From Neurons to Behavior

Doctoral Thesis

From Neurons to Behavior:
Investigating Auditory Information Processing across Multiple Scales

Isma Zulfiqar
Maastricht Centre for Systems Biology (MaCSBio)
Maastricht University
2021

ISBN: 978-94-6423-278-3
Printed by ProefschriftMaken || www.proefschriftmaken.nl
Cover design by Isma Zulfiqar and Marian Sloot || www.proefschriftmaken.nl

From Neurons to Behavior:
Investigating Auditory Information Processing across Multiple Scales

Dissertation

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus, Prof.dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on Wednesday the 9th of June 2021, at 13:00 hours

by

Isma Zulfiqar

Supervisors
Prof. Dr. Elia Formisano
Prof. Dr. Peter De Weerd

Co-supervisor
Dr. Michelle Moerel

Assessment Committee
Prof. Dr. Ilja C. W. Arts (chair)
Prof. Dr. Uta Noppeney, Radboud University, Nijmegen
Dr. Mario Senden
Prof. Dr. Kâmil Uludağ, University of Toronto, Canada

The work in this thesis was supported by the Dutch Province of Limburg.

میرے پیارے والدین کے نام

Table of Contents

Chapter 1

Introduction

In his novel *East of Eden*, John Steinbeck writes, *"Maybe the knowledge is too great and maybe men are growing too small... Maybe kneeling down to atoms, they're becoming atom-sized in their souls. Maybe a specialist is only a coward, afraid to look out of his little cage. And think what any specialist misses—the whole world over his fence."* Albeit probably unintended, Steinbeck makes an excellent argument for research that crosses the borders of a single discipline. The ideas against the fragmentation of science into discrete disciplines can be traced to philosophers as early as the sixteenth century. By the late twentieth century, the realization that specialist research misses relevant parallels and developments outside its own scope, fueled the use of interdisciplinary approaches to tackle the key problems facing society (Ledford, 2015). The key to mobilizing collaboration across the table has been to develop a shared sense of research goals and take practical steps towards bridging the cross-disciplinary gaps (Brown et al., 2015). In biology, these efforts have translated into research that links phenomenological observations obtained via experimental methods to physiological mechanisms; modeled together through computational and mathematical tools [referred to as systems biology, (Kitano, 2002)].

The need for research beyond a discipline holds especially true for brain research. The human brain is the most complex biological system in the universe. In order to understand it, we must study it from the cellular (micro) level to the behavioral (macro) level (Figure 1). Neuroscience has therefore evolved as a multidisciplinary "science", encompassing among others molecular biology, psychology, physiology, medicine, mathematics, and computer science (Grant, 2003). Neuroscientists face the vital challenge of relating insights from these different subfields of neuroscience to each other. Within neuroscience, research on the sensory systems and the neural processing of sensory information is of primary interest as it is essential to understand how the brain generates representations of objects and events in the environment, and thereby makes sense of the external world. The present thesis focuses on *audition*. The ability to hear and interpret the sounds around us is not only necessary for survival but also enriches our life with interpersonal communication. However, how our brain makes sense of the auditory information remains unknown. Modern systems neuroscience of audition attempts to unify the understanding of hearing by linking different scales of research on auditory processing, ranging from cellular (micro) recordings in animal models to neuroimaging and behavioral (macro) observations in humans. However, it remains a challenge to meaningfully integrate the observations and results obtained in different species, with different methods and at different resolutions (spatial and temporal). A possible solution to tackle the lack of integration across subdivisions of neuroscience is the use of computational modeling. Depending on specific modeling goals and available computational resources, these models range from single neuron models (micro) to models of population-level responses (meso), and large-scale networks across brain regions or behavior (macro).

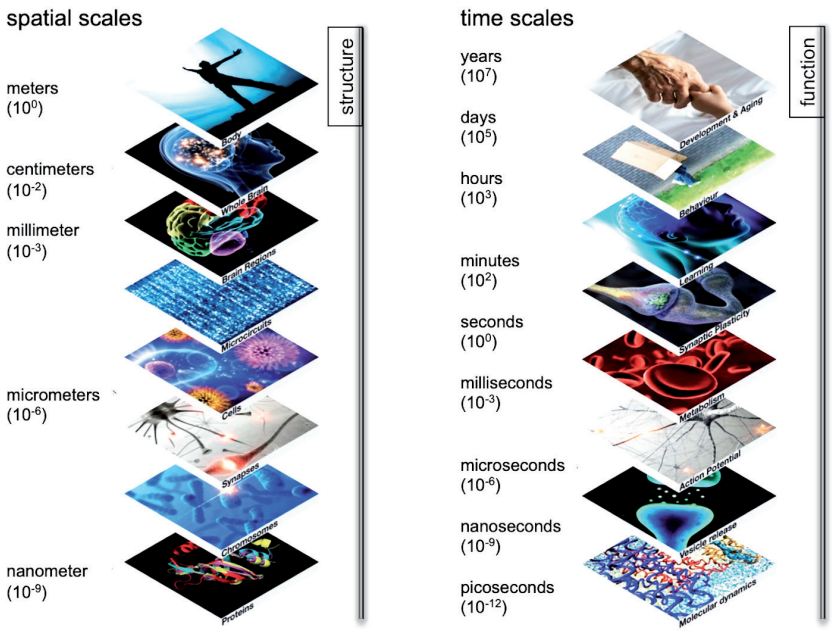


Figure 1: Spatial and time scales of neuroscience. Relevant brain processes occur at different spatial and temporal scales. While specialized methodology exists to study each of these different scales, the challenge lies in integrating available information and create a holistic view of brain functioning. Reproduced with permission from (Frackowiak and Markram, 2015).

This thesis tackles the challenge of creating a unified view of auditory processing in the auditory cortex (AC) using computational tools and empirical data originating from diverse techniques across different spatial and temporal scales. To that effect, we developed a computational model of sound processing in the AC, which integrates existing knowledge from electrophysiological and psychophysical observations. The model was then employed to investigate the neuronal underpinnings of the neuroimaging data. Furthermore, the thesis explores the multisensory (i.e., visual) influences on the information processing in the AC. Collected multisensory data may be integrated into the proposed AC model, thereby inching towards a holistic view of information processing in the AC. With this approach, we merged results from the small spatial scale of neuronal firing (as observed in animal models) to the large scale of human behavior.

The current chapter introduces the fundamentals of auditory information processing; it describes the hierarchical auditory pathway, focusing on the distinct neuronal and functional characteristics of sound processing streams at the level of the cortex. This chapter also introduces multisensory processing in the AC, as recent evidence from anatomical and functional studies suggests a possible role of the early auditory regions in multisensory processing (Falchier et al., 2002, 2010; Kayser et al., 2008, 2010; Schroeder

and Lakatos, 2009; Atilgan et al., 2018; Gau et al., 2020). These findings drive the experimental studies reported in the second part of the thesis.

1 The Auditory Cortex

After sounds reach the ear, processing of incoming sound signals in the auditory periphery, subcortical auditory structures, the AC, and subsequent higher-order cortical regions allows us to perceive, recognize, and respond to sound sources in our environment. At the periphery, the information processing begins at the level of the outer ear. The vibrations in the air are converted to mechanical reverberations in the middle ear and are then transduced to electrical signals in the inner ear where the cochlea maps the frequencies of the vibrations onto a spatial axis. That is, different spatial locations along the spiral of the cochlea preferentially respond to specific frequencies of the sound waveform (with low to high frequencies being coded from the apex to the base of the cochlea). This spatial representation of sound frequency is referred to as tonotopy. The tonotopically transformed information is passed on to the eighth cranial nerve and processed by a series of subcortical nuclei [including the cochlear nucleus, superior olivary complex, inferior colliculus, and medial geniculate body (MGB) of the thalamus] before it reaches the AC.

Over the past five decades, advances in research techniques have enabled researchers to collect and analyze a tremendous amount of data on the anatomy and function of the AC. These techniques vary from cyto- and myeloarchitectural and tract-tracing studies, to intracranial recordings of a single cell and small neuronal populations in animal models, to investigations in humans using both non-invasive techniques [such as magnetoencephalography (MEG), electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), and positron emission tomography (PET)] and invasive techniques [such as electrocorticography (ECoG) recordings from epilepsy patients]. These data act as resources to understand the anatomical architecture of the AC, the processing hierarchy, and the connectivity amongst subcortical-to-cortical and cortico-cortical auditory processing stages.

The AC is located on the superior temporal plane, and - in the human brain - is largely hidden within the lateral sulcus (Hackett et al. 2011; Hackett et al., 1998; Sweet, Dorph-Petersen, and Lewis 2005; Kaas and Hackett 2000). The information arrives at the AC through three distinct types of projections originating from MGB, namely lemniscal, non-lemniscal, and multisensory pathways (Rouiller et al., 1991; de la Mothe, 2016). These projections originate from different divisions of the MGB [ventral: lemniscal, dorsal: non-lemniscal, medial: multisensory (Aitkin et al., 1972; Calford and Aitkin, 1983)]. While the connectivity between the lemniscal ventral MGB and the AC is well

described, the thalamocortical non-lemniscal and multisensory pathways remain less understood (de la Mothe, 2016).

The hierarchical organization of the human AC into three regions (core – belt – parabelt) is rooted in non-human primate models (Kaas and Hackett, 2000; Hackett et al., 2011; Hackett et al., 1998; Rauschecker et al., 1995). A homologous organization has been replicated in humans (Sweet et al., 2005). The core regions are the first stage of auditory cortical processing. The core receives the majority of input from the ventral division of the MGB (Andersen et al., 1980; Calford and Aitkin, 1983). There is evidence of two subdivisions of the core areas in humans; primary auditory cortex (A1) and a rostral (R) core area (Galaburda and Sanides, 1980; Rivier and Clarke, 1997; Wallace et al., 2002). The core regions project to surrounding belt regions [six subdivisions reported in humans (Wallace et al., 2002)], which in turn project to the parabelt regions [two subdivisions (Hackett et al., 2011)].

Through its lemniscal input from the ventral subdivision of the MGB, the auditory core areas receive tonotopically-organized input (Andersen et al., 1980; Calford and Aitkin, 1983). This tonotopic organization seems, at least in part, to be preserved throughout the AC hierarchy, resulting in multiple topographic maps of frequency preference as established using a variety of stimuli and imaging methods in humans (Formisano et al., 2003; Moerel et al., 2012; Su et al., 2014) and non-human primates (Bendor and Wang, 2008; Merzenich and Brugge, 1973; Kuśmierek and Rauschecker, 2009). Frequency preference shows a columnar organization [i.e., it is preserved throughout the cortical depth of the AC (Abeles and Goldstein, 1970; Shamma et al., 1993; De Martino et al., 2015; Tischbirek et al., 2019)].

1.1 Information Processing Pathways

The core – belt – parabelt hierarchy processes auditory information sequentially. That is, the belt regions receive their input from the core and project heavily to parabelt, while the parabelt does not receive major input from the core regions (Hackett et al., 1998; Rauschecker et al., 1997). This connectivity-based hierarchy is reflected in the neuronal responses to sounds, which grow increasingly complex when moving through the auditory cortical stages. Neurons in the core regions show sharper frequency tuning and faster temporal dynamics in comparison with the belt regions (which display broader tuning and slower temporal dynamics) (Rauschecker et al., 1997; Recanzone et al., 2000). The parabelt shows even slower temporal dynamics (Camalier et al., 2012).

Apart from serial processing, the information in the AC is also processed in parallel by two anatomically distinct streams. The *rostral* or *ventral* stream originates in areas located rostrally to the primary auditory core and projects via the anterior temporal

lobe to the ventral regions of the frontal cortex. The *caudal* or *dorsal* stream originates in areas located caudally to the primary core and projects via the parietal cortex to dorsal frontal regions (Scott et al., 2017). These two processing streams show distinct neuronal properties (Jasmin et al., 2019; Zulfiqar et al., 2020). Compared to primary and surrounding auditory areas, neurons in the rostral field exhibit longer response latencies and narrower frequency tuning (Recanzone et al., 2000; Tian et al., 2001; Bendor and Wang, 2008; Camalier et al., 2012). Instead, neurons in the caudal fields respond with shorter latencies, comparable to or even shorter than those in A1, and have broader frequency tuning (Recanzone et al., 2000; Kuśmierek et al., 2014). These streams are hypothesized to process the incoming sound in parallel, with each stream representing the sound at a different spectro-temporal resolution (Schönwiesner and Zatorre, 2009; Santoro et al., 2014).

This parallel information processing along the rostral-caudal axis is hypothesized to underlie auditory cognition. The rostral and caudal stream have been hypothesized to reflect specialized mechanisms of sound analysis for deriving semantic information ("what" processing) and processing sound location and sound movement ("where" processing), respectively (Kaas et al., 1999; Romanski et al., 1999b; Belin et al., 2000; Kaas and Hackett, 2000; Rauschecker and Tian, 2000; Tian et al., 2001; Arnott et al., 2004; Jasmin et al., 2019). Recent human neuroimaging studies reported evidence in support of the existence of parallel processing streams in the human AC. Along the rostral-caudal axis of the human AC, a spectro-temporal trade-off has been observed. That is, the fine-grained spectral properties of sounds were shown to be analyzed optimally in rostral auditory regions, whereas the fine-grained temporal properties were analyzed optimally in caudal regions (Schönwiesner and Zatorre, 2009; Santoro et al., 2014). However, the hemodynamic blood oxygenation level-dependent (BOLD) signals measured with fMRI are an indirect measure of neuronal activity. BOLD signals originate from vascular changes (i.e., changes in blood oxygenation, blood flow, and blood volume) in response to neuronal activity. While the resulting fMRI signal is correlated to the underlying neuronal activity (Logothetis et al. 2001; Logothetis et al. 1999; Rees et al., 2000), it does not directly measure the neuronal activity. Thus, it remains to be determined how the observed spectro-temporal preferences along the rostral-caudal streams relate to neuronal mechanisms.

1.2 Multisensory Processing in the Auditory Cortex

Considering sounds independently of the other sensory signals only provides a limited view of information processing in the AC. Our environment is bursting with multisensory information that forms our percept of the world around us. Traditional models of late cortical integration suggest that multisensory integration is a function of higher-order association cortices which combine the information processed by early sensory sites.

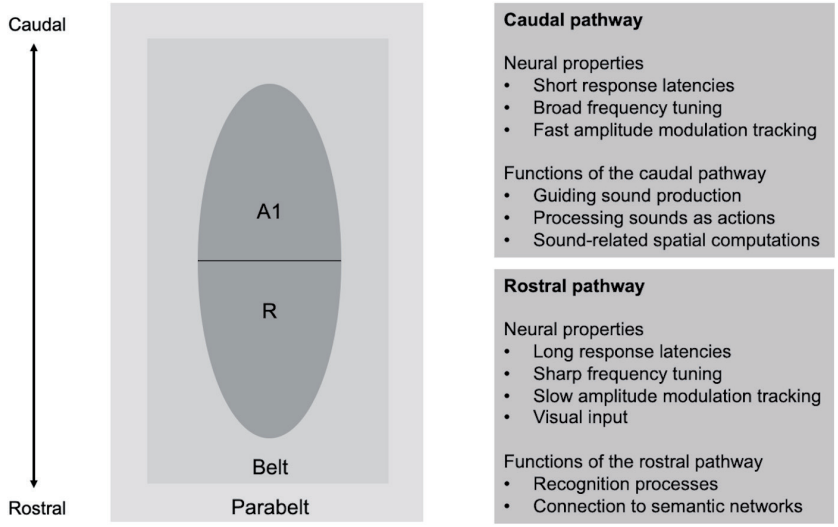


Figure 2: Parallel information processing in the AC. AC information processing is characterized by distinct neuronal and functional responses along the rostral-caudal axis. These streams are hypothesized to underlie the specialized processing of "where" (caudal) and "what" (rostral) pathways. Adapted with permission from (Jasmin et al., 2019).

However, the early sensory areas have also been shown to play a role in multisensory processing (Hackett et al., 2007; Driver and Noesselt, 2008; Koelewijn et al., 2010). The multisensory responses in the AC can be driven by thalamic (feed-forward) and/or lateral cortico-cortical (feedback) connectivity. The multisensory thalamocortical inputs originate from the medial division of the MGB and target all regions of the AC (Aitkin et al., 1972; Calford and Aitkin, 1983).

The lateral cortico-cortical connections targeting the AC can be originating from early sensory cortices and/or higher association cortices (Cappe et al., 2009; Lewis and van Essen, 2000). There is ample functional evidence showing visual influences on responses in the human non-primary auditory cortex (Calvert et al., 2000; Calvert and Campbell, 2003; van Atteveldt et al., 2004; Noesselt et al., 2007; Stevenson et al., 2010; Laing et al., 2015) and even at the level of primary auditory cortex as shown in animals models (Kayser et al., 2007, 2008, 2010; Bizley and King, 2009) and human studies [electrophysiology: (Besle et al., 2008), neuroimaging (Werner and Noppeney, 2010; Gau et al., 2020)]. At the level of single units, the multisensory influences are reflected in changes in the phase of auditory local field potential (Kayser et al., 2008, 2010). These changes in local field potentials have been shown to amplify sensory inputs (Schroeder and Lakatos, 2009) and, more recently, to provide cross-modal cues in auditory scene analysis (Atilgan et al., 2018). The early onset of observed multisensory effects supports the role of early sensory cortical connectivity in multisensory interactions (Wang et al. 2008; Besle et al. 2008).

Through the use of anterograde and retrograde tracers, Falchier et al. (2002) showed direct projections from primary and secondary auditory areas to the early visual areas in rhesus monkeys as well as reciprocal connections from the secondary visual area (V2) and pro-striata to the auditory cortex (Falchier et al., 2010). Multisensory effects in the AC can also be driven by higher-order association areas implicated in multisensory processing such as the posterior superior temporal sulcus and the middle temporal gyrus (Beauchamp et al., 2004; Starke et al., 2017; van Atteveldt et al., 2004; von Kriegstein et al., 2005; Perrodin et al., 2014; Tanabe, 2005), the intraparietal sulcus (Lewis and van Essen, 2000; Cate et al., 2009) and the frontal areas (Gaffan and Harrison, 1991; Romanski et al., 1999a). Overall, the functional implications of the early connections, top-down influence from higher regions, and the role of thalamic input in the multisensory processing in the AC remain to be explored.

1.3 Computational Modeling of the AC

Driven by empirical observations, the mathematical formulations of the neural dynamics have been around for decades and vary from action potentials in single neurons [e.g., (Hodgkin and Huxley, 1952)], to average firing-rate in neuronal populations [mesoscale, e.g., (Wilson and Cowan, 1973)] to large-scale cortical networks [e.g., (Kuramoto, 1984)]. Recent technological advances in computing have made it possible to realize the potential role of these models as an integrative tool. Computational models in general have been shown to provide clear advantages over experimental approaches in understanding biological systems by their ability to test an arbitrary number of simulations, make inferences without disturbing the system and manipulate parameters in a controlled way (Brodland, 2015). In particular for the AC, models have been used to computationally characterize auditory cortical receptive fields (Lindeberg and Friberg, 2015; Chambers et al., 2019), plasticity in the frequency representation in primary auditory areas (de Pinho et al., 2006), the role of inhibition in encoding of temporal information in an auditory cortical neuron (Bendor, 2015), homeostatic plasticity as a compensatory mechanism of hearing loss-induced abnormal activities in A1 (Chrostowski et al., 2011), and stimulus-specific adaptation (Yarden and Nelken, 2017). The choice of the model is dependent on specific modeling goals, data available for validation, and previous applications. Generally, the best modeling endeavors follow the "minimal model approach" as simpler models allow for more constraints, making it easier to estimate parameters from the data and generate inferences about parameter space. Overall, these computational techniques can integrate existing knowledge about the AC, test hypotheses, and generate not only new insights into experimental observations but propositions for new and improved experiments as well. However, one has to remain cautious of the limitations of the models, as they represent a simplification of a complex system and their link to empirical observations must be maintained.

In this thesis, we examine if and how characteristics in neuronal response properties, as resulting from animal electrophysiology, are compatible with findings in human neuroimaging, and psychophysics. For example, perceptual modulation detection thresholds have characteristic dependence on carriers (Bacon and Viemeister, 1985; Kohlrausch et al., 2000; Simpson et al., 2013) but how these are driven by general principles of information processing in the AC, remains unknown. Are the mechanisms underlying spectro-temporal tradeoff in the auditory belt measured using neuroimaging techniques reflective of underlying neuronal dynamics? How does multisensory input influence the sound processing at various stages of information processing in the AC? To study these research questions that range across the levels of neuronal dynamics, human neuroimaging, and psychophysical observations, a computational model that incorporates the serial processing along the AC hierarchy and differences in neuronal response properties underlying parallel information processing streams is required. As we are interested in general mechanisms of sound processing, the modeling approach should be designed to reflect topographic processing (i.e., model neuronal units that vary in frequency preference), and capture the temporal dynamics that are a key element of sound structure. Given that the human observations largely come from neuroimaging (mesoscale) and behavior (macroscale), the model needs to be at a level of abstraction that can successfully link these observations to neuronal dynamics. Thus, a model that captures the population level neuronal dynamics, such as the Wilson Cowan Cortical Model (Wilson and Cowan, 1972, 1973) as employed in this thesis, can be used to produce predictions of meso- and macroscopic observations.

2 Thesis Outline

The current chapter acts as a backdrop for this thesis and reviews the fundamentals of information processing along the auditory cortical hierarchy. Chapter 2 presents a recurrent neuronal model built on simple and established assumptions on general mechanisms of auditory cortical hierarchy and neuronal processing (rostral-caudal differences). Despite its simplicity, the model mimics results from (animal) electrophysiology and links these results to those of psychophysics and neuroimaging studies in humans. Additionally, the model shows a "division of labor" between the simulated rostral-caudal processing streams, providing predictions regarding cortical speech processing mechanisms. The model is valuable for generating hypotheses on how the different cortical areas/streams may contribute towards behaviorally relevant aspects of acoustic signals. In Chapter 3, the proposed neuronal model is used along with a model of the hemodynamic coupling and response (Havlicek et al., 2015) to estimate the neuronal underpinnings of the rostral (caudal) preferences for fine spectral (temporal) features of the sounds, as measured in existing fMRI datasets (Santoro et al., 2014, 2017). Chapter 4 describes a psychophysics

study designed to investigate the multisensory interactions between audiovisual stimuli in the far periphery. The results show visual-to-auditory effects only for specific phase-differences between the modulated audiovisual stimuli. In Chapter 5, the neural correlates of the behavioral observations of the psychophysics study (Chapter 4) are investigated in an fMRI study. Using high-resolution fMRI and peripheral audiovisual stimuli, we present evidence for multisensory processing across the auditory cortical hierarchy, with attentional modulation of multisensory responses in the deep layers of the belt regions. The data reported in Chapters 4 and 5 can be used to inform the existing model, thus completing the necessary loop across methodologies. Finally, Chapter 6 provides an integrative outlook for our findings along with the prospects of systems neuroscience of audition.

Chapter 2

Spectro-temporal Processing in a Two-Stream Computational Model of Auditory Cortex

Zulfiqar, I., Moerel, M., and Formisano, E. (2020). Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex. *Front. Comput. Neurosci.* 13, 95. doi: 10.3389/fncom.2019.00095

Abstract

Neural processing of sounds in the dorsal and ventral streams of the (human) auditory cortex is optimized for analyzing fine-grained temporal and spectral information, respectively. Here we use a Wilson and Cowan firing-rate modeling framework to simulate spectro-temporal processing of sounds in these auditory streams and to investigate the link between neural population activity and behavioral results of psychoacoustic experiments. The proposed model consisted of two *core* (A1 and R, representing primary areas) and two *belt* (*Slow* and *Fast,* representing rostral and caudal processing respectively) areas, differing in terms of their spectral and temporal response properties. First, we simulated the responses to amplitude modulated (AM) noise and tones. In agreement with electrophysiological results, we observed an area-dependent transition from a temporal (synchronization) to a rate code when moving from low to high modulation rates. Simulated neural responses in a task of amplitude modulation detection suggested that thresholds derived from population responses in *core* areas closely resembled those of psychoacoustic experiments in human listeners. For tones, simulated modulation threshold functions were found to be dependent on the carrier frequency. Second, we simulated the responses to complex tones with missing fundamental stimuli and found that synchronization of responses in the *Fast* area accurately encoded pitch, with the strength of synchronization depending on the number and order of harmonic components. Finally, using speech stimuli, we showed that the spectral and temporal structure of the speech was reflected in parallel by the modeled areas. The analyses highlighted that the *Slow* stream coded with high spectral precision the aspects of the speech signal characterized by slow temporal changes (e.g., prosody), while the *Fast* stream encoded primarily the faster changes (e.g., phonemes, consonants, temporal pitch). Interestingly, the pitch of a speaker was encoded both spatially (i.e., tonotopically) in the *Slow* area and temporally in the *Fast* area. Overall, performed simulations showed that the model is valuable for generating hypotheses on how the different cortical areas/streams may contribute towards behaviorally relevant aspects of auditory processing. The model can be used in combination with physiological models of neurovascular coupling to generate predictions for human functional MRI experiments.

1 Introduction

The processing of sounds in primate auditory cortex (AC) is organized in two anatomically distinct streams: a *ventral* stream originating in areas located rostrally to the primary auditory core and projecting to the ventral regions of the frontal cortex, and a *dorsal* stream originating in areas located caudally to the primary core and projecting to dorsal frontal regions. Processing in these separate streams is hypothesized to underlie auditory cognition and has been linked respectively to specialized mechanisms of sound analysis for deriving semantic information ("what" processing) or processing sound location and sound movement ("where" processing) (Arnott et al., 2004; Belin and Zatorre, 2000; Kaas and Hackett, 2000; Kaas et al., 1999; Rauschecker and Tian, 2000; Romanski et al., 1999b; Tian et al., 2001). Interestingly, the basic response properties (e.g., frequency tuning, latencies, temporal locking to the stimulus) of neurons in areas of dorsal and ventral auditory streams show marked differences (Bendor and Wang, 2008; Nourski et al., 2013, 2014; Oshurkova et al., 2008; Rauschecker et al., 1997), and differences have been reported even for neurons from areas within the same (dorsal) stream (Kuśmierek and Rauschecker, 2014). A consistent observation is that neurons in the rostral field, in comparison to primary and surrounding auditory areas, exhibit longer response latencies and narrower frequency tuning (Bendor and Wang, 2008; Camalier et al., 2012; Recanzone et al., 2000; Tian et al., 2001), whereas neurons in the caudal fields respond with shorter latencies, comparable to or even shorter than those in A1, and have broader frequency tuning (Kuśmierek and Rauschecker, 2014; Recanzone et al., 2000). How this organization of neuronal properties within AC contributes to the processing of spectro-temporally complex sounds remains unclear and poses an interesting question for computational endeavors (Jasmin et al., 2019).

Recent results of neuroimaging studies in humans have put forward the hypothesis that fine-grained spectral properties of sounds are analyzed optimally in ventral auditory regions, whereas fine-grained temporal properties are analyzed optimally in dorsal regions (Santoro et al., 2014; Schönwiesner and Zatorre, 2009). It is, however, unlikely that the neural processing of spectral and temporal properties of sounds is carried out through completely independent mechanisms. Several psychophysical phenomena such as pitch perception based on temporal cues (Bendor et al., 2012; Houtsma and Smurzynski, 1990) or the frequency dependence of amplitude modulation (AM) detection thresholds (Kohlrausch et al., 2000; Sek and Moore, 1995) suggest an interdependence between neural processing mechanisms for spectral and temporal properties.

Therefore, in this study, we aim to introduce a simple, stimulus-driven computational framework for modeling the spectral and temporal processing of sounds in AC and examine the role of the different processing streams. We use the firing rate model of

Wilson and Cowan (Wilson Cowan Cortical Model, WCCM; Wilson and Cowan 1972, 1973; Cowan et al., 2016) which simulates complex cortical computations through the modeling of dynamic interactions between excitatory and inhibitory neuronal populations. Over the years, WCCM has been successfully implemented for simulating neuronal computations in the visual cortex (Wilson and Kim 1994; Wilson 1997; Ermentrout and Cowan, 1979). More recently, WCCM has been applied to the AC as well to describe the propagation of activity in the interconnected network of cortical columns and to generate predictions about the role of spontaneous activity in the primary AC (Loebel et al., 2007), the role of homeostatic plasticity in generating traveling waves of activity in the AC (Chrostowski et al., 2011). Furthermore, WCCM has been proposed for modeling stimulus-specific adaptation in the AC (May et al., 2015; Yarden and Nelken, 2017) and to generate experimentally verifiable predictions on pitch processing (Tabas et al., 2019), etc. While WCCMs are less detailed than models of interconnected neurons, they may provide the right level of abstraction to investigate functionally relevant neural computations, probe their link with psychophysical observations, and generate predictions that are testable using invasive electrocorticography (ECoG) as well as non-invasive electro- and magnetoencephalography (EEG, MEG) and functional MRI (fMRI) in humans.

Here, we used the WCCM to simulate the dynamic cortical responses (population firing rates) in the AC to both synthetic and natural (speech) sounds. After filtering from the periphery, the proposed model processes the spatiotemporally structured (i.e., tonotopic) input in two primary auditory *core* areas. The output of the core areas is then fed forward to two secondary auditory *belt* areas, which differ in terms of their processing of spectral and temporal information and thereby represent the dorsal and ventral auditory processing streams. In a number of simulations, we used this model to examine the coding of amplitude modulated (AM) broadband noise and tones using metrics derived from the electrophysiology (firing rate and temporal synchronization with the stimulus). We also simulated three psychoacoustic experiments to study the role of the multiple information streams that may underlie behavioral AM detection thresholds observed for noise (Bacon and Viemeister, 1985) and tones (Kohlrausch et al., 2000), as well as pitch perception with missing fundamental stimuli (Houtsma and Smurzynski, 1990). Lastly, we investigated the processing of speech stimuli in the model to generate predictions on how this cortical spectro-temporal specialization (represented by the four areas) may encode the hierarchical structure of speech.

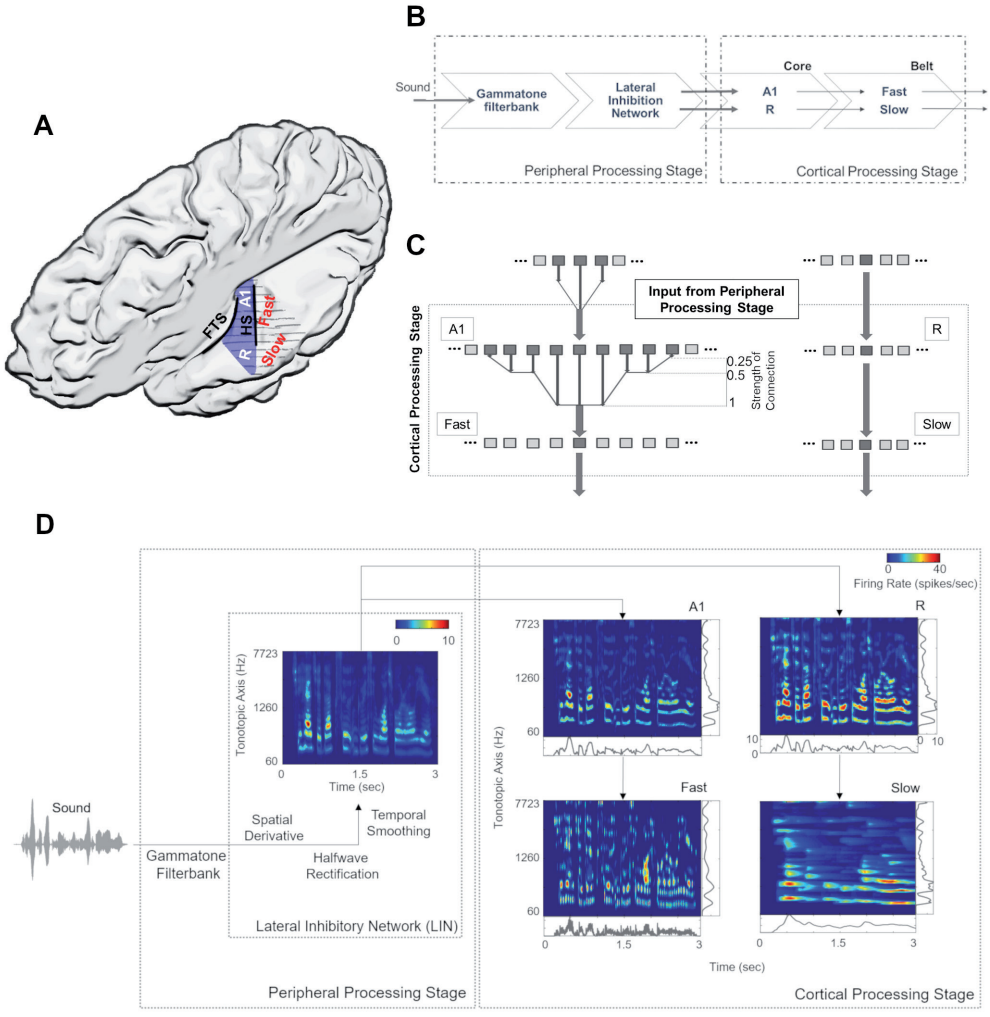


Figure 1: Model design and architecture. (A) Anatomical schematic of the modeled areas shown on top view of the left supratemporal plane (with the parietal cortex removed). Heschl's sulcus (HS) and first transverse sulcus (FTS) are marked to provide anatomical references while Heschl's Gyrus is highlighted in blue. (B) The sound waveform is filtered with a Gammatone filterbank and passed through a Lateral Inhibitory Network (LIN) in the peripheral processing stage, which serves as input to the cortical stage. The neural responses of the simulated core areas (A1, R) are fed forward as input to two simulated belt areas (*Slow* and *Fast*), which differ from each other in their spectral and temporal properties. (B) Connections between model stages are shown. The output of Lateral Inhibitory Network (LIN) projects to excitatory units of A1 and R, which in turn project to excitatory units of *Fast* and *Slow*, respectively. While the convergence through A1 to the *Fast* area is high (i.e., many excitatory units of A1 provide input to a single unit of the *Fast* area), convergence through R to the *Slow* area is low (i.e., the units in areas R and *Slow* receive input from only one unit). (C) Model output for a sample speech sound is shown at different stages of processing as a spectrogram. The panels at the right and bottom of the output of cortical processing stage show mean firing rates across time and tonotopic axis respectively.

2 Methods

2.1 Model Design and Architecture

Figure 1A provides an anatomical schematic of the modeled cortical areas with approximate locations shown on the left supratemporal plane. Figure 1B illustrates the overall architecture of the model, consisting of a *peripheral* processing stage and a *cortical* processing stage. The *peripheral* processing stage simulates the peripheral auditory processing in two steps. First, the tonotopic response of the cochlea is estimated using a set of band-pass filters (Gammatone filterbank, $N = 100$) (Patterson, 1986; Patterson et al., 1992). The gains of the filters represent the transfer function of the outer and middle ear (4^{th} order Gammatone filterbank implementation by Ma et al., 2007). Following the results from psychoacoustics, the center frequencies of the filters are equally spaced on an ERB_N number scale and their bandwidth increases with center frequency, so as to have a constant auditory filter bandwidth (Glasberg and Moore, 1990). Thus, the bandwidth of the 100 rectangular filters is set as 1 ERB [Equivalent Rectangular Bandwidth, based on psychoacoustic measures; for a review of critical bandwidth as a function of frequency see (Moore, 2003)]. The filter frequencies are centered from 50 to 8000 Hz, equally spaced with a distance of 0.3 Cams (on the ERB_N number scale, ERB_N is the ERB of the auditory filters estimated for young people with normal hearing; Glasberg and Moore, 1990).

Second, the basilar response of the Gammatone filterbank is spectrally sharpened using a Lateral Inhibitory Network (LIN) implemented in three steps by taking a spatial (tonotopic) derivative, half-wave rectification, and temporal integration (Chi et al., 2005). The output of extreme filters (i.e., first and last filter) is removed to avoid any boundary effects of filtering, thus reducing the output of the *peripheral* processing stage to 98 units (60 – 7723 Hz).

For the *cortical* processing stage, the filtered tonotopic cochlear input is processed in two primary auditory *core* areas (A1 and R) and then fed forward to two secondary auditory *belt* areas (*Slow* and *Fast*; Figure 1). These four areas approximate the known architecture of human (Galaburda and Sanides, 1980; Rivier and Clarke, 1997; Wallace et al., 2002) and non-human primates (Hackett et al., 1998; Kaas and Hackett, 2000; Read et al., 2002) AC. Simulated areas primarily differ in their temporal and spectral (spatial) response properties. Specifically, neuronal units in the *Fast* area (approximating caudomedial-caudolateral areas) are characterized by fast temporal dynamics and coarse spectral tuning, whereas units in the *Slow* area (approximating middle lateral-anterolateral areas) are characterized by slow temporal dynamics and fine spectral tuning. It is important to note that these units represent an abstraction at the level of neural population behavior and are not always indicative of single-neuron properties.

In addition, we introduce an interdependence between temporal and spatial (tonotopic) processing within the two *belt* areas, as the variable that determines the temporal dynamics of the responses varies with frequency. Consequently, the units corresponding to lower frequencies in the tonotopic axis respond more slowly than those corresponding to higher frequencies (see Heil and Irvine, 2017; Scott et al., 2011; Simpson et al., 2013). Each simulated area comprises 98 units, which are modeled by excitatory and inhibitory unit pairs. Each of the excitatory core units receives tonotopic input from the corresponding frequency-matched *peripheral* stage. This input only targets the excitatory units of A1 and R. Excitatory responses of A1 and R act as tonotopic input for *Fast* and *Slow* areas, respectively (Figure 1C). The output (excitatory responses) at different stages of the model is shown in Figure 1D.

2.2 The WCCM

Neuronal units of the cortical areas were simulated using the WCCM in MATLAB (The MathWorks, Inc.). The WCCM is a recurrent firing rate model where neural population processes are modeled by the interaction of excitatory and inhibitory responses. The model dynamics are described by Wilson (1999):

$$\tau \frac{dE_n(t)}{dt} = -E_n(t) + S_E\left(\sum_m w_{EE_{mn}} E_n(t) - \sum_m w_{IE_{mn}} I_n(t) + P_n(t)\right) \quad (1)$$

$$\tau \frac{dI_n(t)}{dt} = -I_n(t) + S_I\left(\sum_m w_{EI_{mn}} E_n(t) - \sum_m w_{II_{mn}} I_n(t)\right) \quad (2)$$

where E_n and I_n are the mean excitatory and inhibitory firing rates at time t at tonotopic position n, respectively. P_n is the external input to the network and τ is the time constant. The sigmoidal function S, which describes the neural activity (Sclar et al., 1990), is defined by the following Naka-Rushton function:

$$S(P) = \frac{MP^2}{\theta^2 + P^2} \quad (3)$$

θ is the semi-saturation constant and M is the maximum spike rate for high-intensity stimulus P. The excitatory and inhibitory units are connected in all possible combinations (E – E, E – I, I – E, I – I). The spatial spread of synaptic connectivity between the units m and n is given by the decaying exponential w_{ij} (i,j = E, I) function:

$$w_{ij_{mn}} = b_{ij}\, exp\left(\frac{-|m-n|}{\sigma_{ij}}\right) \quad (4)$$

In equation 4, B_{ij} is the maximum synaptic strength and σ_{ij} is a space constant controlling the spread of activity. The equations were solved using Euler's method with a time step of 0.0625 ms.

