in the dissertation use ecologically relevant paradigms, where the subjects are only instructed to listen to narrated stories and answer comprehension questions. Therefore, the results obtained here probably underlie the neural processing during everyday listening.

The reliable neural synchronization to the speech envelope implies several important computational properties of the auditory cortex. First, the human auditory cortex is sensitive to slow temporal modulations below 10 Hz (cf. Ding & Simon, 2009; Wang et al., 2012). The modulation transfer function (MTF) has a low-pass shape (Chapter 3) and the cut-off frequency shifts towards lower frequencies when the speech is corrupted by stationary noise (Chapter 5). Second, posterior association auditory cortex carries out object-based analysis. It selectively encodes the auditory object of the listener's interest rather than the raw acoustic scene (Chapter 4). Furthermore, the strength of the neural response to an auditory object is normalized: It is independent of the intensity of the encoded auditory object and the intensity of the interfering auditory objects, when the stimulus is comfortably loud (Chapter 4-5).

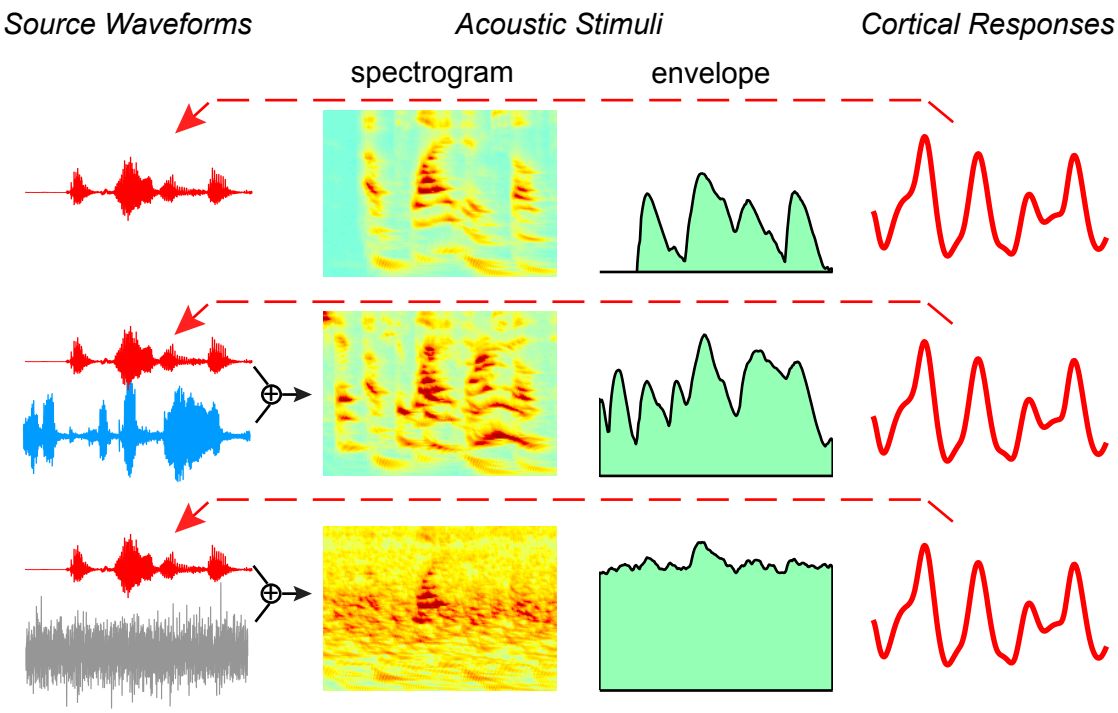

Figure 6.1. Illustrations of interfere-resilient neural synchronization to speech. Left panel: A speech signal (red) is presented either alone, or together with another speech signal (blue) or stationary noise (gray). Middle panels: When speech is mixed with other sounds, its spectro-temporal features are severely degraded. Right panel: *Top*, cortical activity is synchronized to the temporal envelope of speech, when it is presented alone (Chapter 3). *Middle and bottom*, cortical activity is selectively synchronized to the temporal envelope of the speech stream the listener attends to, in the presence of either a competing speech stream (Chapter 4) or a stationary noise (Chapter 5). In sum, cortical activity is reliably synchronized to the attended speech stream, even in complex listening environments.