2.3 Parameter Selection and Optimization

Model parameters were selected and optimized based on the following procedure. First, the stability constraints of the model, as derived and implemented by Wilson (1999), were applied. Second, parameters range was chosen so that the model operates in active transient mode, which is appropriate to simulate activity in sensory areas (Wilson and Cowan, 1973). In active transient mode, recurrent excitation triggers the inhibitory response, which in turn reduces the network activity. The balance of excitation and inhibition was achieved by fixing the parameters as described in Table 1 (for the derivation of these parameters see Wilson, 1999). As shown in previous modeling endeavors (Loebel et al., 2007; May et al., 2015), it is crucial to understand the behavior generated through the interaction of various model properties rather than the exact values of the parameters. In our case, we are interested in the interaction of spectral selectivity and temporal dynamics in neural populations constrained by known physiological response properties of the AC. Thus, while most of the parameters were fixed, further tuning was performed to find the combination of spatial spread (σ), connectivity between areas and time constant (τ),such that the areas reflected the general spectral and temporal constraints, as derived from the electrophysiology literature (see following subsections).

2.3.1 Spatial Resolution of the Model

Model parameters: spatial spread (σ), and connectivity between areas, were determined by matching the sharpness of the model's resulting frequency tuning curves (FTCs) with values reported in the literature. FTCs represent the best frequency of auditory cortical neurons as well as their frequency selectivity (i.e., the sharpness of frequency tuning; Schreiner et al., 2000). In primate AC, the sharpness of neuronal FTCs varies from sharp to broad. Quality factor (Q) has been used to express the sharpness of the FTCs:

$$Q = \frac{Best\ Frequency}{Bandwidth}$$

The Q values for sharply and broadly tuned auditory cortical neurons have been reported to be around 12 and 3.7, respectively (Bartlett et al., 2011). Also, the core areas have been described as having narrower tuning bandwidths than belt regions (Recanzone et al., 2000). In order to generate narrow FTCs of A1, R and *Slow* areas and broad FTCs for *Fast* area, we iteratively changed the spread of activity within the simulated area (final values are listed in Table 2). When changing the spread of activity (σ) within an area did not affect the Q of the area, the connectivity across the areas was manipulated. It should be noted that the projections act as a filter, which is then convolved with the spatial input per unit time. To avoid any boundary effects, symmetric kernel filters (odd number of elements) were used and the central part of the convolution was taken as a result. Final connectivity across regions (i.e., distribution of input units projecting from one area to another) is shown in Figure 1B.

Table 1: Fixed parameters of the model. M is the maximum spike rate, θ the is semi-saturation constant. Parameters b_{EE}, b_{II}, b_{EI} and b_{IE} represent the maximum synaptic strength between excitatory units, between inhibitory units, from excitatory to inhibitory units, and vice versa, respectively. All the listed parameter values are the same across the four simulated areas.

Parameters	Values
M	100
θ Inhibition	80
θ Excitation	60
b_{EE}	1.5
$b_{EI} = b_{IE}$	1.3
b_{II}	1.5
σ_{II}	10

Table 2: Model parameters across the four simulated areas. For the four simulated areas, the values for varying parameters, time constant τ (reported over the tonotopic axis from low to high best frequencies of the units), spatial spread parameter σ (EE, EI/IE) are listed.

Parameters	Values			
	A1	R	*Slow*	*Fast*
τ (ms)	10	20	300 – 200	3 – 1
σ_{EE}	40	40	20	200
$\sigma_{EI} = \sigma_{IE}$	160	160	80	300

The narrower tuning in the *Slow* area results from the smaller spread of excitation (σEE, see Table 2), and from the one-to-one projection from R units (Figure 1C). The broader tuning in the *Fast* area is simulated by a many-to-one projection from the Gammatone filterbank to a single unit of A1 (three to one) and from A1 to the *Fast* areas (nine to one). The strength of these connections is shown in Figure 1C. The FTCs across areas are quantified using Q at half-maximum bandwidth. The units tuning in the simulated A1 and R areas have mean Q = 6.32, (std = 1.43), units in the *Fast* area have mean Q = 4, (std = 0.87), while units in the *Slow* have Q = 8.35, (std = 2.1). In line with the experimental observations (Kuśmierek and Rauschecker, 2009), the Q values increased with increasing center frequencies, while maintaining the general trend of broad tuning in *Fast* and narrow tuning in *Slow* area. Figure 2 shows FTCs across the four simulated areas for a single unit with best frequency at 4.3 kHz.

2.3.2 Temporal Resolution of the Model

Temporal structure represents an important aspect of natural acoustic signals, conveying information about the fine structure and the envelope of the sounds (Giraud and Poeppel, 2012). In several species, a gradient of temporal responses has been observed in AC,

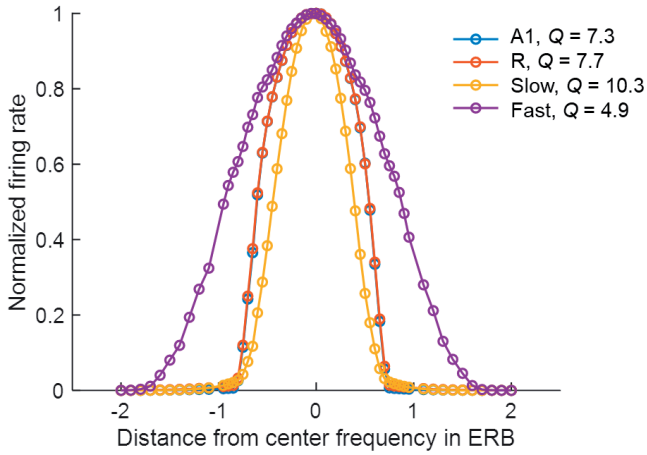


Figure 2: Frequency tuning curves (FTCs) of the unit with best frequency at 4.3 kHz across simulated areas. Areas A1 (blue line) and R (red line) are sharply tuned, with *Q* of 7.3 and 7.7, respectively. The *Slow* area (yellow line) has the sharpest tuning curves with *Q* of 10.3, while *Fast* (purple line) has the broadest tuning with *Q* of 4.9. *Q* is measured as the ratio of the best frequency and the half-maximum bandwidth in Hz.

with higher stimulus-induced phase locking (synchrony) and lower latencies in area AI compared to adjacent areas (AI vs AII in cats: Eggermont 1998; Bieser and Müller-Preuss 1996, AI vs R and RT in monkeys: Bendor and Wang 2008). Correspondingly, model parameters determining the temporal properties of population responses in the simulated areas were adjusted to match such electrophysiological evidence. Table 2 shows the resulting time constant τ for the simulated areas. Note that the values of parameter τ do not represent the latency of the first spike measured for single neurons but affect the response latencies and dynamics at a population level.

2.3.2.1 Temporal Latencies

As neurons in core area R have longer latencies than A1 (Bendor and Wang, 2008), we selected a higher value of τ for simulated R than A1. Based on the evidence of the caudomedial field showing similar latencies to A1 (Kuśmierek and Rauschecker, 2014; Recanzone et al., 2000), we adjusted τ of the *Fast* area so that the area is as fast as A1. In contrast, we set τ of the *Slow* area such that this region generates a more integrated temporal response, with the firing rate taking longer to reach the semi-saturation point. These τ values, in combination with the spatial connectivity constraints, cause the simulated belt area to display a spectro-temporal tradeoff. Additionally, in both *Slow* and *Fast* areas τ decreases linearly along the spatial axis (maximum and minimum values are reported in Table 2) with increasing best frequency, following electrophysiological evidence of interaction of the temporal and frequency axis where shorter latencies have been found to be correlated with high best frequencies in macaques (Scott et al., 2011).

2.3.2.2 Temporal Synchrony

To further refine parameter τ, next we examined stimulus-driven phase locking of the simulated neural activity. Electrophysiological measurements report synchronization in the neural response to the sound carrier and envelope for a limited range of frequencies, and the upper limit of this phase locking has been found to decrease along the auditory pathway (Joris et al., 2004). At the level of cortex, while the strongest synchronization is reported for modulation rates up to 50 Hz (AM stimuli: Liang et al., 2002, Clicks: Nourski et al., 2013), weaker synchronization to even higher rates (up to 200 Hz) has been observed for a subset of units (Steinschneider et al., 1980; Bieser and Müller-Preuss 1996; Lu et al., 2001; Nourski et al., 2013). In light of the evidence above, we adjusted τ to mimic this behavior and have the strongest temporal synchronization for the low range of modulation rates (up to 50 Hz), with some residual synchronization to higher rates.

2.4 Model Evaluation

The model performance was evaluated in three stages. First, we simulated the electrophysiological coding of AM (for both noise and tone carriers). Second, we evaluated the model's ability to predict results of human psycho-acoustical tasks, including the determination of amplitude modulation detection threshold functions, tMTFs, and perception of missing fundamental. Lastly, we used speech stimuli to investigate the representation of pitch and AM features of a complex sound across the simulated areas. All artificial stimuli (AM noise, AM tones and missing fundamental complex tones) were generated using MATLAB with a sampling rate of 16 kHz and 1 s duration). Speech stimuli were taken from the LDC TIMIT database (Garofolo et al., 1993). In all cases, the key readouts of the model were synchronization to stimulus features and firing rates. The pitch estimates matched against model output, where relevant, were computed using the YIN algorithm (de Cheveigné and Kawahara, 2002).

2.4.1 Coding of AM Stimuli: Evidence from Electrophysiology

To evaluate the model's coding of AM, sinusoidally amplitude modulated (sAM) stimuli were used. AM sounds were defined by

$$(1 + m \sin 2\pi g t) * carrier,$$

where m is the modulation depth, g is the modulation rate and t is time. The modulation rates were chosen to be 2 to 9 Hz (linearly spaced), and 10 to 1000 Hz (logarithmically spaced). Broadband noise was used as carrier to study the response of all units working together while pure tones (500 Hz, 3 kHz, and 5 kHz) were employed to evaluate carrier-specific effects on amplitude modulation coding.

To quantify synchronization of responses to the temporal structure of AM sounds, we employed two measures from the electrophysiology literature (Joris et al., 2004; Eggermont 1991; Bendor and Wang 2008): vector strength VS where:

$$VS = \frac{Strength\ of\ Fourier\ Component\ at\ the\ Modulation\ Rate}{Average\ Firing\ Rate}$$

(Goldberg and Brown, 1969), and rate modulation transfer function (rMTF), which is the average firing rate as a function of the modulation rate. VS was computed for all modulation rates (and three harmonics), for both tone and noise carriers, across the four simulated areas. We considered a simulated area as being synchronized to a modulation rate when VS was greater than 0.1 (this is an arbitrary threshold chosen to compare phase-locking across conditions and areas).

rMTFs were calculated from the average firing rates (i.e., the Fourier component at 0 Hz) and normalized for all areas. For the computation of rMTFs, the modulation depth is fixed at 100% across all AM stimuli. For noise carriers, the computation of the VS and rMTF is based on the mean across all 98 excitatory channels. For the tone carriers, only the channel maximally tuned to the carrier frequency is considered.

2.4.2 Simulating Psychoacoustical Observations

The model was tested using three paradigms approximating human psychoacoustic studies. The first two experiments simulated temporal modulation transfer functions (tMTFs: quantifying the modulation depth required to detect different modulation rates) for broadband noise (Bacon and Viemeister, 1985) and tones (Kohlrausch et al., 2000). The third experiment simulated pitch identification with missing fundamental stimuli (Houtsma and Smurzynski, 1990).

For the simulated tMTFs, AM sounds with incremental modulation depths (from 1 to 100%) were presented to the model and the oscillations in the model's output were measured. In the psychoacoustic measurements, the lowest modulation depth at which subjects can detect the modulation is considered the detection threshold. In the model, using synchronization as output measure, the lowest value of modulation depth at which the output is synchronized to the modulation rate (i.e., the strongest Fourier component was at the modulation rate) is considered as the detection threshold for that AM rate. This procedure was repeated for all the modulation rates and, for all simulated areas. For noise carriers, the mean across the excitatory units across each area is analyzed and compared to data collected by Bacon and Viemeister (1985). The model response was simulated for modulation rates at 2 to 9 Hz (linearly spaced), and 10 to 1000 Hz (logarithmically spaced).

For AM tones, the analysis of the waveform shows spectral energy at the carrier frequency, and the carrier frequency ± modulation rate. These accompanying frequency components are called "spectral sidebands" of the carrier frequency. If the modulation rate is high enough, these sidebands activate distinctively different auditory channels than the carrier frequency and can be detected audibly apart from the carrier frequency. Thus, for the tone carriers (1k and 5k) the active part of the population (comprising the best frequency channel and spectral sidebands) was used to compute tMTFs based on temporal synchronization to the modulation rate (temporal code) and detection of sidebands (spatial code). As before, for the temporal code, the lowest value of modulation depth at which the output is synchronized to the modulation rate (i.e., the strongest Fourier component was at the modulation rate) is considered as the detection threshold for that AM rate. For the spatial code, the modulation depth at which the side-band amplitude (mean firing rate over time) is at least 5%, 10%, 15% or 20% of the peak firing rate (firing rate of the channel with CF closest to carrier frequency) are calculated. The best (lowest) value of modulation depth is chosen from both coding mechanisms. The combination of these coding mechanisms is then compared to tMTFs (at 30 dB loudness) reported by Kohlrausch et al. (2000). The modulation rates tested were 10 to 1600 Hz (logarithmically spaced).

Pitch of missing fundamental complex tones has been shown to be coded by temporal and spatial codes, depending on the order of harmonics and frequency of missing fundamental (Bendor et al., 2012). Here we replicated this finding by simulating the model response to complex tones with low order (2 – 10) and high order harmonics (11 – 20) and varying missing fundamental frequency from 50 to 800 Hz. The synchronization to the missing F_0, measured in VS, is computed from the mean responses over time in each of the four simulated areas. Furthermore, to evaluate the role of synchronization in pitch perception, we simulated model responses to complex tones with unresolved harmonics of a missing fundamental frequency by approximating a pitch identification experiment by Houtsma and Smurzynski (1990). The missing fundamental tone complexes vary in two aspects: the number of harmonic components (2-11) and the lowest harmonic component (10 and 16) while the fundamental frequency (F_0) is fixed at 200 Hz. For each combination of the lowest harmonic component and the number of components in the harmonic complex, we computed the synchronization to the F_0 (in VS) and mean firing rates for all four regions.

2.4.3 Model Responses to Speech

Model responses to the speech stimuli were analyzed in two stages. The speech stimuli (630 sentences, all spoken by different speakers; mean duration 3.4s) were randomly selected from LDC TIMIT database (Garofolo et al., 1993). To study how key temporal features of speech waveforms are represented in the modeled areas, we compared

the temporal modulations in the output of all four simulated areas to the temporal modulations of the input signals. To this end, we computed the input-output magnitude spectrum coherence (*mscohere* in MATLAB with a 2048 point symmetric hamming window and overlap of 1500 samples) between the input speech signal (after LIN) and the output of all four areas. The coherence values are then scaled across the four areas using the mean spatial activity along the tonotopic axis (i.e., the mean firing rate over time for all sounds). To highlight the difference in spectrum coherence between the spectro-temporal processing streams in the model, the difference between the scaled input-output coherence is computed to compare the two *core* (R – A1) regions to each other and the two *belt* areas (*Slow* – *Fast*).

3 Results

3.1 Coding of AM Stimuli

We investigated the model's AM coding using both broadband noise and tone carriers. By using broadband noise as carrier, we simulated general responses for each of the four areas, and then use pure tone carriers to study the dependence of the synchronization and rate coding on the tonotopic location (i.e., the best frequency of the units).

3.1.1 Sinusoidal AM Noise

Figure 3 shows the response of the four simulated cortical areas (A1, R, *Fast*, and *Slow*) as a function of the modulation rate of sinusoidally amplitude modulated (sAM) noise. We analyzed the mean response of all units for each area. Across regions, the response synchronization (measured as VS) decreases with increasing modulation rate (solid lines in Figure 3 A, B, C, and D for A1, R, *Fast,* and *Slow* areas respectively). The decrease in synchronization is observed to be rapid above an area-specific modulation rate (8 Hz for A1, R, and *Fast* areas; 2 Hz for *Slow*). Taking the lower limit for synchronization as VS = 0.1, the highest AM rate to which the areas synchronize is 54 Hz in A1, 33 Hz in R, 4 Hz in *Slow* and 54 Hz in *Fast*. Overall, the observed responses to modulation rates show a low-pass filter profile.

Instead, the firing rate (rate Modulation Transfer Functions (rMTFs), dash-dotted lines) shows different behavior across the four areas in response to AM noise. For A1, R and *Fast* areas (Figure 3 A-C respectively), the firing rate does not change for lower modulation rates (until 10 Hz for A1 and *Fast*, until 6 Hz for R) and then rapidly increases until a maximum limit (54 Hz for A1, R, and *Fast*) and does not further change in response to higher modulation rates. In contrast, the firing rate in the *Slow* area (Figure 3D) shows a band-pass profile between 6 and 100 Hz, peaking at ~ 20 Hz.

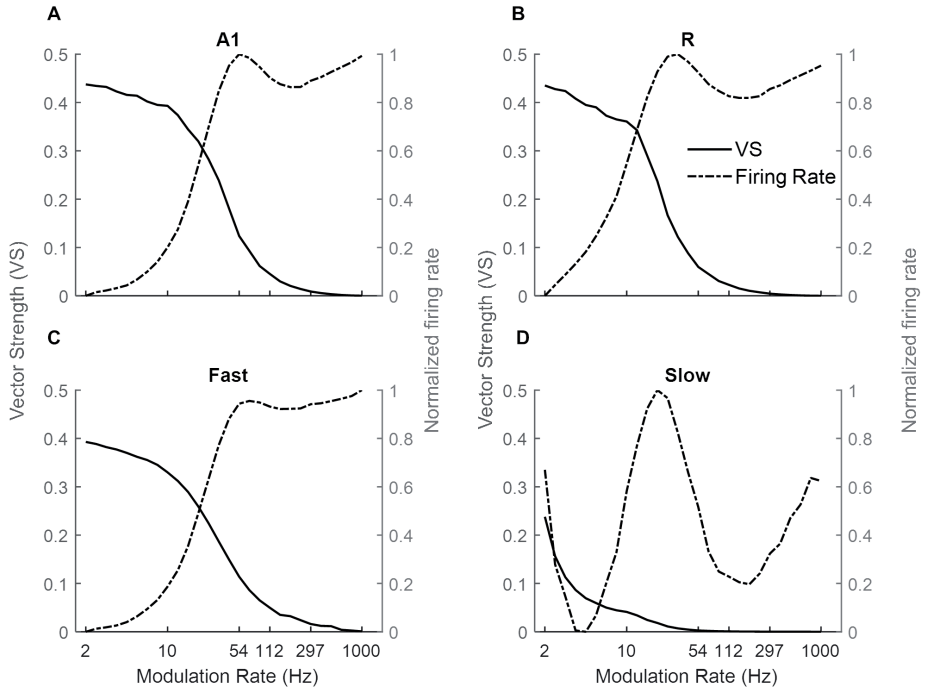


Figure 3: Model responses to sAM noise across simulated areas. A dual coding mechanism for modulation rates, i.e., temporal (measured as Vector Strength, VS, solid lines) and rate codes (quantified as the rate Modulation Transfer Functions, rMTFs, dash-dotted lines), are shown for A1, R, *Fast* and *Slow* areas in panels A, B, C, and D respectively. In A1, R and *Fast* areas, the synchronization decreases for higher modulation rates and is complimented by increasing firing rate. While very little synchronization is observed in the *Slow* area, the respective rMTF shows an interesting band-pass profile.

3.1.2 Sinusoidal AM Tones

Next, we explored the frequency dependence of AM processing. As the use of broadband noise as a carrier provides no information about the temporal properties of different frequency channels along the tonotopic axis, we simulated model responses to AM pure tone carriers. Figure 4 shows response synchronization (VS, left column) and firing rate (rMTFs, right column) across cortical areas as a function of AM rate, separately for units best responding to a low (solid lines), middle (dashed lines), and high (dash-dotted lines) frequency pure tone carriers (500, 1k and 3k Hz respectively). For each area, the responses in the model's frequency channel matching the tone carrier are shown. The synchronization shows a low-pass filter profile consistently for all three carriers. With increasing carrier frequency, the A1, R, and *Slow* areas (Figure 4A, C, and E) are synchronized (VS cut-off at 0.1) to higher modulation rates (A1: 33 Hz for 500 Hz, 54 Hz for 1 kHz and 3 kHz, R: 26 Hz for 500 Hz, 33 Hz for 1 kHz and 3 kHz, *Slow*: 3 Hz for 500 Hz, 4 Hz for 1 kHz and 3 kHz). This behavior is a consequence of the relationship between the temporal and spatial axis (a property of the model), with

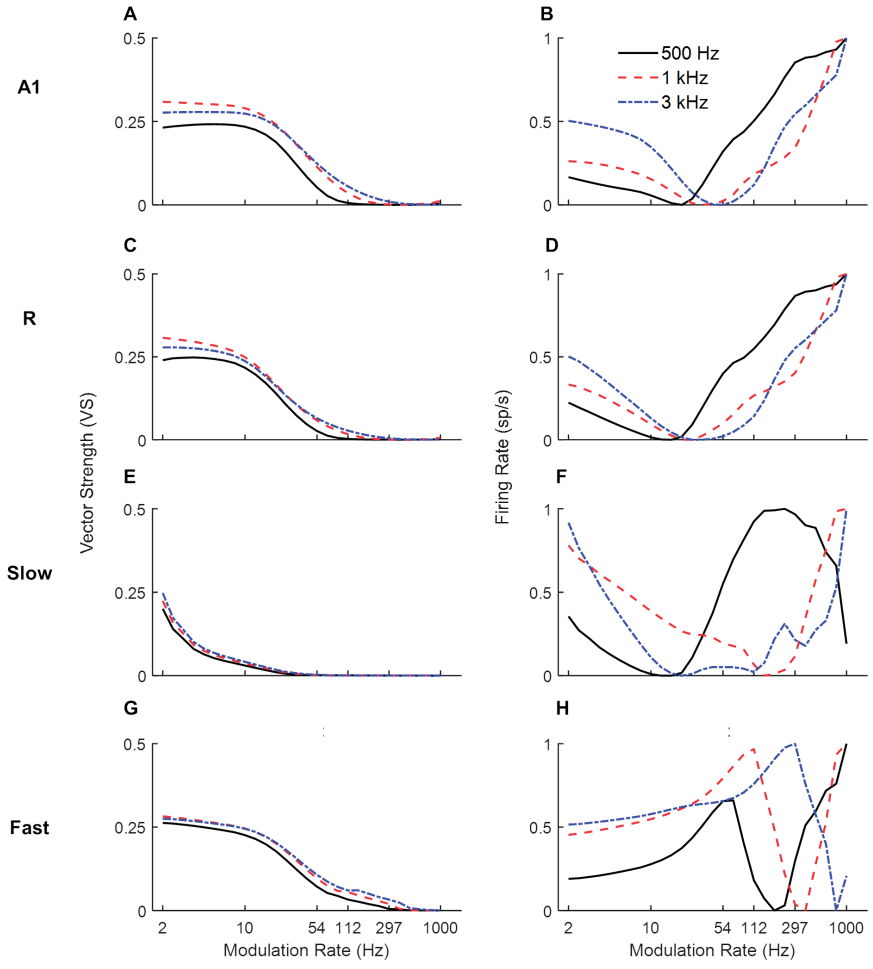


Figure 4: Model responses to sAM tones across simulated areas. A dual coding mechanism for modulation rates, i.e., temporal (measured as Vector Strength, VS, left panels) and rate codes (quantified as the rate Modulation Transfer Functions, rMTFs, right panels), are shown for A1, R, *Fast* and *Slow* areas in respective panels (A1: A-B, R: C-D, *Slow*: E-F, *Fast*: G-H). For the three different carriers, synchronization to higher modulation rates is observed with increasing carrier frequencies across areas (panels A, C, E, and G). Rate coding, however, shows more varied profiles with different carriers (panels B, D, F, and H).

temporal latencies reducing with increasing center frequencies of the units allowing phase-locking to higher modulation. The *Fast* area (Figure 4G) shows a similar cutoff for all carriers at 54 Hz. The rMTFs (Figure 4B, D, F and H for areas A1, R, *Slow* and *Fast* respectively), however, show more complex and varied behavior for different carriers (including monotonically increasing, band-pass, and band-stop behavior). This behavior is in line with rMTFs from electrophysiological studies, where instead of singular behavior (like low-pass filter profile reported for tMTFs), rMTFs show a variety of response profiles

(Bendor and Wang, 2008; Bieser and Müller-Preuss, 1996; Liang et al., 2002; Schreiner and Urbas, 1988).

3.2 Simulating Psychoacoustic Observations

Next, the model was tested using three experimental paradigms similar to those employed in human behavioral studies. The first two experiments tested the temporal modulation transfer functions (tMTFs characterizing the modulation depth required to detect different modulation rates) for broadband noise (Bacon and Viemeister, 1985) and tones (Kohlrausch et al., 2000). The third experiment examined the effects of the number of harmonics in pitch identification with missing fundamental stimuli (Houtsma and Smurzynski, 1990).

3.2.1 Temporal Modulation Transfer Functions for Broadband White Noise

Similar to the behavioral task of Bacon and Viemeister (1985), we measured responses of the model to AM sounds with variable modulation depth and record the minimum modulation depth where the output signal was synchronized to the modulation rate (i.e., the strongest Fourier component was at the modulation rate) of the AM noise. Figure 5 illustrates the simulation results (solid colored lines), along with human psychoacoustic data (dash-dotted black lines with circles, adapted from Bacon and Viemeister, 1985). Lower values depict higher sensitivity to the modulation rates. A1 and R show lower thresholds for slower than faster modulation rates. In the Fast area, the detection profile is similar to A1 and R, but the minimum detection depth is higher than in the other areas. The broad tuning of the Fast area reduces the precision of the temporal structure of the input signal. Thus, the Fast area performs worse than the other areas across modulation rates. In the *Slow* area, modulation detection is observed to be limited to rates below 10 Hz. Thus, the *core* areas outperformed the *belt* areas in the detection of amplitude modulations. The modulation depth detection profile of the core areas resembles the results from human psychophysics suggesting that primary auditory cortical processing may underlie tMTFs reported in psychophysics. In comparison with synchronization, rate coding is difficult to quantify as observed before with varying response profiles for rMTFs along the frequency axis (Figure 4F and H). The difference between our simulations and psychophysical findings at faster rates may be explained by the fact that our simulations only considered coding through response synchronization and ignored the contribution of rate coding contributing to the detection of higher modulation rates.

3.2.2 Temporal Modulation Transfer Functions of Sinusoidal Carriers

We then investigated the model's detection threshold function of sAM tones. Psychoacoustic studies have shown that human performance does not change across the lower modulation rates, becomes worse for a small range and then improves after the sidebands introduced by the modulation become detectable (Kohlrausch et al.,

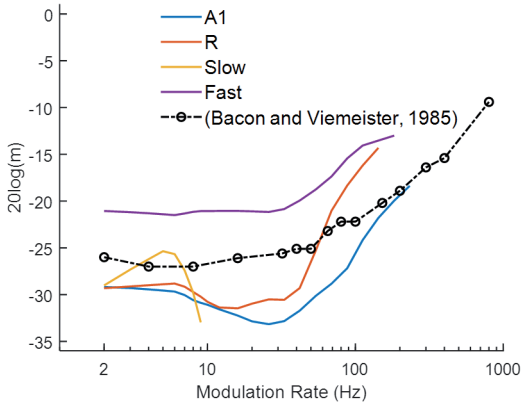


Figure 5: Modulation detection with sAM noise. The temporal Modulation Transfer Functions (tMTFs), illustrating the minimum depth required to detect the amplitude modulation in sAM noise, are shown for the four model areas (in colored lines) and for a psychoacoustic study (black line and circles; adapted from Bacon and Viemeister, 1985). Lower values depict higher sensitivity to modulation rate. Modulation depth, m (dB) of the signal is plotted on the y-axis.

2000; Moore and Glasberg, 2001; Sek and Moore, 1995; Simpson et al., 2013). We obtained model responses to sAM tones as a combination of temporal and spatial codes. To characterize an area's modulation detection threshold represented by temporal code, the lowest modulation depth at which the best frequency unit or the spectral sideband synchronized to the modulation rate was chosen. Additionally, the spatial code was quantified by detection of spectral sideband. Figure 6 shows the lowest modulation depth for which A1 (solid lines in panel A, C) and R (solid lines panel B, D) code modulation rates of sAM tones and the psychoacoustic data for 1 kHz and 5 kHz sinusoidal carriers at 30 dB (dash-dotted lines with circles, Kohlrausch et al., 2000). The initial increase in depth values indicates the contribution of temporal coding of the modulation rates that gets worse with higher modulation rates. With increasing modulation rates, however, the spectral sidebands dissociate from the carrier channel and the contribution of spectral coding is observed. The modulation depths at which the sideband amplitude (mean firing rate over time) is detectable (multiple threshold cut-offs are shown where sideband activity is 5%, 10%, 15%, and 20% of the firing rate of the channel with CF closest to carrier frequency) are also shown in Figure 6. No synchronization is observed in the *Slow* and *Fast* areas. Overall, model results show a clear frequency dependence as detection of higher rates was observed for the higher carrier (maximum for A1: 500 Hz for 1 kHz carrier, 1.2 kHz for 5 kHz carrier; R: 1.2 kHz for 1 kHz carrier, 1.6 kHz for 5 kHz carrier). The modulation detection by the model slightly worsened with increasing modulation rate but improved (lower m values) as the sidebands introduced by the modulation became detectable (after 100 Hz for the 1 kHz carrier in A1 and R, after 400 Hz for 5 kHz carrier in A1). This improvement of AM detection threshold for high AM

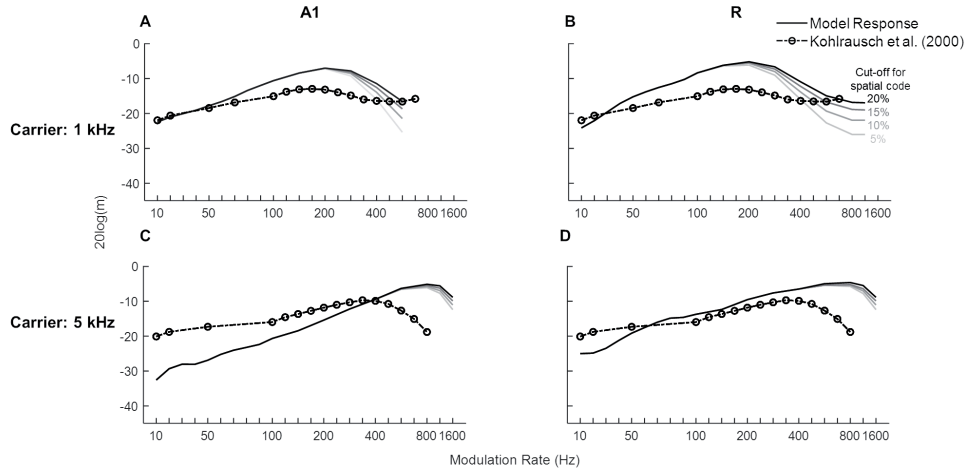


Figure 6: Modulation detection with sAM tones. The solid lines show the temporal Modulation Transfer Functions (tMTFs), illustrating the minimum depth required to detect the amplitude modulation in sAM tones (1 kHz in top panels, 5 kHz in bottom panels), are shown for the two core areas (A1 in panels A and C, R in panels B and D). The model output is a combination of temporal and spatial codes for modulation detection. Variation in the spatial code is shown at four different cut-off values, represented by the solid lines in different gray-scales. Data from a psychoacoustic study are shown in dash-dotted lines with circles (adapted from Kohlrausch et al., 2000). Lower values depict higher sensitivity to modulation rate. Modulation depth, m (dB) of the signal is plotted on the y-axis.

rates is in accordance with human psychophysics, where observations show a decrease in performance with increasing modulation rates is followed by a performance increase accompanied with side-band detection (Kohlrausch et al., 2000; Moore and Glasberg, 2001; Sek and Moore, 1995; Simpson et al., 2013). Additionally, matching the model results, human psychophysics show improved performance (i.e., detection of higher rates) with increasing carrier frequencies.

3.2.3 Pitch of Missing Fundamental Sounds

Missing fundamental sounds are harmonic complexes that, despite lacking energy at the fundamental frequency (F_0), induce the percept of a pitch corresponding to F_0 (Oxenham, 2012; Yost, 2010). If the harmonic components in the missing fundamental sound are resolved (i.e., each component produces a response on the basilar membrane that is distinct from that of neighboring harmonic components), the pitch information can be extracted through a spectral (spatial) mechanism, or a temporal mechanism if harmonics are unresolved, or a combination of the two (Yost, 2009). Bendor et al. (2012) have shown that low F_0 sounds with higher-order harmonics are primarily represented by temporal mechanisms. Thus, we tested the effect of harmonic order on the detection of missing F_0 through temporal synchrony across simulated areas. Figure 7 shows synchronization (temporal code, measured as VS) to missing F_0 of complex tones with lower-order

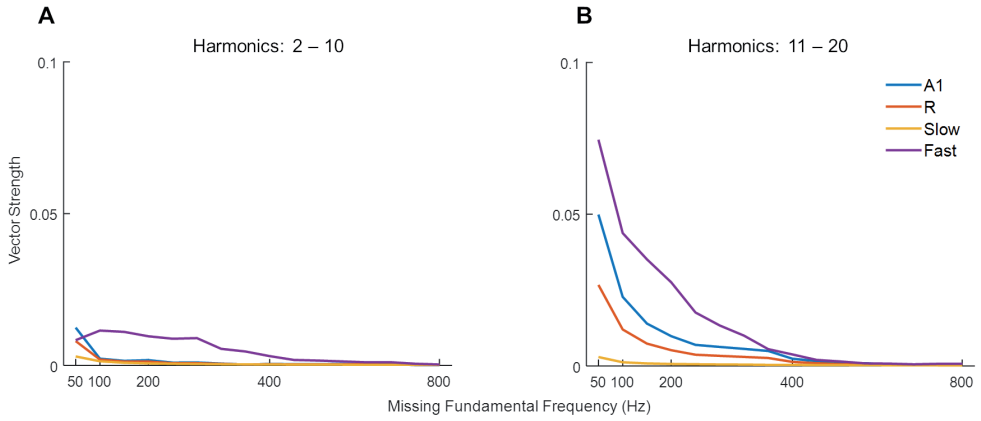


Figure 7: Synchronization to missing fundamental frequency across harmonic order. The model performance in detecting missing fundamental of complex tones (measured as vector strength) with (A) low-order harmonics, and (B) high-order harmonics. Simulated responses in the four areas are shown in different colors.

and higher-order harmonics in panels A and B respectively. Stronger synchronization is observed for higher-order harmonics compared to lower-order harmonics for lower missing F_0 complex tones in A1, R, and *Fast* areas. The effect is most pronounced in the *Fast* area. However, the synchronization drops with increasing missing F_0, and very little to none synchronization is observed after 400 Hz irrespective of the order of harmonics in the complex tone.

For low pitch missing fundamental sounds, psychophysics experiments employing sounds with unresolved harmonics have shown that humans are better at identifying a missing fundamental pitch when the sound consisted of lower (lowest harmonic = 10) compared to higher unresolved harmonics (lowest harmonic = 16), yet the performance reaches a plateau as more harmonic components are included for the sound consisting of lower but not higher-order harmonics (Houtsma and Smurzynski, 1990). To evaluate whether temporal mechanisms play a role in these findings we simulated a pitch identification experiment (Houtsma and Smurzynski, 1990) and explored the effects of the number of harmonic components and lowest order harmonic in the missing fundamental complex tone on the model's behavior. As already established, simulated populations could only successfully synchronize to lower missing F_0 (Figure 7), thus the task employed complex tones with low missing F_0 (200 Hz). Figure 8 shows the model's synchronization (VS) to the missing F_0 (200 Hz and the first three harmonics) across the simulated regions (in blue lines), along with the results from the psychophysics experiment (in red lines, data adapted from Houtsma and Smurzynski, 1990).

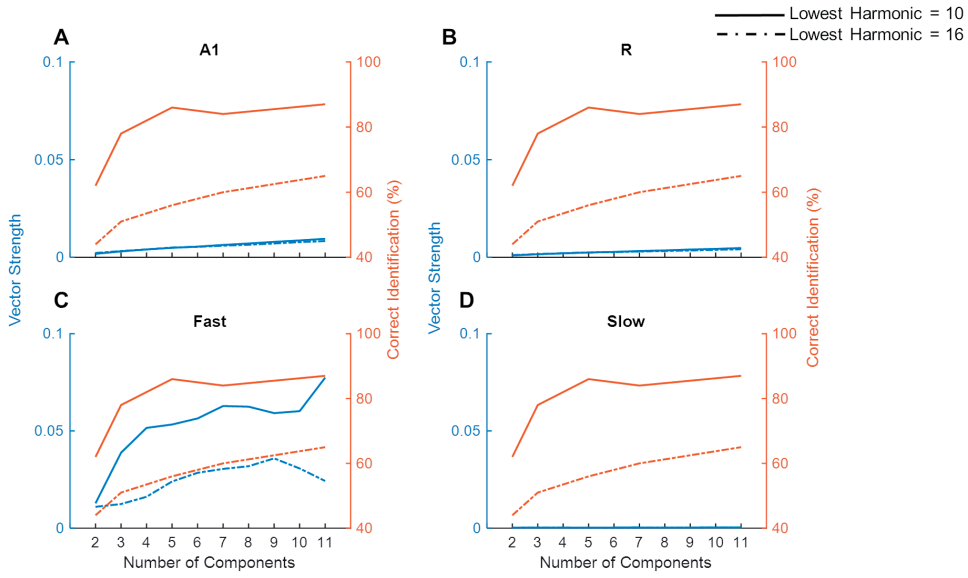


Figure 8: Model performance on a missing fundamental task. The model performance in detecting missing fundamental of complex tones (synchronization to missing fundamental frequency at 200 Hz, measured as Vector Strength) is shown for areas A1, R, *Fast* and *Slow* (blue lines in panels A, B, C, and D respectively). Human behavioral data on pitch identification (%) task (Houtsma and Smurzynski, 1990) is plotted in orange lines. Solid lines show complex tones with the lowest harmonic at 10 while the dash-dotted lines show the lowest harmonic component at 16.

While we did not observe any differences due to harmonic order in VS measured in A1, R, and *Slow* areas (Figure 8A, B, and D), the *Fast* area (Figure 8C) showed clear dissociation in synchronization code when the lowest order harmonic changed from 10 to 16. That is, the synchronization to the missing F_0 in the *Fast* area was stronger when the lowest order harmonic was 10. Additionally, for both complex tones, the performance of the *Fast* area improved with an increasing number of components. The improvement in synchronization was rapid when the number of components changed from 2 to 4 for the lowest order harmonic at 10. These observations are in line with the pitch identification data shown in the red lines. Thus, neural response properties similar to those of the *Fast* area are optimized to temporally detect the F_0 from missing fundamental sounds, and responses in the *Fast* area follow human behavior.