Hierarchical Processing in Auditory Cortex

In Chapter 4 and 5, it is demonstrated that the $M50_{STRF}$ and $M100_{STRF}$ show differential behavior in tracking the speech envelope. In the two-speaker auditory scene, the $M50_{STRF}$ is not modulated by selective attention but the $M100_{STRF}$ is. Neither is affected by the level of the interfering speaker. In contrast, for speech embedded in noise, the $M50_{STRF}$ is weakened as the level of noise increases, while the $M100_{STRF}$ is not. These facts clearly distinguish the roles of the $M50_{STRF}$ and $M100_{STRF}$. The $M50_{STRF}$ probably reflect the audibility of a sound stream and does not reflect the selection of audible sound streams. Therefore, it is likely to be a neural representation of the physical properties of an acoustic scene. The $M100_{STRF}$, however, is robust to acoustic degradations and is involved in the top-down driven selection of auditory streams. Therefore, it is likely to represent the perceptually dominant auditory stream. Moreover, in Chapter 5, as the level of noise increases, the weakening of the $M50_{STRF}$ is accompanied by a lowering of the cutoff frequency of the MTF. Therefore, it is possible that the neural source of the $M50_{STRF}$ has a higher cutoff frequency than the neural source of the $M100_{STRF}$. This hypothesis, if true, provides further evidence for the hierarchical relationship between the $M50_{STRF}$ and $M100_{STRF}$, since the cut-off frequency of the MTF generally decreases along the ascending auditory pathway (Giraud et al., 2000; Lerner et al., 2011). Consistent with their functional hierarchy, the $M50_{STRF}$ has shorter latency and is localized to roughly core auditory cortex, and the $M100_{STRF}$ has longer latency and is localized to posterior auditory cortex. These results suggest that, in between shorter latency (~50 ms) activity in core auditory cortex and longer latency (~100 ms) activity in

posterior auditory cortex, the neural representation of an acoustic scene is transformed into a neural representation of the attended auditory stream.

Cortical Synchronization to Speech and Speech Recognition

Chapter 4 and 5 demonstrate that cortical synchronization to the speech envelope is robust to acoustic maskers, whether speech or noise, at least when the target-to-masker ratio (TMR) is above -6 dB. At -6 dB TMR, however, speech intelligibility drops to about 50% for a speech masker, and only about 10% for a noise masker. Therefore, cortical synchronization to speech is even more robust to acoustic interference than speech intelligibility, which is already known for its robustness. This is especially remarkable since speech-synchronized activity is precisely phase-locked and is generated form the auditory cortex, which is commonly supposed to encode the raw acoustics of the stimulus.

The robustness of speech-synchronized neural activity gives new insights into how speech is recognized in the human brain. In short, I would argue that speech recognition involves two fundamental processes, the *detection* of auditory elements form the target speech stream and the *recognition* of them, and that speech-synchronized cortical activity reflects the first process. According to this hypothesis, the robustness of speech-synchronized cortical activity suggests that, in adverse listening environments, the listeners can detect auditory elements from the target speech stream but may have difficulty retrieving linguistic information, e.g. phonetic categories, from them. This idea is further elaborated below.

First, I discuss the importance of parsing continuous speech into auditory elements that occur at a rate of a few Hertz, in line with the rhythm of speech-synchronized cortical activity. The recognition of speech converts a continuous sound signal into a string of discrete symbols. Such a process is not trivial, and the decoding of each symbol or each string of symbols requires integration of acoustic information over time, or, in other words, packaging acoustic features into auditory elements. The auditory element I propose here is the elementary unit that the auditory system analyzes integrally and interfaces with the language and memory systems to retrieve the linguistic information from (cf. Poeppel et al., 2008). The auditory element is an intermediate representation between the continuous acoustic stream of speech and the discrete symbols the brain decodes. It is not necessarily a discrete representation but it varies slowly, at the rate that discrete, categorical linguistic information is decoded.

The auditory elements must have the appropriate temporal granularity. On the one hand, they cannot be too long, to be processed by the physical circuitry of auditory cortex, which integrates information over hundreds of milliseconds (Eggermont, 2002; Wang et al., 2012). On the other hand they cannot be too short, to correspond to reliable linguistic information (Greenberg et al., 2003; Plomp, 2002). Because of these constraints, the auditory element should have a length of a few hundred milliseconds, or equivalently, a rate few hertz (Plomp, 2002; Poeppel, 2003). This time scale may ultimately originate from the rhythmic open-close alternation of the mouth and reflect the physical properties of human articulators (MacNeilage, 1998). Acoustically, this time scale corresponds to the slow temporal modulations of speech (Chi et al., 1999; Elliott

and Theunissen, 2009) and, linguistically, it corresponds to syllables or short phrases centered with a stressed syllable (Greenberg et al., 2003).

Second, I discuss the dissociation between the detection of auditory elements from the target speech stream and the recognition of them. The detection of auditory elements from only the target speech stream requires correctly identifying the sound source of a potential auditory element, i.e. whether it is from the target speech stream or the interfering streams. This identification of sound sources is distinctive from, and precedes the recognition, e.g. the decoding of phonetic information, of speech. Whereupon an auditory element from the target speech stream is detected, large populations of neurons are activated and devoted to the subsequent processing of that element, which gives rise to speech-synchronized MEG activity (cf. Chait et al., 2007 for MEG evidence of the detection of non-speech auditory elements). The subsequent processing of an element, including the decoding of phonemic information, however, may not be successful even if it consumes activity from millions of neurons, since it may require a high-fidelity spectro-temporal representation that is lost due to acoustic degradations. The distinction between auditory element detection and recognition has also been supported by psychoacoustical studies. It has been shown that in very challenging listening conditions, listeners maintain the ability to detect the boundaries between words in a sentence, but have difficulty correctly recognizing the words (Woodfield and Akeroyd, 2010).

Furthermore, from an ecological perspective, the detection and recognition of auditory elements are also distinct processes. Animals always need to detect meaningful auditory elements in their environments, and identify whether they are from a predator, a

prey, or a mate. (The time scale of the auditory element might be species dependent.) Although the detection of auditory elements and identification of the sound sources are critical for an animal's survival, fine-grained decoding of spectro-temporal information is seldom necessary until sophisticated vocalizations or speech is evolved. Most human languages have tens of phonemes and speech carries tens of phonemes per second. Therefore, the decoding of phonemic categories requires very fine spectro-temporal information and is naturally much more challenging than the detection of auditory elements from the target sound stream.

Last, I give a conceptual model of the generation of MEG activity: Individual neurons in the midbrain, thalamus, and possibly primary auditory cortex encode acoustic features (Nelken, 2008). Based on these microscopic, feature-based neural representations, a collective, mesoscopic neural representation specific to the attended speech stream is constructed. The construction of the mesoscopic representation is strongly influenced by top-down attention, bottom-up neural adaptation, and probably the context of the auditory stream (Holt, 2005; Jones et al., 2002), as well as the intrinsic oscillations of neuronal excitability (Schroeder and Lakatos, 2009). One computational strategy for the neural construction of such a stream-specific representation is the following. The auditory system keeps track of the neuronal representations of the acoustic features unique to the target speech stream, e.g. pitch, and then selectively routes neural activity temporally coherent with those neuronal representations (in the range of a few Hertz) into higher-level cortical networks (Shamma et al., 2011). This way, in the higher-level cortical networks, the spatial-temporal dynamics encode uniquely and collectively the attended speech stream. At each moment, the spatial activation pattern of

the neural networks encodes all available attributes of an auditory element, which is the basis for the recognition of that element (Chang et al., 2010; Formisano et al., 2008). The temporal dynamics of the spatial activation pattern, on the other hand, reflects the rate of the auditory elements being processed. This mesoscopic, spatial-temporal neural representation, when integrated over space, gives rise to the macroscopic speech-synchronized response measured by MEG. This model is summarized in Fig. 6.2.

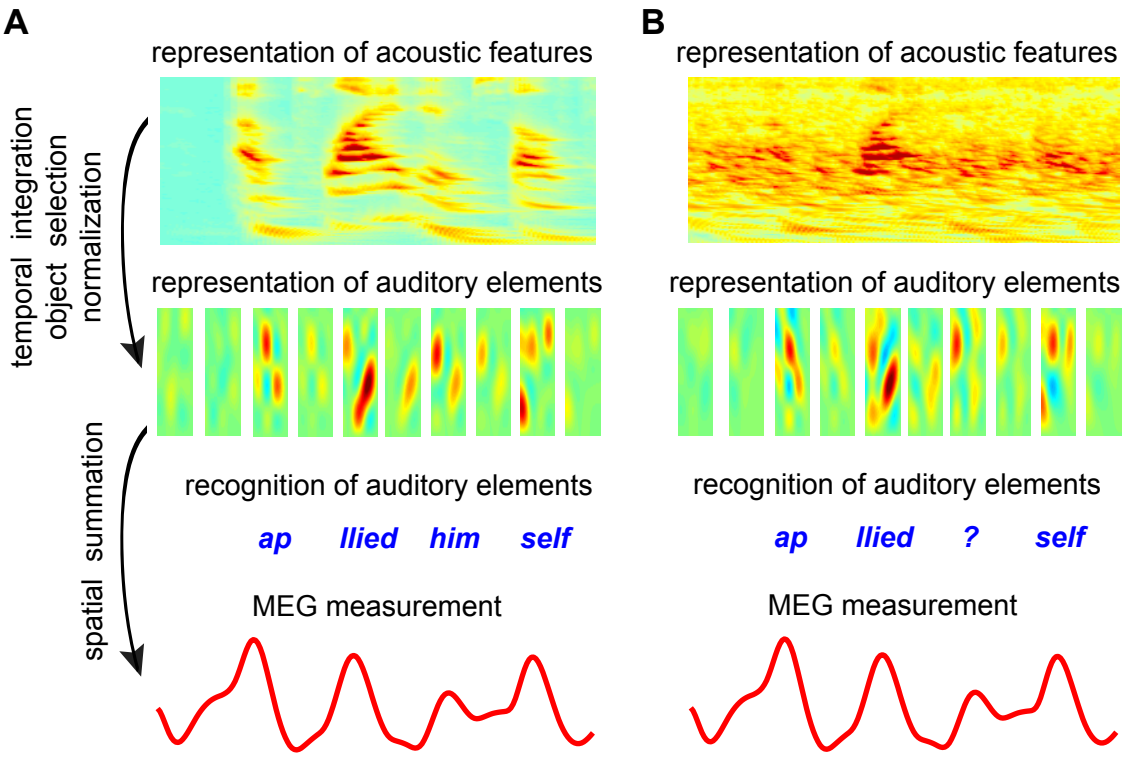

Figure 6.2. A model for human speech recognition and the generation of MEG activity. (A) *Top row*: In sub-cortical nuclei and possibly core auditory cortex, the spectro-temporal features of the stimulus are faithfully represented. In the figure, each row can be viewed as the time course of the response of a neuron. *Second row*: In the superior temporal gyrus (STG), which includes association