Unlike synchronization, the simulated firing rate (Supplementary Figure 1) did not show a pattern that matched the behavioral data. Specifically, the simulated firing rate increased monotonically as a function of the number of components in the complex tone, irrespective of the lowest order harmonic.

3.3 Model Responses to Speech

Speech signals encode information about intonation, syllables, and phonemes through different modulation rates. We explored the processing of speech sounds across simulated cortical areas to study the importance of simple spectro-temporal cortical properties, as reported by electrophysiology and represented by the model, in coding these temporal features of speech. To this end, we analyzed model output in response to 630 speech stimuli by computing the magnitude spectrum coherence between these sounds (the output of the LIN stage) and the simulated model responses for each of the four areas. Figure 9 shows the normalized coherence plots (scaled by the normalized time-averaged activity). In all regions, we observed model synchronization to slow changes in the stimuli (<20 Hz).

Next, in order to highlight differences in the temporal response properties between regions, we computed difference plots for the simulated core and belt areas. While we observed no differences in coding of temporal features between A1 and R, Figure 10 shows that differences are present in the *belt* stream (comparing the coding of temporal features in the *Fast* to those in the *Slow* area). The difference between the coherence (*Slow* – *Fast*)

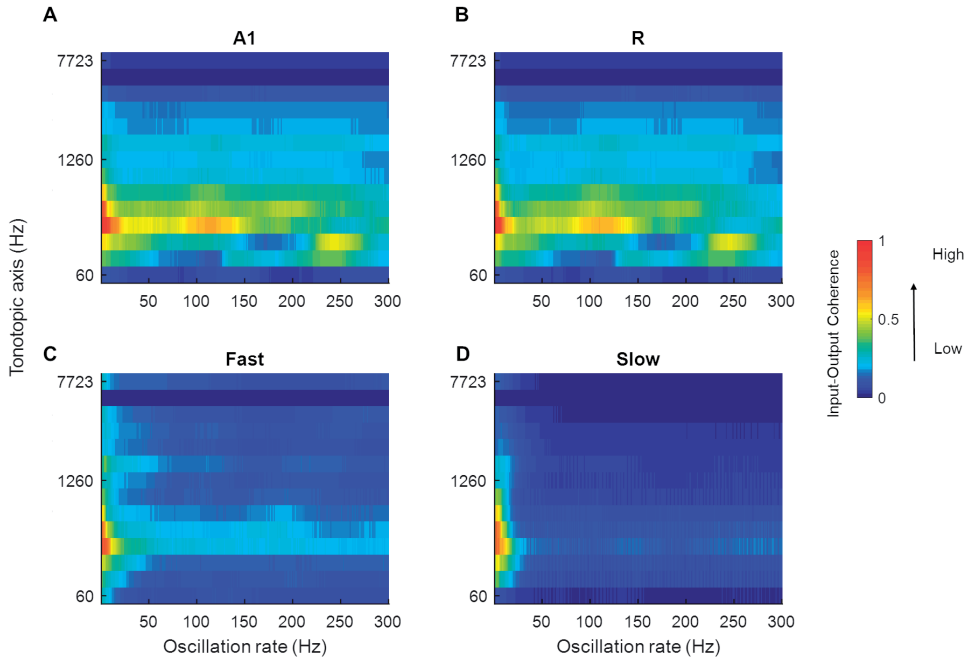


Figure 9: Mean magnitude spectrum coherence between speech sounds and model output. The coherence values in A1, R, *Fast* and *Slow* areas are shown in panels A, B, C, and D respectively (scaled by the normalized mean spatial response of the model to 630 speech sounds). All areas show high coherence with the slow oscillations present in the input signal (indicated by red and yellow colors).

across 630 stimuli (mean: -0.0332, SEM: 0.0041) was used to compute data distribution in four percentiles (65, 75, 85, and 95%). These percentiles are shown along the color bar in Figure 10 (with the distribution) to provide a threshold for the significance to the difference between input-output coherence of the *Slow* and *Fast* area. Shades of blue show stronger input-output coherence in the *Slow* area, while the warmer colors indicate stronger input-output coherence in the *Fast* stream. The *Slow* area represents the slower changes (4 – 8 Hz) in the speech envelope better than the *Fast* area. The *Fast* area, on the other hand, highlights faster changes in the temporal structure of speech in two frequency ranges (30 – 70 Hz, and around 100 – 200 Hz).

We hypothesized that the higher of these two frequency ranges (100 – 200 Hz) may reflect the presence of temporal pitch information in the *Fast* area. The temporal code for pitch in the simulated areas was estimated by computing short-time Fourier Transform (window length: 300 ms, overlap: 200 ms) over the length of the signal. The resulting power spectral density estimates showed temporal synchronization to the frequencies approximating the pitch in A1, R and *Fast* areas over time. For the purpose of comparison across simulated areas, the pitch estimates and contour obtained for voiced portions of the sounds (using the YIN algorithm) were correlated with the oscillatory activity of individual simulated areas for all 630 speech stimuli. Mean correlation values were A1: 0.46 (SEM: 0.02), R: 0.47 (SEM: 0.02), *Slow*: -0.14 (SEM 0.01), *Fast*: 0.59 (SEM 0.01)

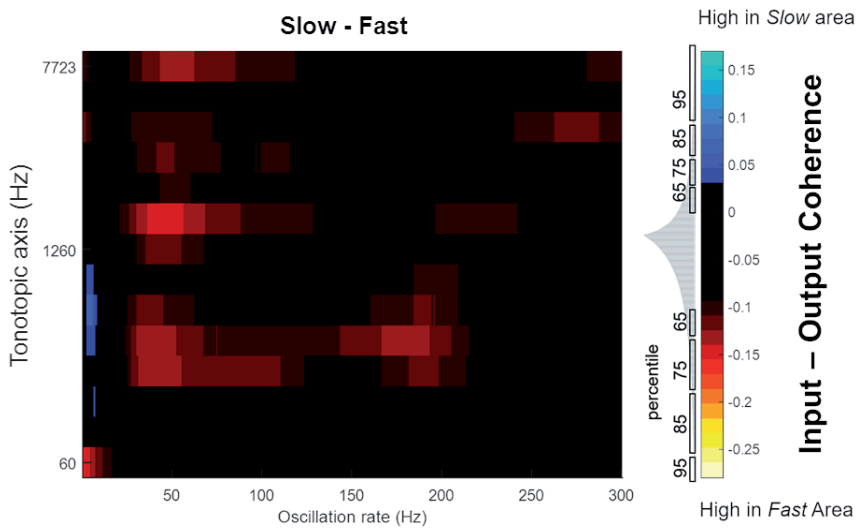


Figure 10: Mean difference in magnitude spectrum coherence between belt regions. The *Slow* area showed higher coherence with slow oscillations in speech (4 – 8 Hz, indicated by blue colors). Instead, the *Fast* area showed greater coherence to faster oscillations of speech (30 – 70 Hz, around 100 Hz, and 150 – 200 Hz, indicated by the warmer colors). The distribution of the difference in magnitude spectrum coherence between the *Slow* and *Fast* area for all 630 sounds is shown in gray, adjacent to the color bar, with percentiles marked to indicate the statistical significance.

showed that the *Fast* area best represented the pitch information through synchronization to instantaneous $F_{0.}$

Figure 11 highlights the presence of a dual mechanism for coding pitch, as pitch information is present in both spectral (i.e., spatially, by different units) and temporal (by different oscillatory activity) model responses for a sample sound (male speaker, sentence duration 3.26 s; selected from LDC TIMIT database, Garofolo et al. (1993)). In Figure 11A, the time-averaged response to the speech sentence across the tonotopically-organized channels in the four simulated areas is shown. In all the areas, a peak in the response profile can be observed in those frequency channels that matched the F_0 of the speaker (best estimate computed using YIN algorithm: 109 Hz). This spectral (i.e., spatial) representation of the speech signal's pitch is strongest in the *Slow* area and weakest in the *Fast* area. A1 and R show similar profiles with respect to each other. Contour tracking of pitch in the *Fast* area with the sample sound (correlation 0.74) is shown in Figure 11B (pitch contour of the speech signal measured by YIN algorithm is shown as the white boxes). The simulated *belt* regions show functional specialization to represent pitch spectrally (in the *Slow* area) and temporally (in the *Fast* area) in parallel streams.

Overall, the model responses to speech sounds highlight the presence of a distributed code for representing different temporal features of speech signals at the level of *belt* regions, but not for the *core* regions. Each *belt* area showed a functionally relevant specialization, as the temporal features highlighted by *Slow* and *Fast* areas are key structures of speech signals.

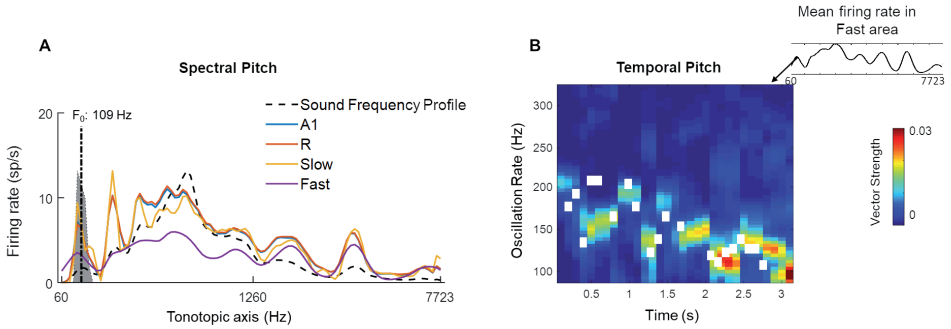


Figure 11: A dual code for pitch estimation. For a sample sound, (A) Mean firing rate of all units in the four simulated areas (A1, R, *Slow* and *Fast*, colored lines) is shown. Sound frequency profile (scaled) is plotted in a black dashed line for reference. The gray highlighted portion of the plot indicates estimates of pitch by YIN algorithm (distribution over time, with the best estimate of F_0 plotted with a dash-dotted line, de Cheveigné and Kawahara (2002)). A spectral code is observed in model outputs with firing rate peaks overlapping with YIN estimates. (B) Temporal code for pitch is observed as weak synchronization to pitch contour in oscillatory activity (measured as Vector Strength) of the *Fast* area unit corresponding to spectral peak corresponding to best pitch estimate by YIN algorithm. The pitch contour estimates over time computed by YIN algorithm are depicted by white boxes. The correlation between YIN estimates the temporal profile of *Fast* area is 0.74.

4 Discussion

In this study, we presented a computational model of the AC that consists of information processing streams optimized for processing either fine-grained temporal or spectral information. The model is employed to investigate the contribution of the different cortical streams in the representation and processing of basic acoustic features (i.e., temporal modulation, pitch) in the context of artificial and natural (speech) stimuli.

We started by simulating responses to artificial AM sounds. Electrophysiological studies have characterized AM coding by a dual mechanism of temporal (synchronization) and rate coding (Joris et al., 2004). In comparison with the phase-locking in the auditory nerve [reported up to 1.5 – 8kHz in humans, Verschooten et al. (2019)], the synchronization code has been measured to be comparatively diminished at the level of the cortex for human and non-human primates. The preferred AM rates have been reported as ranging from 1-50 Hz in monkeys (Steinschneider et al., 1980; Bieser and Müller-Preuss 1996; Lu et al., 2001), despite neurons have been shown to synchronize as high as 200 Hz in monkeys (Steinschneider et al., 1980) and similar weak synchronization could be detected in humans with electrocorticography (Nourski et al., 2013). In agreement with these electrophysiology studies, our model exhibited a dual coding mechanism. While the contribution of a temporal code (synchronization) was strong up to a maximum of 50 Hz, synchronizations became weaker for higher modulation rates and were complemented with a rate code mechanism.

Furthermore, in electrophysiology, the maximum AM rate for which a temporal code is present has been reported to differ across fields of the AC (Liang et al., 2002). Caudal fields (i.e., regions belonging to the dorsal processing stream) are reported to be as fast as or even faster than the primary AC and synchronize with the stimulus envelope up to high AM rates. Instead, the rostral field (i.e., part of the ventral processing stream) does not show a temporal code for AM sounds but codes AM with changes in firing rate (i.e., a rate code) (Bieser and Müller-Preuss, 1996). In the simulated responses, the relative contribution of the temporal and rate coding mechanisms also varied across the simulated cortical areas, depending upon the areas' temporal and spectral processing properties. While the temporal code displayed a low-pass filter profile, the shape of the rate code varied from low-pass to band-pass and band-stop patterns. Evidence for such variation in rate coding pattern has been reported in electrophysiological studies as well with sAM stimuli (Bendor and Wang, 2008; Bieser and Müller-Preuss, 1996; Liang et al., 2002; Schreiner and Urbas, 1988). In our model, this observation was highlighted when the firing rate was examined within carrier-matched frequency channels. The interaction of spectral and temporal response properties underlies these observations.

In order to assess the relationship between neural population activity (i.e., synchronization and firing rate) with human behavior, we next used the model to simulate psychoacoustic experiments. We were able to successfully predict psychoacoustically-determined modulation detection thresholds (i.e., modulation detection transfer functions, tMTFs) for AM noise and tones (Bacon and Viemeister, 1985; Kohlrausch et al., 2000). The model suggested a role for auditory *core* areas, rather than *belt* areas, in coding modulation detection with simple AM stimuli. The tMTF for AM noise was replicated by computing temporal synchronization. However, for AM tones, we observed the best prediction of the psychoacoustical tMTF by using a combination of synchronization and spatial (sideband detection) code. Additionally, we observed that compared to low-frequency carriers, high carriers allowed modulation detection up to faster rates. This replicated psychoacoustic observations of detection up to faster modulation rates with a higher carrier frequency (Kohlrausch et al., 2000; Moore and Glasberg, 2001; Sek and Moore, 1995; Simpson et al., 2013). Our simulations indicate that these frequency-specific responses, which arise at the periphery, are inherited by the cortex, especially in the *core* areas.

We further evaluated the contribution of temporal coding mechanisms to psychoacoustical phenomena. While current views on pitch perception suggest that the role of synchronization is limited to the auditory periphery and cortex might use information from individual harmonics (Plack et al., 2014), there is evidence of temporal cues being used especially for unresolved harmonics for low pitch sounds (Bendor et al., 2012). The model successfully decoded the low frequency missing fundamentals of complex tones and showed a dependence of the strength of synchronization on the order of harmonics. By simulating a psychoacoustic task employing missing fundamental complex tones with varying unresolved harmonics, we further investigated the role of synchronization and its dependence on the number and order of harmonics. The model output matched the previously reported human behavior performance through synchronization in the simulated neural responses, but not by a rate coding mechanism. That is, we could successfully replicate three key findings from Houtsma and Smurzynski (1990). First, the synchronization to the missing F_0 was stronger for the lower compared to higher-order harmonic sounds and second, it improved with an increasing number of components of complex tone. Third, only for the lower order harmonic sounds, the improvement in model performance was sharp when the number of components was increased from two to four and displayed a plateau when further components were added. Interestingly, the match between psychoacoustics and the model output was limited to the *Fast* area, suggesting a role for this fine-grained temporal processing stream in the extraction of the pitch using temporal cues. Additionally, using speech sounds, we further observed a strong spatial (spectral) pitch correlate (observed in all areas, strongest in *Slow* area) along with weaker oscillations tracking pitch contour (only in *Fast* area). However, the spatial code is not observable in model output for pitch with missing fundamental complex tones

and suggests the need for a more complex network to effectively detect pitch just from harmonic information in space. Moreover, the temporal code for pitch can benefit from feedback connectivity (Balaguer-Ballester et al., 2009) while precise interspike intervals can shed light on phase sensitivity of pitch perception (Huang and Rinzel, 2016). Thus, future model modifications can move from general (current) to more specific hypotheses of auditory processing.

Coding of pitch in the AC has been extensively investigated with fMRI, resulting in somewhat conflicting findings. While some studies pointed to lateral Heschl's Gyrus (HG) as a pitch center (De Angelis et al., 2018; Griffiths and Hall, 2012; Norman-Haignere et al., 2013), other studies showed that pitch-evoking sounds produced the strongest response in human planum temporale (PT) (Garcia et al., 2010; Hall and Plack, 2009). This disagreement may be due to differences between studies in experimental methods and stimuli. Our computational model provides an opportunity to merge these fMRI-based findings, as it allows for the efficient and extensive testing of model responses to a broad range of sounds. Based on the sounds we tested, observations of a pitch center in PT, part of the *Fast* stream, may be dominated by temporal pitch. Instead, human fMRI studies reporting a pitch area in lateral HG (De Angelis et al., 2018; Griffiths and Hall, 2012; Norman-Haignere et al., 2013), which is part of the *Slow* stream), maybe reflecting the spectral rather than the temporal processing of pitch. Our simulations suggest a functional relevance for temporal representations albeit through weak synchronization. These predictions are in line with evidence of synchronization in the AC contributing to the percept of pitch (up to 100 Hz) observed with MEG (Coffey et al., 2016) and require future studies with both high spectral and temporal precision data from the AC.

The distributed coding pattern shown by the different regions (i.e., coding of modulation detection thresholds by the *core* regions, coding of temporal pitch by the *Fast* area and spectral acuity by the *Slow* area of the *belt* stream) reflected a hierarchical processing scheme based on varying spectro-temporal properties of the neural populations. We then applied this modeling framework to the analysis of (continuous) speech with the aim of exploring the influence of basic neural processing properties on the representation and coding of speech. All modeled areas represented the slow oscillations present in speech (<20 Hz). In the belt areas, an additional distributed coding of temporal information was observed. That is, the optimization for coding slow temporal changes with high spectral precision in the *Slow* stream resulted in the coding of temporal oscillations in the lower 4–8 Hz frequency range. Processing properties similar to those of the *Slow* stream may thus be suited for coding spectral pitch and prosody in speech signals. Instead, optimization for processing fast temporal changes with low spectral precision in the *Fast* stream resulted in the coding of temporal oscillations in the higher 30 – 70 Hz and 100 – 200 Hz frequency ranges. Processing properties similar to those of the *Fast* stream

may therefore instead be optimal for coding phonemes (consonants), and temporal pitch. In sum, we showed that the hierarchical temporal structure of speech may be reflected in parallel and through distributed mechanisms by the modeled areas, especially by simulated belt areas. This is in line with the idea that the temporal response properties of auditory fields contribute to distinct functional pathways (Jasmin et al., 2019).

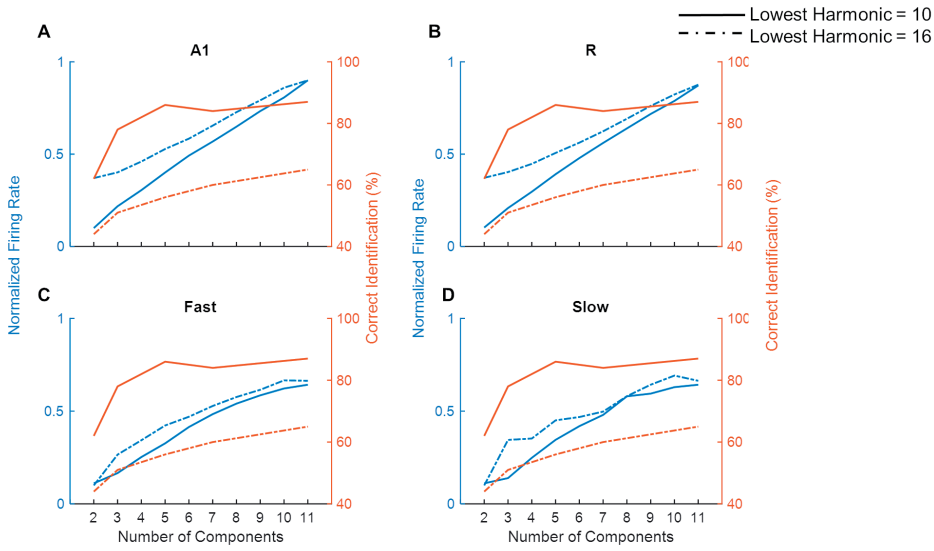
The "division of labor" observed between the simulated processing streams provides predictions regarding cortical speech processing mechanisms. Specifically, the slowest oscillations, representing the speech envelope, were coded in parallel across regions with different processing properties and may serve to time stamp the traces of different speech aspects belonging to the same speech utterance across streams. This may serve as a distributed clock: A binding mechanism that ensures the unified processing of different components of speech (Giraud and Poeppel, 2012; Yi et al., 2019) that are instead coded in a distributed fashion. Such a temporal code can also underlie the binding of auditory sources in stream segregation (Elhilali et al., 2009). While in the current implementation of the model the responses are driven by stimuli, the model could be extended to include stimulus-independent oscillatory cortical activity. As the oscillations inherent to AC processing that occur on multiple timescales are known to decode complimentary informational structures in speech processing (Overath et al., 2015) and auditory scene analysis, such a model extension may in the future be used to study the effects on these 'inherent' oscillations on responses to speech and other structured inputs.

To summarize, we have presented a recurrent neural model built on simple and established assumptions on general mechanisms of neuronal processing and the auditory cortical hierarchy. Despite its simplicity, the model was able to mimic results from (animal) electrophysiology and was useful to link these results to those of psychophysics and neuroimaging studies in humans. As the response properties of the AC (tonotopic organization, phase-locking, etc.) are inherited from the periphery, it remains possible that the model in actuality depicts earlier stages in the auditory pathway rather than AC. In future implementations of the model, the distinction between peripheral and cortical stages can benefit from a more detailed peripheral model (Meddis et al., 2013, Zilany et al., 2014). Ultimately, establishing a clear distinction between peripheral and cortical contribution would require simultaneous high-resolution (spatial and temporal) recordings across multiple locations of the auditory pathway and cortex. Furthermore, how the model dynamics shape up in presence of intrinsic noise in the system can also provide interesting insights into sound processing.

Nonetheless, the model is valuable for generating hypotheses on how the different cortical areas/streams may contribute towards behaviorally relevant aspects of acoustic signals. The presented model may be extended to include a physiological model of neurovascular

coupling (Havlicek et al., 2017) and thus generate predictions that can be directly verified using functional MRI. Such a combination of modeling and imaging approaches is relevant for linking the spatially resolved but temporally slow hemodynamic signals to dynamic mechanisms of neuronal processing and interaction.

Supplementary Materials



Supplementary Figure 1: Model performance on a missing fundamental task. The model performance in detecting missing fundamental of complex tones (average firing rate, normalized across all tones) is shown for areas A1, R, Fast and Slow (blue lines in panels A, B, C, and D respectively). Human behavioral data on pitch identification (%) task (Houtsma and Smurzynski, 1990) is plotted in orange lines. Solid lines show complex tones with lowest harmonic at 10 while the dash-dotted lines show lowest harmonic component at 16.

Chapter 3

Predicting Neuronal Response Properties from Hemodynamic Responses in the Auditory Cortex

---

Zulfiqar I.*, Havlicek M.*, Moerel M., and Formisano E. (in revision).
Predicting Neuronal Response Properties from Hemodynamic Responses in the Auditory Cortex.
*equal contribution

Abstract

Recent functional MRI (fMRI) studies have highlighted differences in responses to natural sounds along the rostral-caudal axis of the human superior temporal gyrus. However, due to the indirect nature of the fMRI signal, it has been challenging to relate these fMRI observations to actual neuronal response properties. To bridge this gap, we present a forward model of the fMRI responses to natural sounds combining a neuronal model of the auditory cortex with physiological modeling of the hemodynamic BOLD response. Neuronal responses are modeled with a dynamic recurrent firing rate model, reflecting the tonotopic, hierarchical processing in the auditory cortex and the spectro-temporal tradeoff in the rostral-caudal axis of its belt areas. To link rostral-caudal differences in neuronal response properties with human fMRI data in the auditory belt regions, we generated a space of neuronal models, which differed parametrically in spectral and temporal specificity of neuronal responses. Then, we obtained predictions of fMRI responses through a biophysical model of the hemodynamic BOLD response (P-DCM). Using Bayesian model comparison, results showed that the hemodynamic BOLD responses of the caudal belt regions in the human auditory cortex were best explained by modeling faster temporal dynamics and broader spectral tuning of neuronal populations, while rostral belt regions were best explained through fine spectral tuning combined with slower temporal dynamics. These results support the hypotheses of complementary neural information processing along the rostral-caudal axis of the human superior temporal gyrus.

1 Introduction

Auditory information in the primate auditory cortex (AC) is processed hierarchically from core to belt and then to the parabelt region (Kaas and Hackett, 2000). The hierarchical organization allows for efficient sequential information processing, stemming from the dense connectivity between core and belt areas, and then belt and parabelt areas. Already at the early stages of information processing, at the level of belt areas, neuronal populations along the rostral-caudal axis show distinct response property profiles (Scott et al., 2017). In comparison with core and other surrounding areas of the AC, neurons in the rostral stream show longer latencies and narrow frequency tuning (Recanzone et al., 2000; Bendor and Wang, 2008; Tian et al., 2001; Camalier et al., 2012). On the other hand, neurons in the caudal areas show broader frequency tuning and latencies that are comparable or shorter to the primary AC (Recanzone et al., 2000; Kuśmierek and Rauschecker, 2014). These differences in neuronal properties are thought to support the specialized functions of *'what'* and *'where'* (or *'how'*) processing in the rostral and caudal streams, respectively (Jasmin et al., 2019; Kaas et al., 1999; Belin et al., 2000; Rauschecker and Tian, 2000).

Apart from different neuronal properties measured with electrophysiology in non-human primates, differences in responses across the rostral-caudal pathway have also been measured using functional MRI (fMRI) in humans. Recent neuroimaging studies have reported a spectro-temporal trade-off of responses, with a preference for fine spectral structures of sounds in rostral regions, in comparison with preference to fine temporal features of sounds in the caudal regions (Schönwiesner and Zatorre, 2009; Santoro et al., 2014). However, the hemodynamic blood oxygenation level-dependent (BOLD) signals measured with fMRI originate from (nonlinear) vascular changes (i.e., changes in blood oxygenation, blood flow, and blood volume) in response to neuronal activity. While the resulting fMRI signal is correlated to the underlying neuronal activity (Logothetis et al., 2001; Logothetis et al., 1999; Rees et al., 2000), it does not directly measure the neuronal activity. Thus, it remains to be determined whether the spectro-temporal preferences along the rostral-caudal streams inferred from the modeling of fMRI data (Santoro et al., 2014) are a direct result of fundamental neuronal mechanisms and response properties as observed in electrophysiology. The relationship between neuronal and hemodynamic responses is commonly (but less accurately) approximated with a simple linear convolution model. Nonlinear biophysical generative models provide an improvement over linear convolution approaches by more accurately capturing the nonlinear neuronal dynamics and have been shown to describe the causal relation between neuronal activity and the data measured from different modalities (Friston et al., 2003). Within the dynamic causal modeling (DCM) framework, forward models have been used to generate predictions about observed responses using a combination of neuronal

models (for a review of different neuronal models, see Moran et al., 2013) and modality-specific measurement models [e.g., hemodynamic model for fMRI: Havlicek et al., 2015; Stephan et al., 2008, lead field model for E/MEG: Kiebel et al., 2006)].

In this article, we adopt a DCM framework to study the causal link between the properties of neuronal populations in auditory belt regions and fMRI responses to natural sounds. We hypothesize that, just as evidenced in electrophysiology, the rostral-caudal axis of the human superior temporal plane will cluster into two distinct populations; a caudal region that is spectrally less specific (i.e., broader frequency tuning) with faster temporal responses, and a rostral region with narrow frequency tuning but longer temporal responses. In the forward model, we use a dynamic recurrent neuronal model of the AC (Zulfiqar et al., 2020), which incorporates the hierarchical organization of the AC where information flows from the primary core to secondary belt auditory regions (Kaas and Hackett, 2000), as well as a rostral-caudal organizational axis along which neuronal units differ in their spectro-temporal properties (i.e., their frequency tuning width and response latency) as reported in animal electrophysiology (Scott et al., 2017; Bendor and Wang, 2008; Camalier et al., 2012; Recanzone et al., 2000; Tian et al., 2001; Kuśmierek and Rauschecker, 2014). To simulate the BOLD responses, we use a generative hemodynamic model of the BOLD signal presented in a recent DCM extension (physiological DCM [P-DCM], Havlicek et al., 2015) which provides several advances over standard approaches (Buxton et al., 2004; Friston, Mechelli et al., 2000; Sotero and Trujillo-Barreto, 2007). P-DCM uses feedforward neurovascular coupling in the measurement model, which allows the dynamic changes in neuronal activity to be better reflected in the BOLD response, whereas standard DCM uses feedback-based neurovascular coupling, which due to its oscillatory behavior can diminish some dynamics of neuronal response. Furthermore, passive vascular uncoupling between cerebral blood flow and cerebral blood volume allows us to explain hemodynamic transients that are not explainable by changes in neuronal activity, i.e., it accounts for the vascular source of variability in the hemodynamic BOLD response (Havlicek et al., 2015).

To test our hypotheses at the voxel level, we first generate a space of neuronal models, which differ parametrically in spectral and temporal specificity of neuronal responses. We then simulate the BOLD responses and perform a voxel-wise model comparison using Bayesian Model Selection (BMS). Results showed that the hemodynamic BOLD responses of the caudal belt regions in the human auditory cortex were best explained by modeling faster temporal dynamics and broader spectral tuning of neuronal populations, while rostral belt regions were best explained through fine spectral tuning combined with slower temporal dynamics.

2 Methods

An overview of the implemented methodology is shown in Figure 1. In the workflow of P-DCM (Havlicek et al., 2015), we substitute an adaptive two-state neuronal model with a dynamic recurrent firing rate model, specifically tailored to sound processing in the AC (Zulfiqar et al., 2020). The model space is populated with the AC models that simulate multiple spectro-temporal response properties for the *belt* region (with spectral specificity varying from broad to narrow, and temporal latencies ranging from very fast i.e., few milliseconds, to slow i.e., hundreds of milliseconds). The simulated responses are fitted to the BOLD responses from the measured fMRI dataset (from the belt regions)

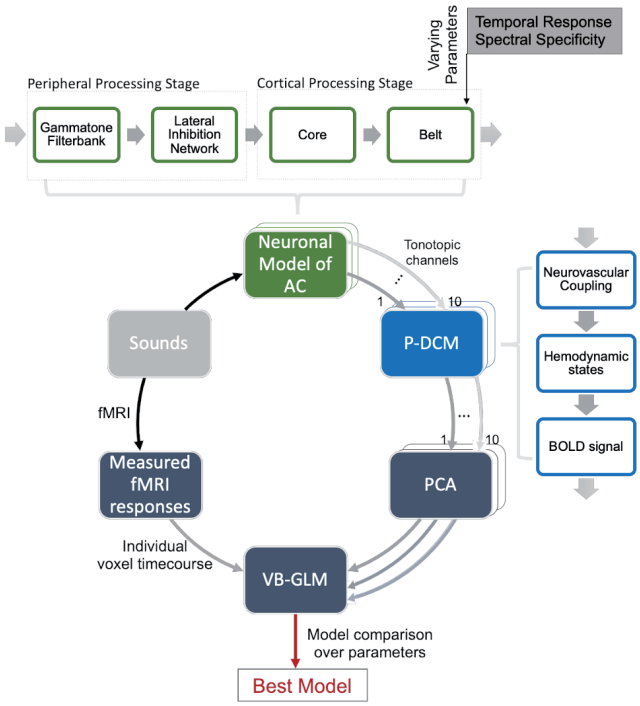


Figure 1: Procedural workflow for predicting the neuronal response properties from the BOLD signals. Sounds are passed through a hierarchical, tonotopic neuronal model of the AC. In different iterations of the cortical processing stage, the *belt* area is modeled with a distinctive spectro-temporal response profile. The output of each of the belt simulations is processed through the three remaining stages of P-DCM to generate simulated BOLD responses. To retain maximum model-specific information, the output of the P-DCM is reduced to three principal components using Principal Component Analysis (PCA). These principal components of the predicted timecourses from all models are fitted to the measured fMRI responses (in the auditory belt regions) for each voxel using VB-GLM. By model comparison, the best model prediction for each voxel is generated. This prediction is linked to the neuronal model properties, resulting in characteristic temporal response and spectral specificity of the neuronal population underlying the voxel activation measured using fMRI.

using a Variational Bayesian optimization for General Linear Model (VB-GLM, Penny, 2012). By model comparison using the Free Energy metric (Friston and Stephan, 2007), we predict the best neuronal model (with a specific spectral and temporal preference) for each voxel, thereby assigning neuronal response properties to each voxel based on its hemodynamic responses. These stages are discussed in detail in the following sections.

2.1 Measured fMRI Dataset

The measured fMRI responses are from an existing 7 Tesla fMRI dataset (Santoro et al., 2017), where responses to 288 natural sound stimuli were measured in 5 subjects. These sounds included samples of human vocalizations (including speech), animals, musical instruments, tool sounds, and sounds of nature, and were of 1 s duration. Responses to these sounds were measured in a fast event-related design (TR = 2.6 s, average inter-stimulus interval = 7.8 s) with 1.5 mm isotropic voxel size (1.5 mm isotropic; Santoro et al., 2017) in 12 runs. For this study, the measured fMRI responses for each voxel were averaged across sounds irrespective of sound categories.

The neuronal model described below is used to generate responses to the 288 natural sounds, which were processed in the same order by the model as they were presented to the participants in the aforementioned study. The simulations were performed for all 12 runs and then they were concatenated. Some parameters used in P-DCM were also updated based on the specifics of the current dataset and are described below.

2.2 Neuronal Model of the Auditory Cortex

The auditory processing pathway is modeled by modifying an existing hierarchical two-stream computational model of the AC (Zulfiqar et al., 2020). The model consists of two stages; a *peripheral processing stage* and a *cortical processing stage*, and approximates neural sound processing using recurrent excitatory and inhibitory populations.

2.2.1 Peripheral Processing Stage

In the *peripheral processing stage* (Figure 1), the sounds are filtered using a set of band-pass filters that simulate the tonotopic cochlear filtering (100 filters, 4^{th} order Gammatone filterbank implementation by Ma et al., 2007). The filter bandwidths are fixed at 1 ERB (Equivalent Rectangular Bandwidth, Glasberg and Moore, 1990). The center frequencies of the filters are equally spaced on the ERB_N number scale (between 50 – 8000 Hz) with a distance of 0.3 Cams (ERB_N is the ERB of the auditory filters estimated for young people with normal hearing). After cochlear filtering, the response of the filters is tonotopically sharpened by the Lateral Inhibition Network, LIN (Chi et al., 2005) (Figure 2). The LIN is implemented by taking a spectral derivative followed by half-wave rectification and temporal integration of the output. To remove any boundary effects of filtering,

the output of the first and the last filters is removed, resulting in 98 units (with center frequencies between 60 – 7723 Hz).

2.2.2 Cortical Processing Stage

The *cortical processing stage* consists of two auditory areas. The first area approximates a primary *core* region while the second area represents a secondary *belt* region of the auditory cortex. These areas are simulated using an adaptation of the Wilson-Cowan excitatory and inhibitory units (Wilson and Cowan, 1973; 1972) as reported in Zulfiqar et al. (2020). The key response properties of the simulated *core* and *belt* regions are frequency tuning curves (FTCs) and temporal dynamics. The FTC defines the spatial resolution of a simulated area and is quantified using the Quality factor (Q) where:

$$Q = \frac{Best\ Frequency}{Half\ Maximum\ Bandwidth},$$

while the temporal latencies of the population regulate the temporal dynamics of the model. The response properties of the simulated *core* area are set to model the primary auditory cortex (A1) while the *belt* region is varied systematically to generate the model space.

2.2.3 Model Space

As the basic response properties in the belt region have been found to change along its rostral-caudal axis, we generate 28 models of the belt approximating the variety of spectro-temporal responses. That is, the model space ranges from broad to narrow spectral specificity, and from fast to slow temporal responses (Figure 2). The spectral specificity (measured as Q) is modified (in 4 steps, from narrow to broad) by varying spatial spread (modeled by parameter σ, see (Zulfiqar et al., 2020)) within *belt* models (smaller spread of excitation and larger inhibition leads to narrow frequency tuning). When a change in the spatial spread of activity did not further affect spectral resolution Q), the connectivity kernel (see (Zulfiqar et al., 2020) for further details) between *core* and *belt* units was iteratively increased in space (many units projecting to single unit results in broad frequency tuning). The connectivity kernels that are convolved with the spatial input, are reported in Supplementary Table 1. As the spatial and temporal axes are not independent of each other, the procedure of selecting σ and connectivity kernel was iteratively repeated for all separate values of time constant [modeled by parameter τ, see (Zulfiqar et al., 2020)]. The resulting Q values are reported (mean and standard deviation) for the whole tonotopic axis for each model, along with Q for a unit with the best frequency 1 kHz to provide reference across models. For simplification, the Q values are discretized to 3, 6, 9, and 12 (smaller values indicate broader responses) to simplify the model space.

Figure 2 shows the final model space (each model in the model space is generated by parameters reported in Supplementary Table 1) and the hypothesized response properties of rostral-caudal regions in reference to the model space. The temporal latencies, controlled by parameter τ, are varied in 7 steps (3, 10, 50, 100, 200, 300, and 400 ms). Additionally, based on electrophysiological evidence (Scott et al., 2011), the values of τ decrease linearly with increasing best frequency along the tonotopic axis in all of the individual *belt* models (extreme values are reported in Supplementary Table 1). Neuronal responses from all the models of the model space were generated in response to sounds, as presented in the measured fMRI dataset.

2.3 Measurement Model for fMRI

To reduce the number of computations, for each of the neuronal models, the output is downsampled in space to 10 tonotopic units (mean over channels) and in time to 10 Hz. This downsampled output acts as input to the measurement model comprising of a feed-forward neurovascular coupling (NVC) model along with a hemodynamic model including viscoelastic properties of venous vessels and a physical BOLD signal model [as reported by Havlicek et al. (Havlicek et al., 2015)]. The current neuronal model, in contrast with the neuronal model proposed by Havlicek et al. (2015), simulates faster temporal dynamics but no neuronal contributions to post-stimulus undershoot. Thus, the model parameters of NVC and hemodynamic models differed slightly from the default parameterization.

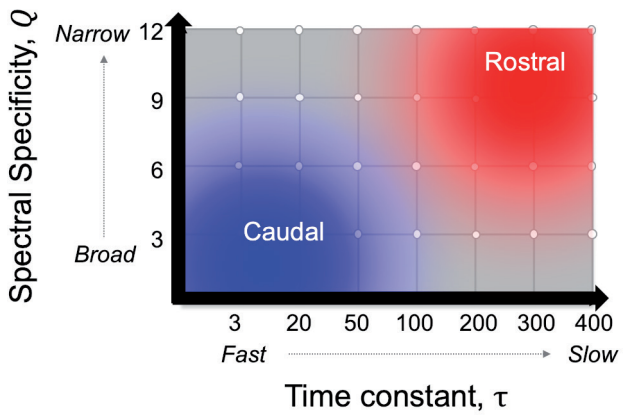


Figure 2: The model space showing 28 neuronal models of varying temporal and spectral resolution for the simulated *belt* region. Spectral specificity is quantified by the Quality Factor (*Q*). Higher values of *Q* indicate narrow frequency tuning and vice versa. The characteristic temporal dynamics of each model are indicated by the time constant (τ). Higher values of τ indicate slower temporal dynamics. The blue (red) region highlights the hypothesis that neuronal models with broad (narrow) tuning curves and faster (slower) responses will be the best fit model for caudal (rostral) belt AC.