auditory cortex, auditory elements from the target speech stream are represented by a spatial-temporal code. Each plot in this row represents the activation pattern of a patch of auditory cortex, and the series of plots show the time evolution of the spatial cortical activation pattern. *Third row*: Linguistic content decoded from the auditory elements. *Last row*: MEG is sensitive to the neural representation of auditory elements. Nevertheless, attributable to the low spatial resolution of MEG, what is measured is the neural representation integrated over a large cortical volume and therefore only reflects the temporal dynamics of the neural representation. (B) The neural processing of noise-corrupted speech. The feature-based neural representation is strongly corrupted by noise (*top row*). The spatial but not the temporal pattern of the neural representations of auditory elements is corrupted by noise (*second row*). The syllabic structure of the speech stimulus is correctly identified, but, in this illustration, one syllable is not successfully recognized. The decoded message is therefore "applied ?self" (*third row*). The MEG response, which reflects the temporal dynamics of the neural representation of auditory elements, is not affected by the background noise and is a neural correlate to the robust perception of the syllabic and phrasal structure of speech (*last row*).

6.2 Future Work

The work in this dissertation provides a new paradigm to investigate cortical processing of speech. It reveals that cortical synchronization to speech is robust to acoustic interference in normal hearing human subjects. A future direction is to

investigate the cortical processing in hearing impaired listeners and elderly listeners, who have difficulty understanding speech in adverse listening environments (Bernstein and Grant, 2009; Pichora-Fuller and Souza, 2003), and to diagnose the possible central deficits of these listeners. Another future direction is to investigate the neural origin of MEG activity synchronized to speech. This requires ideally simultaneous MEG and high-density intracranial recording. Nevertheless, it could also be probed using ordinary single unit recording, by investigating what kind of neural measures, form which part of the cortex, show properties similar to those of the MEG response (Ding et al., 2012).

A distinct but also interesting future direction is to use the phenomena observed here to develop a brain-computer-interface (BCI) system, which has broad applications. For example, current hearing aid devices cannot recover the listeners' ability to recognize speech in complex auditory scenes (Bernstein and Grant, 2009). One way to facilitate speech/sound recognition in complex auditory scenes is to segregate the speech stream of the listener's interest and enhance it. To some extent, the segregation of concurrent sound sources can be achieved by directional microphone arrays. Nevertheless, it is not yet possible to determine which sound source is of the listener's interest and should be enhanced. One promising solution to this is to develop an EEG-based BCI system that decodes the listener's attentional focus, which is feasible as shown by this dissertation, and use it to guide a directional microphone array to selectively amplify the attended sound source.

Bibliography

Abrams, D.A., Nicol, T., Zecker, S., and Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. J. Neurosci. *28*.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M.M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. U. S. A. *98*, 13367-13372.

Ahveninen, J., Hämäläinen, M., Jääskeläinen, I.P., Ahlfors, S.P., Huang, S., Lin, F.-H., Raij, T., Sams, M., Vasios, C.E., and Belliveau, J.W. (2011). Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proc. Natl. Acad. Sci. U. S. A. *108*, 4182-4187.

Aiken, S.J., and Picton, T.W. (2008). Human cortical responses to the speech envelope. Ear Hear. *29*, 139-157.

Alaerts, J., Luts, H., Hofmann, M., and Wouters, J. (2009). Cortical auditory steady-state responses to low modulation rates. Int. J. Audiol. *48*, 582-593.

Baillet, S., Mosher, J.C., and Leahy, R.M. (2001). Electromagnetic brain mapping. IEEE Signal Processing Magazine *18*, 14-30.

Bell, A.J., and Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Comput. *7*, 1129-1159.

Bernstein, J.G.W., and Grant, K.W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. *125*, 3358-3372.

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

Biermann, S., and 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.

Billings, C.J., Tremblay, K.L., Stecker, G.C., and Tolin, W.M. (2009). Human evoked cortical activity to signal-to-noise ratio and absolute signal level. Hearing Res. *254*, 15-24.

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

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Muller, K.-R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Processing Magazine *25*, 41-56.

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

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

Brugge, J.F., Nourski, K.V., Oya, H., Reale, R.A., Kawasaki, H., Steinschneider, M., and Howard, M.A., III (2009). Coding of repetitive transients by auditory cortex on heschl's gyrus. J. Neurophysiol. *102*, 2358-2374.

Brungart, D.S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. *109*, 1101-1109.

Carandini, M., and Heeger, D.J. (2012). Normalization as a canonical neural computation. Nat. Rev. Neurosci. *13*, 51-62.

Chait, M., Poeppel, D., Cheveigné, A.d., and Simon, J.Z. (2007). Processing asymmetry of transitions between order and disorder in human auditory cortex. J. Neurosci. *27*, 5207-5214.

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

Chang, E., Rieger, J., Johnson, K., Berger, M., Barbaro, N., and Knight, R. (2010). Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. *13*, 1428-1432.

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

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

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

Cooke, M. (2006). A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. *119*, 1562-1573.

Cooke, M., Hershey, J.R., and Rennie, S.J. (2010). Monaural speech separation and recognition challenge. Computer Speech & Language *24*, 1-15.

Cover, T.M., and Thomas, J.A. (1991). Elements of information theory (New York: Wiley).

Cunningham, J., Nicol, T., Zecker, S.G., Bradlow, A., and Kraus, N. (2001). Neurobiologic responses to speech in noise in children with learning problems: Deficits and strategies for improvement. Clin. Neurophysiol. *112*, 758-767.

David, S.V., Mesgarani, N., Fritz, J.B., and Shamma, S.A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J. Neurosci. *29*, 3374-3386.

David, S.V., Mesgarani, N., and Shamma, S.A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. Network: Comput. Neural Syst. *18*, 191 - 212.

de Cheveigné, A. (2005). Pitch perception models. In Pitch, C. Plack, R. Fay, A. Oxenham, and A. Popper, eds. (New York: Springer).

de Cheveigné, A., and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. *111*, 1917-1930.

de Cheveigné, A., and Simon, J.Z. (2007). Denoising based on time-shift PCA. J. Neurosci. Methods *165*, 297-305.

de Cheveigné, A., and Simon, J.Z. (2008). Denoising based on spatial filtering. J. Neurosci. Methods *171*, 331-339.

Dean, I., Harper, N.S., and McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. J. Neurosci. *8*, 1684-1689.

deCharms, R.C., Blake, D.T., and Merzenich, M.M. (1998). Optimizing sound features for cortical neurons. Science *280*, 1439 - 1444.

Depireux, D.A., Simon, J.Z., Klein, D.J., and Shamma, S.A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. *85*, 1220-1234.

Ding, N., Shamma, S.A., Simon, J.Z., and David, S.V. (2012). Breaking down the cortical representations of speech in LFP and MUA. In Research in Otolaryngology MidWinter Meeting (ARO) (San Diego, CA).

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

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

Durlach, N.I., Mason, C.R., Gerald Kidd, J., Arbogast, T.L., Colburn, H.S., and Shinn-Cunningham, B.G. (2003). Note on informational masking (L). J. Acoust. Soc. Am. *113*, 2984-2987.

Eggermont, J.J. (2002). Temporal modulation transfer functions in cat primary auditory cortex: Separating stimulus effects from neural mechanisms. J. Neurophysiol. *87*, 305-321.

Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z., and Shamma, S.A. (2004). Dynamics of precise spike timing in primary auditory cortex. J. Neurosci. *24*, 1159-1172.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., and Shamma, S.A. (2009a). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron *61*, 317-329.

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

Elliott, T., and Theunissen, F. (2009). The modulation transfer function for speech intelligibility. PLoS Comp. Biol. *5*.

Escabí, M.A., Miller, L.M., Read, H.L., and Schreiner, C.E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J. Neurosci. *23*, 11489-11504.

Festen, J.M., and 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.

Fishman, Y.I., and Steinschneider, M. (2010). Formation of auditory streams. In The oxford handbook of auditory science: The auditory brain, A. Rees, and A. Palmer, eds. (New York), pp. 215–245.

Formisano, E., Martino, F.D., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. Science *322*, 970-973.

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

Fritz, J.B., Elhilali, M., David, S.V., and Shamma, S.A. (2007). Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? Hearing Res. *229*, 186-203.

Fuentemilla, L.I., Marco-Pallarés, J., and Grau C. (2006). Modulation of spectral power and of phase resetting of EEG contributes differentially to the generation of auditory event-related potentials. NeuroImage *30*, 909-916.

Fukunaga, K. (1972). Introduction to statistical pattern recognition (New York: Academic Press).

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S.J., and Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron *56*, 1127-1134.

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

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. Nat. Neurosci. *15*, 511-517.

Gold, B., and Morgan, N. (2000). Speech and audio signal processing (New York: John Wiley & Sons, Inc.).

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535 -574.

Greenberg, S. (1999). Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. Speech Commun. *29*, 159-176.

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

Greenwood, D.D. (1990). A cochlear frequency-position function for several species - 29 years later. J. Acoust. Soc. Am. *87*, 2592-2605.

Griffiths, T.D., and Warren, J.D. (2002). The planum temporale as a computational hub. Trends Neurosci. *25*, 348-353.