The NVC is modeled as reported by Havlicek et al. (2015) with few changes in parameter values (see Supplementary Table 2). The output of the neuronal model i.e., the excitatory activity modulated by inhibitory activity, is transformed to blood flow in a strict feedforward fashion, via vasoactive signal. The feedforward NVC ensures that neuronal dynamics are conveyed to blood flow response, albeit in a smooth version. Decay and delay of the cerebral blood flow response with respect to the neuronal response are regulated with fixed constants reported in Supplementary Table 2.

The hemodynamic model is represented by the balloon model (Buxton et al., 1998). It models the mass balance of normalized changes in blood volume and deoxyhemoglobin content as they pass through the venous compartment. Their changes are driven by changes in blood inflow and oxygen metabolism respectively. It is assumed that blood inflow and oxygen metabolism are linearly coupled, with *n*-ratio equals 3 (Buxton et al., 2004). The steady-state relationship between blood outflow and volume is given by the power-law relationship (Grubb et al., 1974). During transient periods, the viscoelastic time constant venous controls uncoupling between blood inflow and volume. Given the short stimulus duration used in the measured fMRI dataset (1 s), we assume only very small uncoupling during the response inflation phase but larger uncoupling during the deflation phase. This larger uncoupling is responsible for modeling post-stimulus BOLD response undershoot. The updated parameter values are reported in Supplementary Table 3.

The BOLD signal is implemented as a function of deoxyhemoglobin content and blood volume, which comprise a volume-weighted sum of extra- and intra-vascular signal components (Havlicek et al., 2015; Havlicek and Uludağ, 2020). The parameters of the BOLD signal model adjusted for the magnetic field strength (7 T) of the measured fMRI dataset are reported in Supplementary Table 4.

2.4 Channel Reduction, Bayesian Model Fitting and Model Selection

As the values selected for parameter *Q*, and those for τ, are quite close to each other across the model space, spatial spread of activity within simulated region and connectivity between simulated regions resulted in correlated activity in neighboring tonotopic channels. Similarly, as τ and *Q* values between adjacent models did not vary drastically, model responses were correlated across the model space as well. This was further exacerbated by the low temporal resolution of the BOLD signal. Thus, to capture maximum variation in BOLD responses, the PCA analysis was used to reduce 10 tonotopic channels (for each of the models) into three principal components capturing >95% of simulated data variance. Linear combination of three principal components (PCs) and set of independent intercepts for each functional run were fitted to the single voxel timecourse using VB-GLM. This VB-GLM approach optimizes the weights of the

linear model by maximizing the log-model evidence that is approximated by the free energy, *F*. By fitting all 28 neuronal models to each voxel, the free energy can be used to perform to Bayesian model selection (BMS) of the most favored model given the data (Penny, 2012). In particular, models were compared by taking the log-evidence difference with respect to the model with the lowest log-evidence, which equals the log Bayes factor $\ln B_i = F_i - \min(F)$ and can be also expressed in terms of posterior model probability $p_j = B_j / \sum_j B_j$ (Kass and Raftery, 1995).

2.5 Model Validation

Using fMRI encoding, Santoro et al. (2014) reported regions most sensitive to fast temporal changes and broader spectral features, and fine spectral features and slower temporal changes in the stimuli, respectively. However, due to the limitations of fMRI, we cannot conclude that these functional differences are due to specific underlying neuronal properties. With a dataset similar to Santoro et al. (2014), the aforementioned modeling approach can help us test the hypothesis that the activity in the caudal regions would best be described by models with low spectral specificity (lower *Q*) and high temporal precision (smaller τ) in their neuronal responses and vice versa for rostral regions (higher *Q* and larger τ).

To test the aforementioned hypothesis, we first replicated these findings in the fMRI dataset (Santoro et al., 2017) following the same procedures as in Santoro et al. (2014). Three maps were generated per participant, per hemisphere (see Figure 3A, B, and C for maps for a representative participant, right hemisphere) detailing the frequency preference (tonotopic map), temporal-feature preference (temporal modulations, shown in rate map), and spectral-feature preference (spectral modulations, shown in scale map) across the AC. Tonotopy and anatomical markers were used to identify Heschl's gyrus to indicate the location of the core regions. Next, we identified a *Caudal* and *Rostral* region posterior to the Heschl's gyrus, as a region most sensitive to fast temporal changes and broader spectral features, and fine spectral features and slower temporal changes in the stimuli, respectively (Figure 3D).

The timecourses of the voxels in the two labeled streams, irrespective of their label, were used for model fitting. Through model comparison, we selected a single best neuronal model from the model space for each voxel. As each of the models was characterized by different neuronal response properties, this resulted in a prediction of the optimal *Q* and τ across the voxel space. These predicted properties were then compared across the labeled *Caudal* and *Rostral* region using a non-parametric statistical test.

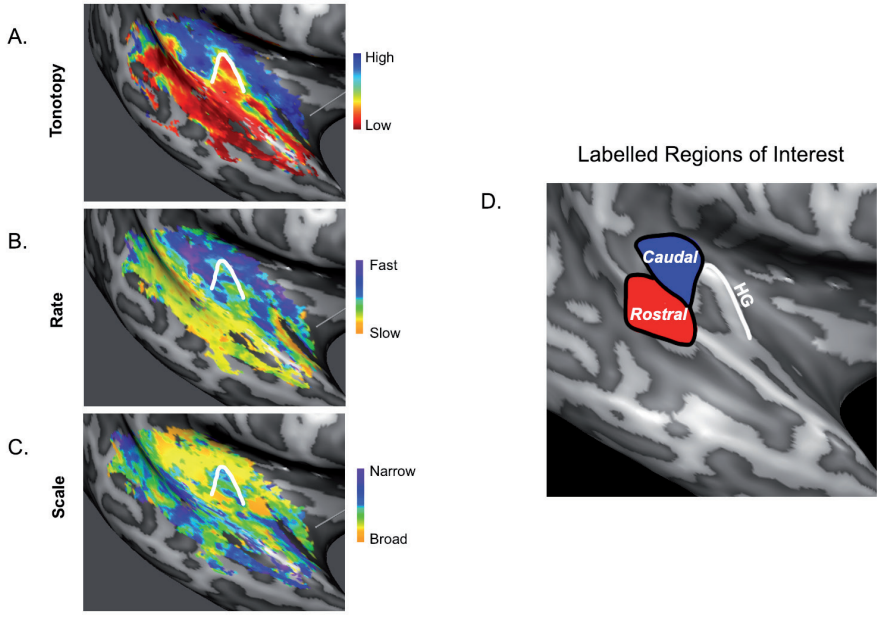


Figure 3: Labeled regions of interest. The *Caudal* and *Rostral* areas were labeled based on their responses to spectro-temporal changes in the sounds as reported by Santoro et al. (2014). The regions are displayed for a single subject (right hemisphere) along the superior temporal plane. The Heschl's Gyrus (HG) is marked in white.

2.6 Predicting the Best Frequency per Voxel

In order to evaluate the frequency channel contributing most to the voxel timecourse, given the earlier selected best neuronal model, we constructed a feature vector, w_j (j = 1 ... 10), by projecting each channel (simulated BOLD signal) timecourse, x_j, into subspace spanned by three PCs (i.e., features are weights of a linear combination of the PCs contributing to specific channels). Similarly, we created a feature vector, w^*, representing the contribution of the PCs into model prediction given by VB-GLM estimate, $\hat{y}$, described above. The similarity between the feature vector based on voxel timecourse prediction and the feature vector of a specific frequency channel was evaluated by fitting a linear model using the VB-GLM approach. Bayesian Model Selection (BMS) based on free energy, previously evaluated for 10 channels, was then used to select the best channel per voxel. Note that a simpler approach to evaluate the similarity between two vectors, e.g., based on Euclidean distance, would yield comparable results. Tonotopic channels of the model present a range of 60 to 7723 Hz range. This procedure (summarized in Figure 4) predicted the best frequency for each voxel in the labeled regions.

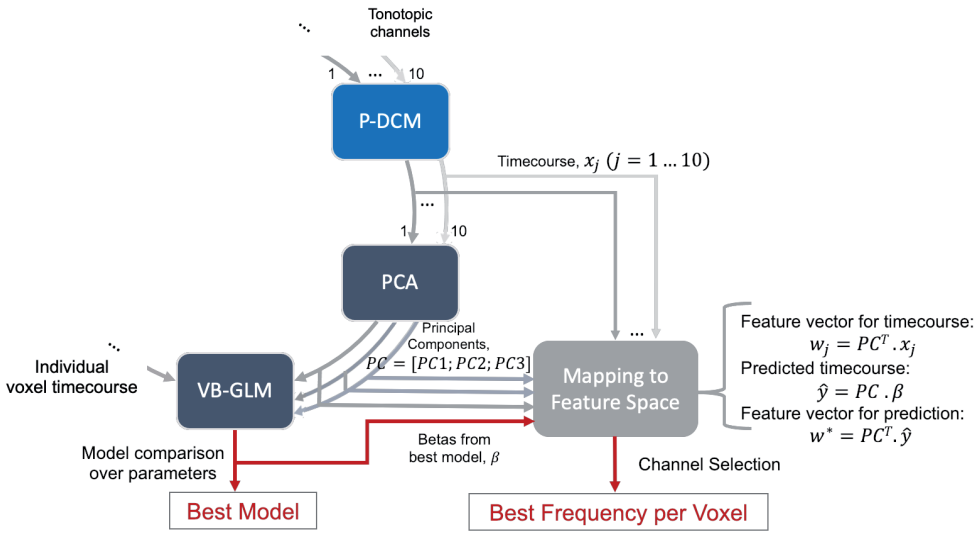


Figure 4: Predicting the best frequency per voxel. The simulated BOLD timecourses (for all 10 tonotopic channels), their principal components, and best model prediction were used to generate feature vectors. Similarity analysis between the feature vector of all timecourses and feature vector of the best prediction (using VB-GLM) was used to compare the 10 tonotopic channels (as 10 models) using Bayesian Model Selection and select the best channel providing the best frequency per voxel.

3 Results

3.1 Simulated Neuronal and Hemodynamic Responses

We investigated the characteristics of the model space by comparing the simulated neuronal responses and the simulated BOLD responses across models. Figure 5 indicates the results of the simulations for four models along the diagonal of the model space. The neuronal responses (Figure 5A and C) are shown normalized by the mean firing rate at 1 s (duration of each sound stimulus) and the two adjacent time points (before and after 1 s mark). The BOLD responses (Figure 5B, D, and E) are plotted normalized by peak response at the time point at 5 s. The displayed BOLD responses are average responses obtained with linear deconvolution from simulated time-courses.

Figures 5A and C show that the models across the diagonal of the model space depict a variety of neuronal behaviors. The effects of temporal integration are visible in the time course profile of the neuronal model with the largest τ in the model space (indicated by red lines in Figure 5A and C). Figures 5A and C also show that the differences in responses are maintained within the tonotopic channels, however; the responses vary between tonotopic channels (as sound stimuli used do not have equally distributed energy across the spectrum). The respective BOLD estimations, however, have more comparable

profiles for different models; except for the model with the largest τ, which generates a delayed peak response (indicated by a shift in slope between 5s and 10s time points) in comparison with the other three models. This trend holds across the tonotopic channels.

Figure 5E shows, for a single representative voxel, the weighted combination of the three PCs (combining information across tonotopic channels) of the simulated BOLD response for the four models. These responses show that moving across the model space diagonally, the BOLD responses are characterized by a shift in the time taken to reach peak amplitude (indicated by a shift in slope between 5 s and 10 s time points). Overall, the model space captures diverse hemodynamic responses (through modeling of varying neuronal dynamics).

3.2 Model Predictions

After confirming that the variation in neuronal responses (Figure 5A and C) across the model space can be, to some extent, reflected in hemodynamic BOLD response (Figure 5E), we next examined the results of BMS in terms of estimated posterior probability across the model space. Figures 6A and B, show posterior probabilities of the models

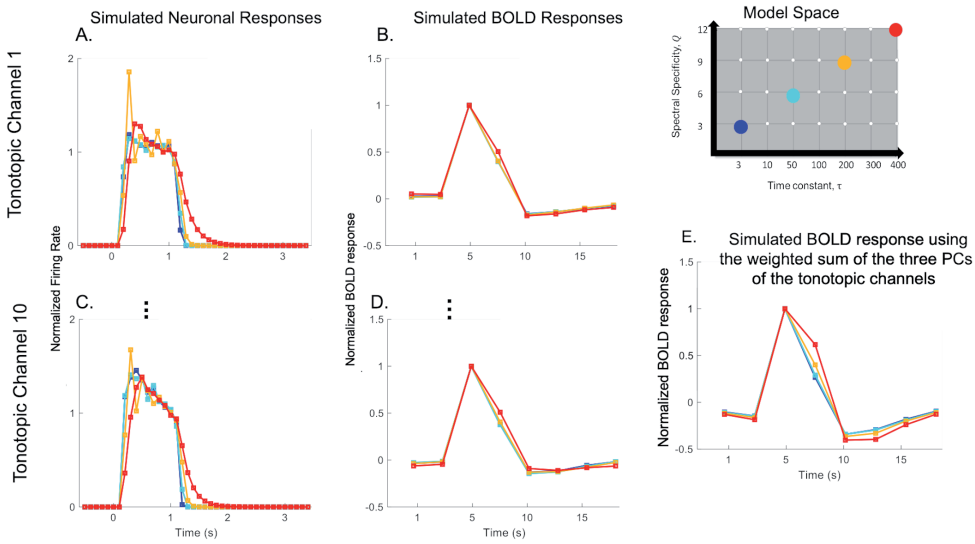


Figure 5: Simulated neuronal responses and consequent BOLD responses. (A-D) The simulated responses are shown for four models chosen across the diagonal of the model space. (E) To highlight the differences between model responses across the model space, the BOLD response for the four models of the model space is reconstructed individually based on the weighted sum of the three PCs of the tonotopic channels. The responses are shown for a single representative voxel. With increasing τ and Q, the peak amplitude of the BOLD is seen to shift forward in time. The neuronal responses are shown normalized by the mean firing rate at 1 s (duration of each sound stimulus) and the two adjacent time points (before and after the 1s mark). The BOLD responses (Figure 5B, D, and E) are plotted normalized by peak response at the time point at 5 s.

for two example voxels respectively. For voxel 1 (Figure 6A), the highest probability is attributed to the model with $\tau = 50$ *ms* and $Q = 6$ while the competing models are adjacent to the best model (i.e., model with the highest probability) in the model space (ranging between Q of 3 – 9 and τ of 20 – 100 ms). On the other hand, for voxel 2 (Figure 6B), the best model is characterized by $\tau = 300$ *ms* and $Q = 9$ surrounded by competing models of the model space (ranging between Q of 6 – 12 and τ of 200 – 400 ms). This shows that models closer to each other in the model space behave alike, and different from the rest of the model space.

Figure 6C shows an example of the measured and predicted BOLD responses for the same two voxels as in Figure 6A and B, along with the underlying neuronal responses of the voxel's best models (Figure 6D). The displayed BOLD responses are average responses

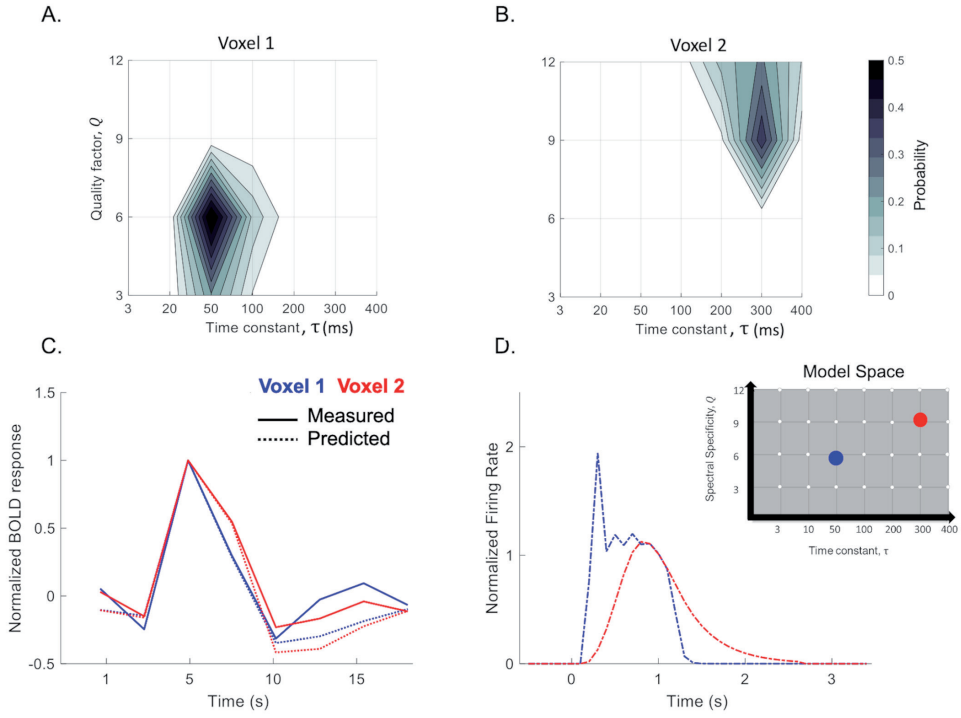


Figure 6: Model Predictions. (A, B) Model fit across the model space for two sample voxels. (A) For voxel 1, the best model predictions are grouped in neuronal models of broad spectral specificity and fast temporal dynamics. (B) For voxel 2, the best model predictions cluster on the opposite (to voxel 1) spectrum of the model space (narrow frequency tuning and slow temporal responses). (C) A comparison of the best model predicted BOLD responses and measured BOLD responses. The BOLD responses are plotted normalized by peak response at the time point at 5 s. (D) The simulated neuronal responses underlying the best models' prediction of the BOLD responses for both voxels are indicated in the model space. The neuronal responses are plotted normalized by the mean firing rate at 1 s and the two adjacent time points (before and after the 1s mark).

obtained with linear deconvolution from measured and simulated time courses and are normalized by peak response at the time point at 5 s (Figure 6C). The neuronal responses (Figure 6D) are shown normalized by the mean firing rate at 1 s (duration of each sound stimulus) and the two adjacent time points (before and after the 1s mark). Interestingly, within the measured BOLD responses, differences can be observed with a delayed peak shown for voxel 2 (solid line in red) compared to the peak in the BOLD response for voxel 1 (solid line in blue). The neuronal model with longer τ and narrower frequency tuning (red circle on the model space) successfully captures the delayed peak. On the other hand, the model with faster τ and broader tuning (blue circle on the model space) effectively models the BOLD response with an earlier peak. This shift in BOLD responses for the two voxels suggests that neuronal response properties might indeed be a contributing factor to the differences in measured BOLD responses. The differences between the predicted and measured BOLD responses increase after 10 s mark. As we apply the same deconvolution to both measured and simulated BOLD responses, these differences might reflect variability in the voxel responses that we are not capturing with model dynamics. Also, as we are not optimizing hemodynamic parameters in the measurement model, it is expected that we would not be able to explain post-stimulus undershoot in all voxels.

3.3 Predicting Neuronal Response Properties across the Belt Regions

3.3.1 Individual Responses

Computing the best model prediction per voxel in the labeled regions allowed assigning underlying neuronal response properties to each voxel. Figure 7 shows the maps of the temporal parameter τ (panels A, B) and the spectral specificity parameter Q (panels C and D) of the best-fitting neuronal models for all voxels in labeled *Caudal* and *Rostral* regions, for a representative participant (see Supplementary Figure 1 for all other individual participant maps).

In Figure 7A and B, a gradient of characteristic temporal constants can be observed, moving from fast to slow along the caudal to the rostral axis. In terms of spectral specificity (panels C and D), the model predicts that spectrally broad response properties underlie the measured BOLD responses in the *Caudal* region, while fine-grained frequency tuning properties best explain activity in the *Rostral* region of the lateral belt.

Moreover, using the PCA-based back-projection methodology (Figure 4), we also predicted the voxels' frequency preference (panels E and F). The best tonotopic channel (from a total of 10 tonotopic channels) for the best model is computed to predict frequency preference (on a low [red] to high [blue] scale). The models indicate a higher frequency region in the *Caudal* area while lower frequencies dominate the *Rostral* area.

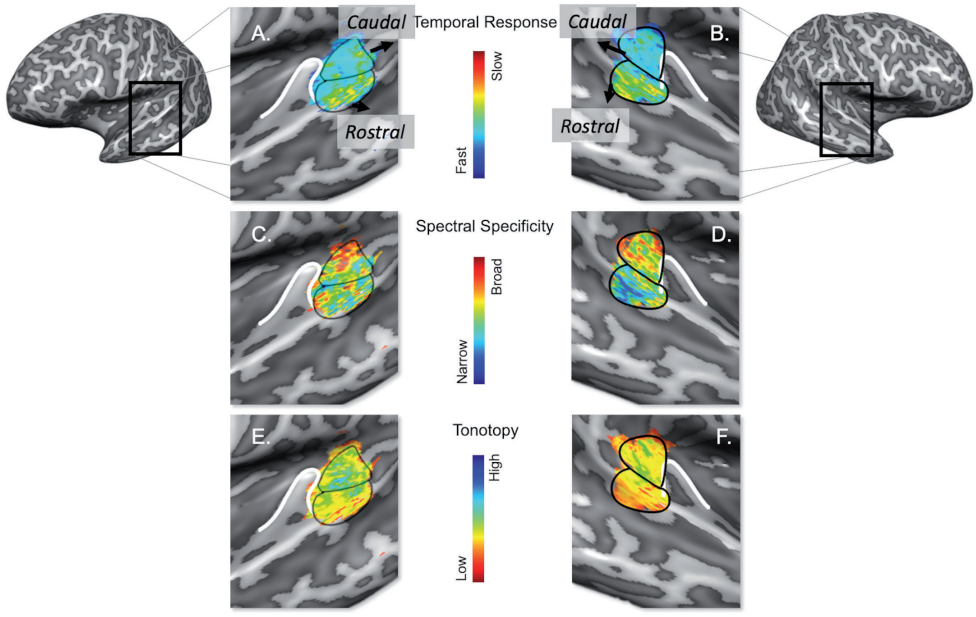


Figure 7: Predicted neuronal response properties along the rostral-caudal axis on belt regions for a single subject. Panels A and B show the temporal response characterized by parameter τ of the neuronal model while spectral specificity (characterized by *Q* of the FTCs) is indicated in panels C and D. Panels E and F show estimation of the best frequency channel. The *Rostral* and *Caudal* regions are labeled in solid black lines based on their BOLD responses to characteristic spectro-temporal features of the sounds. Heschl's Gyrus (HG) is marked by the white solid line to provide a reference for the estimated location of core areas.

3.3.2 Group Responses

In order to test if the results observed in individual maps were stable across the subjects of the fMRI study, we compared the accuracy of the predicted frequency per voxel, the mean temporal constant (τ), and the mean spectral specificity (*Q*) values for all subjects in the labeled *Caudal* and *Rostral* areas. The predicted tonotopy (frequency per voxel) was compared with the measured tonotopy for all subjects. The measured frequencies were first binned (between 60 and 7723 Hz, frequency bins based on Gammatone filterbank) to correspond to the 10 tonotopic channels. For the *Rostral* area, best frequencies of approximately 28% (std: 6, mean and std values reported over five subjects) voxels were predicted accurately while 41% (std: 3) were predicted with one channel difference from; and 16% (std: 4) with two channel difference. For the *Caudal* region, 14% (std: 6) best frequencies were predicted accurately; 28.9% (std: 3) predicted with one channel difference and; 26.4% (std: 6) with two channel difference. Overall, for the *Caudal* and *Rostral* areas, the best frequencies of 69.3% and 85% voxels were predicted proximally (less than or equal to difference of 2 frequency bins) of the measured tonotopic preferences. The lower percentage of prediction for the *Caudal* compared to the *Rostral* area might stem from the overall distribution of best models that best represent

Caudal areas. These models are concentrated in the bottom left corner of the model space, indicating broad spectral and fast temporal dynamics. To simulate broad spectral responses, the connectivity with-in the *simulated belt* region, and connectivity between A1 and the simulated *belt* was widened (see Supplementary Table 1). This widening might have resulted in tonotopic reorganization (i.e., shifted center frequencies) of the *belt* units, thus rendering the frequency bin labels dissimilar to the ones from the Gammatone filterbank (used for the 10 tonotopic channels).

A non-parametric Wilcoxon signed-rank test is used to test the differences between the two areas for τ and Q (Figure 8). The median spectral specificity (Q) was significantly higher (Figure 8A, $p = 0.007$) in the *Rostral* (10.11) compared to the *Caudal* area (7.8); the median τ was significantly longer (Figure 8B, $p = 0.05$) in the *Rostral* (171 ms) compared to the *Caudal* area (100).

We further elucidated these differences between *Rostral* and *Caudal* areas by counting the number of voxels whose dynamics were best described by each of the models in the model space (normalized by the size of the labeled region) for all subjects (both hemispheres). Figure 9A shows that, across the labeled regions, the majority of voxels cluster in different regions of the model space.

Taking the difference between the normalized number of voxels per model further highlighted our results. The hemodynamics of the majority of voxels in the *Caudal* area were best explained by models with faster temporal dynamics (short τ) and broad spectral

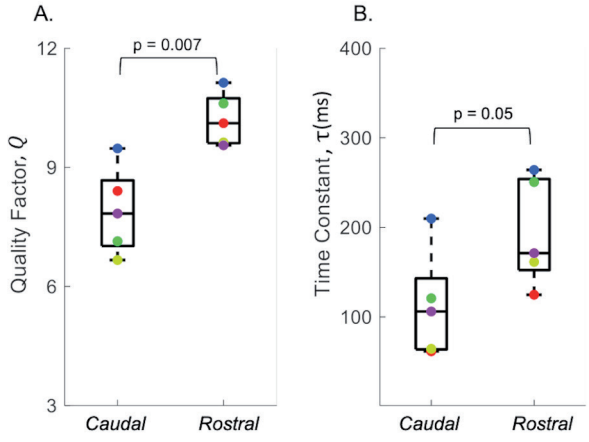


Figure 8: Mean temporal response and spectral specificity in the belt regions. Across the subjects, the hemodynamic activity of the labeled belt regions is best represented by neuronal models with fast temporal and broad spectral responses in the *Caudal* area, and slow temporal dynamics and fine spectral tuning in the *Rostral* area. The dots of different colors indicate the five participants.

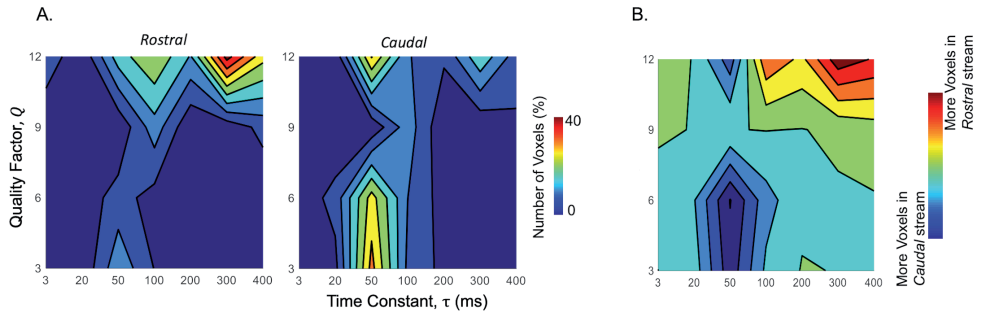


Figure 9: Distribution of voxels best represented by each of the models in the model space. (A) For the two labeled regions *Rostral* and *Caudal*, the distribution of all voxels across subjects is shown on the model space. (B) Difference between predicted neuronal response properties of the labeled *Caudal* and *Rostral* areas. The hemodynamics of the majority of voxels in the *Caudal* area (shown in blue) were best explained by models with faster temporal dynamics (short τ) and broad spectral properties (low *Q*). For the majority of voxels in the *Rostral* area, the dynamics were best predicted by models with fine spectral tuning (high *Q*) and slower temporal dynamics (longer τ).

properties (low *Q*; blue region in Figure 9B). For the majority of voxels in the *Rostral* area, the dynamics were best predicted by models with fine spectral tuning (high *Q*) and slower temporal dynamics (red region in Figure 9B).

4 Discussion

In this article, we investigated the role of neuronal dynamics in the functional observations from the rostral and the caudal streams of the AC. Recent fMRI studies suggested a spectro-temporal tradeoff within the auditory belt: regions caudal to primary areas were found to be most sensitive to fast temporal changes and broader spectral features; rostral regions, however, preferred fine spectral features and slower temporal changes in the stimuli (Santoro et al., 2014). A similar dichotomy in rostral-caudal belt processing has been observed in neuronal response properties using invasive electrophysiological studies, which reported shorter neuronal latencies (i.e., similar to those of the primary auditory cortex) and broader frequency tuning in caudal areas (Recanzone et al., 2000; Kuśmierek and Rauschecker, 2014) and longer latencies accompanied by sharp frequency tuning in the rostral areas (Recanzone et al., 2000; Tian et al., 2001; Bendor and Wang, 2008; Camalier et al., 2012). To link the spectro-temporal tradeoff observations from neuroimaging studies to neuronal response properties reported in electrophysiology, we presented a forward model combining neuronal model specifically catered to model sound processing in the AC with a physiological model of the hemodynamic BOLD response. The neuronal dynamics were generated by a dynamic recurrent firing rate model of the auditory cortex that reflected the tonotopic, hierarchical processing in the auditory cortex

and focused on the spectro-temporal tradeoff in the rostral-caudal axis of its belt areas. The fMRI signals were computed using a nonlinear physiological model to simulate the BOLD responses [P-DCM]. In contrast with simple convolutional models, the nonlinear physiological model captures the non-linear transformation between neuronal responses and the fMRI signals. After confirming that the model captured a diverse set of BOLD responses to sounds, we fitted the simulated BOLD responses to a previously acquired dataset of fMRI BOLD responses to natural sounds. The current approach did not involve model inversion.

We observed that the hemodynamics of a *Caudal* belt region in the human auditory cortex were best explained by models with faster temporal dynamics and broader spectral properties, while that of a *Rostral* belt region were best explained through fine spectral tuning combined with slower temporal dynamics. As we modeled and fitted average responses to all sounds, the assignment of neuronal response properties to a voxel was based on the overall shape of its hemodynamic response function. These voxel properties are thus based on fundamentally different characteristics of the BOLD fMRI data than those used to study spectro-temporal tuning in previous studies (Santoro et al., 2014; Schönwiesner and Zatorre, 2009). The tonotopic-specific responses of the model space suggest that the faster responses are correlated with comparatively higher best frequency regions and the slower responses with the lower best frequency units. This interaction of temporal response properties with the tonotopic axis has been shown through electrophysiological experiments as well (Scott et al., 2011). All in all, our results along with the existing evidence suggest that the response properties of the neuronal populations along the rostral-caudal axis in the belt areas of the human AC are optimized to simultaneously process complementary sound features in parallel streams (Jasmin et al., 2019; Kaas et al., 1999; Belin et al., 2000; Rauschecker and Tian, 2000).

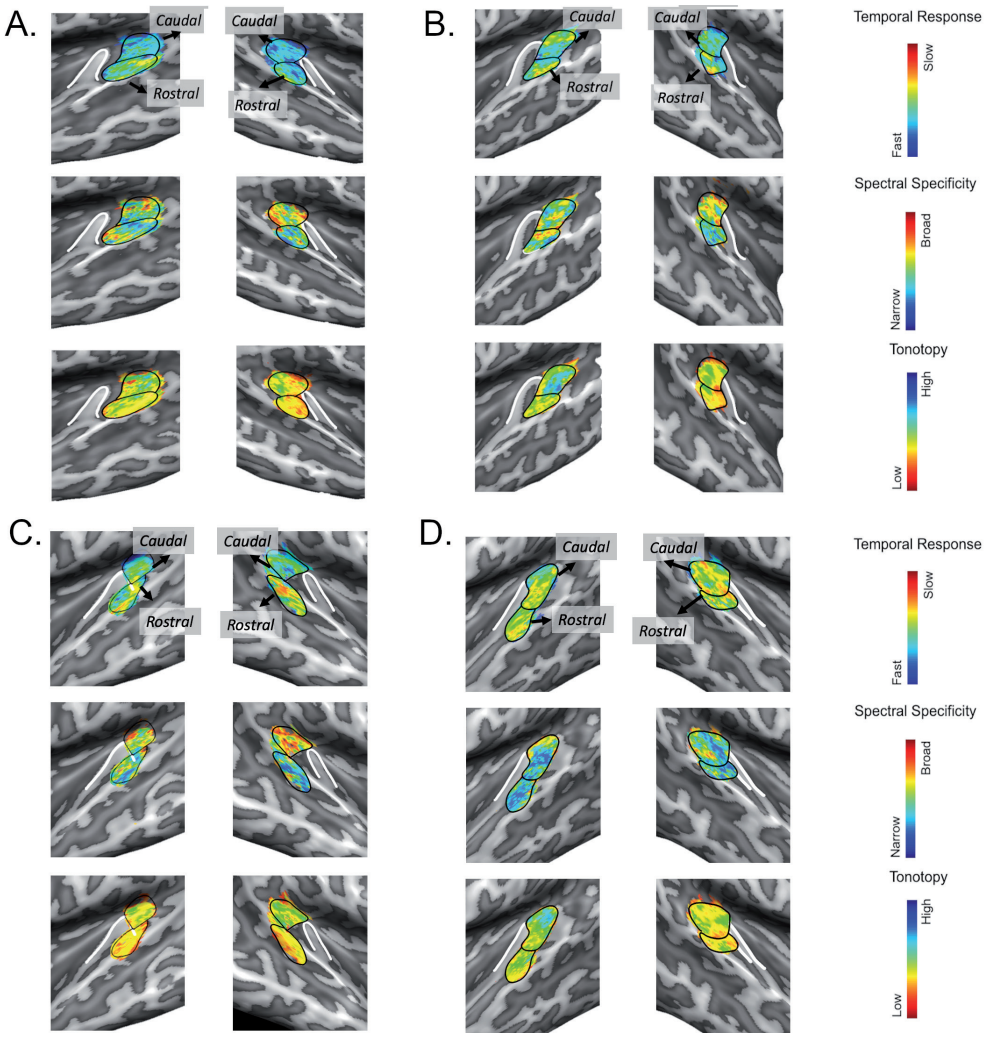
The neuronal model presented here streamlines sound processing in the AC as it employs simplistic models of peripheral and cortical processing. Currently, the model is limited by a small number of simulated regions, by its exclusively feed-forward nature, and by tonotopic-specific connectivity. Future endeavors can improve the sound processing model with better models of the periphery, feedback connectivity, and the addition of non-tonotopic and multisensory projections in the simulated cortical areas. Additionally, in the BOLD model of P-DCM, we have not optimized hemodynamic parameters to the current dataset to capture the variability of vascular properties across voxels. Additionally, we used a simple GLM approach where the tonotopic channels were reduced to only three PCs at the level of BOLD response. Overall, we are accounting only for neuronal but not vascular variability in the hemodynamic response across voxels. For example, the BOLD response originating from larger pial veins often exhibits slower dynamics, characterized by a longer time to peak. It could be that to some extent we are explaining

it with neuronal variability. Therefore, in the future, one could try to apply the full nonlinear model inversion where both neuronal and hemodynamic parameters are optimized simultaneously. However, to do this effectively, it requires data acquired with an experimental design that makes disentangling of neuronal and vascular parameters possible, e.g. using mixed design with blocks of faster events to explore neuronal variability interleaved with longer resting periods to fully observe hemodynamic transients reflecting passive vascular properties or using multimodal fMRI data such as arterial spin labeling (ASL), where both blood flow and BOLD responses are acquired simultaneously (Havlicek et al., 2017; Gardumi et al., 2017).

In the future, it will be interesting to study neuronal response properties underlying BOLD responses across the AC to highlight their contribution to functional streams of information processing. Such modeling frameworks may then be used to study the neuronal underpinnings of other fMRI-based observations. The standard DCM generally focuses on region-level (few) dynamics while our approach models voxel-level (many) responses, which makes model inversion computationally infeasible. Thus, modified approaches, such as regression DCM (Frässle et al., 2017), which provide efficient model inversion solutions can be used to model effective connectivity in individual voxels. Furthermore, apart from fMRI, the neuronal models can also be used in conjunction with models of other non-invasive measures of neural activity (electro- and magneto-encephalography) in a multimodal DCM framework to constrain model inversion and improve the quality of model predictions (Wei et al., 2020). In such efforts, the overall modeling framework acts as an integrative tool, combining the existing knowledge while also generating predictions for future modeling undertakings.

Supplementary Materials

1 Supplementary Figures



Supplementary Figure 1: The simulated maps of neuronal response properties. The maps are shown for four individual subjects in Panels A, B, C, and D. For each subject, predicted temporal responses, spectral specificity, and the best frequency maps are shown. Responses for the fifth participant are shown in Figure 7.

2 Supplementary Tables

Supplementary Table 1: Modified parameter values used in the modeling of the neuronal model space for simulating the auditory *belt* area. For each model, the time constant (τ) is reported in ms. Spectral specificity measured as Q is shown for a single tonotopic unit (best frequency at 1 kHz), along with mean across tonotopic space (Q_{mean}). The within area spatial spread between excitatory-excitatory and excitatory-inhibitory units is shown by σ_{EE} and σ_{EI} respectively. The connectivity from the simulated *core* to belt units is controlled by the connectivity kernels that are convolved with input across tonotopic space. The parameters not listed in the table are fixed as described in the original implementation (Zulfiqar et al., 2020).

	τ (ms)	Q_{1k}	Q_{mean}	σ_{EE}	σ_{EI}	Connectivity kernel
1	**3 – 1**	**14**	12.8 ± 3.8	20	260	One-to-one
2		**8.7**	7.9 ± 1.8	50	150	One-to-one
3		**6**	5.7 ± 1.5	200	300	0.25 – 0.5 – 1 – 0.5 – 0.25
4		**3.2**	3.17 ± 0.8	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5
5	**20 – 18**	**13.4**	11 ± 3.1	25	200	One-to-one
6		**8.7**	7.7 ± 2.1	50	150	One-to-one
7		**6**	5.7 ± 1.5	200	300	0.25 – 0.5 – 1 – 0.5 – 0.25
8		**3.2**	3.1 ± 0.85	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5
9	**50 – 48**	**12.3**	10.8 ± 3.1	25	200	One-to-one
10		**8.7**	7.5 ± 2	50	150	One-to-one
11		**6**	5.6 ± 1.5	200	300	0.25 – 0.5 – 1 – 0.5 – 0.25
12		**3.2**	3.1 ± 0.8	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5
13	**100 – 98**	**13.4**	11 ± 3.1	20	80	One-to-one
14		**8.7**	7.5 ± 2.1	50	150	One-to-one
15		**6.4**	5.8 ± 1.6	200	300	0.5 – 1 –0.5
16		**3.2**	3.2 ± 0.8	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5
17	**200 – 170**	**12.3**	10.1 ± 3.1	20	100	One-to-one
18		**8.7**	7.5 ± 2.1	50	150	One-to-one
19		**6.1**	5.5 ± 1.5	200	300	0.25 – 1 –0.25
20		**3.2**	2.9 ± 0.8	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5
21	**300 – 270**	**12.3**	10.1 ± 2.8	15	200	One-to-one
22		**9**	8.1 ± 2.4	20	60	One-to-one
23		**6.2**	5.5 ± 1.5	200	300	One-to-one
24		**3.2**	2.9 ± 0.7	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5
25	**400 – 370**	**11.3**	10.1 ± 2.7	15	200	One-to-one
26		**9.2**	7.4 ± 2	20	60	One-to-one
27		**6.1**	5.6 ± 1.5	150	300	One-to-one
28		**3**	2.8 ± 0.7	200	300	0.5 – 1 – 1 – 1 – 1 – 1 – 1 – 1 – 0.5

Supplementary Table 2: Modified parameter values used in the modeling of Neurovascular Coupling. The parameters not listed in the table are fixed as described in the original implementation (Havlicek et al., 2015).