Griffiths, T.D., and Warren, J.D. (2004). What is an auditory object? Nat. Rev. Neurosci. *5*, 887–892.

Gutschalk, A., Micheyl, C., and Oxenham, A.J. (2008). Neural correlates of auditory perceptual awareness under informational masking. PLoS Biol. *6*.

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

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

Hämäläinen, M., and Ilmoniemi, R. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. Med. Biol. Eng. Comput. *32*, 35-42.

Hari, R., Hämäläinen, M., Kaukoranta, E., Mäkelä, J., Joutsiniemi, S.L., and Tiihonen, J. (1988). Selective listening modifies activity of the human auditory cortex. Exp. Brain Res. *74*.

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

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

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

Hillyard, S.A., Hink, R.F., Schwent, V.L., and Picton, T.W. (1973). Electrical signs of selective attention in the human brain. Science *182*, 177 - 180.

Holt, L.L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol. Sci. *16*, 305-312.

Hudspeth, A.J. (2008). Making an effort to listen: Mechanical amplification in the ear. Neuron *59*, 530-545.

Jones, M.R., Moynihan, H., MacKenzie, N., and Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. Psychol. Sci. *13*, 313-319.

Jorgensen, S., and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. *130*, 1475-1487.

Joris, P.X., Carney, L.H., Smith, P.H., and Yin, T.C. (1994). Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency. J. Neurophysiol. *71*, 1022-1036.

Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. U. S. A., 11793-11799.

Kaplan-Neeman, R., Kishon-Rabin, L., Henkin, Y., and Muchnik, C. (2006). Identification of syllables in noise: Electrophysiological and behavioral correlates. J. Acoust. Soc. Am. *120*, 926-933.

Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352-355.

Kerlin, J.R., Shahin, A.J., and Miller, L.M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. *30*, 620-628.

Koskinen, M., Viinikanoja, J., Kurimo, M., Klami, A., Kaski, S., and Hari, R. (2012). Identifying fragments of natural speech from the listener's MEG signals. Hum. Brain Mapp., doi: 10.1002/hbm.22004.

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

Lalor, E.C., and Foxe, J.J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. *31*, 189-193.

Lalor, E.C., Power, A.J., Reilly, R.B., and Foxe, J.J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. J. Neurophysiol. *102*, 349-359.

Lazzouni, L., Ross, B., Voss, P., and Lepore, F. (2010). Neuromagnetic auditory steady-state responses to amplitude modulated sounds following dichotic or monaural presentation. Clin. Neurophysiol. *121*, 200-207.

Lerner, Y., Honey, C.J., Silbert, L.J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. *31*, 2906-2915.

Lesica, N., and Grothe, B. (2008). Efficient temporal processing of naturalistic sounds. PLoS ONE *3*, e1655. doi:1610.1371/journal.pone.0001655.

Liang, L., Lu, T., and Wang, X. (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. J. Neurophysiol. *87*, 2237-2261.

Liegeois-Chauvel, C., Lorenzi, C., Trebuchon, A., Regis, J., and Chauvel, P. (2004). Temporal envelope processing in the human left and right auditory cortices. Cereb. Cortex *14*, 731-740.

Linden, R.D., Picton, T.W., Hamel, G., and Campbell, K.B. (1987). Human auditory steady-state evoked potentials during selective attention. Electroencephalogr. Clin. Neurophysiol. *66*, 145-159.

Lippmann, R.P. (1997). Speech recognition by machines and humans. Speech Commun. *22*, 1-15.

Logothetis, N.K. (2003). The underpinnings of the bold functional magnetic resonance imaging signal. J. Neurosci. *23*, 3963-3971.

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

Lütkenhöner, B., and Mosher, J.C. (2006). Source analysis of auditory evoked potentials and fields. In Auditory evoked potentials: Basic principles and clinical application, R.F. Burkard, M. Don, and J.J. Eggermont, eds. (Philadelphia: Lippincott Williams & Wilkins).

Lütkenhöner, B., and Steinsträter, O. (1998). High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiol. Neurootol. *3*, 191-213.

Lyon, R., and Shamma, S. (1996). Auditory representations of timbre and pitch. In Auditory computation, H.L. Hawkins, T.A. McMullen, A.N. Popper, and R.R. Fay, eds. (Springer).

MacNeilage, P.F. (1998). The frame/content theory of evolution of speech production Behav. Brain Sci. *21*, 499–546.

Mallat, S. (1999). A wavelet tour of signal processing (London: Academic Press).

Malone, B.J., Scott, B.H., and Semple, M.N. (2010). Temporal codes for amplitude contrast in auditory cortex. J. Neurosci. *30*, 767-784.

Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2008). Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am. *123*, 899-909.

Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J. Neurophysiol. *102*, 3329-3339.

Millman, R.E., Woods, W.P., and Quinlan, P.T. (2011). Functional asymmetries in the representation of noise-vocoded speech. NeuroImage *54*, 2364-2373.

Moore, B.C.J. (2003). Introduction to the psychology of hearing, 5 edn (Boston: Academic Press).

Mosher, J.C., Baillet, S., and Leahy, R.M. (2003). Equivalence of linear approaches in bioelectromagnetic inverse solutions. In IEEE Workshop on Statistical Signal Processing (St. Louis).

Mosher, J.C., Leahy, R.M., and Lewis, P.S. (1999). EEG and MEG: Forward solutions for inverse methods. IEEE Trans. Biomed. Eng. *46*, 245-259.

Näätänen, R., and 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.

Nagel, K.I., and Doupe, A.J. (2006). Temporal processing and adaptation in the songbird auditory forebrain. Neuron *51*, 845-859.

Narayan, R., Best, V., Ozmeral, E., McClaine, E., Dent, M., Shinn-Cunningham, B., and Sen, K. (2007). Cortical interference effects in the cocktail party problem. Nat. Neurosci. *10*, 1601 - 1607.

Nelken, I. (2008). Processing of complex sounds in the auditory system. Curr. Opin. Neurobiol. *18*, 413-417.

Nelken, I., and Bar-Yosef, O. (2008). Neurons and objects: The case of auditory cortex. Frontiers in Neuroscience *2*, 107-113.

Nelken, I., Rotman, Y., and Yosef, O.B. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. Nature *397*, 154-157.

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Matthew A. Howard, I., and Brugge, J.F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. *29*, 15564-15574.

Okamoto, H., Stracke, H., Bermudez, P., and Pantev, C. (2011). Sound processing hierarchy within human auditory cortex. J. Cognit. Neurosci. *23*, 1855-1863

Pantev, C., Lütkenhöner, B., Hoke, M., and Lehnertz, K. (1986). Comparison between simultaneously recorded auditory-evoked magnetic fields and potentials elicited by ipsilateral, contralateral and binaural tone burst stimulation. Int. J. Audiol. *25*, 54-61.

Panzeri, S., Brunel, N., Logothetis, N.K., and Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. Trends Neurosci. *33*, 111-120.

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

Pichora-Fuller, M.K., and Souza, P.E. (2003). Effects of aging on auditory processing of speech. Int. J. Audiol. *42*, 11-16.

Pickles, J.O. (1988). An introduction to the physiology of hearing (Academic Press).

Picton, T.W., Skinner, C.R., Champagne, S.C., Kellett, A.J.C., and Maiste, A.C. (1987). Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone. J. Acoust. Soc. Am. *82*, 165-178.

Plomp, R. (2002). The intelligent ear: On the nature of sound perception (Mahwah, NJ: Lawrence Erlbaum Associates).

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. Speech Commun. *41*, 245-255.

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

Poeppel, D., Yellin, E., Phillips, C., Roberts, T.P.L., Rowley, H.A., Wexler, K., and Marantz, A. (1996). Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. Cognitive Brain Research *4*, 231-242.

Power, A.J., Lalor, E.C., and Reilly, R.B. (2010). Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. Cereb. Cortex.

Qiu, A., Schreiner, C.E., and Escabí, M.A. (2003). Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. J. Neurophysiol. *90*, 456-476.

Rabinowitz, N.C., Willmore, B.D.B., Schnupp, J.W.H., and King, A.J. (2011). Contrast gain control in auditory cortex. Neuron *70*, 1178-1191.

Raichle, M.E. (1983). Positron emission tomography. Annu. Rev. Neurosci. *6*, 249 - 267.

Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. Nat. Neurosci. *12*, 718-724.

Rees, A., Green, G., and Kay, R. (1986). Steady-state evoked responses to sinusoidally amplitude-modulated sounds recorded in man. Hearing Res. *23*, 123-133.

Rif, J., Hari, R., Hämäläinen, M.S., and Sams, M. (1991). Auditory attention affects two different areas in the human supratemporal cortex. Electroencephalogr. Clin. Neurophysiol. *79*, 464 - 472.

Robinson, B.L., and McAlpine, D. (2009). Gain control mechanisms in the auditory pathway. Curr. Opin. Neurobiol. *19*, 402-407.

Robinson, S.E., and Vrba, J. (1999). Functional neuroimaging by synthetic aperture magnetometry (sam). In Recent advances in biomagnetism, T. Yoshimoto, M. Kotani, S. Kuriki, H. Karibe, and N. Nakasato, eds. (Sendai: Tohoku University Press), pp. 302-305.

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

Ross, B., Borgmann, C., Draganova, R., Roberts, L.E., and Pantev, C. (2000). A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones. J. Acoust. Soc. Am. *108*, 679-691.

Ross, B., Herdman, A.T., and Pantev, C. (2005). Right hemispheric laterality of human 40 Hz auditory steady-state responses. Cereb. Cortex *15*, 2029-2039.

Ross, B., Hillyard, S.A., and Picton, T.W. (2010). Temporal dynamics of selective attention during dichotic listening. Cereb. Cortex *20*, 1360-1371.