Parameter	*Value*
Scaling constant for input, c	$\frac{1}{16}$
Regulatory constant, φ	1
Regulatory constant, χ	1
Regulatory constant, ϕ	1.5

Supplementary Table 3: Modified parameter values used in the modeling of Hemodynamics States. The parameters not listed in the table are fixed as described in the original implementation (Havlicek et al., 2015).

Parameter	*Value*
Power law constant, α	0.35
Viscoelastic time constant venous, τ_v	During response inflation phase = 2 s
	During response deflation phase = 10 s
Mean transit time, t_0	2 s

Supplementary Table 4: Modified parameter values used in the BOLD signal model. The parameters not listed in the table are fixed as described in the original implementation (Havlicek et al., 2015) and (Havlicek and Uludağ, 2020).

Parameter	*Value*
Resting blood volume, V_0	0.03
Oxygen extraction fraction, E_0	0.4
Echo-time, TE	28 ms
Frequency offset at the surface of a blood vessel, ϑ_0	188 s^{-1} (at 7T)
Sensitivity of changes in intravascular signal relaxation rate, r_0	125 s^{-1} (at 7T)
Ratio of intra- to extra-vascular fMRI signal contribution, ε	0.25

Chapter 4

Audiovisual Interactions among Near-threshold Oscillating Stimuli in the Far Periphery are Phase-dependent

Zulfiqar I., Moerel M., Lage-Castellanos A., Formisano E., and De Weerd P. (under review). Audiovisual Interactions among Near-threshold Oscillating Stimuli in the Far Periphery are Phase-dependent

Abstract

Recent studies have highlighted the possible contributions of direct connectivity between early sensory cortices to audiovisual integration. Anatomical connections between the early auditory and visual cortices are concentrated in visual sites representing the peripheral field of view. Here, we aimed to engage early sensory interactive pathways with simple, far-peripheral audiovisual stimuli (auditory noise and visual gratings). Using a modulation detection task in one modality, we investigated the multisensory interactions by simultaneously presenting a barely-detectable stimulus (at 55% and 65% detection threshold, modulated or static) in the unattended modality. Furthermore, we manipulated the temporal congruence between the cross-sensory streams. We found evidence for an influence of barely-detectable visual stimuli on the response times for auditory stimuli, but not for the reverse effect. These visual-to-auditory influences only occurred for specific phase-differences (at onset) between the modulated audiovisual stimuli. We discuss our findings in light of a possible role of direct interactions between early visual and auditory areas, along with contributions from the higher-order association cortex. In sum, our results extend the behavioral evidence of audio-visual processing to the periphery, and suggest – within this specific experimental setting – an asymmetry between the auditory influence on visual processing and the visual influence on auditory processing.

1 Introduction

Multisensory information is ubiquitous in our environment. Our brain is adept at pooling information from multiple modalities to form a unified view of our surroundings, thus guiding perception and behavior. The relationship between sensory stimuli (e.g., spatial, temporal, contextual, attentional, etc.) and the task at hand (Spence, 2013; Odegaard and Shams, 2016) helps unify or disassociate binding between senses, leading to changes in behavior [as indexed by discriminability, response times, accuracy, etc. (Bizley et al., 2016; Chen and Vroomen, 2013; Odegaard et al., 2015, 2016)]. These cross-modal interactions can also affect subsequent unisensory processing (Wozny and Shams, 2011; Barakat et al., 2015).

Traditionally, multisensory anatomical and functional processing pathways in human and non-human primates have been credited to converging inputs in higher-order association cortex (Ghazanfar and Schroeder, 2006; Cappe et al., 2009). More specifically, evidence shows that audiovisual (AV) integration regions include posterior superior temporal sulcus and middle temporal gyrus (Beauchamp et al. 2004; Starke et al. 2017; van Atteveldt et al. 2004; von Kriegstein et al. 2005; Perrodin et al., 2014; Tanabe, 2005). The intraparietal sulcus (Lewis and van Essen, 2000; Cate et al., 2009) and frontal areas (Gaffan and Harrison, 1991; Romanski et al., 1999b) have also been implicated in AV integration.

More recently, however, early sensory areas have also been shown to play a role in multisensory processing (Hackett et al., 2007; Driver and Noesselt, 2008; Koelewijn et al., 2010). Through the use of anterograde and retrograde tracers, Falchier et al. (2002) showed direct projections from primary and secondary auditory areas to the early visual areas in rhesus monkeys as well as reciprocal connections from secondary visual area (V2) and prostriata to the auditory cortex (Falchier et al., 2010). Evidence for a role of these early cortico-cortical connections in multisensory effects has been functionally established as well. Auditory influences on primary visual areas (V1) have been shown across species (Wang et al., 2008; Ibrahim et al., 2016). The responses in the auditory areas are also directly influenced by the visual cortex (Besle et al. 2008, 2009), for example through changes in the phase of auditory local field potential and single unit activity (Kayser et al., 2008, 2010). These changes in local field potentials have been shown to amplify sensory inputs (Schroeder and Lakatos, 2009) and, more recently, to provide cross-modal cues in auditory scene analysis (Atilgan et al., 2018). The early onset of observed multisensory effects supports the role of early sensory cortical connectivity in multisensory interactions (Wang et al., 2008; Besle et al., 2008).

Interestingly, the direct connections between early visual and auditory cortices are not uniform. Instead, neurons with peripheral visual fields (> 30° visual angle) receive

and project the majority of these connections (Falchier et al., 2002, 2010; Rockland and Ojima, 2003; Eckert et al., 2008). In accordance, recent human neuroimaging and behavioral studies showed that AV integration is different between centrally and peripherally located stimuli (Charbonneau et al., 2013) and that this difference is influenced by stimulus modality (i.e., auditory-to-visual vs. visual-to-auditory influences). For example, the double flash illusion (induced by sound) was found to be stronger in the peripheral visual cortex compared to the foveal regions using fMRI (Zhang and Chen, 2006) and similar results were observed behaviorally (Shams et al., 2002; Chen et al. 2017). In addition to the influence of spatial location, AV integration also depends on the temporal characteristics (Chen and Vroomen, 2013) and salience of the stimuli (Meredith and Stein, 1983; Stein and Stanford, 2008; Stein et al., 2009). To understand how the brain uses temporal features in integrating information from multiple sources, one key approach has been to manipulate the temporal congruency between both naturalistic (McGurk and MacDonald, 1976) and artificial oscillating stimuli (Laing et al., 2015). The temporal characteristics of the stimuli and their salience have also been observed to interact and collectively affect AV integration. In a recent study, the lowest contrast detection thresholds for oscillating visual stimuli were observed when accompanied by in-phase auditory stimuli of weak salience (Chow et al., 2020). While the effects of different stimulus features on audiovisual integration have been extensively studied for centrally presented stimuli (Chen and Vroomen, 2013; Kayser et al., 2008; Spence and Squire, 2003; ten Oever et al. 2014; Denison et al., 2013; Shams et al., 2000; Soto-Faraco et al., 2004; Frassinetti et al., 2002), the influence of these audiovisual stimulus features on multisensory integration at peripheral locations [beyond 10^{o} degrees visual angle (Chen et al., 2017; Chow et al., 2020)] is largely unknown.

Therefore, in this study, we aimed to characterize AV interactions in the far periphery (at 28.5° eccentricity) by using simple AV stimuli. Specifically, we aimed to study the temporal criteria for successful AV interactions at far-peripheral locations. The temporal features of the AV stimuli were varied in a two-fold structure. First, to study the relevant contribution of the temporal structure of the stimuli in the process of integration in far periphery, the presented AV stimuli were either modulated or static. Second, keeping the stimulus onset time the same for AV stimuli, we manipulated the onset phases of the modulated stimuli to analyze the role of temporal structure in AV integration. During the modulation detection task (either auditory or visual), the stimulus of the unattended modality was presented at a barely-detectable intensity (at 55% and 65% detection threshold) in either modulated or static state. We hypothesized that by manipulating the phase of the modulated signal of either modality (the attended or unattended stream), the temporal synchrony conditions would be optimized for auditory-to-visual and visual-to-auditory interactions.

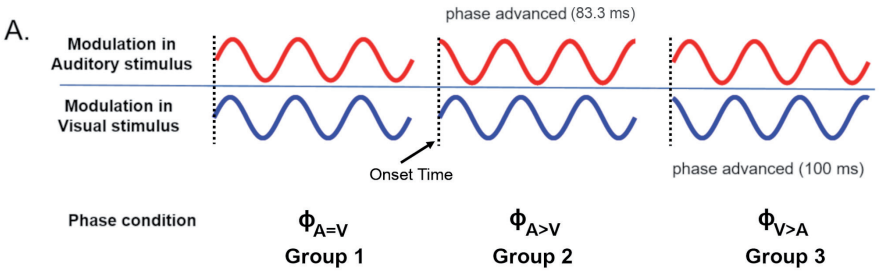
2 Methods

The key features of the experimental design are shown in Figure 1. The participants were divided into three groups. Each of these groups took part in one of three conditions (N = 9 per condition) that differed from each other in the onset phase (ϕ) of the modulations in the auditory and visual stimuli (Figure 1A). There was no difference in stimulus onset times. For each participant, the experiment consisted of six sessions of two-hour duration each, divided over six days (Figure 1B). We used a staircase experimental design to measure the modulation detection thresholds for the visual and auditory stimuli in either a unisensory or a multisensory setting. Within each session, we used a two-alternative forced choice task where participants had to indicate by a button press if a visual or auditory stimulus was modulated or static (Figure 1C). In the multisensory condition, the stimuli were presented in congruent (both modulated or both static) and incongruent (one modulated while other is static) manner. Apart from detection thresholds, response times of the participants were also recorded during the staircases.

In Group 1 (9 participants), the auditory and visual stimuli were modulated sinusoidally (both starting with the default onset phase of a sinusoid being 0). This condition will be referred to as the $\phi_{A=V}$ condition. In Group 2 (9 participants), the auditory stimulus modulation started with an advanced phase of $\frac{\pi}{2}$ (83.3 ms) while the visual stimulus modulation started at the default onset phase of 0 ($\phi_{A>V}$ condition). In Group 3 (9 participants), the modulated visual stimulus was phase-advanced by $1.2\frac{\pi}{2}$ (100 ms), hence leading in phase compared to the modulated auditory stimulus with no phase-shift at the onset. This condition will be referred to as the $\phi_{V>A}$ condition. All static stimuli remained the same across the three phase conditions. Throughout this manuscript, the term "threshold" refers to the modulation detection threshold (i.e., the stimulus intensity at which participants can discriminate modulated from static stimuli). The term "modulation" always refers to the oscillatory feature of the stimuli rather than to a cross-sensory influence.

2.1 Participants

All twenty-seven participants (mean age 22.8 ± 3.2, including 8 males) had normal or corrected-to-normal vision. A pure-tone audiogram was obtained before the first session to exclude participants with hearing loss (using 25 dB hearing level as a threshold). Prior to the first session, each participant was informed about the procedures, and verbal and written consent was obtained. Participants were compensated with either monetary reward alone or a combination of monetary reward and credit for their course requirements. Following the last session, all participants were debriefed about the purpose of the experiment, and they filled out a questionnaire about their impression of low-intensity stimuli of the unattended modality presented during the tasks. The experiment was

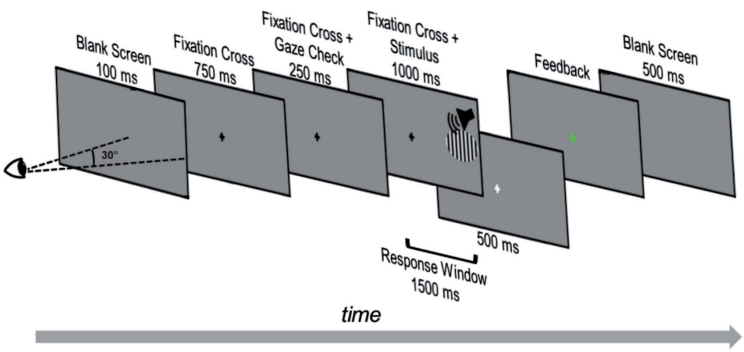


B.

Stimuli: Unisensory / Multisensory

	Staircase Type	Threshold level	Auditory Task	Visual Task
Session 1	Single	84%	1 → 2 → 3	1 → 2 → 3
Session 2	2-Interleaved	84%	1,2 → 3,4	1,2 → 3,4
Session 3,4	Single	50%	1 → 2 → 3	
	2-Interleaved	84%		1,2 3,4 5,6
Session 5,6	Single	50%		1 → 2 → 3
	2-Interleaved	84%	1,2 3,4 5,6	

C.



approved by the Ethics Review Committee of the Faculty of Psychology and Neuroscience at Maastricht University.

2.2 Apparatus

Participants sat in a soundproof, dimly lit room with their heads supported by a chin and head rest affixed 42 cm in front of an LCD monitor (24" Iiyama Prolite B2481HS LED monitor, Iiyama Corporation, Tokyo, Japan; 60 Hz refresh rate, 1920 x 1080

< Figure 1. Experimental design. (A) The phase of sinusoidally modulated stimuli across the three phase conditions for the three participant groups. For phase condition $\phi_{A=V}$ (Group 1), the modulated stimuli have no phase-shift. For phase condition $\phi_{A>V}$ (Group 2), the phase of the modulated auditory stimulus was phase-advanced by 83.3 ms. For phase condition $\phi_{V>A}$ (Group 3), the modulated visual stimulus was phase-advanced by 100 ms. The onset time of all stimuli is the same (vertical dotted lines). (B) Experimental design of the study. Over six sessions (each session conducted on a separate day), participants performed a modulation detection task in a staircase design. Each executed staircase is represented by a numbered box. The number indicates the staircase number, and the box size corresponds to the staircase duration. Each staircase measurement results in a modulation detection threshold. Depending on the staircase settings, either a 50% or 84% modulation detection threshold is measured ('Threshold level'). White outlined and black filled boxes represent unisensory and multisensory conditions, respectively. In session 1, three repetitions (indicated by numbers 1-2-3) of single 84% detection threshold staircases are performed on unisensory auditory and visual stimuli. In session 2, participants execute 2-interleaved 84% detection threshold staircases, whose longer duration is indicated by larger boxes, twice each for unisensory auditory and visual stimuli. We then used two types of sessions to collect the multisensory data and associated unisensory control data. In the first type (repeated twice, designated Session 3, 4), three unisensory 50% correct staircases for the auditory task were administered followed by three 2-interleaved 84% correct visual staircases. Auditory stimuli were presented in two of these 2-interleaved 84% correct visual staircases (i.e., they were multisensory), and these staircases were used to measure the influence of auditory stimuli on performance in the visual task. In the second session type (repeated twice, designated Session 5, 6), three unisensory 50% correct staircases for the visual task were administered followed by three 2-interleaved 84% correct auditory staircases. Visual stimuli were presented in two of these 2-interleaved 84% correct auditory staircases (i.e., they were multisensory), and these staircases were used to measure the influence of visual stimuli on performance in the auditory task. Note that the order of the days for sessions 3-6 was randomized over participants. In addition, for sessions 3-6, the order of the 2-interleaved staircases (two multisensory and one unisensory) was varied over participants, but kept the same for an individual participant. (C) Experimental design of a single trial. Participants fixated on the fixation cross at the center of the screen. The stimulus was only presented if the participant maintained fixation (fixation error < 2.5°) during the 250 ms gaze check prior to the stimulus presentation. The stimulus was auditory (white noise burst), visual (vertical grating, 6° in diameter) or audiovisual, and was presented at 28.5° azimuth in the participants' right hemifield. The task of the participants was to indicate whether the attended stimulus was modulated or static. Feedback was given by a change in the color of the fixation cross (green = correct, red = incorrect).

resolution). Fixation during the task was checked using ViewPoint Eyetracker (MIU03 Monocular, Arrington Research, Inc., Scottsdale, Arizona, USA; 220 Hz) which was mounted towards the left side of the chin and head rest. All stimuli were generated at runtime in MATLAB (The MathWorks, Inc.) using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). The stimulus PC interfaced with the Eyetracker PC via an Ethernet connection using the ViewPoint Client App and ViewPoint MATLAB toolbox (v2.8.5), providing runtime access to the Eyetracker data.

2.3 Stimuli

The static visual stimulus consisted of a circular sinusoidal grating (vertical orientation, 1.6 grating cycles/degree at a screen resolution of 1920 x 1080 pixels, diameter = 6.2°) that was presented at 28.5° eccentricity to the right on the azimuthal plane. A lower modulation rate (3 Hz, prevalent in speech (Overath et al., 2015)) was chosen, in line

with previous research where modulated stimuli have been used to study AV interactions in the brain (Laing et al., 2015). To create the modulated visual stimulus, the Michelson contrast of the static grating was sinusoidally modulated over time at 3 Hz with a modulation depth of 80%. The contrast of the static stimulus was set at the peak contrast of the corresponding modulated stimulus.

The static sound was created as a normally distributed white noise stimulus (generated using *randn* in MATLAB with mean = 0, std = 0.5, sampling rate 44.1 kHz). To create the modulated sound, the sound pressure level (SPL) of the static sound (central SPL fixed at ~32.2 dB) was varied sinusoidally at 3 Hz with a modulation depth of 80%. Sounds were presented using a headphone set (AKG K72). Sound location was matched to the visual stimulus location by adjusting the sounds' interaural level difference (ILD). ILD was set based on subjective measurements from authors IZ and PDW, and confirmed by each participant at the beginning of the experiment. The resulting ILD of 3 dB is slightly smaller than expected (Shaw, 1974). This difference between our subjective measurements and values reported in the literature may be explained by the overall low intensity of the employed sounds. The intensity of the static stimulus was set at the peak intensity of the corresponding modulated stimulus.

2.4 Experimental Design

The following sections detail the staircase design along with specifics of measurements taken during each session for all three phase conditions. The task, along with the stimuli specifications is also described.

2.4.1 Staircase Design

We used separate staircases to measure the detection thresholds for modulations in the visual and auditory stimuli. In each staircase, the presentation order of modulated and static stimuli was randomized while ensuring an equal number of modulated and static stimuli for each block of 10 trials.

Three different staircase designs were used during the experiment. The 50% auditory (visual) detection thresholds were measured by a staircase where for each wrong/correct response, the auditory intensity (visual contrast) increased/decreased by the respective step size. To measure 84% detection thresholds, the intensity/contrast decreased for every four consecutive correct answers and increased for every wrong response (Wetherill and Levitt, 1965). For staircases with the auditory task, the sound amplitude was varied by 20% for each step. To compute the contrast steps for the visual staircase, the Michelson contrast was measured. By fitting a polynomial to the measured contrast values, the luminance values of the screen were converted to corresponding contrast values. The highest contrast value of the grating was capped at 30% Michelson contrast and reduced

by a step size of 20%. Each staircase finished either after 14 reversal points were acquired, or upon completion of 120 trials. Supplementary Figure 1 shows an example of the staircase procedure used to measure the 84% (Supplementary Figure 1A) and 50% (Supplementary Figure 1B) correct thresholds for a single subject.

We also created interleaved staircases by merging two independent staircases (84% detection threshold), such that trial blocks of two independent staircases were presented in an interleaved fashion. In blocks of 10 trials, the staircase switched pseudo-randomly (to ensure that long stretches of the same staircase did not occur) between the congruent (i.e., auditory and visual stimulus are either static, or both modulated) and incongruent (i.e., one of the multisensory stimuli is static, and the other is modulated) conditions. In order to compare these multisensory conditions to unisensory thresholds, participants also performed unisensory interleaved staircases. Note that the staircases remained independent: responses to trials in one staircase did not affect stimulus presentation in the other staircase. Interleaved staircases finished when both staircases completed either 14 reversal points or 120 trials.

2.4.2 Sessions

The six sessions were spread over a period of two weeks, with every session at the same time of the day for each participant. Session 1 was designed to familiarize participants with the task, the chin and head rest, and the Eyetracker setup. During session 1, the participants performed three auditory and three visual unisensory staircases in order to determine their 84% modulation detection thresholds. Each staircase took approximately 8 minutes, and participants were given a break of approximately 5-10 minutes between staircases (Figure 1B).

In Session 2, the participants completed two interleaved 84% staircases of the unisensory auditory and visual conditions. The duration of an interleaved staircase was approximately 20 minutes, and participants were given 5-10 minute breaks between staircases. The purpose of this session was to provide a baseline behavior for unisensory thresholds in interleaved staircases, as these staircases were then repeated in the next four sessions as discussed below.

Session 3 to 6 started with the estimation of unisensory 50% detection thresholds in three staircases. Next, participants performed two multisensory interleaved staircases. Per session, they performed either the auditory or the visual task, while the stimulus of the other modality was presented at a barely detectable intensity/contrast (an estimated 55% or 65% modulation detection threshold). The barely detectable intensities were selected as they allow above chance identification of the modulated stimuli yet should not act as a distractor during task execution on the other stimulus modality. The order

of these four measurement days ((multisensory auditory or visual task) x (55% or 65%)) was randomized across participants. The 55% and 65% modulation detection thresholds were estimated by z-scoring the contrast steps between 84% (from session 2) and 50% measurements and then linearly interpolating the intermediate steps (from 50% to 84%). Supplementary Figure 1C shows an example of how the 65% and 55% correct thresholds were estimated using the two measured thresholds (84% and 50% correct). The resulting psychometric curve is not a straight line because of the conversion from screen luminance to Michelson contrast.

Per session (sessions 3 to 6), two multisensory staircases were conducted. The two interleaved staircases were used to test the effects of (in)congruence of the auditory and visual stimulus on detection thresholds and response times. Participants also performed an interleaved staircase for the unisensory task condition. The order in which participants performed unisensory and multisensory interleaved staircases was balanced across participants to minimize fatigue effects but was kept the same for an individual participant across sessions. Each session took two hours to complete and participants were actively encouraged to take breaks during sessions.

2.4.3 Task

Participants performed a two-alternative forced choice task on the visual or auditory stimuli. That is, they pressed either the right or left arrow key indicating a modulated or static stimulus, respectively (Figure 1C). Each trial began with a gray screen for 100 ms, followed by a black fixation cross that was presented in the center of the screen for 750 ms. For the next 250 ms, the fixation cross remained on the screen while a steady fixation check was performed (with 2.5° freedom from the fixation point). In case of a failed fixation, the trial was aborted, and a new trial began. If the participant passed the gaze check, a stimulus was presented (1 s) while the fixation cross remained on the screen. The response window began at the onset of stimulus presentation and extended 500 ms after stimulus offset (indicated by a light gray fixation cross). Feedback was provided as soon as the participant responded, by a change in the color of the fixation cross to red or green for incorrect and correct responses, respectively. In case the participant did not respond within the response window, the trial condition was appended to the trial list and the next trial was initiated. The inter-trial interval was 500 ms (gray screen). Following every tenth trial, the center for the gaze check was readjusted to correct for drifts of the Eyetracker setup and/or subject motion. The apparatus-induced propagation delay between auditory and visual stimuli was estimated to be ~20ms.

In the 250 ms time window before stimulus onset, trials were aborted if the gaze position was more than 2.5° away from fixation. Due to a programming error, the eye movements were only recorded in a 5 ms time window before the stimulus presentation

(that is if the fixation was maintained in the previous 245 ms), and at the moment of the response. As saccade execution requires ~200 ms (Şentürk et al., 2016), it is highly likely that participants fixated during the initial part of the stimulus. While we only have eye position recordings at the instance of response, the large number of trials with widely varying response latencies allowed us to sample eye position from ~250 ms after stimulus presentation until the end of stimulus presentation (1s). Figure 2 shows the distribution of the absolute distance of eye gaze from the fixation point recorded across trials (at the instance of response) during both multisensory tasks (panels A and B show a representative participant; panels C and D show combined data of all participants). Irrespective of the latency of the response (and hence the time since stimulus onset) eye position was within 10° of fixation in 98% of the ~100,000 samples, within 5° of fixation

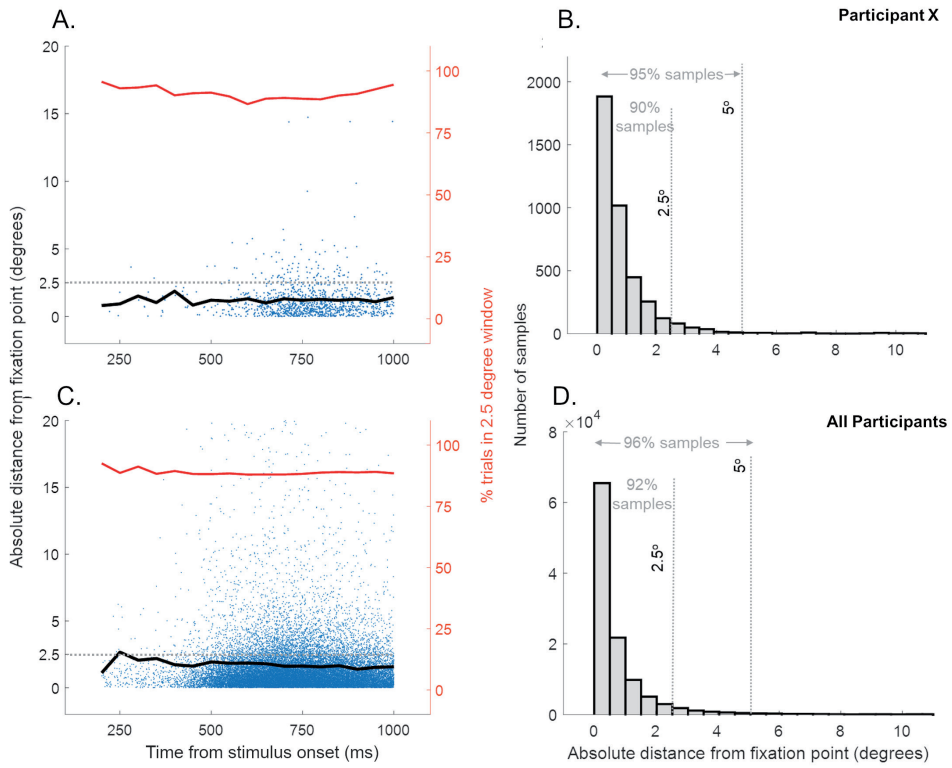


Figure 2: Distance of eye gaze location from the fixation point at the instance of response across trials. Panels A and B show the data for a single representative participant while Panels C and D show data for all 27 participants, for both multisensory tasks. Each blue dot in panels A and C represents the eye gaze distance from fixation for a single trial while the black line shows the mean eye gaze distance from fixation over trials. The red line indicates the percentage of trials, at the response time, where the eye gaze location was within 2.5o distance from the fixation point. Panels C and D show the distribution of eye gaze distance from fixation and confirm that eye gaze position was within a few degrees from the fixation point for a high percentage of trials.

in 96% of the trials, within 2.5° of fixation in 92% of the trials in 27 participants (Figure 2C and D). Supplementary Figure 2 shows that, across the three participant groups, the eye locations away from fixation did not specifically target the stimulus location. In addition, the data suggest similar fixation performance in the two tasks across the three participant groups. This is supported by a two-way ANOVA over fixation accuracy (the percentage of trials with eye position within 2.5° of the fixation) with between-subject factor "Phase condition" (3 levels: Group 1 – Group 2 – Group 3) and within-subject factor "Task" (2 levels: Auditory Task – Visual Task) showing neither a significant interaction ($F(2,24) = 2.43$, $p = 0.1$) nor any significant main effects (Phase Condition: $F(2,24) = 0.03$, $p = 0.9$; Task: $F(1,24) = 0.1$, $p = 0.7$). Altogether, this evidence supports consistent fixation in our participants and suggests that large fixation errors were present only in a very small minority of trials.

2.5 Statistical Analysis

Modulation detection thresholds for congruent, incongruent, and unisensory stimuli were computed as the average intensity/contrast of the last 10 reversal points and were averaged over repeated staircases. The response times were computed based on all trials spanning the last 10 reversal points and were averaged over repeated staircases as well. Mixed ANOVA analyses (Caplette, 2017), conducted in MATLAB, were used to test for changes in response time and modulation detection thresholds across the three phase conditions, with auditory and visual stimulation, and as driven by the congruency and intensity of multisensory stimulation. After observing significant interactions, we performed follow-up analyses per level of one of the interacting factors while correcting the F-ratio of these follow-up analyses by using error term and degree of freedom of significant interaction error term (indicated as $F_{\alpha\text{-corrected}}$) (Hedayat and Kirk, 2006). Bonferroni-corrected pairwise comparison testing was used to further evaluate significant main effects.

3 Results

In the following section, results from the experiment are shown in order to analyze the effects of (a) temporal (in)congruence between coincident static and modulated AV streams and (b) the phase relation between modulated stimuli in cross-sensory facilitation. While the participants performed the modulation detection task in the visual or auditory task (in 84% correct staircase), a barely-detectable stimulus (modulated and static) of the other modality was presented at an estimated 55% and 65% detection threshold intensity. During the task, detection thresholds and response times were obtained for AV congruent (both static or both modulated) and AV incongruent (one static and the other modulated) conditions, as well as in the unisensory conditions. Additionally, in both the auditory and visual modulation detection task, the modulated AV streams were presented in three

relative phase conditions ($\phi_{A=V}$, $\phi_{A>V}$, $\phi_{V>A}$; see Figure 1A), each of which was tested in a different participant group. No significant effect of the experimental conditions was found on the modulation detection thresholds (see Supplementary Data 1.1 and Supplementary Figure 3). The effects on response times are explored below.

At the end of the experiment, participants reported on their perception of the barely-detectable stimulus of the other modality during the auditory and the visual task performed on AV stimuli. Overall, 10 out of 27 participants reported they were unaware of the low-intensity stimulus in the other modality in both tasks. Others reported being moderately (12 out of 27 participants) to largely aware (5 out of 27 participants) of the low-intensity stimulus. From these 17 participants, eight further highlighted that they primarily noticed the presence of the low-intensity visual stimulus (only at the estimated 65% detection threshold intensity) during the auditory task but not vice versa. None of the participants reported being aware of (in)congruence between the AV streams.

3.1 Faster Unisensory Response Times for Modulated than Static Stimuli

The 84% unisensory measurements were made during three sessions, i.e. on Day 2, and then twice during the multisensory sessions. We studied the effects on unisensory response times using only the data collected during the multisensory sessions. Figure 3 shows unisensory response times measured during auditory (A-B) and visual (C-D) tasks. In the context of unisensory stimuli, the labels refer to the assignment of the three participant groups to the phase-related manipulation of the unimodal modulated visual (Group 1 and Group 2 have no phase-shift, Group 3 is phase advanced and indicated by red outline) or unimodal modulated auditory (Group 1 and Group 3 have no phase-shift, Group 2 is phase advanced and indicated by blue outline) stimuli. Static stimuli are the same across groups. Overall, in both tasks, modulated stimuli (gray bars) yielded faster response times than static stimuli, and phase advancing the modulated stimulus (bars with colored outlines) provided an extra response time advantage.

Figure 3A and B show the response times (averaged across the two sessions) during the unisensory auditory task for the two main stimulus dynamics conditions (static vs modulated) split over the three participant groups. Figure 3A shows an overall trend for modulated auditory stimuli to yield faster response times (gray bars) compared to static stimulus (dark bars). Additionally, when the modulated stimulus was also phase-advanced (red-outlined gray bar), the response time was the fastest. Figure 3B1 visually illustrates that a phase-advanced auditory stimulus provided a response time advantage. Figure 3B2 visually illustrates a sizeable response time advantage for modulated auditory stimuli over static stimuli.

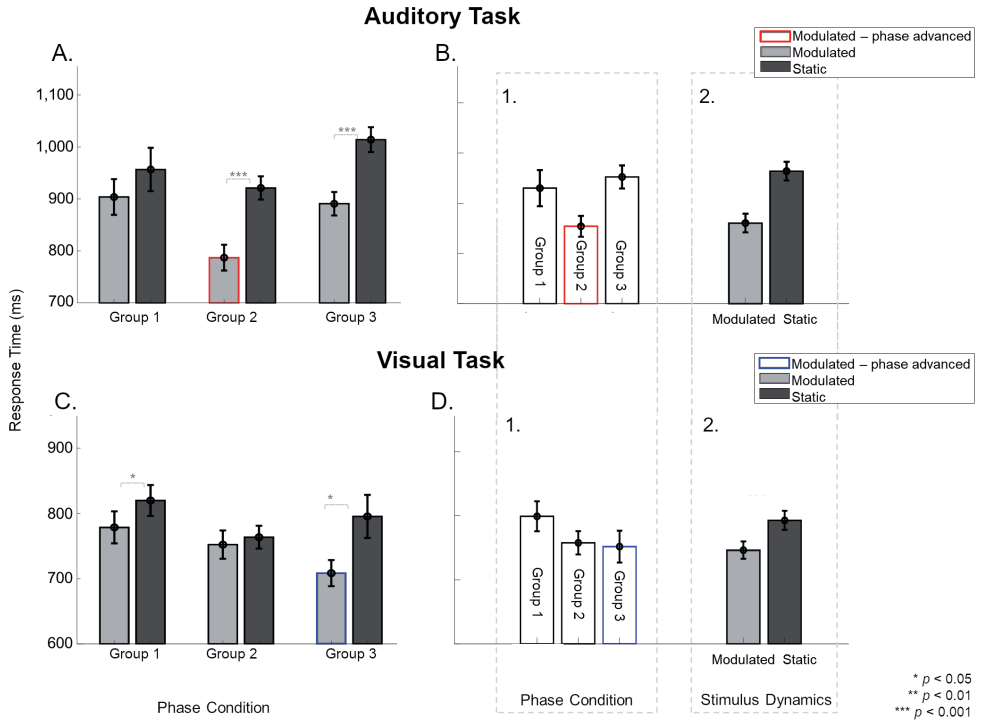


Figure 3. Effect of modulating unisensory stimuli on response times in modulation detection tasks across the three participant groups, averaged over two sessions. Light gray and dark gray bars indicate response times for modulated and static stimuli respectively. Red and blue borders indicate the modulation conditions with phase shifts. A) In the auditory task, participants in Group 2 and 3, but not Group 1, were significantly faster to identify modulated than static sounds. Response times to auditory stimuli not only varied between the three groups (B1) but also for stimulus dynamics (B2). C) In the visual task, participants in Group 1 and 3, but not Group 2, were significantly faster to identify modulated than static sounds. D1) Group had no overall effect on response times. Only stimulus dynamics (D2) significantly affected the response times for visual stimuli. Error bars indicate ± 1 SEM.

In the task on visual unisensory stimuli (Figure 3C and D), analogous to findings with the auditory unisensory task, the overall trend towards a response time advantage for modulated stimuli appeared to be strengthened when phase advancing the visual modulated stimulus (Figure 3D, blue-outlined bar). Again, there was no physical difference between the stimuli shown in the three groups of participants, except that the modulated visual stimulus was phase-advanced in the $\phi_{V>A}$ group compared to the modulated stimuli in the two other groups. There was no clear difference in response time among the different groups (Figure 3D1), but as with the auditory task, response times were faster for modulated than static auditory stimuli (Figure 3D2).

The above description is supported by the following statistical analyses. To analyze the unisensory auditory response times, a three-way ANOVA was performed in which participants were assigned to the same conditions as used for the multisensory part of the experiments. Hence, the unisensory data were analyzed to test effects of the "Phase conditions" (3 levels: Group 1 – Group 2 – Group 3), as well as the "Stimulus dynamics" (2 levels: modulated – static) and "Session" (2 levels: unisensory measurements from multisensory conditions with non-task stimulus at an estimated 55% intensity, and 65% intensity). The three-way interaction was not significant. There were, however, two significant two-way interactions between the factors "Stimulus dynamics" and "Session" ($F(1,24) = 5.45$, $p = 0.02$), and "Phase condition" and "Stimulus dynamics" ($F(2,24) = 4.79$, $p = 0.01$). The significant two-way interaction between "Stimulus dynamics" and "Session" was further explored with a pairwise comparison between modulated and static sounds per session. These analyses showed that responses to modulated stimuli were faster than to static sounds in both sessions (55%: $t(8) = -8.06$, p[corrected] < 0.001; 65%: $t(8) = -6.49$, p[corrected] < 0.001).

The interaction between "Phase condition" and "Stimulus Dynamics" was further explored for each phase condition (Figure 3A). Pairwise comparisons showed that the response times for modulated stimuli were statistically faster than static stimuli in two of three cases (for Group 2: $t(8) = -5.94$, p[corrected] = 0.001 and Group 3: $t(8) = -16.45$, p[corrected] < 0.001, but not Group 1: $t(8) = -2.07$, p[corrected] = 0.2). Because Group 3 and Group 1, in the context of the unisensory task, have physically identical stimuli, the presence of a statistical difference between static and modulated conditions for Group 3 and not for Group 1 could reflect an influence of the multisensory context in which the unisensory task was embedded, or a group difference. As the order in which unisensory measurements were taken was randomized across subjects (measured before any multisensory exposure in one-third of participants), our data is limited in the ability to shed light on this observed difference.

For unisensory visual response times, a three-way ANOVA with between-subject factor "Phase condition" (3 levels: Group 1 – Group 2 – Group 3) and within-subject factors "Stimulus dynamics" (2 levels: modulated – static) and "Session" (2 levels: unisensory measurements from multisensory conditions with non-task stimulus at 55% intensity, and 65% intensity) showed only a significant two-way interaction between "Phase condition" and "Stimulus Dynamics" ($F(2,24) = 4.94$, $p = 0.015$). The three-way interaction and all other two-way interactions were insignificant. The significant interaction was explored for each phase condition (Figure 3C). The response times for modulated stimuli were faster than static stimuli for Group 1 ($t(8) = -3.52$, p[corrected] = 0.02) and Group 3 ($t(8) = -3.85$, p[corrected] = 0.01) but not for Group 2 ($t(8) = -0.74$, p[corrected] > 0.99). Here, again the physical conditions for the unisensory task were identical in Group 1 and Group

2 but led to different outcomes of statistical testing. This shows that also for the visual unisensory task, the main finding is an overall response time advantage for the modulated stimulus, which is strengthened by phase advancing the visual modulated stimulus.

3.2 Phase-dependent Response Time Reduction in Auditory Rask due to Visual Influences

Next, we evaluated the multisensory influence of weak (static or modulated) stimuli in one modality on the response time for stimuli in the other modality. We focused first on the influence of visual stimuli on responses to auditory stimuli. We were interested in the effect of the different onset phases for the modulated auditory and modulated visual stimuli, and also in the inter-sensory interactions between modulated and static stimuli of the auditory and visual modalities.

Figure 4 shows all the effects observed for auditory response times. The different factors contributing to the observations are shown in different panels. In panel A, the response times are grouped by the auditory stimulus (modulated and static) for the three phase conditions. Panel B shows the same response time data as in panel A, but categorized by the different multisensory visual influences (modulated, static or no visual stimulus) during the auditory task. In panels C, D, and E, the different effects of visual influence (modulated, static or no visual stimulus) on modulated and static auditory stimuli due to the three phase conditions are shown individually. Overall, we found that modulated auditory stimuli (irrespective of the visual influence) show a shorter response time than the static auditory stimuli. In addition, compared to the absence of visual influence (i.e., unisensory auditory stimuli), response times were shorter when the visual influence was phase advanced ($\phi_{V>A}$). On the other hand, the static visual influence increased response times compared to the unisensory condition (as if the static visual stimulus acted as a distracter). These effects are discussed below in detail along with the statistical analysis of the data.

A mixed four-way ANOVA of between-subject factor "Phase condition" (3 levels: $\phi_{A=V}$ – $\phi_{A>V}$ – $\phi_{V>A}$) and the three within-subject factors "Auditory stimulus" (2 levels: modulated – static), "Visual influence" (3 levels: modulated – static – none) and "Intensity" of the visual influence (2 levels: 55% – 65%) showed a significant four-way interaction ($F(4,48) = 2.957$, $p = 0.029$). The level 'none' indicates the absence of a visual stimulus and thus refers to unisensory auditory response time measurements from the two multisensory sessions. Further analysis of the interaction showed that the effect of "Intensity" of the visual influence was insignificant (see Supplementary Data 1.2 and Supplementary Figure 4 for details). Thus, to simplify the interpretation of the effects, we averaged over the data for the two intensities before exploring a mixed ANOVA with the between-subject factor "Phase condition" and two within-subject factors: "Auditory stimulus" and "Visual

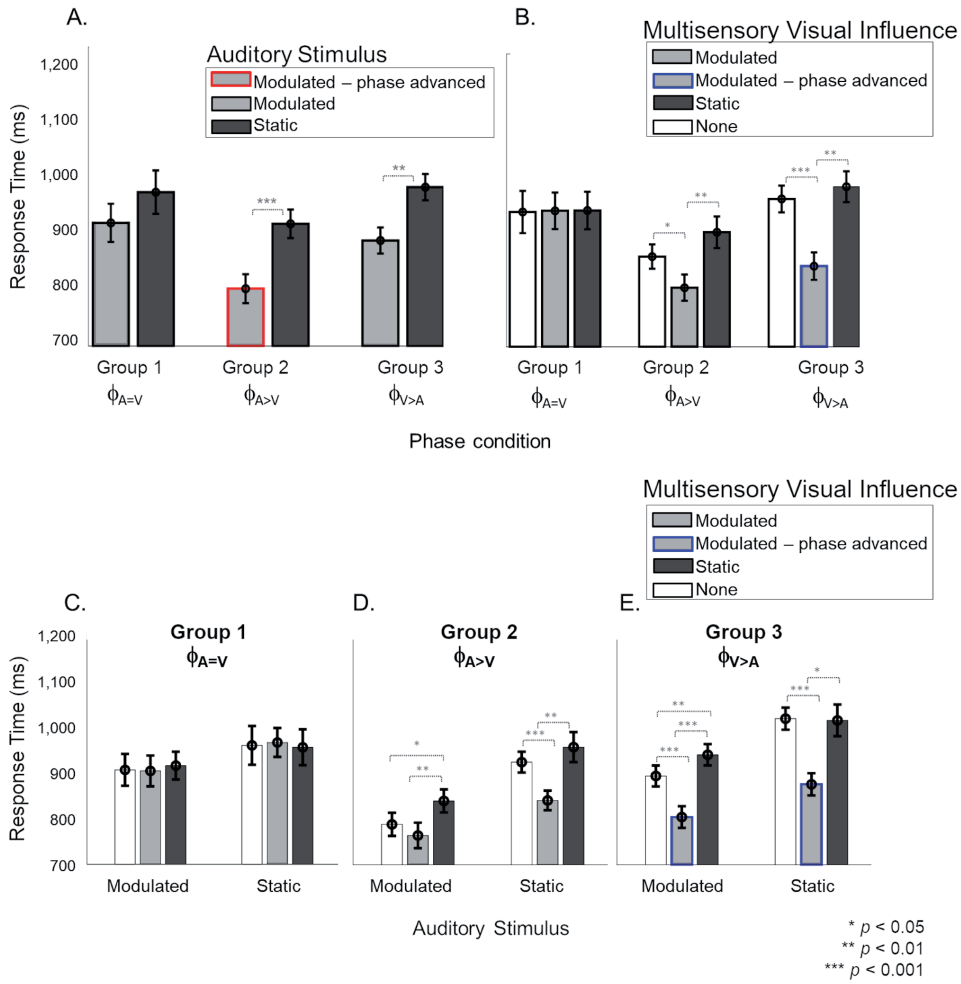


Figure 4. Response times during the auditory task with visual influences. In panels B-E, light and dark gray bars represent the presence of barely-detectable modulated and static visual influences respectively, while white bars show the unisensory condition. Red and blue lines indicate phase-advanced auditory and visual conditions respectively. Error bars represent ± 1 SEM. Panel A and B show the main effects of "Auditory stimulus" (2 levels: modulated – static) and "Visual influence" (3 levels: none – modulated – static) respectively. Panels C-E the interaction of "Auditory stimulus" and "Visual influence" plotted separately for phase conditions $\phi_{A=V}$, $\phi_{A>V}$, and $\phi_{V>A}$, respectively. While no effect of the visual influence on the auditory stimulus was observed in phase condition $\phi_{A=V}$ (C), responses to sounds were significantly faster when a modulated compared to static visual influence was present in phase conditions $\phi_{A>V}$ and $\phi_{V>A}$ (D-E). (C) For phase condition $\phi_{A=V}$, when the auditory stimulus and visual influence were in-phase (no phase shift for modulated stimuli), no significant interaction between the auditory stimulus and visual influence was observed. (D) For phase condition $\phi_{A>V}$ (modulation phase of the auditory stimulus was leading with respect to that of the visual influence), we observed an overall distraction effect of the static visual influence and no advantageous effect of the visual influence for modulated auditory stimuli. Response times for the static auditory stimuli became faster due to the modulated visual influence. (E) For phase condition $\phi_{V>A}$ (the modulation phase of the visual influence was leading with respect to that of the auditory stimulus), an advantage of the modulated visual influence, as well as a disadvantage in case of a static visual influence, were observed compared to unisensory sounds.

influence". Results showed a significant three-way interaction (Figure 4, $F(4,48) = 3.11$, $p = 0.023$), which we further explored by analyzing the data per "Phase condition" using a repeated measures ANOVA (with factors "Auditory stimulus" and "Visual influence"). The main effects of "Auditory stimulus" and "Visual influence" on response time for the auditory stimulus are shown separately in Figure 4A and 4B respectively, for the three phase conditions ($\phi_{A=V}$, $\phi_{A>V}$, $\phi_{V>A}$). The interaction effect is broken down into effects of visual influence on modulated and static auditory stimuli for the three phase conditions in Figure 4C, D, and E respectively.

In the $\phi_{A=V}$ condition (Group 1), Figure 4A, B, shows that there were no main effects, neither of the factors "Auditory stimulus" (Figure 4A, $F(1,8) = 6.16$, p[corrected] = 0.11) nor of "Visual influence" (Figure 4B, $F(2,16) = 0.03$, p[corrected] > 0.999), and the two factors also did not interact (Figure 4C, $F_{0.016}(2,48) = 0.83$, $p = 0.44$).

In the $\phi_{A>V}$ condition (Group 2), the main effects of "Auditory stimulus" (Figure 4A, $F(1,8) = 28.54$, p[corrected] < 0.001) and "Visual influence" (Figure 4B, $F(2,16) = 18.28$, p[corrected] < 0.001), and their interaction (Figure 4D, $F_{0.016}(2,48) = 6.11$, $p = 0.004$) were significant. The interaction was further explored with a separate one-way ANOVA for modulated and static auditory stimuli. For modulated auditory stimuli (Figure 4D, left), there was a significant effect of the "Visual influence" ($F_{0.008}(2,48) = 19.67$, $p <$ 0.001). Pairwise comparisons showed that the presence of a modulated visual influence significantly sped up response times in comparison with a static visual stimulus (compare gray to dark bar, $t(8) = -4.31$, p[corrected] = 0.007), but a modulated visual influence did not give a significant advantage in comparison with the unisensory condition (compare gray to white bar, $t(8) = -1.78$, p[corrected] = 0.33). However, the presence of a static visual influence significantly slowed down response times as compared with a unisensory auditory stimulus (compare white to dark bar, $t(8) = 3.55$, p[corrected] = 0.02). For static auditory stimuli (Figure 4D, right), there was a significant effect of the visual influence as well ($F_{0.008}(2,48) = 47.69$, $p < 0.001$). Responses to static auditory stimuli were faster when accompanied by a modulated, non-phase-advanced visual influence. This was true when comparing to a visual static influence (compare gray to dark bars, $t(8) = -5.44$, p[corrected] = 0.001) and when comparing to a situation in which there was no visual influence at all (compare gray with white bar, $t(8) = -6.34$, p[corrected] < 0.001). The response times for the static auditory stimulus were the same irrespective of whether it was paired with a visual static influence or with no visual stimulus at all (compare dark and white bars, ($t(8) = 1.89$, p[corrected] = 0.28).

In the $\phi_{V>A}$ condition (Group 3), the main effects of "Auditory stimulus" (Figure 4A, $F(1,8) = 87.50$, p[corrected] < 0.001) and "Visual influence" (Figure 4B, $F(2,16) = 27.19$, p[corrected] < 0.001) were significant. There was also a significant interaction between

factors "Auditory stimulus" and "Visual influence" ($F_{0.016}(2,48) = 5.94$, $p = 0.005$; Figure 4F). Further investigation of the interaction (Figure 4E) showed that the presence of a visual influence significantly changed response times for both modulated and static auditory stimuli (modulated: $F_{0.025}(2,48) = 62.98$, $p < 0.001$; static: $F_{0.025}(2,48) = 87.98$, $p < 0.001$). For modulated auditory stimuli (Figure 4E left), there was a response time advantage when there was a modulated rather than a static visual influence (compare gray to dark bars, $t(8) = -7.7$, p[corrected] <0.001), and also when there was a modulated rather than no visual influence (compare gray to white bars, $t(8) = -8.55$, p[corrected] < 0.001). Response times for the modulated auditory stimulus were slower when there was a visual static influence when compared to absence of visual influence (compare dark and white bars, $t(8) = 5.77$, p[corrected] = 0.001). A similar data pattern was present for response times for static auditory stimuli (Figure 4E right, visual modulated vs visual static influences: $t(8) = -3.41$, p[corrected] = 0.02; visual modulated influences vs no visual influence at all: $t(8) = -7.45$, p[corrected] < 0.001; visual static vs no visual influence at all: $t(8) = -0.15$, p[corrected] > 0.999).

Overall, in the auditory task, we found no effect of visual influences for the phase-aligned ($\phi_{A=V}$) condition. We observed that the visual phase-advanced modulatory stimuli ($\phi_{V>A}$) sped up the detection of static and modulated (not phase-advanced) auditory stimuli (Figure 4E). For a modulated auditory stimulus in phase condition $\phi_{A>V}$, the presence of a modulated (not phase-advanced) visual influence did not provide an advantage over the unisensory condition (Figure 4D). The static visual stimuli had a distracting effect on modulated auditory stimuli in phase conditions $\phi_{A>V}$ and $\phi_{V>A}$ (Figure 4D and E).

3.3 Response Times in Visual Task do not Benefit from Auditory Influences

We then explored the reciprocal effect of weak auditory stimuli on response times for visual stimuli (Figure 5). In panels A and B, the response times are grouped by the visual stimulus (modulated and static) and multisensory auditory influence (modulated, static, or no visual stimulus) during the visual task, respectively. Panels C-E show the individual effects of auditory influence on modulated and static visual stimuli in the three phase conditions. Overall, we observed that modulated visual stimuli show a shorter response time than the static visual stimuli, but no multisensory auditory influence on the response time for the visual stimuli. These findings are supported by the statistical analysis of the data, as reported below.

A mixed four-way ANOVA of between-subject factor "Phase condition" (3 levels: $\phi_{A=V}$ – $\phi_{A>V}$ – $\phi_{V>A}$), and the three within-subject factors "Visual stimulus" (2 levels: modulated – static), "Auditory influence" (3 levels: modulated – static – none) and "Intensity" of the auditory influence (2 levels: 55% – 65%) did not show a significant four-way interaction ($F(4,48) = 2.44$, $p = 0.059$, see Supplementary Figure 5). However, the three-

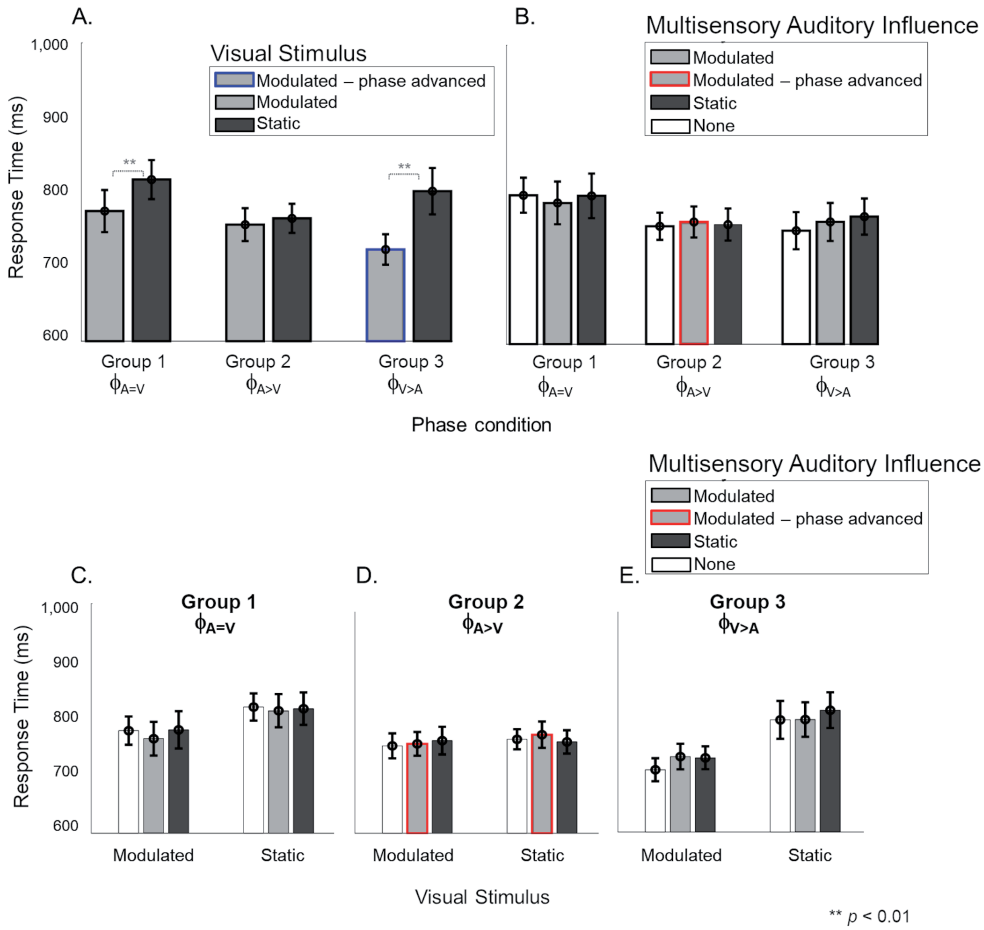


Figure 5. Response times during the multisensory visual task. In panels B-E, light and dark gray bars show the presence of barely-delectable modulated and static auditory influences, respectively, while white bars show the unisensory condition. Red and blue lines indicate phase-advanced auditory and visual conditions respectively. (A)There was a significant main effect of "Visual stimulus" (2 levels: modulated – static) where responses for modulated stimuli were faster than static stimuli for phase conditions $\phi_{A=V}$ and $\phi_{V>A}$. There was no main effect of "Auditory influence" (3 levels: none – modulated – static) (B) nor any significant interaction between the auditory stimulus and visual influence for any phase conditions (C–E). Error bars represent ± 1 SEM.

way interaction between "Phase condition", "Visual stimulus" and "Auditory influence" was significant ($F(4,48) = 2.687$, $p = 0.042$), and we therefore further analyzed the data per phase condition.

The effect of "Visual stimulus" (Figure 5A) was significant for the $\phi_{A=V}$ condition (Group 1,$F(1,16) = 14.08$, p[corrected] = 0.015) and the $\phi_{V>A}$ condition (Group 3, $F(1,8) = 18.44$, $p = 0.006$) but not for the $\phi_{A>V}$ condition (Group 2, $F(1,16) = 0.45$, p[corrected]

> 0.999). The effect of "Auditory influence" (Figure 5B) failed to reach significance for all phase conditions. There was no significant interaction between factors "Visual stimulus" and "Auditory influence" for $\phi_{A=V}$ (Figure 5C), $\phi_{A>V}$ (Figure 4D), and $\phi_{V>A}$ (Figure 5E).

To summarize, we observed no effect of low-intensity auditory influences on the response times for visual stimuli. The responses to modulated visual stimuli were faster than to static visual stimuli for $\phi_{A=V}$ and $\phi_{V>A}$, as was observed already in the unisensory measurements as well. This further confirms the lack of effective auditory influences while performing the visual task.

4 Discussion

In the present work, we detailed the effect of auditory-to-visual and visual-to-auditory interactions in the far periphery using simple stimuli (gratings and noise bursts). For both an auditory and a visual task, we studied the influence of multisensory temporal (in)congruence on modulation detection threshold and response time by using static and modulated stimuli and also by manipulating the relative phase of the modulated AV stimuli.

We report three main sets of findings. First, in the unisensory conditions, we found that the response times were generally faster for modulated stimuli compared to static stimuli for both auditory and visual modalities. This finding is in line with the advantage of having a temporal modulation in a peripheral visual stimulus (Hartmann et al., 1979) and with the human sensitivity to temporally structured stimuli in audition (Joris et al., 2004). We also found that advancing the phase of the modulated stimulus to a sharp intensity/contrast change (from maximum to minimum) at the onset of the stimulus, further shortened the response times. The phase advancement creates both a stronger onset and a maximal intensity change from maximal to minimal at the beginning of the stimulus. Our observations hence show the key role of both factors in the detection of modulated stimuli. Overall, visual response times were found to be faster than auditory response times. While generally auditory reaction times have been reported to be faster than visual reaction times (Ng and Chan, 2012; Shelton and Kumar, 2010; Arrighi et al., 2005), the opposite trend has also been observed showing the dependence of this effect on a specific task and stimulus features (Shams et al., 2010).

Second, for the visual task, we found that a weak auditory influence (at an estimated 55% or 65% detection threshold) did not affect visual detection thresholds or response times. The lack of auditory influences on the visual task may be caused by weak auditory stimuli being incapable of capturing attention, as most of the participants reported

being oblivious to the low-intensity auditory stimuli. While previous studies have shown auditory influences on responses in the visual cortex (Wang et al., 2008; Bolognini et al., 2010; Ibrahim et al., 2016) and have also shown behavioral (dis)advantages (Di Russo et al., 2002; Shams et al., 2002), differences in task and stimuli with our study may have played a role in these divergent results. The primary explanatory factor can be that all these studies presented their stimuli more centrally [foveal and parafoveal between 0-8^{o} (Shams et al., 2002; Bolognini et al., 2010),10^{o} (Chen et al., 2017) in humans] or at a maximal peripheral location of 20^{o} for monkeys (Wang et al., 2008). Additionally, we are studying the influences of low-intensity stimuli. Thus, compared to more centrally presented stimuli as used in previous studies, the added auditory stimuli might need to be at a higher intensity for significant cross-modal effects on the detection of peripheral visual stimuli to occur. This hypothesis will, however, require further testing.

Third, we observed a cross-sensory effect of visual stimuli on response times for auditory stimuli. Depending on temporal (in)congruence and synchrony between modulated AV streams, we observed that visual influences could not only speed up (facilitation effects) the response times for auditory stimuli but could also slow them down (degradation effects). We first consider the facilitation effects of modulated visual influences on modulated auditory stimuli during the auditory task (see the left halves of Figure 4C, D, and E). These effects depended on the phase relations between the visual and auditory streams. When the phase of the visual modulation led the modulated sound by 100 ms in the auditory task ($\phi_{V>A}$), a multisensory benefit (i.e., faster response times for both modulated and static sounds) due to the modulated visual influence was observed (gray bar Figure 4E, left). However, when auditory and visual modulations were in phase ($\phi_{A=V}$), no multisensory interaction was observed (gray bars Figure 4C, left). This finding of a visually-driven benefit on response time for modulation detection in the peripheral sounds when the concurrent visual stimulus is phase-advanced by 100ms may indicate a role of the direct influences from early visual to early auditory cortex. The response time advantage cannot be attributed to increased salience at the onset of the phase-advanced visual stimulus, as the static visual stimuli have the same salience at onset yet provide no advantage. Thus, the temporal dynamics of the visual stimulus must play a role. In the phase-advanced visual stimulus, the maximum-to-minimum intensity sweeps in the visual stream precede the analogous intensity sweeps in the auditory stimulus by 100 ms. Taking into account that neuronal response latencies are longer for visual stimuli than sounds [55 ms (Schroeder et al., 2008) in V1, and 23 ms (Besle et al. 2008) in A1], the visual intensity sweeps would have ~75 ms to carry cross-modal information to the early auditory areas that could facilitate auditory neural activity in response to the auditory sweeps. This is short enough to be compatible with direct interactions between early cortical sites and shows the prominent role of stimulus features at onset in driving the cross-modal advantages. Such early advantages may provide a benefit to multisensory

information processing in higher-order cortical regions. Note that when the modulated auditory stimulus itself was phase advanced $\phi_{A>V}$, the visual modulated influence did not provide a response time benefit (gray bars Figure 4D, left).

The underlying mechanisms and pathways for the observed multisensory interaction cannot be disentangled based on the present study and would require future neuroimaging and electrophysiological studies. However, the current findings can be put into perspective based on existing evidence of mechanisms that underlie cross-sensory effects. For example, "oscillatory phase-resetting" has been shown to play a part in multisensory interactions among early sensory cortices (Lakatos et al., 2007; Doesburg et al., 2008; Schroeder and Lakatos, 2009; Atilgan et al., 2018). More specifically, visual stimuli may influence auditory processing by resetting the phase of ongoing oscillatory auditory cortical activity. Cross-sensory phase-resetting has been observed in early auditory areas with influences coming from somatosensory (Kayser and Kayser, 2018) and visual input (Kayseret al., 2010). Facilitation or suppression effects have been shown to be dependent on the temporal relationship between the onsets of stimuli (Kayser et al., 2010), in line with the lead in onset for visual compared to auditory stimuli in the $\phi_{V>A}$ condition in the present study. Additionally, these effects are more pronounced at near-threshold levels (Schroeder and Lakatos, 2009; ten Oever et al., 2014) compatible with the low-contrast visual stimuli we have used. Based on our observations for phase condition $\phi_{V>A}$, where a leading modulated visual stimulus provided a response time benefit to static and modulated sounds, the sharp intensity changes at the first part of the visual stimuli (from peak to trough contrast), as well as the recurring intensity changes in further cycles of the visual oscillation, may have caused phase-resets in local oscillatory activity in the early visual cortices. These changes, in turn, might have led to an enhanced representation of the auditory information, engaging sensory integration between early cortical sites. While its underlying mechanism remains speculative, our results may provide a basis for future experiments.

An additional facilitatory effect of modulated visual influence was observed for static auditory stimuli in the $\phi_{A>V}$ and $\phi_{V>A}$ conditions, but not the $\phi_{A=V}$ condition (gray bars in right-hand parts of Figs. 3C, D, E). The facilitatory effect of the modulated visual influence in the $\phi_{A>V}$ condition is remarkable because the AV stimuli in that condition and the $\phi_{A=V}$ condition were identical (i.e., the same static auditory stimulus combined with the same visual influence). Therefore, the advantages in the $\phi_{A>V}$ condition (and possibly also the $\phi_{V>A}$ condition) for the static auditory stimulus somehow were acquired indirectly from the advantages experienced by the modulated auditory stimulus from the modulated visual influences, thus implying cross-trial effects. It is not clear how these cross-trial influences occur, but in a broad sense, they are in line with the idea that audiovisual interactions can occur at multiple stages of sensory processing (Cappe et

al., 2009; Koelewijn et al., 2010). Hence, whereas a large portion of the observed data shows that direct visual-to-auditory influences at peripherical locations might play a role in multisensory processing, the observed cross-trial dependencies of visual-to-auditory benefits to trials with static stimuli might rely on contributions of higher association cortices in the brain (Covic et al., 2017). As in our experiment design, the (in)congruent modulated and static stimuli are presented randomly in a staircase design with varying intensity of stimuli, we are unable to comment on the nature of serial interactions extending over trials. These observations pose interesting questions for further research.

We also observed a degradation effect of static visual influence. That is, only in the presence of a phase difference between AV streams, the static visual influence slowed down the response time for both modulated and static auditory stimuli compared to the unisensory and congruent modulated conditions. This effect, however, was present only for the modulated auditory (Figure 4D-E left) and not for the static auditory stimuli (Figure 4D-E right). Our findings may represent a distraction effect of the static visual stimulus. A possible explanation for the absence of this effect in phase condition $\phi_{A=V}$ might be found in the overall longer response times for that condition. As participants already took a long time to respond, the presence of the static visual stimulus may not have further slowed the responses down. That is, the static visual stimuli can only provide a disadvantage in case of a comparative advantage driven by phase-advanced modulated auditory stimuli ($\phi_{A>V}$) or visual stimuli ($\phi_{V>A}$). A slightly different view on these degradation effects is the idea that, especially in the cases where the modulated visual influence is integrated with the modulated auditory stimulus (as witnessed by a response time benefit), a static visual influence will be harmful. Hence, the observed degradation effects also support a form of audiovisual interaction.

To summarize, in our paradigm studying audiovisual interactions in far periphery, we found evidence for barely-detectable visual stimuli influencing (facilitation and degradation) the response times for auditory stimuli in a modulation detection task, but not for the reverse. Due to a programming error, eye movements were only recorded before the stimulus presentation and at the response. However, fixation samples at response time in each trial strongly support that participants fixated accurately (in 96% of the ~100,000 trials fixation samples fell within 2.5° of the fixation center, see Methods). Although our conclusions would have been stronger without our programming error, the fixation data we do have make it unlikely that the observed asymmetrical nature of multisensory interaction would be due to a confounding effect of inaccurate fixation. The observed visual-to-auditory influences only occurred for appropriate phase-differences between the modulated AV stimuli. Our data support a role of direct interactions between early visual and auditory areas through manipulation of AV synchrony (Lakatos et al., 2007). The involvement of early sensory regions in multisensory processing of stimuli at peripheral

locations does not exclude a probable major role for higher-order cortices. Multisensory integration is a multifaceted process, and higher-order cortices are likely involved in among others directing attention, object recognition and cross-trial effects. Hence, our findings support a view where both the early auditory and visual cortices as well as higher-order auditory and visual cortex contribute to multisensory integration (Ghazanfar and Schroeder, 2006). This research extends the behavioral evidence of the importance of cross-sensory temporal cues for auditory processing (Besle et al. 2008; Doesburg et al. 2008; Stevenson et al. 2010) to the far periphery. By combining temporally and spatially high-resolution neuroimaging techniques, future studies may provide insight into the precise temporal mechanisms as well as locate the cortical sites driving these cross-modal observations. Future work may also detail cross-sensory interactions ranging from far peripheral to foveal visual space.

Supplementary Materials

1 Supplementary Data

1.1 Congruency does not influence Auditory and Visual Modulation Detection Thresholds

We first explored the effects of the congruency of AV streams (modulated and static) and the phase relation between modulated AV stimuli on modulation detection thresholds. Supplementary Figure 3 shows auditory (A-C) and visual (D-F) modulation detection thresholds obtained in the unisensory condition, as well as those obtained in the audiovisual congruent and incongruent conditions for each phase-condition. No effect of (in)congruence was observed on either the auditory or the visual detection thresholds for any phase condition as shown by the following statistical analysis.

Auditory detection thresholds are shown in Supplementary Figure 3 (A-C). A mixed ANOVA showed neither significant main effects nor interactions among between-subject factor "Phase condition" (3 levels: $\phi_{A=V} - \phi_{A>V} - \phi_{V>A}$) and within-subject factors "Sensory condition" (3 levels: unisensory – congruent – incongruent) and "Intensity" (65% – 55%).

Visual detection thresholds are shown in Supplementary Figure 3 (D-F). A nearly significant 3-way interaction ($F(4,48) = 2.48$, $p = 0.056$) was found between "Phase condition" (3 levels: $\phi_{A=V} - \phi_{A>V} - \phi_{V>A}$), "Sensory condition" (3 levels: unisensory – congruent – incongruent), and Intensity (2 levels: 65% – 55%). All other interactions and main effects failed to reach significance.

1.2 Detailed statistical analysis of response times during Auditory task

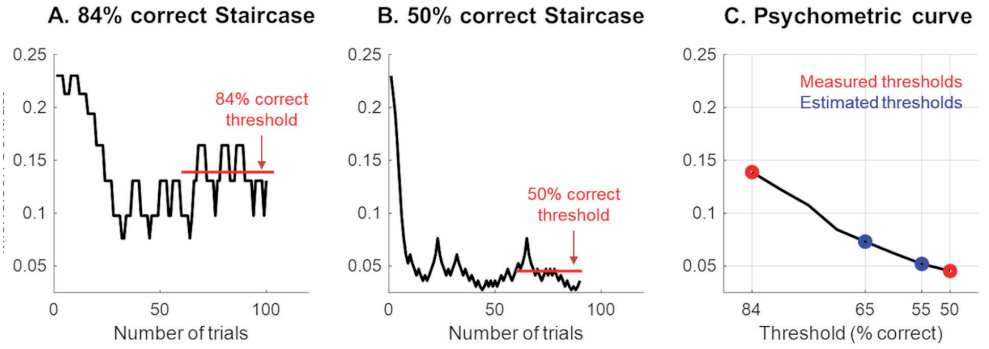
A mixed four-way ANOVA analysis of between-subject factor: "Phase condition" (levels: $\phi_{A=V} - \phi_{A>V} - \phi_{V>A}$) and the three within-subject factors "Auditory stimulus" (levels: modulated – static), "Visual influence" (levels: modulated – static – none) and "Intensity" of the visual influence (levels: 55% – 65%) showed a significant four-way interaction ($F(4,48) = 2.957$, $p = 0.029$). The interaction was analyzed for each level of phase condition for further analysis (Supplementary Figure 4). As concluded by the following analysis, the effects across "Intensity" levels were not significantly different across phase conditions and thus the observations were simplified in the main text by combining both levels.

For phase condition $\phi_{A=V}$, there was no significant three-way interaction between factors "Auditory stimulus", "Visual influence" and "Intensity". The two-way interactions and main effects also failed to reach significance.

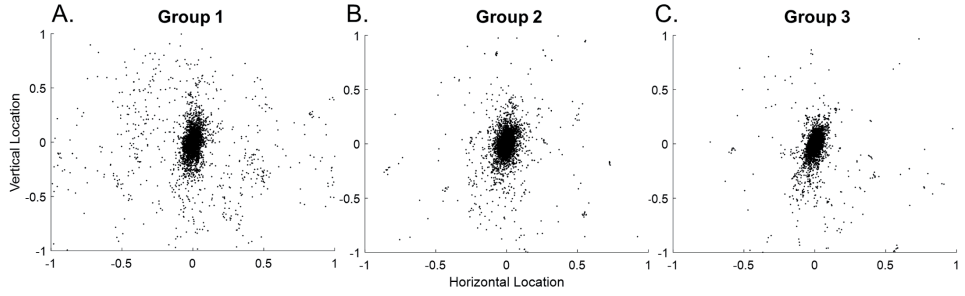
For phase condition $\phi_{A>V}$, the three-way interaction between "Auditory stimulus", "Visual influence" and "Intensity" was significant ($F_{0.016}(2,48) = 6.267$, $p = 0.003$). This interaction was explored for modulated and static auditory stimuli separately. For modulated auditory stimuli, there was no significant interaction between factors "Visual influence" and "Intensity". The main effect of "Visual influence, however, was significant ($F(2,16) = 12.7411$, $p < 0.001$, $\alpha = 0.008$). Post hoc comparisons showed that modulated visual influence sped up response times for static auditory stimuli, while static visual influence slowed down the responses (modulated vs static $t(8) = -4.31$, p[corrected] = 0.007, modulated vs none $t(8) = -1.78$, p[corrected] = 0.337, static vs none $t(8) = 3.54$, p[corrected] = 0.02). For static auditory stimuli, the interaction between factors "Visual influence" and "Intensity" was significant ($F_{0.008}(2,48) = 5.38$, $p = 0.007$) and was further explored for the two levels of Intensity (55% and 65%). At 55% intensity, there was a significant effect of "Visual influence" ($F_{0.003}(2,16) = 14.366$, $p < 0.001$; modulated vs static: $t(8) = -4.14$, p[corrected] = 0.009; modulated vs none: $t(8) = -5.217$, p[corrected] = 0.002, static vs none $t(8) = 0.232$, p[corrected] > 0.999). "Visual influence" also significantly affected the response times at 65% intensity ($F_{0.003}(2,16) = 19.08$, $p < 0.001$; modulated vs static $t(8) = -5.178$, p[corrected] = 0.002; modulated vs none $t(8) = -4.32$, p[corrected] = 0.007; static vs none $t(8) = 2.87$, p[corrected] = 0.062). Overall, the presence of a modulated visual influence improved the response times for static sounds irrespective of its intensity.

For phase condition $\phi_{V>A,}$ there was a significant three-way interaction between "Auditory stimulus", "Visual influence" and "Intensity" ($F_{0.016}(2,48) = 6.64$, $p = 0.002$). This interaction was broken down for each level of "Auditory stimulus". For modulated auditory stimuli, the main effect of "Visual influence" was significant ($F_{0.008}(2,16) = 58.99$, $p < 0.001$; modulated vs static $t(8) = -7.698$, p[corrected] < 0.001; modulated vs none $t(8) = -8.55$, p[corrected] < 0.001; static vs none $t(8) = 5.778$, p[corrected] = 0.001). The two-way interaction between "Visual influence" and "intensity" and the main effect of "Intensity" was insignificant. In case of static auditory stimuli, the interaction between "Visual influence" and intensity was significant ($F_{0.008}(2,48) = 10.174$, $p < 0.001$). This interaction was further explored for each level of "Intensity". At 55% intensity, "Visual influence" had a significant effect on response times of static sounds ($F_{0.003}(2,16) = 18.5$, $p < 0.001$; modulated vs static $t(8) = -3.27$, p[corrected] = 0.03; modulated vs none $t(8) = -11.89$, p[corrected] < 0.001; static vs none $t(8) = -1.468$, p[corrected] = 0.54). The "Visual influence" at 65% also showed significant effects ($F_{0.003}(2,16) = 10.67$, $p = 0.001$; modulated vs static $t(8) = -3.40$, p[corrected] = 0.02; modulated vs none $t(8) = -3.86$, p[corrected] = 0.01; static vs none $t(8) = 1.33$, p[corrected] = 0.65).

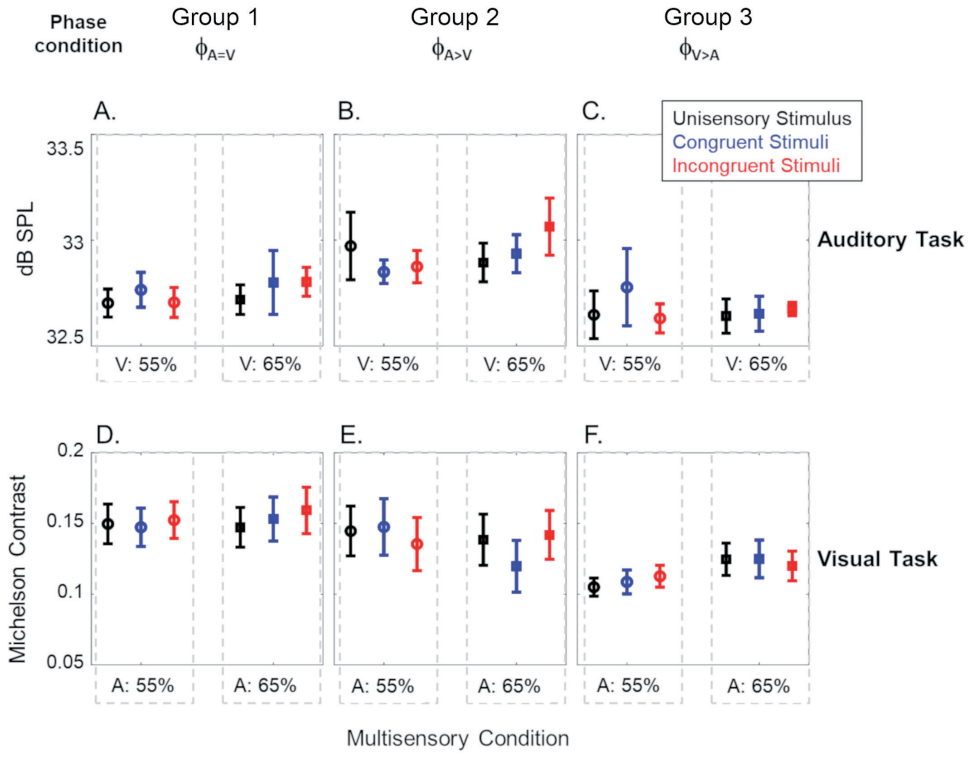
2 Supplementary Figures



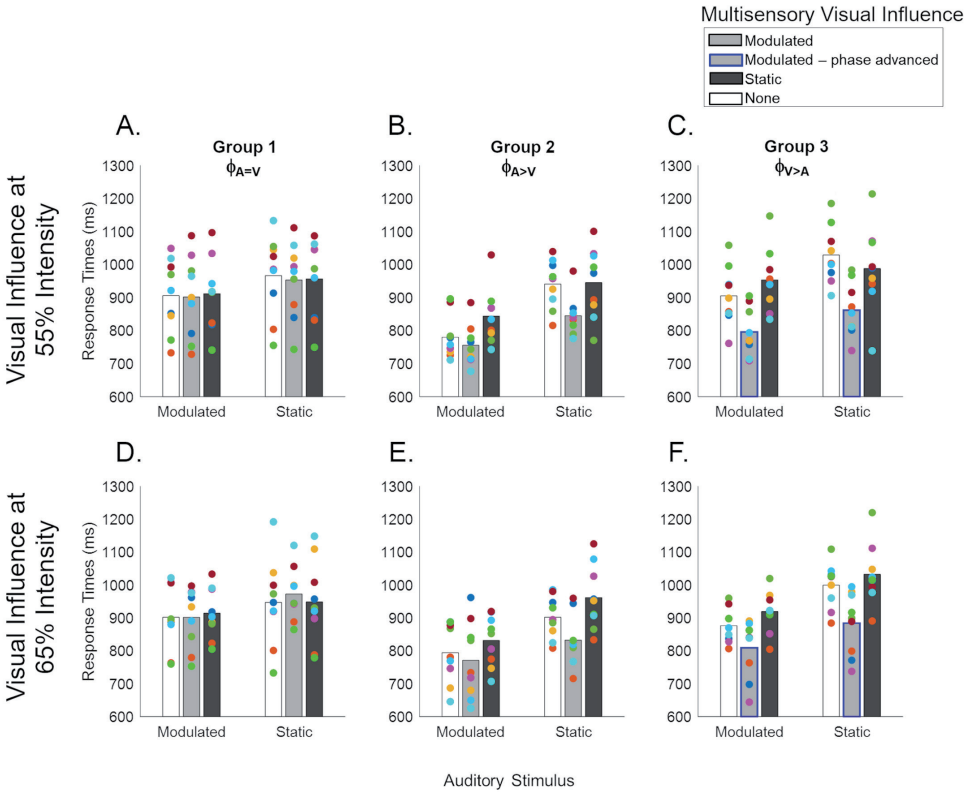
Supplementary Figure 1: Estimation of unisensory modulation detection thresholds for a single participant executing the visual task. (A) Measurements (black line) in an 84% correct detection threshold staircase (4 correct responses: contrast down, 1 incorrect response: contrast up). The final threshold is shown by the red line and is computed as the mean of the last 10 reversal points. (B) The 50% detection threshold measurements (black line, 1 correct response: contrast down, 1 incorrect response: contrast up) are shown, with the red line indicating chance detection level (the mean of the last 10 reversal points). (C) The measured 84% and 50% correct contrast thresholds (in red) are used to compute the 65% and 55% correct contrast thresholds (in blue) used in the multisensory conditions. Specifically, the 65% and 55% detection steps are estimated by z-scoring the contrast steps between the 84% and 50% measurements, and then interpolating the intermediate steps (from 50% to 84%).