Särelä, J., and Valpola, H. (2004). Denoising source separation. The Journal of Machine Learning Research *6*, 233 - 272.

Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fmri. Proc. Natl. Acad. Sci. U. S. A. *106*, 14611-14616.

Schroeder, C.E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. *32*, 9-18.

Scott, S.K., Rosen, S., Beaman, C.P., Davis, J.P., and Wise, R.J.S. (2009). The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. J. Acoust. Soc. Am. *125*, 1737-1743.

Scott, S.K., Rosen, S., Wickham, L., and Wise, R.J.S. (2004). A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. J. Acoust. Soc. Am. *115*, 813-821

Shamma, S. (2006). Analysis of speech dynamics in the auditory system. In Dynamics of speech production and perception: Life and behavioural sciences, D. P, G. S, and M. G, eds. (Washington DC: IOS Press), pp. 335–342.

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

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science *270*, 303-304.

Sheft, S. (2007). Envelope processing and sound-source perception. In Auditory perception of sound sources, W.A. Yost, A.N. Popper, and R.R. Fay, eds. (New York: Springer).

Shinn-Cunningham, B.G. (2008). Object-based auditory and visual attention. Trends Cog. Sci. *12*, 182-186.

Shomstein, S., and Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. J. Neurosci. *26*, 435-439.

Simon, J.Z., Depireux, D.A., Klein, D.J., Fritz, J.B., and Shamma, S.A. (2007). Temporal symmetry in primary auditory cortex: Implications for cortical connectivity. Neural Comput. *19*, 583-638.

Smith, K.R., Hsieh, I.-H., Saberi, K., and Hickok, G. (2010). Auditory spatial and object processing in the human planum temporale: No evidence for selectivity. J. Cognit. Neurosci. *22*, 632-639

Snyder, J.S., Gregg, M.K., Weintraub, D.M., and Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. Frontiers in Psychology *3*.

Stickney, G.S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. J. Acoust. Soc. Am. *116*, 1081-1091.

Stone, M.A., Fullgrabe, C., Mackinnon, R.C., and Moore, B.C.J. (2011). The importance for speech intelligibility of random fluctuations in "steady" background noise. J. Acoust. Soc. Am. *130*, 2874-2881.

Tervaniemi, M., and Hugdahl, K. (2003). Lateralization of auditory-cortex functions. Brain Res. Rev. *43*, 231-246.

Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Comput. Neural Syst. *12*, 289-316.

Uutela, K., Hamalainen, M., and Salmelin, R. (1998). Global optimization in the localization of neuromagnetic sources. IEEE Trans. Biomed. Eng. *45*, 716-723.

Van Veen, B.D., Van Drongelen, W., Yuchtman, M., and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. *44*, 867-880.

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

Wang, X., Lu, T., and Liang, L. (2003). Cortical processing of temporal modulations. Speech Commun. *41*, 107-121.

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

Watkins, P.V., and Barbour, D.L. (2009). Specialized neuronal adaptation for preserving input sensitivity. Nat. Neurosci. *11*, 1259 - 1261.

Wen, B., Wang, G.I., Dean, I., and Delgutte, B. (2009). Dynamic range adaptation to sound level statistics in the auditory nerve. J. Neurosci. *29*, 13797-13808.

Whiting, K.A., Martin, B.A., and Stapells, D.R. (1998). The effects of broadband noise masking on cortical event-related potentials to speech sounds /ba/ and /da/. Ear Hear. *19*, 218-231.

Woldorff, M.G., Gallen, C.C., Hampson, S.A., Hillyard, S.A., Pantev, C., Sobel, D., and Bloom, F.E. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proc. Natl. Acad. Sci. U. S. A. *90*, 8722-8726.

Woodfield, A., and Akeroyd, M.A. (2010). The role of segmentation difficulties in speech-in-speech understanding in older and hearing-impaired adults. J. Acoust. Soc. Am. *128*, EL26-EL31.

Woolley, S.M.N., Fremouw, T.E., Hsu, A., and Theunissen, F.E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat. Neurosci. *8*, 1371 - 1379.

Woolley, S.M.N., Gill, P.R., and Theunissen, F.E. (2006). Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. J. Neurosci. *26*, 2499-2512.

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

Yang, X., Wang, K., and Shamma, S.A. (1992). Auditory representations of acoustic signals. IEEE Trans. Info. Theory *38*, 824-839.

Young, E.D. (2008). Neural representation of spectral and temporal information in speech. Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci. *363*, 923-945.

Zatorre, R.J., Bouffard, M., Ahad, P., and Belin, P. (2002). Where is 'where' in the human auditory cortex? Nat. Neurosci. *5*, 905-909.

Zeng, F.-G., Nie, K., Stickney, G.S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (2005). Speech recognition with amplitude and frequency modulations. Proc. Natl. Acad. Sci. U. S. A. *102*, 2293-2298.

Zilany, M.S.A., Bruce, I.C., Nelson, P.C., and Carney, L.H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. J. Acoust. Soc. Am. *126*, 2390-2412.