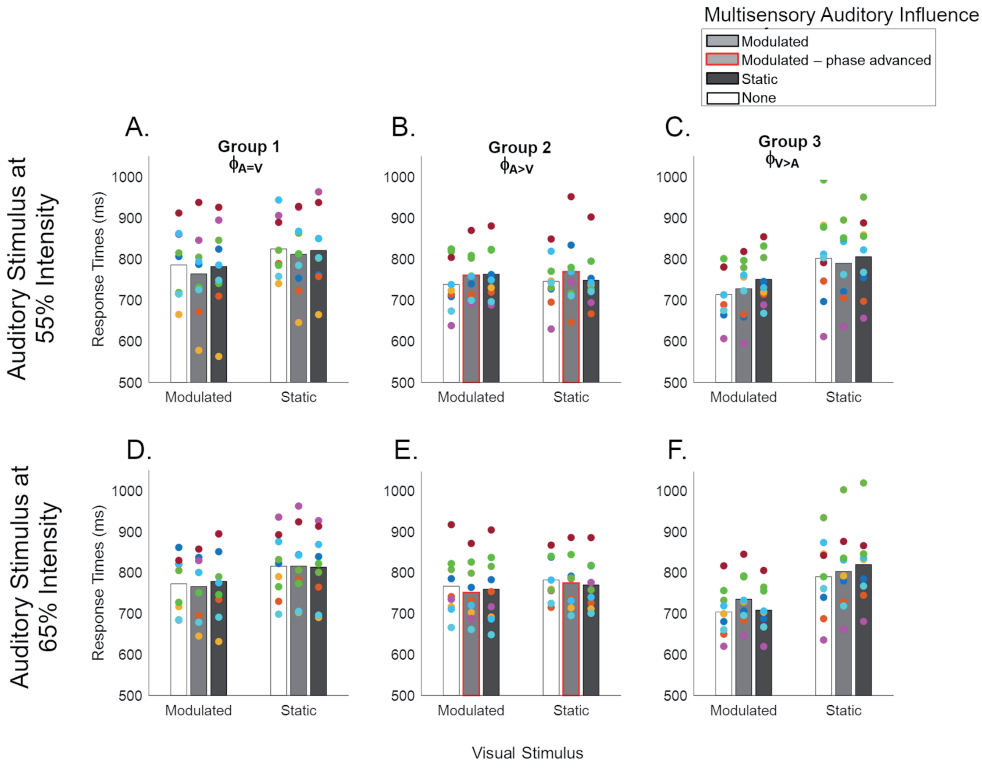
Supplementary Figure 2: Eye location at the instance of response for all trials shown for the three participant groups. The axes indicate the location along the screen, where the participants were required to fixate at the center (0,0) and the visual stimulus was presented at the farthest right location along the azimuth. Participants across groups were fixating close to the fixation center in the vast majority of the trials and showed no particular bias towards the stimulus location.



Supplementary Figure 3. Auditory and visual modulation detection thresholds. Auditory detection thresholds are shown in (A-C) for all phase conditions ($\phi_{A=V}$, $\phi_{A>V}$, $\phi_{V>A}$ respectively). The SPL values are shown for the peak intensity of the stimuli. Visual detection thresholds are shown in (D-F) for all phase conditions. The detection thresholds for all unisensory (black), congruent (blue) and incongruent (red) conditions are plotted for two intensities of the unattended modality (at estimated 55% and 65% detection threshold). Error bars represent ± 1 SEM.



Supplementary Figure 4: Response times during the auditory modulation detection task with visual influences. The figure shows the interaction of "Auditory stimulus" and "Visual influence" plotted separately for the three phase conditions $\phi_{A=V}$ (Group 1), $\phi_{A>V}$ (Group 2) and $\phi_{V>A}$ (Group 3), with "Intensity" of visual influence at 55% (A-C) and 65% (D-F). Light and dark gray bars represent the presence of near-threshold modulated and static visual influences respectively, while white bars show the unisensory condition. Gray bars with a blue outline indicate the phase-advanced visual condition. Individual participant data is shown using colored dots (different groups refer to different participant samples). Bar height reflects the mean of the data.



Supplementary Figure 5: Response times during the visual modulation detection task with auditory influences. How response times for "Visual stimulus" are affected by "Auditory influence", is plotted separately for the three phase conditions $\phi_{A=V}$ (Group 1), $\phi_{A>V}$ (Group 2) and $\phi_{V>A}$ (Group 3), with "Intensity" of auditory influence at 55% (A-C) and 65% (D-F). Light and dark gray bars represent the presence of near-threshold modulated and static visual influences respectively, while white bars show the unisensory condition. The gray bars with a red outline indicate the phase-advanced auditory condition. Individual participant data is shown using colored dots (different groups refer to different participant samples). Bar height reflects the mean of the data.

Chapter 5

Cortical Depth-dependent Multisensory and Attentional Influences on Peripheral Sound Processing

Zulfiqar I., Formisano E., Kashyap S., De Weerd P., and Moerel M. (in preparation). Cortical Depth-dependent Multisensory and Attentional Influences on Peripheral Sound Processing

Chapter 6

Summary and General Discussion

The present thesis investigated information processing in the human auditory cortex (AC) and the use of computational modeling to bridge information obtained across methods (physiological to behavioral), scales (from single neuron to behavior), and species (human and non-human primates). Furthermore, the presented work generated new datasets and empirical results that can inform, extend, and improve the AC computational models. In the first part of the thesis, we constructed a computational model of the AC that incorporates the parallel information processing pathways along the rostral-caudal axis of the AC. This model links neuronal response properties at the microscale to functional observations at the meso- and macroscale. The model was validated against existing data and then employed to construct hypotheses on the neural correlates of experimental (i.e., behavioral and neuroimaging) observations in human sound perception. In the second part of the thesis, behavioral and neuroimaging techniques were used to detail the visual influences on auditory processing in the AC. Overall, our results suggested distinct roles of the parallel information processing pathways for sound processing and provided evidence for the role of the AC beyond uniquely unisensory processing. Across these studies, both the simulated and observed responses showed interesting variations along the auditory cortical hierarchy, and suggest a prominent role for belt regions in auditory processing of complex sounds and audiovisual processing. In this chapter, we integrate results reported in the individual chapters and discuss follow-up research along with potential future applications.

1 Bridging the Scales: From Neurons to Imaging and Behavior

The computational modeling approach taken in Chapters 2 and 3 primarily intended to link the different scales of empirical observations to each other. Specifically, we focused on the differences in sound processing that exist along the rostral-caudal axis of the AC. This research was fueled by evidence that the areas located caudally and rostrally to the primary auditory cortex exhibit distinct neuronal response properties (Recanzone et al., 2000; Tian et al., 2001; Bendor and Wang, 2008; Camalier et al., 2012; Kuśmierek et al., 2014) which have been proposed to underlie their functional specialization, thus forming the starting points of "what" (rostral) and "where" (caudal) pathways (Kaas et al., 1999; Romanski et al., 1999b; Belin et al., 2000; Kaas and Hackett, 2000; Rauschecker and Tian, 2000; Tian et al., 2001; Arnott et al., 2004). How the differences in neuronal dynamics lead to specific roles in auditory perception, has remained an open question in auditory neuroscience (Jasmin et al., 2019).

The computational model presented in Chapter 2 was built on physiological and electrophysiological observations primarily recorded from non-human primates (Kaas and Hackett, 2000). The model was employed to investigate the contribution of the

different cortical streams in the representation and processing of basic acoustic features (i.e., temporal modulation, pitch) in the context of artificial and natural (speech) stimuli. The model, simulating neuronal populations (operating at mesoscale), replicated human performance in simple psychophysical tasks. Thereby it provided insight on how human auditory perception may be shaped by underlying neuronal responses and which cortical sites might underlie said behavior. The simulations showed more complex computations when moving higher in the auditory cortical hierarchy. This is consistent with the role of belt areas in increasingly complex auditory tasks. That is, while the detection of amplitude modulations in simple artificial stimuli was primarily coded by the simulated core areas of the AC, testing with more complex stimuli showed that the simulated auditory belt (but not core) provided a distributed coding mechanism for temporal and spectral pitch (in the caudal and rostral regions of the simulated belt, respectively). Further analysis with speech stimuli strengthened the idea that the neuronal response properties may be optimized along the rostral-caudal belt to process different acoustical features in parallel, with different simulated regions preferentially coding different oscillatory components of the signal. Interestingly, the slowest oscillations, representing the speech envelope, were coded in parallel across simulated regions and may serve to "timestamp" the traces of different speech aspects belonging to the same speech utterance across streams. This might serve as a binding mechanism that ensures the unified processing of different components of speech (Giraud and Poeppel, 2012; Yi et al., 2019), which may be coded in a distributed fashion. Such a temporal code can also underlie the binding of auditory sources in stream segregation (Elhilali et al., 2009).

Despite being simplistic, the proposed computational model of the auditory cortex offered a general framework for information processing along the rostral-caudal axis in the AC. The model was then used to gain new insights into existing experimental data in Chapter 3. Recent neuroimaging studies have reported a spectro-temporal trade-off along the rostral-caudal belt, i.e., a preference for fine spectral structures of sounds in the rostral regions, in comparison with partiality to fine temporal features of sounds in the caudal regions (Schöwiesner and Zatorre, 2009; Santoro et al., 2014). While the hemodynamic blood oxygenation level-dependent (BOLD) signals measured with functional MRI (fMRI) are correlated to the underlying neuronal activity (Logothetis et al. 2001; Logothetis et al. 1999; Rees et al., 2000), it does not directly measure the neuronal activity. Thus, a forward modeling approach was put forth in Chapter 3 to determine whether the spectro-temporal preferences for sound features along the rostral-caudal streams, inferred from the modeling of fMRI data (Santoro et al., 2014, 2017), could be a direct consequence of the fundamental neuronal mechanisms and response properties. The applied approach combined the computational model of the AC presented in Chapter 2 with a biophysical model of the hemodynamic BOLD response (Havlicek et al., 2015). Our simulations showed that the hemodynamics of

a caudal belt region in the AC were best explained by the neuronal models with faster temporal dynamics and broader spectral tuning, while that of a rostral belt region were best explained through fine spectral tuning combined with slower temporal dynamics. These simulations provided a direct link between observations of neuronal dynamics from electrophysiological recordings (microscale) upon which the model was built, to the BOLD responses (mesoscale). Whereas the observation of BOLD responses provided information about distinct preferences for sound features along the rostral-caudal belt regions, the proposed modeling approach provided insights into the neuronal dynamics that may cause the observed experimental effects.

The modeling endeavors of Chapters 2 and 3 have shown how computational modeling can meaningfully integrate experimental observations, generate new insights into existing datasets, and produce hypotheses for future research. The primary focus of the model was to establish evidence for the role of neuronal dynamics in meso- and macroscale level observations. The model, however, represents a simplification of a complex system and one has to remain cautious of its limitations. Models cannot replace data and the link to empirical observations must be maintained. Also, the models can only suggest or disprove a certain mechanism as a root cause for an observation and will always be reliant on the experimental findings for definitive proof. The simplicity of the current model, which allowed us to manipulate parameters in a well-controlled manner, ignored other key contributors to information processing in the AC. We explored the processing in tonotopic channels, but the influences of non-tonotopic connectivity and multisensory information were essentially disregarded in the current model implementation. Furthermore, the model was strictly feed-forward and modeled no cortico-cortical connectivity beyond the AC. Thus, for the model to grow towards a more "realistic" view of the information processing in the AC, we required deeper exploration of other information arriving in the AC (modulatory or driving influences, feed-forward and/or feedback sources) and how that information interacts with sound processing. Thus, the latter half of the thesis specifically focused on collecting datasets that may shed light on multisensory influences on auditory processing.

2 Visual Influences in the Auditory Cortex

Our environment is highly multisensory, and sounds are almost always accompanied by information from other senses. Recent studies show direct anatomical connections between the early auditory and visual cortices that are concentrated in visual sites representing the far peripheral field of view (Falchier et al., 2002, 2010). The behavioral and cortical correlates of this spatially specific connectivity have, to-date, remained unexplored in humans and were the focus of the research presented in Chapters 4 and 5.

To establish evidence of cross-sensory influences between peripherally-presented audiovisual stimuli, we employed a psychophysical approach in Chapter 4. The bidirectional audiovisual interactions were explored in the far periphery using simple stimuli (gratings and noise bursts; modulated and static) in a modulation detection task. We found evidence of multisensory influences of visual stimuli on auditory reaction times during the modulation detection task, but no reciprocal effects of audition on vision. By manipulating the congruency and the phase of the modulated stimuli (auditory and visual) at the onset of the stimuli, we found that the observed effects were highly sensitive to the temporal structure of the stimuli. That is, depending on temporal (in)congruence and synchrony between modulated audiovisual streams, the visual influences not only sped up (facilitation effects) the response times for auditory stimuli but also slowed them down (degradation effects). These results showed successful multisensory integration but painted a complex picture of underlying neuronal mechanisms, which could rely on direct communication between the early auditory and visual cortices but also influences from higher-order cortical sites.

The study presented in Chapter 5 was driven by the two key results reported in Chapter 4 i.e., the visual influence on audition with no reciprocal effects and, the sensitivity of these effects to the temporal relationship (phase, congruency) between the far-peripheral stimuli. To locate the cortical sites driving these cross-modal observations, we investigated the visual influences on the auditory cortex in a cortical depth-dependent manner using high-resolution functional MRI at 7 Tesla in Chapter 5. Due to the setup constraints of the MRI scanner, the stimuli could not be presented as peripherally as in Chapter 4. Thus, we first tested the spatial dependence of previous observations by repeating the modulation detection task measurements at a less peripheral location. We found evidence of a visual benefit for the auditory modulation detection thresholds even without a cross-sensory phase shift, while this shift was essential for observing a multisensory benefit at the more peripheral location. These results suggested that the exact nature of the audiovisual interactions varies with respect to the location of the stimuli, something that sets up precedence for future research beyond this thesis.

Driven by the task-dependence observed in the behavioral study, these multisensory interactions were explored in two different attention conditions with the hypothesis that by directing attention to the auditory stream, the multisensory effect would be enhanced in the auditory regions. The depth-dependent analysis of high-resolution fMRI data exploits the fact that neuronal populations at different cortical depths have distinct anatomical connectivity and properties. While the sensory input arrives at the middle layers, feedback signals shape predominantly the responses of deep and superficial layers (Felleman and Van Essen, 1991; Winer and Schreiner, 2011). These distinct columnar properties can be reflected in the independent responses across cortical depths and are

measurable because of the sub-millimeter resolution of the fMRI data (De Martino et al., 2015; Moerel et al., 2018; Gau et al., 2020). Our preliminary analyses showed a significant multisensory enhancement of responses across a cortical network including the primary and non-primary auditory cortex, the left primary and non-primary visual cortex (contralateral to stimulus location), bilateral insular cortex, and the ventrolateral prefrontal cortex. In the AC, the multisensory enhancement increased along the auditory cortical hierarchy and was strongest in the superior temporal gyrus, which likely reflects the auditory parabelt. While multisensory influences (enhancement and suppression) were present throughout layers of the AC, the multisensory enhancement was modulated by attention in the deep layers of the auditory belt. This effect was only observed when directing attention towards the auditory stimulus and was absent when the attention was diverted away from both stimulus streams. This modulatory effect of attention in deep layers, rather than middle layers, suggests that this context-dependent multisensory influence originates as a feedback signal. Where this feedback originates, remains to be determined. Some possible candidates could be tertiary auditory regions, visual cortex, or frontal regions. However, the tertiary auditory regions are not a likely source of the observed effects as short-range feedback more strongly targets the superficial than deep layers (Clavagnier et al., 2004). Future analyses will include multivariate pattern analysis to examine the multisensory effects in a multivariate fashion, and cortical depth-dependent connectivity analysis which may help discriminate between the frontal regions and visual cortex as sources of the observed context-dependent multisensory enhancement in deep layers of the auditory belt.

How do the observed cross-sensory influences on AC processing relate to the parallel processing streams of the AC explored in the first half of this thesis? The increased multisensory enhancement in the deep layers of belt areas when attention is directed to the auditory stimulus is of particular interest. The multisensory effect observed in Chapter 4 is driven by congruency in spatial location and temporal features of the auditory and visual stimuli. Our modeling approaches presented in Chapter 3 suggested that the caudal belt regions are optimized for capturing temporal sound dynamics. We, therefore, hypothesize that the attentional influence on multisensory processing targeted caudal instead of rostral belt regions. Moreover, beyond processing the temporal dynamics of auditory stimuli, we hypothesize that caudal belt regions may process the temporal dynamics of their multisensory counterparts as well. In line with these hypotheses, the direct projections from peripheral primary and secondary visual cortex have been shown to target caudal auditory regions (Falchier et al., 2010). These hypotheses may be in part addressed through further analysis of the dataset reported in Chapter 5. Electrophysiological experiments will also be required to fully test these hypotheses, as these measurements are needed in order to shed light on the temporal mechanisms of multisensory integration of peripheral audiovisual stimuli.

3 Future Perspectives

The research presented in this thesis showed how information across spatial and temporal scales, from neuron to behavior, can be integrated to better the understanding of the information processing in the AC. There are multiple avenues to pursue next. The primary focus of the proposed model was to merge the evidence for the role of neuronal dynamics with meso- and macroscale level observations. In future work, our experimental data described in Chapters 4 and 5 will allow extending the computational model of the AC to incorporate the influence of multisensory input. This extension will require additional model mechanisms. Specifically, as the observed multisensory effects were modulating rather than driving, modulatory cortico-cortical and/or thalamo-cortical connections will need to be added to the model. Furthermore, as the multisensory effects varied with cortical depth, a cortical depth-specific model of the AC is needed. To this end, the existing modeling approach can be extended with the neuronal models that capture the laminar-specific processing [e.g., Canonical Microcircuit model (Bastos et al., 2012)]. With such updates, the resulting "multisensory AC model" could be used for several purposes. A forward modeling approach, as used in Chapter 3, where the updated neuronal model can now be paired with laminar models of BOLD signals (Havlicek and Uludağ, 2020), would allow gaining insight into the neuronal dynamics (i.e., the microscale) of the mesoscale data collected in Chapter 5. As such, the model could be used to generate predictions regarding the spatial dependence of the observed effects across the visual field, test the role of cross-sensory temporal dynamics, and model attentional influences. These predictions will, however, be dependent on empirical data to be tested.

Based on data, the neuronal model may also be further optimized in the future. For example, model inversion informed by the high-resolution imaging data could help in further refining the current architecture and connectivity constraints of the model. Similarly, the proposed neuronal model could be used along with other measurement models, such as local field potential models, that simulate electrophysiological recordings and results could be used to refine the current temporal properties of the model. New data would then be required to test the validity and generalizability of these model modifications. Future electrophysiological measurements of the multisensory cortical responses could also be modeled in a multimodal dynamic causal modeling framework (Wei et al., 2020). Here an integrative approach could be taken to maximize the benefits of data from fMRI (spatial) and electrophysiological (temporal) modalities, inform the neuronal dynamics, and thereby improve the quality of model predictions.

Generating and optimizing an AC model that incorporates multisensory influences may prove beneficial for the understanding of aberrant multisensory integration. Diminished

multisensory integration has been hypothesized to underlie a number of psychiatric disorders, including autism and schizophrenia spectrum disorders (Stevenson et al., 2014; Zhou et al., 2018). Through non-invasive imaging (e.g., fMRI, or MEG data), mesoscale information of changes in brain processing with these disorders could be collected. By adapting model parameters to match this mesoscale information, conclusions could be drawn at other spatial scales. This may improve our understanding of the performance of patients on behavioral tests and neuroimaging studies (de Gelder et al., 1991; Surguladze et al., 2001; de Gelder et al., 2003; Smith and Bennetto, 2007; Stone et al., 2011; Brandwein et al., 2013; Stevenson et al., 2014), and allow constructing hypotheses on what underlies these disorders at the microscale level, possibly opening the road to intervention.

A model always provides a simplistic view of a complex system and therefore never captures that system to its full extent. This holds especially true for models of the brain, where the available data only cover a very small fraction – and often at a very limited temporal and spatial resolution–of the modeled system. However, our results have shown that despite their simplicity, computational models can serve a variety of purposes. By linking multiple scales of observations, the use of computational models ranges from hypotheses testing to the generation of new hypotheses, thereby improving our current understanding of the brain. When modeling and data-driven approaches are designed to complement each other, their collaborative advantages benefit the understanding of a system, as shown for auditory processing in this thesis. Thus, by developing data and models together, we avoid the risk of losing sight of the proverbial bigger picture.

Impact Statement

The ability to hear and interpret the sounds around us is not only necessary for survival but also enriches our life with interpersonal communication. In this thesis, we used computational and experimental methods to enhance our understanding of how the human brain processes sounds, and showed how the two approaches reinforce each other. We presented a computational model of the auditory cortex and used it to generate insight into the cortical processes that may underlie a range of experimental observations. The model predictions were used to generate hypotheses on auditory cortical processing as well, which can be tested in future experiments. However, the model is a simplification of a complex system and needs to evolve to better represent the auditory cortex. An avenue for the model to grow was explored by studying multisensory processing, and specifically the effects of visual input on auditory processing. Multisensory processing is important because our environment is full of information from different senses. This multisensory information guides our perception and behavior. In a behavioral study, we found an influence of what we see on what we hear, but not vice versa. We then explored the regions of the brain involved in the process. In the future, we plan to use this data to extend and improve the model of information processing in the auditory cortex. This can help elucidate the brain processes that underlie multisensory processing. As quite a few psychiatric and neurodevelopmental disorders, including schizophrenia and autism, are characterized by abnormalities in multisensory processing, this extended model may in the future also be used to characterize the neuronal sources of multisensory processing deficits.

Bibliography

Abeles, M., and Goldstein, M. H., Jr (1970). Functional architecture in cat primary auditory cortex: columnar organization and organization according to depth. *J. Neurophysiol.* 33(1), 172–187. doi: 10.1152/jn.1970.33.1.172

Aitkin, L. M., Fryman, S., Blake, D. W., and Webster, W. R. (1972). Responses of neurones in the rabbit inferior colliculus. I. Frequency-specificity and topographic arrangement. *Brain Res.* 47(1), 77–90. doi: 10.1016/0006-8993(72)90253-3

Andersen, R. A., Knight, P. L., and Merzenich, M. M. (1980). The thalamocortical and corticothalamic connections of AI, AII, and the anterior auditory field (AAF) in the cat: evidence for two largely segregated systems of connections. *J. Comp. Neurol.* 194(3), 663–701. doi: 10.1002/cne.901940312

Andersson, J. L., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*. 20(2), 870–888. doi: 10.1016/S1053-8119(03)00336-7

Arnott, S. R., Binns, M. A., Grady, C. L., and Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *NeuroImage*. 22(1), 401–408. doi: 10.1016/j.neuroimage.2004.01.014

Arrighi, R., Alais, D., and Burr, D. (2005). Neural latencies do not explain the auditory and audio-visual flash-lag effect. *Vision Res*. 45(23), 2917-25.

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., et al. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. *Neuron*. 97(3), 640–655.e4. doi: 10.1016/j.neuron.2017.12.034

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011a). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*. 54(3), 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., and Gee, J. C. (2011b). An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*. 9(4), 381–400. doi: 10.1007/s12021-011-9109-y

Bacon, S. P., and Viemeister, N. F. (1985). Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners. *Audiology*. 24(2), 117–134. doi: 10.3109/00206098509081545

Balaguer-Ballester E., Clark N. R., Coath M., Krumbholz K., and Denham S. L. (2009). Understanding Pitch Perception as a Hierarchical Process with Top-Down Modulation. *PLoS Comput. Biol.* 5(3), e1000301. doi: 10.1371/journal.pcbi.1000301

Barakat, B., Seitz, A.R., and Shams, L. (2015). Visual rhythm perception improves through auditory but not visual training. *Curr. Biol.* 25(2), R60-R61. doi: 10.1016/j.cub.2014.12.011

Bartlett, E. L., Sadagopan, S., and Wang, X. (2011). Fine frequency tuning in monkey auditory cortex and thalamus. *J. Neurophysiol.* 106(2), 849–859. doi: 10.1152/jn.00559.2010

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*. 76(4), 695–711. doi: 10.1016/j.neuron.2012.10.038

Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*. 41(5), 809–823. doi: 10.1016/s0896-6273(04)00070-4

Belin, P., and Zatorre, R. J. (2000). 'What', 'where' and 'how' in auditory cortex. *Nat. Neurosci.* 3(10), 965–966. doi: 10.1038/79890

Bendor D. (2015). The role of inhibition in a computational model of an auditory cortical neuron during the encoding of temporal information. *PLoS Comput. Biol.* 11(4), e1004197. doi: 10.1371/journal.pcbi.1004197

Bendor, D., Osmanski, M. S., and Wang, X. (2012). Dual-pitch processing mechanisms in primate auditory cortex. *J. Neurosci.* 32(46), 16149–16161. doi: 10.1523/JNEUROSCI.2563-12.2012

Bendor, D., and Wang, X. (2008). Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *J. Neurophysiol.* 100(2), 888–906. doi: 10.1152/jn.00884.2007

Benson, N. C., Butt, O. H., Brainard, D. H., and Aguirre, G. K. (2014). Correction of distortion in flattened representations of the cortical surface allows prediction of V1-V3 functional organization from anatomy. *PLoS Comput. Biol.* 10(3), e1003538. doi: 10.1371/journal.pcbi.1003538

Besle, J., Bertrand, O., and Giard, M.H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hear. Res*. 258(1-2), 143-51. doi: 10.1016/j.heares.2009.06.016

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., et al. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *J. Neurosci.* 28(52), 14301–14310. doi: 10.1523/JNEUROSCI.2875-08.2008

Bieser, A., and Müller-Preuss, P. (1996). Auditory responsive cortex in the squirrel monkey: Neural responses to amplitude-modulated sounds. *Exp. Brain Res*. 108(2), 273–284.

Bizley, J. K., and King, A. J. (2009). Visual influences on ferret auditory cortex. *Hear. Res.* 258(1-2), 55–63. doi: 10.1016/j.heares.2009.06.017

Bizley, J.K., Maddox, R.K., and Lee, A.K.C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. *Trends Neurosci.* 39(2), 74-85. doi: 10.1016/j.tins.2015.12.007

Bolognini, N., Senna, I., Maravita, A., Pascual-Leone, A., and Merabet, L.B. (2010). Auditory enhancement of visual phosphene perception: The effect of temporal and spatial factors and of stimulus intensity. *Neurosci. Lett.* 477(3), 109-14. doi: 10.1016/j.neulet.2010.04.044

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spat. Vis*. 10(4), 433–6.

Brandwein, A. B., Foxe, J. J., Butler, J. S., Russo, N. N., Altschuler, T. S., Gomes, H., et al. (2013). The development of multisensory integration in high-functioning autism: high-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. *Cereb. Cortex*. 23(6), 1329–1341. doi: 10.1093/cercor/bhs109

Brodland G. W. (2015). How computational models can help unlock biological systems. *Semin. Cell Dev. Bio.* 47-48, 62–73. doi: 10.1016/j.semcdb.2015.07.001

Brown, R. R., Deletic, A., and Wong, T. H. (2015). Interdisciplinarity: How to catalyse collaboration. *Nature*. 525(7569), 315–317. doi: 10.1038/525315a

Buxton, R. B., Uludağ K., Dubowitz D. J., and Liu T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*. Suppl 1, S220-33. doi: 10.1016/j.neuroimage.2004.07.013

Buxton, R. B., Wong ,E. C., and Frank, L. R. (1998). Modeling the hemodynamic response to brain activation. *Magn. Reson. Med*. 39(6), 855-64. doi: 10.1002/mrm.1910390602

Calford, M. B., and Aitkin, L. M. (1983). Ascending projections to the medial geniculate body of the cat: evidence for multiple, parallel auditory pathways through thalamus. *J. Neurosci. 3*(11), 2365–2380. doi: 10.1523/JNEUROSCI.03-11-02365.1983

Calvert G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex*. 11(12), 1110–1123. doi: 10.1093/cercor/11.12.1110

Calvert, G. A., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* 15(1), 57–70. doi: 10.1162/089892903321107828

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol*. 10(11), 649–657. doi: 10.1016/s0960-9822(00)00513-3

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol.* 98(1-3), 191–205.doi: 10.1016/j.jphysparis.2004.03.018

Camalier, C. R., D'Angelo, W. R., Sterbing-D'Angelo, S. J., de la Mothe, L. A., and Hackett, T. A. (2012). Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. *Proc. Natl. Acad. Sci. U. S. A.* 109(44), 18168–18173. doi: 10.1073/pnas.1206387109

Caplette, L. (2020). Simple RM/Mixed ANOVA for any design. MATLAB Central File Exchange. Available from: https://www.mathworks.com/matlabcentral/fileexchange/64980-simple-rm-mixed-anova-for-any-design

Cappe, C., Rouiller, E. M., and Barone, P. (2009). Multisensory anatomical pathways. *Hear. Res.* 258(1-2), 28–36. doi: 10.1016/j.heares.2009.04.017

Cate, A. D., Herron, T. J., Yund, E. W., Stecker, G. C., Rinne, T., et al. (2009). Auditory attention activates peripheral visual cortex. *PLoS One*. 4(2), e4645. doi: 10.1371/journal.pone.0004645

Chambers, J. D., Elgueda, D., Fritz, J. B., Shamma, S. A., Burkitt, A. N., et al. (2019). Computational Neural Modeling of Auditory Cortical Receptive Fields. *Front. Comput. Neurosci*. 13, 28. doi: 10.3389/fncom.2019.00028

Charbonneau, G., Veronneau, M., Boudrias-Fournier, C., Lepore, F., and Collignon, O. (2013). The ventriloquist in periphery: Impact of eccentricity-related reliability on audio-visual localization. *J. Vis.* 13(12), 20. doi: 10.1167/13.12.20

Chen, Y. C., Maurer, D., Lewis, T. L., Spence, C., and Shore, D. I. (2017). Central–peripheral differences in audiovisual and visuotactile event perception. *Atten. Percept. Psychophys*. 79(8), 2552-2563. doi: 10.3758/s13414-017-1396-4

Chen, L., and Vroomen, J. (2013). Intersensory binding across space and time: A tutorial review. *Atten. Percept. Psychophys*. 75(5), 790-811. doi: 10.3758/s13414-013-0475-4

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am*. 118(2), 887–906.

Chow, H.M., Leviyah, X., and Ciaramitaro, V.M. (2020). Individual Differences in Multisensory Interactions:The Influence of Temporal Phase Coherence and Auditory Salience on Visual Contrast Sensitivity. *Vision.* 4(1), 12. doi: 10.3390/vision4010012

Chrostowski, M., Yang, L., Wilson, H. R., Bruce, I. C., and Becker, S. (2011). Can homeostatic plasticity in deafferented primary auditory cortex lead to travelling waves of excitation?. *J. Comput. Neurosci*. 30(2), 279–299. doi: 10.1007/s10827-010-0256-1

Clavagnier, S., Falchier, A., and Kennedy, H. (2004). Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cogn. Affect. Behav. Neurosci.* 4(2), 117–126. doi: 10.3758/cabn.4.2.117

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., and Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nat. Commun*. 7, 11070. doi: 10.1038/ncomms11070

Covic, A., Keitel, C., Porcu, E., Schröger, E., and Müller, M.M. (2017). Audio-visual Synchrony and Spatial Attention Enhance Processing of Dynamic Visual Stimulation Independently and in Parallel: A Frequency-Tagging Study. *NeuroImage.* 161, 32-42. doi: 10.1016/j.neuroimage.2017.08.022

Cowan, J. D., Neuman, J., and van Drongelen, W. (2016). Wilson–Cowan Equations for Neocortical Dynamics. *J. Math. Neurosci*. 6(1), 1. doi: 10.1186/s13408-015-0034-5

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*. *9*(2), 179–194. doi: 10.1006/nimg.1998.0395

Dale, A. M., and Sereno, M. I. (1993). Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *J. Cogn. Neurosci.* 5(2), 162–176. doi: 10.1162/jocn.1993.5.2.162

De Angelis, V., De Martino, F., Moerel, M., Santoro, R., Hausfeld, L., and Formisano, E. (2018). Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. *NeuroImage*. 180 (Pt A), 291–300. doi: 10.1016/j.neuroimage.2017.11.020

de Cheveigné, A., and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am*. 111(4), 1917–1930.

de Gelder, B., Vroomen, J., Annen, L., Masthof, E., and Hodiamont, P. (2003). Audio-visual integration in schizophrenia. *Schizophr. Res.* 59(2-3), 211–218. doi: 10.1016/s0920-9964(01)00344-9

de Gelder, B., Vroomen, J. H. M., and Van der Heide, L. (1991). Face recognition and lip-reading in autism. *Eu. J. Cogn. Psychol.* 3(1), 69-86

de la Mothe, L. A. (2016). "Evolution of Auditory Cortex in Primates," in *Evolution of Nervous Systems,* ed. J. Kaas (Academic Press)*,* 331-342. doi:10.1016/B978-0-12-804042-3.00088-9

De Martino, F., Moerel, M., Ugurbil, K., Goebel, R., Yacoub, E., et al. (2015). Frequency preference and attention effects across cortical depths in the human primary auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* 112(52), 16036–16041. doi: 10.1073/pnas.1507552112

De Martino, F., Zimmermann, J., Muckli, L., Ugurbil, K., Yacoub, E., and Goebel, R. (2013). Cortical depth dependent functional responses in humans at 7T: improved specificity with 3D GRASE. *PLoS One*. 8(3), e60514. doi: 10.1371/journal.pone.0060514

de Pinho, M., Mazza, M., and Roque, A. C. (2006). A computational model of the primary auditory cortex exhibiting plasticity in the frequency representation. *Neurocomputing.* 70. doi:10.1016/j.neucom.2006.07.004

Denison, R.N., Driver, J., and Ruff, C.C. (2013). Temporal structure and complexity affect audio-visual correspondence detection. *Front. Psychol.* 3, 619. doi: 10.3389/fpsyg.2012.00619

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*. 31(3), 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Di Russo, F., Martínez, A., Sereno, M.I., Pitzalis, S., and Hillyard, S.A. (2002). Cortical sources of the early components of the visual evoked potential. *Hum. Brain Mapp*. 15(2), 95–111

Doesburg, S.M., Emberson, L.L., Rahi, A., Cameron, D., and Ward, L.M. (2008). Asynchrony from synchrony: Long-range gamma-band neural synchrony accompanies perception of audiovisual speech asynchrony. *Exp. Brain Res*. 185(1), 11–20.

Driver, J., and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron*. 57(1), 11–23. doi: 10.1016/j.neuron.2007.12.013

Eckert, M.A., Kamdar, N.V., Chang, C.E., Beckmann, C.F., Greicius, M.D., and Menon, V. (2008). A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis. *Hum. Brain Mapp*. 29(7), 848-57. doi: 10.1002/hbm.20560

Eggermont, J. J. (1991). Rate and synchronization measures of periodicity coding in cat primary auditory cortex. *Hear. Res*. 56(1–2), 153–167.

Eggermont, J. J. (1998). Representation of spectral and temporal sound features in three cortical fields of the cat. Similarities outweigh differences. J. Neurophysiol., 80(5), 2743–2764.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*. 61(2), 317–329. doi: 10.1016/j.neuron.2008.12.005

Ermentrout, G. B., and Cowan, J. D. (1979). A mathematical theory of visual hallucination patterns. *Biol. Cybern*. 34(3), 137–150.

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *J. Neurosci*. 22(13), 5749–5759. doi: 10.1523/JNEUROSCI.22-13-05749.2002

Falchier, A., Schroeder, C. E., Hackett, T. A., Lakatos, P., Nascimento-Silva, S., et al. (2010). Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey. *Cereb. Cortex*. 20(7), 1529–1538. doi: 10.1093/cercor/bhp213

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex.* 1(1), 1–47. doi: 10.1093/cercor/1.1.1

Fischl, B., and Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci. U. S. A*. 97(20), 11050–11055. doi: 10.1073/pnas.200033797

Fischl, B., Liu, A., and Dale, A. M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans. Med. Imaging*. 20(1), 70–80. doi: 10.1109/42.906426

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. 33(3), 341–355. doi: 10.1016/s0896-6273(02)00569-x

Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., Ségonne, F., Quinn, B. T., and Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*. Suppl 1, S69–S84. doi: 10.1016/j.neuroimage.2004.07.016

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., et al. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*. 40(4), 859–869. doi: 10.1016/s0896-6273(03)00669-x

Frackowiak, R., and Markram, H. (2015). The future of human cerebral cartography: a novel approach. *Philos. Trans. R. Soc. B Biol. Sci.* 370(1668), 20140171. doi: 10.1098/rstb.2014.0171

Frassinetti, F., Bolognini, N., and Làdavas, E. (2002). Enhancement of Visual Perception by Crossmodal Visuo-Auditory Interaction. *Exp. Brain Res.* 147(3), 332-43.

Frässle, S., Lomakina, E. I., Razi, A., Friston, K. J., Buhmann, J. M., and Stephan K. E. (2017). Regression DCM for fMRI. *NeuroImage*. 155, 406-421. doi: 10.1016/j.neuroimage.2017.02.090

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*. 19(4), 1273-302. doi: 10.1016/s1053-8119(03)00202-7

Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. (2000). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*. 12(4), 466-77. doi: 10.1006/nimg.2000.0630

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese*. 159(3), 417-458. doi: 10.1007/s11229-007-9237-y

Gaffan, D., and Harrison, S. (1991). Auditory-visual associations, hemispheric specialization and temporal-frontal interaction in the rhesus monkey. *Brain.* 114 *(Pt 5)*, 2133–2144. doi: 10.1093/brain/114.5.2133

Galaburda, A., and Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. *J. Comp. Neurol*. 190(3), 597–610. doi: 10.1002/cne.901900312

Garcia, D., Hall, D. A., and Plack, C. J. (2010). The effect of stimulus context on pitch representations in the human auditory cortex. *NeuroImage*. 51(2), 808–816. doi: 10.1016/j.neuroimage.2010.02.079

Gardumi, A., Ivanov, D., Havlicek, M., Formisano, E., and Uludağ, K. (2017). Tonotopic maps in human auditory cortex using arterial spin labeling. *Hum Brain Mapp*. 38(3), 1140–1154. doi: 10.1002/hbm.23444

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, et al. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium.

Gau, R., Bazin, P. L., Trampel, R., Turner, R., and Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. *eLife*. 9, e46856. doi: 10.7554/eLife.46856

Ghazanfar, A.A., and Schroeder, C.E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci*. 10(6), 278–85.

Giraud, A. L., and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat. Neurosci*. 15(4), 511–517. doi: 10.1038/nn.3063

Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res*. 47(1–2), 103–138.

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum. Brain Mapp*, 27(5), 392–401. doi: 10.1002/hbm.20249

Gogolla N. (2017). The insular cortex. *Curr. Biol*. 27(12), R580–R586. doi: 10.1016/j.cub.2017.05.010

Goldberg, J. M., and Brown, P. B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. J. *Neurophysiol*. 32(4), 613–636.

Grant S. G. (2003). Systems biology in neuroscience: bridging genes to cognition. *Curr. Opin. Neurobiol.* 13(5), 577–582. doi: 10.1016/j.conb.2003.09.016

Griffiths, T. D., and Hall, D. A. (2012). Mapping Pitch Representation in Neural Ensembles with fMRI. *J. Neurosci*. 32(39), 13343–13347.

Grubb, R. L., Jr, Raichle, M. E., Eichling, J. O., and Ter-Pogossian, M. M. (1974). The effects of changes in PaCO2 on cerebral blood volume, blood flow, and vascular mean transit time. *Stroke*. 5(5), 630–639. doi: 10.1161/01.str.5.5.630

Hackett, T. A., Barkat, T. R., O'Brien, B. M., Hensch, T. K., and Polley, D. B. (2011). Linking topography to tonotopy in the mouse auditory thalamocortical circuit. *J. Neurosci*. 31(8), 2983–2995. doi: 10.1523/JNEUROSCI.5333-10.2011

Hackett, T. A., Smiley, J. F., Ulbert, I., Karmos, G., Lakatos, P., et al. (2007). Sources of somatosensory input to the caudal belt areas of auditory cortex. *Perception*. 36(10), 1419–1430. doi: 10.1068/p5841

Hackett, T. A., Stepniewska, I., and Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *J. Comp. Neurol.* 394(4), 475–495. doi: 10.1002/(sici)1096-9861(19980518)394:4<475::aid-cne6>3.0.co;2-z

Hall, D. A., and Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cereb. Cortex*. 19(3), 576–585. doi: 10.1093/cercor/bhn108

Hartmann, E., Lachenmayr, B., and Brettel, H. (1979). The peripheral critical flicker frequency. *Vision Res*. 19(9), 1019–23.

Havlicek, M., Ivanov, D., Roebroeck, A., and Uludağ, K. (2017). Determining excitatory and inhibitory neuronal activity from multimodal fMRI data using a generative hemodynamic model. *Front. Neurosci*. 11, 616. doi: 10.3389/fnins.2017.00616

Havlicek, M., Roebroeck, A., Friston, K., Gardumi, A., Ivanov, D., et al. (2015). Physiologically informed dynamic causal modeling of fMRI data. *NeuroImage*. 122, 355–372. doi: 10.1016/j.neuroimage.2015.07.078

Havlicek, M., and Uludağ, K. (2020). A dynamical model of the laminar BOLD response. *NeuroImage*. 204, 116209. doi: 10.1016/j.neuroimage.2019.116209

Hedayat, A., and Kirk, R.E. (2006). Experimental Design: Procedures for the Behavioral Sciences. *Biometrics*. 26(3), 590.

Heil, P., and Irvine, D. R. F. (2017). First-Spike Timing of Auditory-Nerve Fibers and Comparison With Auditory Cortex. *J. Neurophysiol*. 78(5), 2438–2454.

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol*. 117(4), 500–544. doi: 10.1113/jphysiol.1952.sp004764

Hou, Z., Huang, S., Hu, Q., and Nowinski, W. L. (2006). A fast and automatic method to correct intensity inhomogeneity in MR brain images. *Med. Image Comput. Comput. Assist. Interv*. 9(Pt 2), 324–331. doi: 10.1007/11866763_40

Houtsma, A. J., and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *J. Acoust. Soc. Am*. 87(1), 304–310.

Huang C., and Rinzel J. (2016). A Neuronal Network Model for Pitch Selectivity and Representation. *Front Comput Neurosci*. 10, 57. doi: 10.3389/fncom.2016.00057

Ibrahim, L.A., Mesik, L., Ji, X.Y., Fang, Q., Li, H.F., Li, Y.T., et al. (2016). Cross-Modality Sharpening of Visual Cortical Processing through Layer-1-Mediated Inhibition and Disinhibition. *Neuron*. 89(5), 1031-45. doi: 10.1016/j.neuron.2016.01.027.

Jain, A., Bansal, R., Kumar, A., and Singh, K.D. (2015). A Comparative Study of Visual and Auditory Reaction Times on the Basis of Gender and Physical Activity Levels of Medical First Year Students. *Int. J. Appl. Basic Med. Res*. 5(2), 124-7. doi: 10.4103/2229-516X.157168

James, T. W., and Stevenson, R. A. (2012). "The Use of fMRI to Assess Multisensory Integration," in *The Neural Bases of Multisensory Processes*, ed M. M. Murray (CRC Press/Taylor and Francis). doi:10.1201/b11092-11

Jasmin, K., Lima, C. F., and Scott, S. K. (2019). Understanding rostral-caudal auditory cortex contributions to auditory perception. *Nat. Rev. Neurosci*. 20(7), 425–434. doi: 10.1038/s41583-019-0160-2

Jones E. G. (2000). Microcolumns in the cerebral cortex. *Proc. Natl. Acad. Sci. U. S. A*. 97(10), 5019–5021. doi: 10.1073/pnas.97.10.5019

Jones, E. G., and Powell, T. P. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain.* 93(4), 793–820. doi: 10.1093/brain/93.4.793

Joris, P. X., Schriener, C. E., and Rees, A. (2004). Neural Processing of Amplitude-Modulated Sounds. *Physiol. Rev.* 84(2), 541–577.

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U. S. A.* 97(22), 11793–11799. doi: 10.1073/pnas.97.22.11793

Kaas, J. H., Hackett, T. A., and Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Curr. Opin. Neurobiol.* 9(2), 164–170. doi: 10.1016/s0959-4388(99)80022-1

Kass, R. E., and Adrian, E. Raftery. (1995). Bayes Factors. *J. Am. Stat. Assoc*. 90(430), 773-795, doi: 10.1080/01621459.1995.10476572

Kayser, S. J., and Kayser, C. (2018). Trial by trial dependencies in multisensory perception and their correlates in dynamic brain activity. *Sc,i Rep.* 8(1), 3742. doi: 10.1038/s41598-018-22137-8.

Kayser, C., Logothetis, N. K., and Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Curr. Biol.* 20(1), 19–24. doi: 10.1016/j.cub.2009.10.068

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *J. Neurosci.* 27(8), 1824–1835. doi: 10.1523/JNEUROSCI.4737-06.2007

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cereb. Cortex.* 18(7), 1560–1574. doi: 10.1093/cercor/bhm187

Kiebel, S. J., David, O., and Friston, K. J. (2006). Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization. *NeuroImage*. 30(4), 1273–1284. doi: 10.1016/j.neuroimage.2005.12.055

Kim, J. J., Crespo-Facorro, B., Andreasen, N. C., O'Leary, D. S., Zhang, B., Harris, G., and Magnotta, V. A. (2000). An MRI-based parcellation method for the temporal lobe. *NeuroImage*. 11(4), 271–288. doi: 10.1006/nimg.2000.0543

Kitano H. (2002). Systems biology: a brief overview. *Science.* 295(5560), 1662–1664. doi: 10.1126/science.1069492

Kleiner, M., Brainard, D.H., and Pelli, D.G. (2007). What's new in Psychtoobox-3? *Perception*. 36(14), 1.

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychol.* 134(3), 372–384. doi: 10.1016/j.actpsy.2010.03.010

Kohlrausch, A., Fassel, R., and Dau, T. (2000). The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *J. Acoust. Soc. Am.* 108(2), 723734. doi: 10.1121/1.429605

Kuramoto Y. 1984. Chemical oscillations, waves and turbulence. New York, NY: Springer.

Kuśmierek, P., and Rauschecker, J. P. (2009). Functional specialization of medial auditory belt cortex in the alert rhesus monkey. *J. Neurophysiol.* 102(3), 1606–1622. doi: 10.1152/jn.00167.2009

Kuśmierek, P., and Rauschecker, J. P. (2014). Selectivity for space and time in early areas of the auditory dorsal stream in the rhesus monkey. *J. Neurophysiol*. 111(8), 1671–1685. doi: 10.1152/jn.00436.2013

Laing, M., Rees, A., and Vuong, Q. C. (2015). Amplitude-modulated stimuli reveal auditory-visual interactions in brain activity and brain connectivity. *Front. Psychol*. 6, 1440. doi: 10.3389/fpsyg.2015.01440

Lakatos, P., Chen, C.M., O'Connell, M.N., Mills, A., and Schroeder, C.E. (2007). Neuronal Oscillations and Multisensory Interaction in Primary Auditory Cortex. *Neuron*. 53(2), 279–92.

Ledford H. (2015). How to solve the world's biggest problems. *Nature*. 525(7569), 308–311. doi: 10.1038/525308a

Lewis, J. W., Beauchamp, M. S., and DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. *Cereb. Cortex*. 10(9), 873–888. doi: 10.1093/cercor/10.9.873

Lewis, J. W., and Van Essen, D. C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *J. Comp. Neurol.* 428(1), 112–137. doi: 10.1002/1096-9861(20001204)428:1<112::aid-cne8>3.0.co;2-9

Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods*. 264, 47–56. doi: 10.1016/j.jneumeth.2016.03.001

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(5), 2237–2261.

Lindeberg, T., and Friberg, A. (2015). Idealized computational models for auditory receptive fields. *PLoS One*. 10(3), e0119032. doi: 10.1371/journal.pone.0119032

Loebel, A., Nelken, I., and Tsodyks, M. (2007). Processing of sounds by population spikes in a model of primary auditory cortex. *Front. Neurosci.* 1(1), 197–209. doi: 10.3389/neuro.01.1.1.015.2007

Logothetis, N. K., Guggenberger, H., Peled, S., and Pauls, J. (1999). Functional imaging of the monkey brain. *Nat. Neurosci.* 2(6), 555–562. doi: 10.1038/9210

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*. 412(6843), 150–157. doi: 10.1038/35084005

Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat. Neurosci*. 4(11), 1131–1138.

Ma, N., Green, P., Barker, J., and Coy, A. (2007). Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Commun*. 49(12), 874–891. doi: 10.1016/j.specom.2007.05.003

May, P. J. C., Westö, J., and Tiitinen, H. (2015). Computational modelling suggests that temporal integration results from synaptic adaptation in auditory cortex. *Eur. J. Neurosci*. 41(5), 615–630. doi: 10.1111/ejn.12820

McGurk, H., and MacDonald, J. (1976). Hearing Lips and Seeing Voices. *Nature*. 264(5588), 746-8. doi: 10.1038/264746a0

Meddis, R., Lecluyse, W., Clark, N. R., Jürgens, T., Tan, C. M., Panda, M. R., et al. (2013). A computer model of the auditory periphery and its application to the study of hearing. *Adv. Exp. Med. Biol.* 787, 11–20. doi: 10.1007/978-1-4614-1590-9_2

Meijer, G. T., Montijn, J. S., Pennartz, C., and Lansink, C. S. (2017). Audiovisual Modulation in Mouse Primary Visual Cortex Depends on Cross-Modal Stimulus Configuration and Congruency. *J. Neurosci.* 37(36), 8783–8796. doi: 10.1523/JNEUROSCI.0468-17.2017

Meredith, M.A., and Stein, B.E. (1983). Interactions Among Converging Sensory Inputs in the Superior Colliculus. *Science.* 221(4608), 389-91. doi: 10.1126/science.6867718

Merzenich, M. M., and Brugge, J. F. (1973). Representation of the cochlear partition of the superior temporal plane of the macaque monkey. *Brain Res*. 50(2), 275–296. doi: 10.1016/0006-8993(73)90731-2

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci*. 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* 32(41), 14205–14216. doi: 10.1523/JNEUROSCI.1388-12.2012

Moerel, M., De Martino, F., Uğurbil, K., Formisano, E., and Yacoub, E. (2018). Evaluating the Columnar Stability of Acoustic Processing in the Human Auditory Cortex. *J. Neurosci*. 38(36), 7822–7832. doi: 10.1523/JNEUROSCI.3576-17.2018

Moore, B. C. (2003). An Introduction to the Psychology of Hearing. Bost. Acad. Press.

Moore, B. C., and Glasberg, B. R. (2001). Temporal modulation transfer functions obtained using sinusoidal carriers with normally hearing and hearing-impaired listeners. *J. Acoust. Soc. Am*. 110(2), 1067–1073.

Moran, R., Pinotsis, D. A., and Friston, K. (2013). Neural masses and fields in dynamic causal modeling. *Front. Comput. Neurosci*. 7, 57. doi: 10.3389/fncom.2013.00057

Morrill, R. J., and Hasenstaub, A. R. (2018). Visual Information Present in Infragranular Layers of Mouse Auditory Cortex. *J. Neurosci.* 38(11), 2854–2862. doi: 10.1523/JNEUROSCI.3102-17.2018

Murray, M. M., Lewkowicz, D. J., Amedi, A., and Wallace, M. T. (2016). Multisensory Processes: A Balancing Act across the Lifespan. *Trends Neurosci.* 39(8), 567–579. doi: 10.1016/j.tins.2016.05.003

Ng, A. Y. W., and Chan, A. H. S. (2012). Finger Response Times to Visual, Auditory and Tactile Modality Stimuli. *Proc. Intl. Multi. Conf. Engineers Computer Scientists*. 2:1449–1454.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27(42), 11431–11441. doi: 10.1523/JNEUROSCI.2252-07.2007

Norman-Haignere, S., Kanwisher, N., and McDermott, J. H. (2013). Cortical Pitch Regions in Humans Respond Primarily to Resolved Harmonics and Are Located in Specific Tonotopic Regions of Anterior Auditory Cortex. *J. Neurosci*. 33(50), 19451–19469. doi: 10.1523/JNEUROSCI.2880-13.2013

Nourski, K. V., Brugge, J. F., Reale, R. A., Kovach, C. K., Oya, H., et al. (2013). Coding of repetitive transients by auditory cortex on posterolateral superior temporal gyrus in humans: An intracranial electrophysiology study. *J. Neurophysiol*. 109(5), 1283–1295. doi: 10.1152/jn.00718.2012

Nourski, K. V., Steinschneider, M., McMurray, B., Kovach, C. K., Oya, H., et al. (2014). Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *NeuroImage.* 101, 598–609. doi: 10.1016/j.neuroimage.2014.07.004

Odegaard, B., and Shams, L. (2016). The Brain's Tendency to Bind Audiovisual Signals Is Stable but Not General. *Psychol. Sci.* 27(4), 583-91. doi: 10.1177/0956797616628860.

Odegaard, B., Wozny, D.R., and Shams, L. (2015). Biases in Visual, Auditory, and Audiovisual Perception of Space. *PLoS Comput Biol*. 11(12), e1004649. doi: 10.1371/journal.pcbi.1004649

Odegaard, B., Wozny, D.R., and Shams, L. (2016). The effects of selective and divided attention on sensory precision and integration. *Neurosci. Lett*. 614:24-8. doi: 10.1016/j.neulet.2015.12.039

Oshurkova, E., Scheich, H., and Brosch, M. (2008). Click train encoding in primary and non-primary auditory cortex of anesthetized macaque monkeys. *J. Neuroscience*, 153(4), 1289–1299. doi: 10.1016/j.neuroscience.2008.03.030

Overath, T., McDermott, J. H., Zarate, J. M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci*. 18(6), 903–911. doi: 10.1038/nn.4021

Oxenham, A. J. (2012). Pitch Perception. *J. Neurosci*. 32(39), 13335–13338.

Patterson, R. D. (1986). Auditory filters and excitation patterns as representations of frequency resolution. *Freq. Sel. Hear*. 123–177.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex Sounds and Auditory Images. *Proc. 9th Int. Symp. Hearing Audit., Physiol. Percept*. 429–446.

Pelli D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat. Vis*. 10(4), 437–42.

Penny W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*. 59(1), 319–330. doi: 10.1016/j.neuroimage.2011.07.039

Penny, W., Friston, K., Ashburner, J., Kiebel, S., and Nichols, T. (2007). Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press. doi: 10.1016/B978-0-12-372560-8.X5000-1

Perrodin, C., Kayser, C., Logothetis, N. K., and Petkov, C. I. (2014). Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. *J. Neurosci.* 34(7), 2524–2537. doi: 10.1523/JNEUROSCI.2805-13.2014

Plack, C. J., Barker, D., and Hall, D. A. (2014). Pitch coding and pitch processing in the human brain. *Hear. Res*. 307, 53–64. doi: 10.1016/j.heares.2013.07.020

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* 97(22), 11800–11806. doi: 10.1073/pnas.97.22.11800

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science.* 268(5207), 111–114. doi: 10.1126/science.7701330

Rauschecker, J. P., Tian, B., Pons, T., and Mishkin, M. (1997). Serial and parallel processing in rhesus monkey auditory cortex. *J. Comp. Neurol*. 382(1), 89–103.

Read, H. L., Winer, J. A., and Schreiner, C. E. (2002). Functional architecture of auditory cortex. *Curr. Opin. Neurobiol*. 12(4), 433–440.

Recanzone, G. H., Guard, D. C., and Phan, M. L. (2000). Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. *J.Neurophysiol.* 83(4), 2315–2331. doi: 10.1152/jn.2000.83.4.2315

Rees, G., Friston, K., and Koch, C. (2000). A direct quantitative relationship between the functional properties of human and macaque V5. *Nat. Neurosci*. 3(7), 716–723. doi: 10.1038/76673

Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., and Rauschecker, J. P. (2009). Multisensory integration of sounds and vibrotactile stimuli in processing streams for "what" and "where." *J. Neurosci.* 29(35), 10950–10960. doi: 10.1523/JNEUROSCI.0910-09.2009

Rivier, F., and Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. *NeuroImage*. 6(4), 288–304. doi: 10.1006/nimg.1997.0304

Rockland, K.S., and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *Int. J. Psychophysiol*. 50(1–2), 19–26.

Romanski, L. M., Bates, J. F., and Goldman-Rakic, P. S. (1999a). Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *J. Comp. Neurol.* 403(2), 141–157. doi: 10.1002/(sici)1096-9861(19990111)403:2<141::aid-cne1>3.0.co;2-v

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., et al. (1999b). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci*. 2(12), 1131–1136. doi: 10.1038/16056

Rouiller, E. M., Simm, G. M., Villa, A. E., de Ribaupierre, Y., and de Ribaupierre, F. (1991). Auditory corticocortical interconnections in the cat: evidence for parallel and hierarchical arrangement of the auditory cortical areas. *Exp. Brain Res*. 86(3), 483–505. doi: 10.1007/BF00230523

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* 10(1), e1003412. doi: 10.1371/journal.pcbi.1003412

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., et al. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci. U. S. A*. 114(18), 4799–4804. doi: 10.1073/pnas.1617622114

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(34), 14611–14616. doi: 10.1073/pnas.0907682106

Schreiner, C. E., Christoph E., and Urbas, J. V. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hear. Res*. 32(1), 49–63.

Schroeder, C. E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci*. 32(1), 9–18. doi: 10.1016/j.tins.2008.09.012

Schreiner, C. E., Read, H. L., and Sutter, M. L. (2000). Modular organization of frequency integration in primary auditory cortex. *Annu. Rev. Neurosci*. 23, 501–529.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci*. 12(3), 106-13. doi: 10.1016/j.tics.2008.01.002

Sclar G., Maunsell J. H., Lennie P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res*. 30(1), 1-10.

Scott, B. H., Leccese, P. A., Saleem, K. S., Kikuchi, Y., Mullarkey, M. P., et al. (2017). Intrinsic Connections of the Core Auditory Cortical Regions and Rostral Supratemporal Plane in the Macaque Monkey. *Cereb. Cortex*. 27(1), 809–840. doi: 10.1093/cercor/bhv277

Scott, B. H., Malone, B. J., and Semple, M. N. (2011). Transformation of temporal processing across auditory cortex of awake macaques. *J. Neurophysiol*. 105(2), 712–730. doi: 10.1152/jn.01120.2009

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., and Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*. 22(3), 1060–1075. doi: 10.1016/j.neuroimage.2004.03.032

Ségonne, F., Pacheco, J., and Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans. Med. Imaging*. 26(4), 518–529. doi: 10.1109/TMI.2006.887364

Sek, A., and Moore, B. C. (1995). Frequency discrimination as a function of frequency, measured in several ways. *J. Acoust. Soc. Am*. 97(4), 2479–2486.

Şentürk, G., Greenberg, A. S., and Liu, T. (2016). Saccade latency indexes exogenous and endogenous object-based attention. *Atten. Percept. Psychophys.* 78(7), 1998–2013. doi: 10.3758/s13414-016-1136-1

Shamma, S. A., Fleshman, J. W., Wiser, P. R., and Versnel, H. (1993). Organization of response areas in ferret primary auditory cortex. *J. Neurophysiol.* 69(2), 367–383. doi: 10.1152/jn.1993.69.2.367

Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature*. 408(6814), 788.

Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Brain Res. Cogn. Brain Res.* 14(1), 147-52.

Shams, L., Kamitani, Y., and Shimojo, S. (2010). Sound modulates visual evoked potentials in humans. *J. Vis*. 1(3), 479–479. doi: 10.1167/1.3.479

Shams, L., and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends Cogn. Sci.* 12(11), 411–417. doi: 10.1016/j.tics.2008.07.006

Shams, L., Wozny, D. R., Kim, R., and Seitz, A. (2011). Influences of multisensory experience on subsequent unisensory processing. *Front. Psychol*. 2, 264. doi: 10.3389/fpsyg.2011.00264

Shaw, E.A. (1974). Transformation of Sound Pressure Level From the Free Field to the Eardrum in the Horizontal Plane. *J. Acoust. Soc. Am*. 56(6), 1848-61. doi: 10.1121/1.1903522

Shelton, J., and Kumar, G.P. (2010). Comparison between auditory and visual simple reaction time. *Neurosci. Med.* 1:30–2. doi: 10.4236/nm.2010.11004

Simpson, A. J., Reiss, J. D., and McAlpine, D. (2013). Tuning of human modulation filters is carrier-frequency dependent. *PLoS One*. 8(8), e73590. doi: 10.1371/journal.pone.0073590

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging.* 17(1), 87–97. doi: 10.1109/42.668698

Smith, E. G., and Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *J. Child Psychol. Psychiatry. 48*(8), 813–821. doi: 10.1111/j.1469-7610.2007.01766.x

Sotero, R. C., and Trujillo-Barreto, N. J. (2007). Modelling the role of excitatory and inhibitory neuronal activity in the generation of the BOLD signal. *NeuroImage*. 35(1), 149–165. doi: 10.1016/j.neuroimage.2006.10.027

Soto-Faraco, S., Spence, C., and Kingstone, A. (2004). Cross-modal Dynamic Capture: Congruency Effects in the Perception of Motion Across Sensory Modalities. *J. Exp. Psychol. Hum. Percept. Perform.* 30(2), 330-45. doi: 10.1037/0096-1523.30.2.330

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Ann. N. Y. Acad. Sci.* 1296(1), 31–49. doi: 10.1111/nyas.12121

Spence, C., and Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Curr. Biol*. 13(13), R519-21.

Starke, J., Ball, F., Heinze, H. J., and Noesselt, T. (2020). The spatio-temporal profile of multisensory integration. *Eur. J. Neurosci*. 51(5), 1210–1223. doi: 10.1111/ejn.13753

Stein, B. E., and Stanford, T. R. (2008). Multisensory Integration: Current Issues From the Perspective of the Single Neuron. *Nat. Rev. Neurosci*. 9(4), 255-66. doi: 10.1038/nrn2331

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T.J., and Rowland, B. A. (2009). Challenges in Quantifying Multisensory Integration: Alternative Criteria, Models, and Inverse Effectiveness. *Exp. Brain Res*. 198(2-3), 113-26. doi: 10.1007/s00221-009-1880-8

Steinschneider, M., Arezzo, J., and Vaughan, H. G. (1980). Phase-locked cortical responses to a human speech sound and low-frequency tones in the monkey. *Brain Res*. 198(1), 75–84.

Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E., Breakspear, M., and Friston, K. J. (2008). Nonlinear dynamic causal models for fMRI. *NeuroImage*. 42(2), 649–662. doi: 10.1016/j.neuroimage.2008.04.262

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B., and James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *NeuroImage*. 49(4), 3308–3318. doi: 10.1016/j.neuroimage.2009.12.001

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., et al. (2014). Evidence for diminished multisensory integration in autism spectrum disorders. *J. Autism Dev. Disord.* 44(12), 3161–3167. doi: 10.1007/s10803-014-2179-6

Stone, D. B., Urrea, L. J., Aine, C. J., Bustillo, J. R., Clark, V. P., and Stephen, J. M. (2011). Unisensory processing and multisensory integration in schizophrenia: a high-density electrical mapping study. *Neuropsychologia*. 49(12), 3178–3187. doi: 10.1016/j.neuropsychologia.2011.07.017

Su, L., Zulfiqar, I., Jamshed, F., Fonteneau, E., and Marslen-Wilson, W. (2014). Mapping tonotopic organization in human temporal cortex: representational similarity analysis in EMEG source space. *Front. Neurosci*, 8, 368. doi: 10.3389/fnins.2014.00368

Surguladze, S. A., Calvert, G. A., Brammer, M. J., Campbell, R., Bullmore, E. T., Giampietro, V., et al. (2001). Audio-visual speech perception in schizophrenia: an fMRI study. *Psychiatry Res.* 106(1), 1–14. doi: 10.1016/s0925-4927(00)00081-0

Sweet, R. A., Dorph-Petersen, K. A., and Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *J. Comp. Neurol.* 491(3), 270–289. doi: 10.1002/cne.20702

Tabas A., Andermann M., Schuberth V., Riedel H., Balaguer-Ballester E., and Rupp A. (2019). Modeling and MEG evidence of early consonance processing in auditory cortex. *PLoS Comput. Biol.* 15(2), e1006820. doi: 10.1371/journal.pcbi.1006820

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration?. *Cereb. Cortex.* 17(3), 679–690. doi: 10.1093/cercor/bhk016

Tanabe, H. C., Honda, M., and Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *J. Neurosci*. 25(27), 6409–6418. doi: 10.1523/JNEUROSCI.0636-05.2005

ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., and Zion-Golumbic, E. (2014). Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychologia*. 63:43-50. doi: 10.1016/j.neuropsychologia.2014.08.008

Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science.* 292(5515), 290–293. doi: 10.1126/science.1058911

Tischbirek, C. H., Noda, T., Tohmi, M., Birkner, A., Nelken, I., et al. (2019). In Vivo Functional Mapping of a Cortical Column at Single-Neuron Resolution. *Cell Rep.* 27(5), 1319–1326.e5. doi: 10.1016/j.celrep.2019.04.007

van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*. 43(2), 271–282. doi: 10.1016/j.neuron.2004.06.025

Van de Moortele, P. F., Auerbach, E. J., Olman, C., Yacoub, E., Uğurbil, K., and Moeller, S. (2009). T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. *NeuroImage*. 46(2), 432–446. doi: 10.1016/j.neuroimage.2009.02.009

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C. J., Joris, et al. (2019). The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. *Hear. Res*. 377, 109–121. doi: 10.1016/j.heares.2019.03.011

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., and Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *J. Cogn. Neurosci*. 17(3), 367–376. doi: 10.1162/0898929053279577

Vroomen, J., Bertelson, P., and de Gelder, B. (2001a). Directing spatial attention towards the illusory location of a ventriloquized sound. *Acta Psychol.* 108(1), 21–33. doi: 10.1016/s0001-6918(00)00068-8

Vroomen, J., Bertelson, P., and de Gelder, B. (2001b). The ventriloquist effect does not depend on the direction of automatic visual attention. *Percept. Psychophys*. 63(4), 651–659. doi: 10.3758/bf03194427

Wallace, M. N., Johnston, P. W., and Palmer, A. R. (2002). Histochemical identification of cortical areas in the auditory region of the human brain. *Exp. Brain Res.* 143(4), 499–508. doi: 10.1007/s00221-002-1014-z

Wang, Y., Celebrini, S., Trotter, Y., and Barone, P. (2008). Visuo-auditory interactions in the primary visual cortex of the behaving monkey: electrophysiological evidence. *BMC Neurosci.* 9, 79. doi: 10.1186/1471-2202-9-79

Wei, H., Jafarian, A., Zeidman, P., Litvak, V., Razi, A., Hu, D., and Friston, K. J. (2020). Bayesian fusion and multimodal DCM for EEG and fMRI. *NeuroImage*. 211, 116595. doi: 10.1016/j.neuroimage.2020.116595

Werner, S., and Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J. Neurosci.* 30(7), 2662–2675. doi: 10.1523/JNEUROSCI.5091-09.2010

Wetherill, G.B., and Levitt, H. (1965). Sequential Estimation of Points on a Psychometric Function. *Br. J. Math. Stat. Psychol*. 18(1), 1–10.

Wilson, H. R. (1997). A neural model of foveal light adaptation and afterimage formation. *Vis. Neurosci*. 14(3), 403–423.

Wilson, H. R. (1999). "Computation by excitatory and inhibitory networks", in Spikes, Decisions & Actions: Dynamical Foundations of Neuroscience (Oxford University Press), 88–115.

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12(1), 1–24. doi: 10.1016/S0006-3495(72)86068-5

Wilson, H. R., and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*. 13(2), 55–80. doi: 10.1007/BF00288786

Wilson, H. R., and Kim, J. (1994). Perceived motion in the vector sum direction. *Vision Res*. 34(14), 1835–1842.

Winer, J. A., and Schreiner, C. E. (2011). The Auditory Cortex. Boston, MA: Springer US doi:10.1007/978-1-4419-0074-6

Wozny, D.R., and Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. *J. Neurosci*. 31(12), 4607-12. doi: 10.1523/JNEUROSCI.6079-10.2011

Yarden, T. S., and Nelken, I. (2017). Stimulus-specific adaptation in a recurrent network model of primary auditory cortex. *PLoS Comput. Biol.* 13(3), e1005437. doi: 10.1371/journal.pcbi.1005437

Yi, H. G., Leonard, M. K., and Chang, E. F. (2019). The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron*. 102(6), 1096–1110. doi: 10.1016/j.neuron.2019.04.023

Yost, W. A. (2009). Pitch Perception. *Atten. Percept. Psychophys*. 71(8), 1701–1715. doi: 10.3758/APP.71.8.1701

Yost, W. A. (2010). Pitch Perception. *Senses Compr. Ref*. 3(8), 807–828.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*. 31(3), 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Zhang, N., and Chen, W. (2006). A dynamic fMRI study of illusory double-flash effect on human visual cortex. *Exp Brain Res*. 172(1), 57–66.

Zhou, H. Y., Cai, X. L., Weigl, M., Bang, P., Cheung, E., and Chan, R. (2018). Multisensory temporal binding window in autism spectrum disorders and schizophrenia spectrum disorders: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* 86, 66–76. doi: 10.1016/j.neubiorev.2017.12.01

Zilany M. S., Bruce I. C., and Carney L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *J Acoust Soc Am*. 135(1), 283–286. doi: 10.1121/1.4837815

Zimmermann, J., Goebel, R., De Martino, F., van de Moortele, P. F., Feinberg, D., et al. (2011). Mapping the organization of axis of motion selective features in human area MT using high-field fMRI. *PLoS One*. 6(12), e28716. doi: 10.1371/journal.pone.0028716

Zulfiqar, I., Moerel, M., and Formisano, E. (2020). Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex. *Front. Comput. Neurosci.* 13, 95. doi: 10.3389/fncom.2019.00095

Acknowledgements

Over the *many* years taken to finish this thesis, I was lucky enough to explore my research interests with freedom and try my hand at many different skills. This was possible, first and foremost, because of my awesome supervisory team who lent their unrelenting support and guidance throughout this journey.

Elia, I cannot begin to express my gratitude for this opportunity. From day one, you made me feel like part of the team where we worked together. You always made work fun and exciting with your contagious enthusiasm.

Peter, you are such a great teacher. I am extremely grateful for your patience and kindness over the years and teaching me the value of being detail-oriented and critical thinking.

And Michelle, words cannot express how much your support has meant to me during this journey. Thank you for answering my relentless questions over the years, celebrating with me when things went right (as rare as those moments were), and supporting me when all hope was lost (all the time). You truly are an inspiration with your dedication to research. I wish you the very best in every aspect of your life.

I would also like to extend my sincere thanks to all the people who contributed invaluably with their guidance: Martin, many thanks for introducing me to a different dimension of computational modeling and for your unfailingly helpful nature. Federico, I genuinely appreciate all your help with setting up the fMRI experiment. Special thanks for being extremely supportive and understanding while I took my time to finish this thesis. Agustin, many thanks for your patience and time in explaining the statistical trickery of a four-way ANOVA. Sri, I very much appreciate your help in the analysis of the fMRI data along with your always friendly nature.

This journey was taken along with many comrades-in-arms over at MaCSBio. Thank you all for innumerable cakes! I wish you all the very best for your careers. To those who volunteered to participate in my experiments (you know who you are), my deepest gratitude for recognizing my desperation and taking pity ☺

Chaitra, I am so happy you came along! Thank you for being a great listener, sharing my love of superheroes, innumerable coffees, cathartic conversations, all the great food, and, above all, your friendship.

Shauna, I miss our conversations across the room from the opposite ends (also long hours spent talking in parking lots after saying goodbye ten times). Going through this process with you made it a whole lot better. Thanks for sharing my love of coffee and always being great company. P.S.: The code fairies didn't follow me either.

Maryam, I am grateful for your kindness, amazing food, and hospitality over the years. Samar, thank you for always being a respectful and kind neighbor. Bob, your readiness for help and love of coffee was always appreciated. To the new lot of the office who came in bright-eyed and bushy-tailed (Balazs, Bart and David), I wish you fruitful research ahead ☺

Claudia, you are such a kind and sweet person, and your support has meant a lot to me ☺

Patricia, shine bright wherever you are.

Many thanks to the kind people at CN (Martha, Amaia, Emily, Teresa among many others), especially the people in the auditory group, I learned a lot from each one of you and am deeply grateful for all your feedback.

To the guys at Banditos, many thanks for the great coffee and your always friendly attitude.

Special thanks to my Master's thesis supervisor Dr. Jon Barker for introducing me to the fascinating world of computational modeling of audition. For the people at Neurolex research group, especially Dr. William Marslen-Wilson and Dr. Su Li, your belief in me gave me the confidence to pursue the next step in research. I am forever grateful.

To my sisters in faith who were always breath of fresh air on long tiring days: Farah, Dina, Nahla, and many other beautiful souls... I am deeply grateful for your company and pray for istiqamah on sirat-ul-mustaqeem for us all.

To Lily, Harry and family, you made me feel welcome from day one and took such good care of me over these years. I cannot thank you enough for your kindness. Best landlords ever! ☺

My deepest love and appreciation for Bhai and family (Umber baji, Abdullah, Fatima), Sana and family (Jahangir bhai, Umaar and Mustafa), and Sarah for following me to whichever city and country life has taken me ☺ I hope we continue our adventures in the post-corona world soon.

And lastly, to the two most beautiful supportive people, Amma and Abbu, if this is an achievement, it stands on your shoulders. I am forever indebted to you for always having my back, for letting me be whoever I wanted to be, and for giving me the confidence to explore my way in the world. Thank you for traveling far and beyond for my sake. I love you both very much.

About the Author

Isma Zulfiqar was born on the 16th of April 1989 in Islamabad, Pakistan. She completed her early education in Rawalpindi and Islamabad. She obtained Bachelor's degree in Computer Engineering in 2009 from the College of Electrical and Mechanical Engineering, NUST, Pakistan. In 2011, she completed her Master's degree in Advanced Computer Science with distinction from The University of Sheffield, UK where her thesis was focused on computational auditory scene analysis.

She continued research work first in speech technology at University of Reading, UK, and later in neurolinguistics at the Neurolex Group, University of Cambridge, UK. It was here under the mentorship of Dr William Marslen-Wilson and Dr Su Li that she was introduced to the application of computer science to the auditory neuroscience and encouraged for a future in research.

To pursue this interdisciplinary interest, she joined the Maastricht Centre for Systems Biology (MaCSBio), Maastricht University in 2015 for doctoral research under the supervision of Dr. Elia Formisano, Dr. Peter De Weerd and Dr. Michelle Moerel. During the course of her PhD, she worked on the computational modeling of the auditory cortex and the prospects of using computational models to improve the understanding of neuroimaging and behavioral data. She also explored the multisensory influences on audition through behavioral and neuroimaging studies.

She is currently pursuing post-doctoral research at the Department of Cognitive Neuroscience, Maastricht University under BRAIN initiative (NIH) with Dr. Federico De Martino, Dr. Martin Havlicek and Dr. Elia Formisano.

Publications

Zulfiqar, I., Moerel, M., and Formisano, E. (2020). Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex. *Front. Comput. Neurosci.* 13, 95. doi: 10.3389/fncom.2019.00095

Su, L., **Zulfiqar, I.**, Jamshed, F., Fonteneau, E., and Marslen-Wilson, W. (2014). Mapping tonotopic organization in human temporal cortex: representational similarity analysis in EMEG source space. *Front. Neurosci*, 8, 368. doi: 10.3389/fnins.2014.00368

Zulfiqar I.*, Havlicek M.*, Moerel M., and Formisano E. (in revision). Predicting Neuronal Response Properties from Hemodynamic Responses in the Auditory Cortex.

Zulfiqar I., Moerel M., Lage-Castellanos A., Formisano E., and De Weerd P. (under review). Audiovisual Interactions among Near-threshold Oscillating Stimuli in the Far Periphery are Phase-dependent

Zulfiqar I., Formisano E., Kashyap S., De Weerd P., and Moerel M. (in preparation). Cortical Depth-dependent Multisensory and Attentional Influences on Peripheral Sound Processing

Conference Contributions

Zulfiqar, I., Moerel, M., and Formisano, E. (2016). A computational model of temporal processing in the human auditory cortex. Society for Neuroscience. San Diego, CA

Zulfiqar, I., Moerel, M., and Formisano, E. (2016). A computational model of temporal processing in the human auditory cortex. Advances and Perspectives in Auditory Neuroscience. San Diego, CA

Zulfiqar, I., Moerel, M., and Formisano, E. (2017). A computational model of temporal processing in the human auditory cortex. 26th Annual Computational Neuroscience Meeting. Antwerp, Belgium

* equal contribution